НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО

ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

2022, ТОМ 22, НОМЕР 1 (ЯНВАРЬ–ФЕВРАЛЬ) ISSN 2226-1494 (PRINT), 2500-0373 (ONLINE)

ОПТИЧЕСКИЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Исследование силиконовой пленки, осажденной на поверхность кварцевого стекла под действием лазерного излучения — *Беликов А.В., Клочков И.С., Алексеев И.В., Капралов С.А.* — *1*

Оптические композиты на основе органических полимеров и полупроводниковых пигментов — *Волынкин В.М., Евстропьев С.К., Булыга Д.В., Морковский А.В., Пашин С.С., Дукельский К.В., Бурдин А.В., Бондаренко И.Б.* — *10*

АВТОМАТИЧЕСКОЕ УПРАВЛЕНИЕ И РОБОТОТЕХНИКА

Новый алгоритм идентификации частоты синусоидального сигнала с постоянными параметрами — *Нгуен Х.Т., Власов С.М., Пыркин А.А., Попков И.В.* — *18*

НОВЫЕ МАТЕРИАЛЫ И НАНОТЕХНОЛОГИИ

Исследование кремниевых *p-n* структур с моно- и мультифоточувствительными поверхностями (на англ. яз.) — *Мирзаалимов А.А., Гуломов Ж., Алиев Р., Мирзаалимов Н., Алиев С.* — *25*

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Детектирование состояния зевоты у водителя транспортного средства при помощи модели сверточной нейронной сети — *Лашков И.Б.* — *33*

Применение теории игр для обеспечения безопасности коммуникации киберфизической системы с использованием механизмов репутации и доверия (на англ. яз.) — *Викснин И.И., Мариненков Е.Д., Чупров С.С.* — *47*

Исследование влияния человеческих факторов на скорость движения рельсового городского транспорта — *Ляховенко Ю.А., Попов И.Ю.* — *60*

Алгоритм обнаружения RFID-дубликатов — *Волошина Н.В., Лавринович А.А.* — *67*

Редукция набора детекторов LSB с заданной достоверностью — *Солодуха Р.А., Перминов Г.В., Атласов И.В.* — *74*

Классификация объектов на изображениях с учетом искажений на основе двухэтапного топологического анализа — *Еремеев С.В., Абакумов А.В.* — *82*

Снижение размерности атрибутов с использованием нечетко оптимизированного независимого компонентного анализа для системы обнаружения вторжений в большие данные (на англ. яз.) — *Асуанандини Р., Дипа Ч.* — *93*

Оптимальная быстрая генерация и распределение квантовых ключей (на англ. яз.) — *Сума М.Р., Мадхумати П.* — *101*

Исследование способов векторизации неструктурируемых текстовых документов на естественном языке по степени их влияния на качество работы различных классификаторов — *Шадский В.В., Сизоненко А.Б., Чекмарев М.А., Шишков А.В., Исакин Д.А.* — *114*

Распознавание эмоционального состояния человека на основе сверточной нейронной сети — *Сома Г.М., Каднова А.М.* — *120*

Интеллектуализация управления развитием персонала высокотехнологичных сервис-ориентированных компаний — *Береснев А.Д.* — *127*

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Исследование эффективности работы системы коррекции магнитного компаса — *Грязин Д.Г., Падерина Т.В.* — *138*

Новая аналитическая модель тока стока и параметров малых сигналов AlGaN-GaN транзисторов с высокой подвижностью электронов (на англ. яз.) — *Фарти А., Тухами А.* — *147*

Вменение и системное моделирование параметров кислотно-основного состояния различных групп пациентов (на англ. яз.) — *Курапеев Д.И., Лушнов М.С., Ман Т., Жукова Н.А.* — *155*

Построение на базе задачи машины Дубинса опорных траекторий движения объектов с учетом постоянных внешних воздействий — *Хабаров С.П., Шилкина М.Л.* — *167*

Математическая модель эпидемии с произвольным законом восстановления — *Семёнов В.К., Иванова Н.Б.* — *179*

Моделирование импульсного истечения смеси воздуха и мелкодисперсного порошка, частично заполняющего выбросной канал — *Садин Д.В.* — *187*

Использование векторизованных структур данных при реализации вычислительных алгоритмов решения задач механики сплошной среды — *Брыков Н.А., Волков К.Н., Емельянов В.Н.* — *193*

Сравнительный анализ алгоритмов вычислительного интеллекта для оценки канала LTE (на англ. яз.) — *Патан С., Нуния А., Тамболи М., Патхак С.* — *206*

КРАТКИЕ СООБЩЕНИЯ

Внедрение системы поддержки принятия решений для повышения качества медицинских данных пациентов с артериальной гипертензией — *Ионов М.В., Болгова Е.В., Звартау Н.Э., Авдонина Н.Г., Балахонцева М.А., Ковальчук С.В., Конради А.О.* — *217*

Главный редактор — В.О. Никифоров, д.т.н., профессор

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Айфантис Э.Ч. — PhD, профессор (Греция, США)
Беззатеев С.В. — д.ф.-м.н., доцент (Россия)
Беликов А.В. — д.ф.-м.н., профессор (Россия)
Бобцов А.А. — д.т.н., профессор (Россия)
Бухановский А.В. — д.т.н. (Россия)
Гунько Ю.К. — PhD, профессор (Ирландия)
Ефимов Д.В. — PhD (Франция)
Кармановский Н.С. (заместитель главного редактора) — к.т.н. доцент (Россия)
Кившарь Ю.С. — к.ф.-м.н., профессор Австралийской академии наук (Австралия)
Козлов С.А. — д.ф.-м.н., профессор (Россия)
Колмановский И.В. — PhD, профессор (США)
Коробейников А.Г. — д.т.н., профессор (Россия)
Лисицына Л.С. — д.т.н., профессор (Россия)
Мельников В.Г. — д.т.н., доцент (Россия)
Никоноров Н.В. — д.т.н., профессор (Россия)
Розанов Н.Н. — д.ф.-м.н., член-корреспондент РАН (Россия)
Романов А.Е. — д.ф.-м.н., профессор (Россия)
Сидоркина И.Г. — д.т.н., профессор (Россия)
Симовский К.Р. — д.ф.-м.н., профессор (Финляндия)
Слоот П.М.А. — PhD, профессор (Нидерланды)
Степанов О.А. — д.т.н., профессор (Россия)
Тамре М. — д.т.н., профессор (Эстония)
Татарникова Т.М. — д.т.н., доцент (Россия)
Тимофеева Т.В. — к.х.н., профессор (США)
Шалыто А.А. — д.т.н., профессор (Россия)

Ответственный секретарь — И.В. Малькова

Издание зарегистрировано в Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций (свидетельство ПИ №ФС77-67990 от 6 декабря 2016 г.).

Языки журнала: русский, английский.

ISSN 2226-1494 (print version), ISSN 2500-0373 (online version)

Англоязычное название: «Scientific and Technical Journal of Information Technologies, Mechanics and Optics». Транслитерированное название: «Nauchno-Tekhnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki». Сокращенное название: «Sci.Tech. J. Inf. Technol. Mech. Opt.»

Журнал включен в каталог периодических изданий Ulrich's Periodical Directory.

Журнал входит в утвержденный Высшей аттестационной комиссией «Перечень ведущих рецензируемых научных журналов и изданий, выпускаемых в Российской Федерации, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени доктора и кандидата наук».

Периодичность издания — 6 выпусков в год.

Плата за публикации и редактирование не взимается.

Редакторы: Ю.В. Михайлова, Л.Н. Точилина, Т.В. Точилина
Перевод — М.В. Хохлова
Компьютерная верстка — Е.С. Егорова

Оф. 2136, Университет ИТМО, ул. Ломоносова., д. 9,
Санкт-Петербург, Российская Федерация, 191002

Телефон (812) 480 02 75 http://ntv.ifmo.ru
E-mail: ntvitmo@itmo.ru

Статьи Журнала доступны по лицензии Creative Commons «Attribution-NonCommercial» 4.0 Всемирная.

Editor-in-Chief — Vladimir O. Nikiforov, D.Sc, Professor

EDITORIAL BOARD

Elias C. Aifantis — PhD, Professor (Greece, USA)
Sergey V. Bezzateev — D.Sc, Associate Professor (Russia)
Andrey V. Belikov — D.Sc, Professor (Russia)
Alexey A. Bobtsov — D.Sc, Professor (Russia)
Alexandr V. Boukhanovsky — D.Sc (Russia)
Yurii K. Gun'ko — PhD, Professor (Ireland)
Denis V. Efimov — PhD (France)
Nikolai S. Karmanovskiy (Deputy Chief Editor) — PhD, Associate Professor (Russia)
Yuri S. Kivshar — PhD, Distinguished Professor (Australia)

Sergey A. Kozlov — D.Sc, Professor (Russia)
Ilya V. Kolmanovsky — PhD, Full Professor (USA)
Anatoly G. Korobeynikov — D.Sc, Professor (Russia)
Lubov S. Lisitsyna — D.Sc, Professor (Russia)
Vitaly G. Melnikov — D.Sc, Associate Professor (Russia)
Nikolay V. Nikonorov — D.Sc, Professor (Russia)
Nikolay N. Rosanov — D.Sc, Professor, corresponding member of the RAS (Russia)
Alexey E. Romanov — D.Sc, Professor (Russia)
Irina G. Sidorkina — D.Sc, Professor (Russia)
Konstantin R. Simovski — D.Sci., Full Professor (Finland)
Peter M.A. Sloot — PhD, Professor (the Netherlands)
Oleg A. Stepanov — D.Sc, Professor (Russia)
Mart Tamre — D.Sc, Professor (Estonia)
Tatiana M. Tatarnikova — D.Sc, Associate Professor (Russia)
Tatiana V. Timofeeva — PhD, Professor (USA)
Anatoly A. Shalyto — D.Sc, Professor (Russia)

Executive secretary — Irina V. Malkova

Publication is registered by the Federal Inspectorate Service for Communication, Information Technologies and Communication Media, certificate PI FS77-67 990 dated December, 6th, 2016.

Languages of the journal: Russian, English

ISSN 2226-1494 (print version), ISSN 2500-0373 (online version) The English title is "Scientific and Technical Journal of Information Technologies, Mechanics and Optics".

Transliterated title is "Nauchno-Tekhnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki".

Abbreviated parallel title: "Sci.Tech. J. Inf. Technol. Mech. Opt."

The journal is catalogued in Ulrich's Periodical Directory.

The journal is included in the "List of leading scientific journals and periodicals under review in the Russian Federation, where the main scientific results of theses for the PhD and doctor of sciences degree must be published" approved by the Higher Attestation Commission.

Publication frequency is 6 times a year.

Publication and editing are free of charge.

Editors: Yulia V. Mikhailova, Lyubov N. Tochilina, Tatiana V. Tochilina
English Language Technical Editor — Maria V. Khokhlova
Computer layout — Ekaterina S. Egorova

of. 2136, ITMO University, Lomonosova Street, 9,
St.Petersburg, 191002, Russia

Phone (812) 480 02 75 http://ntv.ifmo.ru/en
E-mail: ntvitmo@itmo.ru



Подписано к печати 01.02.2022 Тираж 350 экз. Заказ № 85 (137)

Отпечатано в учреждении «Университетские телекоммуникации»
Адрес: 197101, Санкт-Петербург, Кронверкский пр., д. 49, литера А

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ОПТИЧЕСКИЕ СИСТЕМЫ И ТЕХНОЛОГИИ

OPTICAL ENGINEERING

doi: 10.17586/2226-1494-2022-22-1-1-9
УДК 621.373.8: 620.1.08

Исследование силиконовой пленки, осажденной на поверхность кварцевого стекла под действием лазерного излучения

Андрей Вячеславович Беликов[1], Иван Сергеевич Клочков[2]✉, Иван Владимирович Алексеев[3], Сергей Анатольевич Капралов[4]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2,3] АО «Научно-производственная корпорация «Системы прецизионного приборостроения», Санкт-Петербург, 190020, Российская Федерация
[4] АО «Научно-производственная корпорация «Системы прецизионного приборостроения», Великий Новгород, 173003, Российская Федерация

[1] volnagdg@yandex.ru, https://orcid.org/0000-0001-8733-4860
[2] Vany-nov13@yandex.ru✉, https://orcid.org/0000-0001-5461-783X
[3] El.magio@yandex.ru, https://orcid.org/0000-0002-4368-9309
[4] kapralov_sa@npkspp.ru, https://orcid.org/0000-0002-3393-4948

Аннотация

Предмет исследования. Исследованы структура, оптические и эксплуатационные свойства силиконовой пленки, нанесенной на поверхность кварцевого стекла в результате действия лазерного излучения на летучие компоненты, которые выделяются из образца силиконовой резины в замкнутом объеме. **Метод.** Исследование выполнено в рамках лазерного многопараметрического метода с помощью оригинальной установки, в состав которой входит твердотельный неодимовый лазер с длиной волны 1064 нм и параметрами лазерного импульса: энергией — 105 мДж, длительностью 11–14 нс и частотой следования 10 Гц. В эксперименте на выходе лазерного излучателя расположена герметичная тест-кювета, внутри которой размещен образец силиконовой резины. При прохождении лазерного излучения через внутренний объем тест-кюветы происходит взаимодействие летучих компонентов, выделенных из образца, с лазерным излучением. В результате процесса лазерно-индуцированного осаждения на оптических элементах тест-кюветы образуются зоны осаждения. Топология зон осаждения определена с помощью профилометра. Сравнение структурного состава исходного образца с продуктами осаждения выполнено с использованием сканирующего электронного микроскопа. **Основные результаты.** Получены зависимости коэффициентов площади зоны осаждения и ослабления от температуры и количества лазерных импульсов. Исследованы элементный состав, цвет, устойчивость к действию растворителя и толщина зон осаждения. Установлено, что с ростом температуры и количества лазерных импульсов коэффициенты площади и ослабления зон осаждения увеличиваются, цвет не изменяется, а устойчивость к действию растворителя возрастает. С ростом температуры зона осаждения, первоначально состоящая из микрофрагментов, становится сплошной, а с ростом количества лазерных импульсов ее толщина увеличивается. Толщина зоны осаждения неравномерно распределена относительно центра. **Практическая значимость.** Полученные результаты могут быть применены при разработке содержащих силикон биочипов для диагностики и терапии состояния здоровья человека.

Ключевые слова

лазерный многопараметрический метод, герметичный объем, тест-кювета, коэффициент ослабления, зона осаждения, силиконовая резина

Ссылка для цитирования: Беликов А.В., Клочков И.С., Алексеев И.В., Капралов С.А. Исследование силиконовой пленки, осажденной на поверхность кварцевого стекла под действием лазерного излучения // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 1–9. doi: 10.17586/2226-1494-2022-22-1-1-9

© Беликов А.В., Клочков И.С., Алексеев И.В., Капралов С.А., 2022

A study of a silicone film deposited on quartz glass under laser radiation

Andrey V. Belikov[1], Ivan S. Klochkov[2]✉, Ivan V. Alekseev[3], Sergey A. Kapralov[4]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation
[2,3] Joint Stock Company "Research and Production Corporation "Precision Systems and Instruments", Saint Petersburg, 190020, Russian Federation
[4] Joint Stock Company "Research and Production Corporation "Precision Systems and Instruments", Veliky Novgorod, 173003, Russian Federation

[1] volnagdg@yandex.ru, https://orcid.org/0000-0001-8733-4860
[2] Vany-nov13@yandex.ru✉, https://orcid.org/0000-0001-5461-783X
[3] El.magio@yandex.ru, https://orcid.org/0000-0002-4368-9309
[4] kapralov_sa@npkspp.ru, https://orcid.org/0000-0002-3393-4948

Abstract

The paper studies the structure, optical and operational properties of a silicone film deposited on the surface of quartz glass as a result of the action of laser radiation on volatile substances released from a silicone rubber sample in a closed volume. The research was carried out within the framework of the laser multiparameter method and using an original setup, which includes a solid-state neodymium laser with a wavelength of 1064 nm, and laser pulse parameters: energy of 105 mJ, a duration of 11–14 ns, a repetition rate of 10 Hz. A sealed test-cuvette is placed at the output of the laser. A silicone rubber sample is placed inside the test-cuvette. When laser radiation passes through the test-cuvette, volatile substances that are released from the sample over time interact with the laser radiation and create deposition zones on the optical elements of the test-cuvette, which affect the optical characteristics of these optical elements. The topology of the deposition zones was studied using a profilometer. The structural composition of the original silicone rubber and the deposition zone was determined using a scanning electron microscope. The main results show the dependences of the coefficients of the area and attenuation of the deposition zone on the temperature and the number of laser pulses. The elemental composition, color, resistance to the action of the solvent, and the thickness of the deposition zones have been investigated. It was found that with an increase in the temperature and the number of laser pulses, the area and attenuation coefficients of the deposition zones increase, the color does not change, and the resistance to the action of the solvent increases. With an increase in temperature, the deposition zone, initially consisting of micro-fragments, becomes continuous, and with an increase in the number of laser pulses, its thickness increases. The thickness of the deposition zone is unevenly distributed along its diameter. The results obtained can be applied in the development of silicone-containing biochips for health diagnostics and therapy.

Keywords

laser multiparameter method, sealed volume, test-cuvette, attenuation, deposition zone, silicone rubber

For citation: Belikov A.V., Klochkov I.S., Alekseev I.V., Kapralov S.A. A study of a silicone film deposited on quartz glass under laser radiation. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 1–9 (in Russian). doi: 10.17586/2226-1494-2022-22-1-1-9

Введение

В настоящее время полимерные носители широко используются в составе биочипов и биосенсоров в различных биомедицинских приложениях, таких как диагностика, адресная доставка лекарственных средств [1], репродуктивная [2] и регенеративная [3, 4] медицина, антибактериальная [5] и противораковая [6, 7] терапия, тканевая инженерия [8, 9], применяются для контроля экспрессии генов клеточных культур [10] и прочее. Существует широкий спектр систем адресной доставки биологически активных веществ (БАВ), таких как полимерные или липосомальные микрокапсулы, полимерные пленки, позволяющие контролируемо высвобождать инкапсулированные в них агенты посредством лазерного, электрического, магнитного, ультразвукового или биохимического воздействий [11–13].

В контексте задач модификации подложек биочипов и биосенсоров, клеточных подложек, поверхности медицинских имплантов перспективными носителями БАВ являются полимерные трехмерные микроструктурированные пленки. Пленки могут изготавливаться как методами последовательной адсорбции, так и методами микролитографии, что позволяет варьировать их физико-химические свойства [14, 15]. Полимерные носители, к которым можно отнести силиконовые пленки, позволяют герметично инкапсулировать низко- и высокомолекулярные водорастворимые БАВ в пленке. В результате достигается возможность равномерно распределять БАВ по поверхности пленки в виде небольших одинаковых кластеров, которые возможно активировать как поодиночке, так и целыми группами, с помощью лазерного излучения, при локальной таргетной терапии. Вместе с тем поиск новых технологий создания полимерных пленок, расширение температурного диапазона их использования, управление адгезией таких пленок с поверхностью подложки, в том числе из кварцевого стекла, используемого в биосовместимых сенсорах, является актуальной задачей.

Постановка задачи

В процессе эксплуатации лазерных изделий в результате конденсации на поверхности оптических элементов летучих веществ под воздействием лазерного излучения формируются зоны осаждения [16–21]. В герметичном объеме лазерного излучателя формирование зон осаждения приводит к деградации оптических элементов и узлов, что негативно сказывается на выходных параметрах излучателя и может привести к выходу прибора из строя. Процесс осаждения с ростом температуры может ускоряться [22]. Вместе с тем

реализуемый при формировании зон осаждения в лазерных излучателях процесс, при размещении на пути лазерного излучения герметичной кюветы, содержащей образец силиконовой резины и образец кварцевого стекла, может быть применен для создания на поверхности кварцевого стекла полимерных, в том числе силиконовых пленок, которые могут быть использованы в составе биочипов и биосенсоров.

Существуют различные методы исследования продуктов осаждения. Наиболее часто для этих целей проводят дегазацию в вакуумной камере, содержащей исследуемый образец материала. Чтобы определить наличие продуктов газовыделения, производят взвешивание образца материала до и после дегазации. На основе полученных измерений делают выводы о присутствии в материале продуктов осаждения (летучих веществ), о потере массы или газовыделении. Состав и природу продуктов осаждения определяют с помощью спектрофотометров [23] или электронного просвечивающего микроскопа, работающего в режиме электронографа [24]. Данные методы позволяют прецизионно определить состав осажденных веществ, а также вес продуктов осаждения с точностью до одного микрограмма. Недостатками этих методов являются значительное количество времени, которое требуется для исследования, и высокие затраты при их реализации.

Известен лазерный многопараметрический (ЛМП) метод, в рамках которого исследуются оптические, геометрические, цветовые параметры зон осаждения и их стойкость к действию растворителей [16]. В отличие от вакуумных методов, ЛМП метод обладает такими преимуществами как простота и высокая скорость измерений. ЛМП метод разработан и успешно применяется в организации АО «НПК «СПП» (Санкт-Петербург, Россия) при контроле материалов и элементов, используемых в герметичном объеме лазерных излучателей. В рамках данного метода выполнен анализ параметров зоны осаждения, таких как коэффициент площади осаждения, цвет, устойчивость к действию растворителей и коэффициент ослабления, прошедшего через нее излучения [16].

При поиске новых технологий создания силиконовых пленок для биомедицинских применений актуально исследование возможности создания и изучение свойств зон осаждения, формируемых на поверхностях оптических элементов, размещенных на пути распространения лазерного излучения.

Задача исследования состоит в получении новых научных данных о возможности создания и свойствах силиконовых пленок, формируемых на поверхности кварцевого стекла в результате действия лазерного излучения на летучие компоненты, выделяющиеся из образца силиконовой резины в замкнутом объеме. При решении этой задачи в рамках ЛМП метода исследуются: коэффициенты ослабления и площади осаждения, устойчивость к действию растворителей и цвет силиконовых пленок, а также определяется состав и толщина пленок методами электронной микроскопии и контактной профилометрии.

Материалы и методы

В эксперименте использован образец силиконовой резины: современная монолитная силиконовая резина марки EL SST (Silicone Engineering, Англия). Создана экспериментальная установка, реализующая ЛМП метод (рис. 1). В состав установки входят: *1*— поглотитель лазерного излучения; *2* — приемная головка измерителя энергии Gentec QE25LP-S-MB-DO (Gentec Electro-Optics, Inc, Канада); *3* — термостол Магистр С20Т1 (ООО НТЦ «Магистр-С», Россия); *4* — тест-кювета; *5* — YAG:Nd лазерный излучатель (АО «НПК «СПП», Россия).

Силиконовая резина получается путем вулканизации эластомера (на основе соединения кремния) и органики[1]. По внешнему виду она неотличима от резины, изготовленной на основе каучука. Технические параметры таких изделий закладываются при выборе состава компонентов еще на стадии разработки (рецепта). В зависимости от свойств органического заместителя, который может быть представлен метильными, винильными, фенильными или другими группами и его структуры, силиконовая резина может приобретать различные физические свойства. Таким образом, производитель может, подобрав состав компонентов, изготовить силиконовую резину с уникальными свойствами и характеристиками. Силиконовая резина[1] обладает высокой прочностью и эластичностью, вследствие чего может применяться в широком температурном диапазоне от –60 °С до +280 °С, устойчива к ультрафиолетовому излучению и имеет высокую огнестойкость. Также

[1] ГОСТ Р 57399-2017. Прокладки и детали из резины на основе силикона. Технические требования. Введ. 01.01.2018. М.: Стандартинформ, 2017. 11 с.

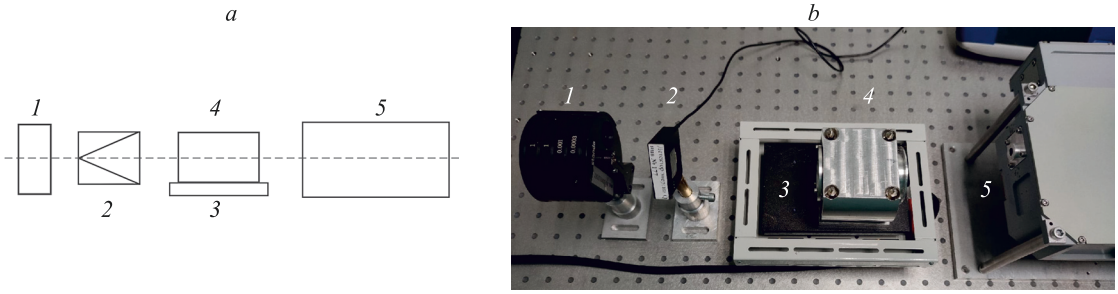


Рис. 1. Схема (*a*) и фото (*b*) установки для реализации лазерного многопараметрического метода
Fig. 1. Scheme (*a*) and photograph (*b*) of the setup for the implementation of the laser multiparameter method

данный материал не токсичен и устойчив к действию широкого круга химическим веществ. Использованная в исследовании резина широко распространена и эксплуатируется при температурах от –60 °С до +200 °С, обладает плотностью 1,15 г/см^2, твердостью (Шор А) 40 усл. ед. и прозрачна в видимом диапазоне электромагнитного спектра. Условная прочность при растяжении составляет 7–11 МПа, удлинение при растяжении порядка 700 %. Внешний вид используемых в исследовании образцов силиконовой резины марки EL SST представлен на рис. 2. Внутренний диаметр кольцевой прокладки составил 24 мм, внешний диаметр 28 мм, толщина кольца — 2 мм, площадь поверхности кольцевой прокладки — 163,4 мм^2.

Подробно ЛМП метод описан в работе [16]. В настоящей работе в состав установки, реализующей ЛМП метод, входит твердотельный неодимовый лазер, работающий в режиме модуляции добротности с длиной волны 1064 нм и параметрами лазерных импульсов: энергией 105 мДж, длительностью 11–14 нс и частотой следования 10 Гц. Диаметр пучка по уровню e^{-2} составил 5 ± 0,25 мм. Излучение многомодовое. Расходимость лазерного пучка составила 1,5 ± 0,1 мрад. Наклон оси лазерного пучка и его диаметр выбраны так, чтобы излучение тестирующего лазерного излучателя проходило сквозь входное и выходное стекла кюветы без виньетирования, и в кювете не возникало дополнительных источников загрязнения. На выходе лазера размещена тест-кювета, имитирующая внутреннее пространство герметичного лазерного излучателя, внутрь которой помещен образец резины марки EL SST. При прохождении лазерного излучения через пространство тест-кюветы летучие вещества, которые выделяются из образца с течением времени, взаимодействуют с лазерным излучением и образуют на оптических элементах тест-кюветы зоны осаждения. Таким образом, лазерное излучение не попадает на силиконовую резину, а взаимодействует с продуктами ее испарения. Кювета герметизирована при помощи индиевой фольги, размещенной между входным и выходным стеклом и корпусом кюветы, а также между основанием кюветы, на которое помещен образец резины, и корпусом кюветы.

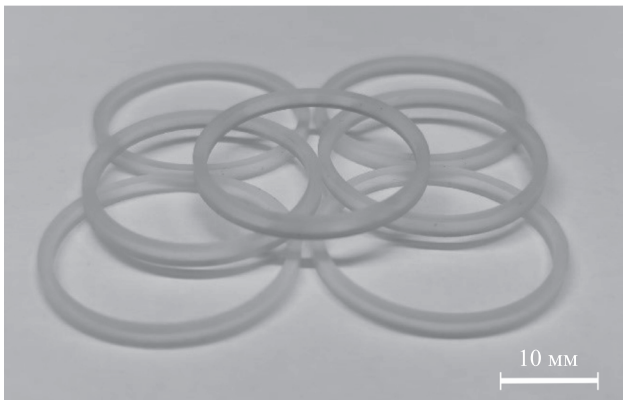


Рис. 2. Фотография образцов уплотнительных колец из силиконовой резины EL SST

Fig. 2. Photo of the samples, O-rings made of silicone rubber EL SST (Silicone Engineering, England)

При исследовании ЛМП методом внутрь тест-кюветы была помещена одна кольцевая прокладка. Исследования выполнены при температурах кюветы 23, 50, 60, 70, 80, 90 и 100 °С. При каждой температуре через кювету пропускалось заранее заданное количество лазерных импульсов. Количество импульсов изменялось от 1 до 150 000, с шагом 10 000. Для каждой комбинации температуры и количества импульсов выполнено 7 реализаций, для каждой реализации использовался новый образец уплотнительной кольцевой прокладки.

В конце каждой реализации вычислен коэффициент ослабления Δk, равный отношению разницы энергий на выходе тест-кюветы в начале и в конце облучения к энергии излучения на входе тест-кюветы [16].

Выполнен анализ окна тест-кюветы на предмет наличия/отсутствия зон осаждения, поверхность кюветы была сфотографирована с помощью камеры микроскопа DCM510 5M pixels, CMOS chip (Scoptek, Китай), и вычислен коэффициент площади зоны осаждения K, который рассчитан как отношение площади зоны осаждения к площади сечения лазерного пучка по уровню e^{-2} [16].

Определялась устойчивость к действию растворителя, для этого в течение 10 мин была предпринята попытка удалить зону осаждения в результате ее протирки ватным тампоном, содержащим 99,7 % этиловый спирт (АО «ЛенРеактив», Россия). В случае если зона не исчезала, то предпринималась попытка удалить ее с помощью протирки ватным тампоном, содержащим 99,8 % ацетона. Если зону осаждения удавалось убрать, то она характеризовалась как неустойчивая к растворителям, в противном случае — как устойчивая.

Цвет зон осаждения (белый или черный) после каждой реализации определялся визуально с помощью микроскопа МС-2 ZOOM (Микромед, Китай) при увеличении 10$^{\times}$.

Элементный состав силиконовой резины EL SST и зон осаждения исследован методом рентгеноструктурного анализа с помощью сканирующего электронного микроскопа Evex Mini-SEM HR-3000 (Evex Analytical Instruments Inc., USA). Исследованы два образца силиконовой резины и два образца зоны осаждения, образованных после воздействия 100 000 импульсов при температуре 100 °С.

Внешний вид поверхности и толщина зоны осаждения исследованы с помощью оптического микроскопа Axio Imager (Carl Zeiss GmbH, Германия) и профилометра Hommel Tester T8000 (Hommelwerke GmbH, Германия) соответственно. Исследование было выполнено при температуре 100 °С. Первый образец зоны осаждения образован после воздействия 50 000 импульсов, а второй — 100 000 импульсов.

Результаты экспериментов обработаны в программном комплексе «StatGraphics Plus 5.0» (Statistical Graphics Corp., США). Выполнена статистическая обработка данных, которая заключалась в определении средних значений и стандартного отклонения измеренных величин.

Результаты и обсуждения

Характерный внешний вид зон осаждения, возникающих на поверхности кварцевого окна тест-кюветы при постоянном количестве лазерных импульсов и различной температуре представлен на рис. 3.

Внешний вид зон осаждения зависит от температуры тест-кюветы. При температурах 23 и 50 °С зоны осаждения не наблюдаются, а при температурах 60, 70 и 80 °С зона представлена набором изолированных друг от друга микрофрагментов. При температурах 90 и 100 °С микрофрагменты образуют сплошную зону осаждения (силиконовую пленку).

Температурную динамику внешнего размера зоны осаждения характеризует коэффициент, который показывает, какую долю занимает площадь зоны осаждения относительно площади лазерного пятна. Зависимость коэффициента площади зоны осаждения K от температуры тест-кюветы при постоянном количестве лазерных импульсов показана на рис. 4, *a*; от количества лазерных импульсов при постоянной температуре тест-кюветы — на рис. 4, *b*.

Видно, что коэффициент площади зоны осаждения K отличен от нуля при превышении температурой тест-кюветы значения 50 °С, что связано с формированием на поверхности окна тест-кюветы микрофрагментов. С ростом температуры коэффициент K увеличивается. Темп роста K замедляется при превышении температурой значения 90 °С, что можно связать с окончанием процесса слияния микрофрагментов и началом процесса увеличения толщины зоны осаждения.

Заметим (рис. 4, *b*), что коэффициент площади зоны осаждения K медленно увеличивается первые 30 000 импульсов, после следует резкий рост до 80 000 импульсов, далее рост замедляется. Подобное поведение коэффициента K так же, как для зависимости

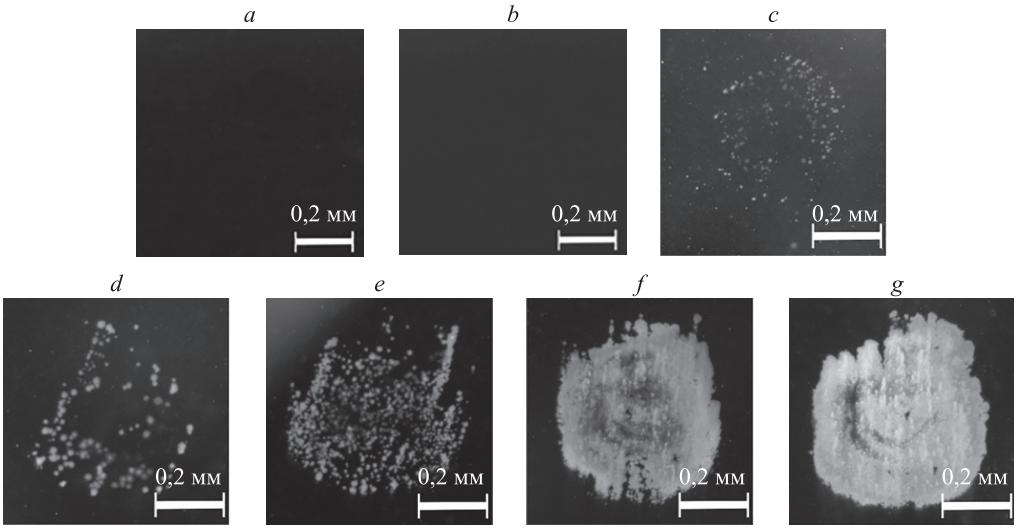


Рис. 3. Фотографии зон осаждения, возникающих на поверхности кварцевого окна тест-кюветы при 100 000 лазерных импульсов и температурах 23 °С (*a*), 50 °С (*b*), 60 °С (*c*), 70 °С (*d*), 80 °С (*e*), 90 °С (*f*) и 100 °С (*g*), силиконовая резина EL SST

Fig. 3. Photographs of deposition zones appearing on the surface of the quartz window of the test-cuvette at 100,000 laser pulses and temperatures of 23 °C (*a*), 50 °C (*b*), 60 °C (*c*), 70 °C (*d*), 80 °C (*e*), 90 °C (*f*) and 100 °C (*g*), silicone rubber EL SST (Silicone Engineering, England)

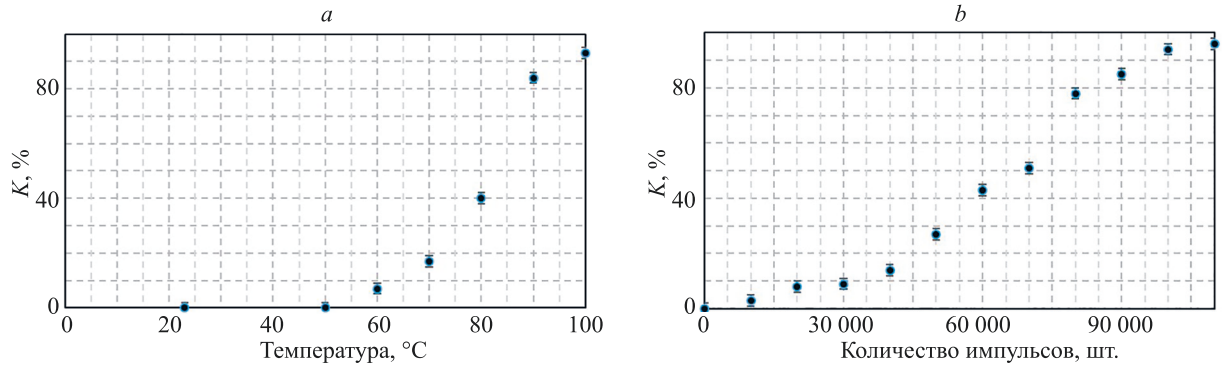


Рис. 4. Зависимость коэффициента площади зоны осаждения K от температуры тест-кюветы при 100 000 лазерных импульсов (*a*) и от количества лазерных импульсов при температуре тест-кюветы 100 °С (*b*), силиконовая резина EL SST

Fig. 4. The dependence of the area coefficient of the deposition zone K on the temperature of the test-cuvette at 100,000 laser pulses (*a*) and on the number of laser pulses at the temperature of the test-cuvette 100 °C (*b*), silicone rubber EL SST

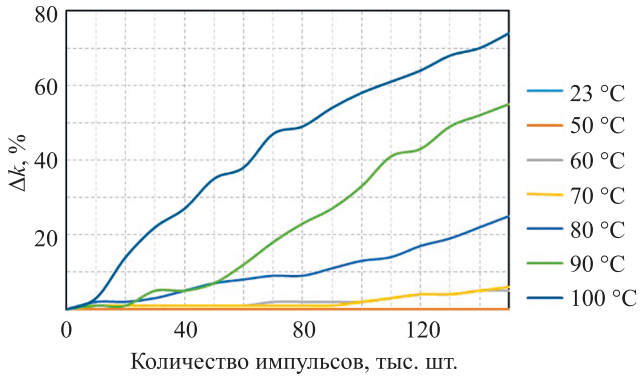


Рис. 5. Зависимости коэффициента ослабления Δ*k* от количества лазерных импульсов при различных температурах тест-кюветы

Fig. 5. The dependences of the attenuation coefficient Δ*k*, on the number of laser pulses at different temperatures of the test cell

K от температуры, можно связать с динамикой формирования и слиянием микрофрагментов.

Результаты измерения коэффициента ослабления Δ*k* представлены на рис. 5.

Отметим, что с ростом температуры тест-кюветы и количества лазерных импульсов коэффициент ослабления Δ*k* увеличивается. При постоянной температуре рост количества импульсов приводит к увеличению Δ*k*. При постоянном количестве лазерных импульсов рост температуры тест-кюветы также приводит к увеличению коэффициента ослабления.

Возрастание значений коэффициентов *K* и Δ*k* связано с ростом температуры, а также диаметра и толщины зоны осаждения, и объясняется повышением скорости выделения летучих веществ из материала образца. На величину Δ*k* оказывает влияние количество лазерных импульсов, которое возрастает пропорционально увеличению диаметра и толщины зоны осаждения.

Увеличение толщины зоны осаждения с ростом количества лазерных импульсов подтверждается при анализе профилограмм, полученных при постоянной температуре тест-кюветы и различном количестве лазерных импульсов, представленных на рис. 6.

Можно заметить, что толщина зоны осаждения неравномерно распределена вдоль ее диаметра. При 50 000 импульсов максимальная толщина зоны осаждения достигает 2,1 мкм, а средняя — 0,54 ± 0,33 мкм. При 100 000 импульсов максимальная толщина зоны осаждения достигает 2,6 мкм, а средняя — 0,90 ± 0,33 мкм.

Результаты исследования цвета и устойчивости зон осаждения к действию растворителя представлены в таблице.

Устойчивость к действию растворителя увеличивается с ростом температуры тест-кюветы, что также связано с ростом геометрических размеров зоны осаждения и в первую очередь с ростом ее толщины.

Цвет зоны осаждения не изменяется с ростом температуры тест-кюветы, что свидетельствует о том, что ее элементный состав постоянен при изменении температуры, а меняется только диаметр и толщина зоны осаждения.

Цвет зоны осаждения очень близок к цвету исследуемого образца силиконовой резины, что очевидно связано с тем, что по элементному составу зона осаждения близка к исходному материалу и состоит из конденсированного на поверхности кварцевого окна. Данный вывод подтверждается результатами рентгеноструктурного анализа, представленными на рис. 7.

Таблица. Цвет и устойчивость к действию растворителя зон осаждения, сформированных при 100 000 лазерных импульсов и различной температуре тест-кюветы

Table. Color and solvent resistance of the deposition zones formed at 100,000 laser pulses and different temperatures of the test-cuvette

Температура, °C	Устойчивость к действию растворителей	Цвет
23	зона осаждения отсутствует	
50		
60	не устойчива	белый
70	устойчива	белый
80	устойчива	белый
90	устойчива	белый
100	устойчива	белый

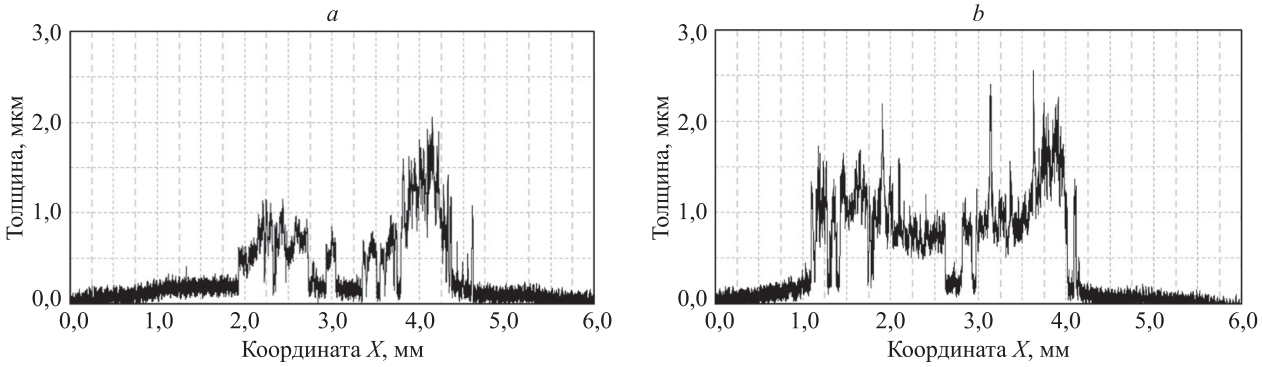


Рис. 6. Характерная профилограмма поверхности зоны осаждения, образованной при температуре 100 °C после воздействия 50 000 (*a*) и 100 000 лазерных импульсов (*b*), силиконовая резина EL SST

Fig. 6. The typical profilogram of the surface of the deposition zone formed at a temperature of 100 °C after exposure to 50,000 (*a*) and 100,000 laser pulses (*b*)

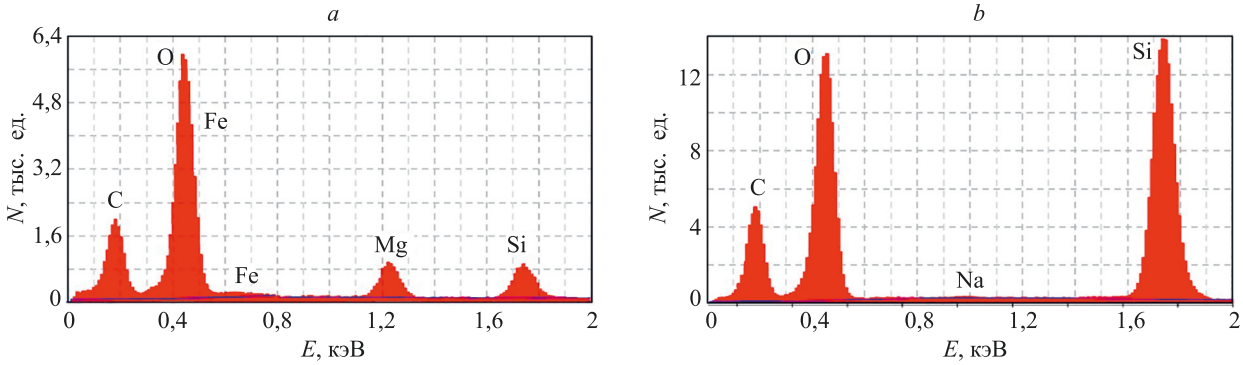


Рис. 7. Результаты рентгеноструктурного анализа образца силиконовой резины EL SST (*a*) и зоны осаждения, сформированной при температуре тест-кюветы 100 °C и 100 000 лазерных импульсов (*b*), где *N* — количество элементов, полученных в результате анализа, *E* — энергия электронов

Fig. 7. Results of X-ray diffraction analysis of a silicone rubber sample EL SST (Silicone Engineering, England) (*a*) and a deposition zone formed at a test cell temperature of 100 °C and 100,000 laser pulses (*b*), where *N* is the number of elements obtained as a result of the analysis, *E* is the electron energy

В результате исследования зоны осаждения на входном стекле тест-кюветы были обнаружены вещества, входящие в состав силиконовой резины EL SST, а именно кислород и кремний.

Заключение

Созданы и исследованы зоны осаждения, возникающие на поверхности кварцевого стекла в результате действия лазерного излучения на летучие компоненты, выделяющиеся из образца силиконовой резины EL SST (Silicone Engineering, Англия) в замкнутом объеме. Зоны осаждения исследованы в рамках лазерного многопараметрического метода. Определено, что динамика формирования зоны осаждения зависит от количества лазерных импульсов и температуры. Вычислены коэффициенты площади и ослабления зон осаждения, возникающих на поверхности оптических элементов герметичной тест-кюветы. Получена зависимость коэффициента площади зоны осаждения от температуры и количества лазерных импульсов. Определена зависимость коэффициента ослабления от количества лазерных импульсов и температуры тест-кюветы. Исследованы элементный состав, цвет, устойчивость к действию растворителя и толщина зон осаждения. Установлено, что с ростом температуры и количества лазерных импульсов коэффициенты площади и ослабления зон осаждения увеличиваются, цвет не изменяется, а устойчивость к действию растворителя увеличивается. С ростом температуры зона осаждения, первоначально состоящая из микрофрагментов, становится сплошной, и формируется силиконовая пленка. С ростом количества лазерных импульсов ее толщина увеличивается, причем изменение количества лазерных импульсов в 2 раза, приводит к увеличению толщины пленки в 1,66 раза. Созданная силиконовая пленка может быть использована в качестве основы при проектировании биочипов и биосенсоров для размещения в ней биоактивных веществ и клеток. Возможность размещения этих активаторов в пленке нуждается в дальнейшем исследовании.

Литература

1. Kamaly N., Yameen B., Wu J., Farokhzad O.C. Degradable controlled-release polymers and polymeric nanoparticles: mechanisms of controlling drug release // Chemical Reviews. 2016. V. 116. N 4. P. 2602–2663. https://doi.org/10.1021/acs.chemrev.5b00346
2. Barkalina N., Charalambous C., Jones C., Coward K. Nanotechnology in reproductive medicine: emerging applications of nanomaterials // Nanomedicine: Nanotechnology, Biology and Medicine. 2014. V. 10. N 5. P. e921–e938. https://doi.org/10.1016/j.nano.2014.01.001
3. Furth M.E., Atala A., Van Dyke M.E. Smart biomaterials design for tissue engineering and regenerative medicine // Biomaterials. 2007. V. 28. N 34. P. 5068–5073. https://doi.org/10.1016/j.biomaterials.2007.07.042
4. Gandhi A., Paul A., Sen S.O., Kumar K.S. Studies on thermoresponsive polymers: Phase behaviour, drug delivery and biomedical applications // Asian Journal of Pharmaceutical Sciences. 2015. V. 10. N 2. P. 99–107. https://doi.org/10.1016/j.ajps.2014.08.010
5. Suhardi V.J., Bichara D.A., Kwok S.J.J., Freiberg A.A., Rubash H., Malchau H., Yun S.H., Muratoglu O.K., Oral E. A fully functional drug-eluting joint implant // Nature Biomedical Engineering. 2017. V. 1. N. 6. P. 0080. https://doi.org/10.1038/s41551-017-0080

References

1. Kamaly N., Yameen B., Wu J., Farokhzad O.C. Degradable controlled-release polymers and polymeric nanoparticles: mechanisms of controlling drug release. *Chemical Reviews*, 2016, vol. 116, no. 4, pp. 2602–2663. https://doi.org/10.1021/acs.chemrev.5b00346
2. Barkalina N., Charalambous C., Jones C., Coward K. Nanotechnology in reproductive medicine: emerging applications of nanomaterials. *Nanomedicine: Nanotechnology, Biology and Medicine*, 2014, vol. 10, no. 5, pp. e921–e938. https://doi.org/10.1016/j.nano.2014.01.001
3. Furth M.E., Atala A., Van Dyke M.E. Smart biomaterials design for tissue engineering and regenerative medicine. *Biomaterials*, 2007, vol. 28, no. 34, pp. 5068–5073. https://doi.org/10.1016/j.biomaterials.2007.07.042
4. Gandhi A., Paul A., Sen S.O., Kumar K.S. Studies on thermoresponsive polymers: Phase behaviour, drug delivery and biomedical applications. *Asian Journal of Pharmaceutical Sciences*, 2015, vol. 10, no. 2, pp. 99–107. https://doi.org/10.1016/j.ajps.2014.08.010
5. Suhardi V.J., Bichara D.A., Kwok S.J.J., Freiberg A.A., Rubash H., Malchau H., Yun S.H., Muratoglu O.K., Oral E. A fully functional

6. Masood F. Polymeric nanoparticles for targeted drug delivery system for cancer therapy // Materials Science and Engineering: C. 2016. V. 60. P. 569–578. https://doi.org/10.1016/j.msec.2015.11.067
7. Ulbrich K., Holá K., Šubr V., Bakandritsos A., Tuček J., Zbořil R. Targeted drug delivery with polymers and magnetic nanoparticles: covalent and noncovalent approaches, release control, and clinical studies // Chemical Reviews. 2016. V. 116. N 9. P. 5338–5431. https://doi.org/10.1021/acs.chemrev.5b00589
8. Soppimath K.S., Aminabhavi T.M., Kulkarni A.R., Rudzinski W.E. Biodegradable polymeric nanoparticles as drug delivery devices // Journal of Controlled Release. 2001. V. 70. N 1-2. P. 1–20. https://doi.org/10.1016/S0168-3659(00)00339-4
9. Mokhtarzadeh A., Alibakhshi A., Hejazi M., Omidi Y., Ezzati Nazhad Dolatabadi J. Bacterial-derived biopolymers: Advanced natural nanomaterials for drug delivery and tissue engineering // TrAC Trends in Analytical Chemistry. 2016. V. 82. P. 367–384. https://doi.org/10.1016/j.trac.2016.06.013
10. Luo D., Gould D.J., Sukhorukov G.B. Local and sustained activity of doxycycline delivered with layer-by-layer microcapsules // Biomacromolecules. 2016. V. 17. N 4. P. 1466–1476. https://doi.org/10.1021/acs.biomac.6b00070
11. Wang C., Zhang G., Liu G., Hu J., Liu S. Photo-and thermo-responsive multicompartment hydrogels for synergistic delivery of gemcitabine and doxorubicin // Journal of Controlled Release. 2017. V. 259. P. 149–159. https://doi.org/10.1016/j.jconrel.2016.11.007
12. Sukhorukov G.B., Donath E., Lichtenfeld H., Knippel E., Knippel M., Budde A., Möhwald H. Layer-by-layer self assembly of polyelectrolytes on colloidal particles // Colloids and Surfaces A: Physicochemical and Engineering Aspects. 1998. V. 137. N 1-3. P. 253–266. https://doi.org/10.1016/S0927-7757(98)00213-1
13. Dhoot N.O., Wheatley M.A. Microencapsulated liposomes in controlled drug delivery: strategies to modulate drug release and eliminate the burst effect // Journal of Pharmaceutical Sciences. 2003. V. 92. N 3. P. 679–689. https://doi.org/10.1002/jps.19104
14. Gai M., Frueh J., Kudryavtseva V.L., Mao R., Kiryukhin M.V., Sukhorukov G.B. Patterned microstructure fabrication: polyelectrolyte complexes vs polyelectrolyte multilayers // Scientific Reports. 2016. V. 6. P. 37000. https://doi.org/10.1038/srep37000
15. Gai M., Frueh J., Tao T., Petrov A.V., Shesterikov E.V., Tverdokhlebov S.I., Sukhorukov G.B. Polylactic acid nano- and microchamber arrays for encapsulation of small hydrophilic molecules featuring drug release via high intensity focused ultrasound // Nanoscale. 2017. V. 9. N 21. P. 7063–7070. https://doi.org/10.1039/C7NR01841J
16. Беликов А.В., Клочков И.С., Алексеев И.В., Капралов С.А. Лазерный многопараметрический метод входного контроля монтажных элементов, используемых в объеме герметичных неодимовых лазерных излучателей // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 2. С. 154–162. https://doi.org/10.17586/2226-1494-2021-21-2-154-162
17. Schneider C.W., Lippert T. Laser ablation and thin film deposition // Springer Series in Materials Science. 2010. V. 139. P. 89–112. https://doi.org/10.1007/978-3-642-13281-0_5
18. Krebs H.-U., Weisheit M., Faupel J., Süske E., Scharf T., Fuhse C., Störmer M., Sturm K., Seibt M., Kijewski H., Nelke D., Panchenko E., Buback M. Pulsed laser deposition (PLD) — a versatile thin film technique // Advances in Solid State Physics. 2003. V. 43. P. 505–518. https://doi.org/10.1007/978-3-540-44838-9_36
19. Dler A. Thin film deposition processes // International Journal of Modern Physics and Applications. 2015. V. 1. N. 4. P. 193–199.
20. Wasa K., Kitabatake M., Adechi H. Thin film materials technology: Sputtering of compound materials. Walliam Andrew Inc, 2004. P. 1.
21. Фоминский В.Ю., Романов Р.И., Неволин В.Н., Гнедовец А.Г. Особенности формирования тонкопленочных слоев металлооксида в газо-сенсорной структуре Pt/WOx/SiC методом импульсного лазерного осаждения // Перспективные материалы. 2011. № 2. С. 5–16.
22. Кулинич С.А., Yamaki T., Miyazoe H., Yamamoto H., Terashima K. Влияние температуры подложки и скорости осаждения на начальный рост тонких пленок ниобата-танталата лития, полученных методом термальной плазмы // Физика твердого тела. 2006. Т. 48. № 5. С. 850–857.
23. Цаплин С.В., Болычев С.А. Исследование динамики осаждения конденсированных продуктов // Вестник Самарского государственного аэрокосмического университета. 2014. Т. 14. № 3.

drug-eluting joint implant. *Nature Biomedical Engineering*, 2017, vol. 1, no. 6, pp. 0080. https://doi.org/10.1038/s41551-017-0080
6. Masood F. Polymeric nanoparticles for targeted drug delivery system for cancer therapy. *Materials Science and Engineering: C*, 2016, vol. 60, pp. 569–578. https://doi.org/10.1016/j.msec.2015.11.067
7. Ulbrich K., Holá K., Šubr V., Bakandritsos A., Tuček J., Zbořil R. Targeted drug delivery with polymers and magnetic nanoparticles: covalent and noncovalent approaches, release control, and clinical studies. *Chemical Reviews*, 2016, vol. 116, no. 9, pp. 5338–5431. https://doi.org/10.1021/acs.chemrev.5b00589
8. Soppimath K.S., Aminabhavi T.M., Kulkarni A.R., Rudzinski W.E. Biodegradable polymeric nanoparticles as drug delivery devices. *Journal of Controlled Release*, 2001, vol. 70, no. 1-2, pp. 1–20. https://doi.org/10.1016/S0168-3659(00)00339-4
9. Mokhtarzadeh A., Alibakhshi A., Hejazi M., Omidi Y., Ezzati Nazhad Dolatabadi J. Bacterial-derived biopolymers: Advanced natural nanomaterials for drug delivery and tissue engineering. *TrAC Trends in Analytical Chemistry*, 2016, vol. 82, pp. 367–384. https://doi.org/10.1016/j.trac.2016.06.013
10. Luo D., Gould D.J., Sukhorukov G.B. Local and sustained activity of doxycycline delivered with layer-by-layer microcapsules. *Biomacromolecules*, 2016, vol. 17, no. 4, pp. 1466–1476. https://doi.org/10.1021/acs.biomac.6b00070
11. Wang C., Zhang G., Liu G., Hu J., Liu S. Photo-and thermo-responsive multicompartment hydrogels for synergistic delivery of gemcitabine and doxorubicin. *Journal of Controlled Release*, 2017, vol. 259, pp. 149–159. https://doi.org/10.1016/j.jconrel.2016.11.007
12. Sukhorukov G.B., Donath E., Lichtenfeld H., Knippel E., Knippel M., Budde A., Möhwald H. Layer-by-layer self assembly of polyelectrolytes on colloidal particles. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 1998, vol. 137, no. 1-3, pp. 253–266. https://doi.org/10.1016/S0927-7757(98)00213-1
13. Dhoot N.O., Wheatley M.A. Microencapsulated liposomes in controlled drug delivery: strategies to modulate drug release and eliminate the burst effect. *Journal of Pharmaceutical Sciences*, 2003, vol. 92, no. 3, pp. 679–689. https://doi.org/10.1002/jps.19104
14. Gai M., Frueh J., Kudryavtseva V.L., Mao R., Kiryukhin M.V., Sukhorukov G.B. Patterned microstructure fabrication: polyelectrolyte complexes vs polyelectrolyte multilayers. *Scientific Reports*, 2016, vol. 6, pp. 37000. https://doi.org/10.1038/srep37000
15. Gai M., Frueh J., Tao T., Petrov A.V., Shesterikov E.V., Tverdokhlebov S.I., Sukhorukov G.B. Polylactic acid nano- and microchamber arrays for encapsulation of small hydrophilic molecules featuring drug release via high intensity focused ultrasound. *Nanoscale*, 2017, vol. 9, no. 21, pp. 7063–7070. https://doi.org/10.1039/C7NR01841J
16. Belikov A.V., Klochkov I.S., Alekseev I.V., Kapralov S.A. Laser multiparameter method for incoming inspection of the mounting elements used in the volume of sealed neodymium laser emitters. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 2, pp. 154–162. (in Russian). https://doi.org/10.17586/2226-1494-2021-21-2-154-162
17. Schneider C.W., Lippert T. Laser ablation and thin film deposition. *Springer Series in Materials Science*, 2010, vol. 139, pp. 89–112. https://doi.org/10.1007/978-3-642-13281-0_5
18. Krebs H.-U., Weisheit M., Faupel J., Süske E., Scharf T., Fuhse C., Störmer M., Sturm K., Seibt M., Kijewski H., Nelke D., Panchenko E., Buback M. Pulsed laser deposition (PLD) — a versatile thin film technique. *Advances in Solid State Physics*, 2003, vol. 43, pp. 505–518. https://doi.org/10.1007/978-3-540-44838-9_36
19. Dler A. Thin film deposition processes. *International Journal of Modern Physics and Applications*, 2015, vol. 1, no. 4, pp. 193–199.
20. Wasa K., Kitabatake M., Adechi H. *Thin Film Materials Technology: Sputtering of Compound Materials*. Walliam Andrew Inc, 2004, pp. 1.
21. Fominski V.Yu., Romanov R.I., Nevolin V.N., Gnedovets A.G. Features of metal oxide thin film formation in Pt/WOx/SiC gas sensor device by pulsed laser deposition. *Perspektivnye materialy*, 2011, no. 2, pp. 5–16. (in Russian)
22. Kulinich S.A., Yamaki T., Miyazoe H., Yamamoto H., Terashima K. Effect of the substrate temperature and deposition rate on the initial growth of thin lithium niobate-tantalate films deposited from a thermal plasma. *Physics of the Solid State*, 2006, vol. 48, no. 5, pp. 904–911. https://doi.org/10.1134/S1063783406050155
23. Tsaplin S.V., Bolychev S.A. Research of the dynamics of condensed product deposition. *Vestnik of Samara University Aerospace and*

С. 109–121. https://doi.org/10.18287/1998-6629-2014-0-3(45)-109-121
24. Пупань Л.И. Изучение структуры материалов методом электронной микроскопии: методические указания к выполнению лабораторной работы по курсу «Введение в нанотехнологии» / НТУ «Харьковский политехнический институт». Харьков, 2011. 35 с.

Mechanical Engineering, 2014, vol. 14, no. 3, pp. 109–121. (in Russian). https://doi.org/10.18287/1998-6629-2014-0-3(45)-109-121
24. Pupan L.I. *The study of material structure by electron microscopy: guidelines for laboratory research on the course "Introduction to nanotechnologies"*. Kharkiv, Kharkiv Polytechnic Institute, 2011, 35 p. (in Russian)

Авторы

Беликов Андрей Вячеславович — доктор физико-математических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 7005600018, https://orcid.org/0000-0001-8733-4860, volnagdg@yandex.ru

Клочков Иван Сергеевич — научный сотрудник, АО «Научно-производственная корпорация «Системы прецизионного приборостроения», Санкт-Петербург, 190020, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57224086250, https://orcid.org/0000-0001-5461-783X, Vany-nov13@yandex.ru

Алексеев Иван Владимирович — ведущий инженер, АО «Научно-производственная корпорация «Системы прецизионного приборостроения», Санкт-Петербург, 190020, Российская Федерация, https://orcid.org/0000-0002-4368-9309, El.magio@yandex.ru

Капралов Сергей Анатольевич — кандидат технических наук, начальник отдела, АО «Научно-производственная корпорация «Системы прецизионного приборостроения», Великий Новгород, 173003, Российская Федерация, https://orcid.org/0000-0002-3393-4948, kapralov_sa@npkspp.ru

Authors

Andrey V. Belikov — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 7005600018, https://orcid.org/0000-0001-8733-4860, volnagdg@yandex.ru

Ivan S. Klochkov — Scientific Researcher, Joint Stock Company "Research and Production Corporation "Precision Systems and Instruments", Saint Petersburg, 190020, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57224086250, https://orcid.org/0000-0001-5461-783X, Vany-nov13@yandex.ru

Ivan V. Alekseev — Lead Engineer, Joint Stock Company "Research and Production Corporation "Precision Systems and Instruments", Saint Petersburg, 190020, Russian Federation, https://orcid.org/0000-0002-4368-9309, El.magio@yandex.ru

Sergey A. Kapralov — PhD, Head of Section, Joint Stock Company "Research and Production Corporation "Precision Systems and Instruments", Veliky Novgorod, 173003, Russian Federation, https://orcid.org/0000-0002-3393-4948, kapralov_sa@npkspp.ru

Статья поступила в редакцию 05.11.2021
Одобрена после рецензирования 19.12.2021
Принята к печати 27.01.2022

Received 05.11.2021
Approved after reviewing 19.12.2021
Accepted 27.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2022 Том 22 Номер 1

doi: 10.17586/2226-1494-2022-22-1-10-17
УДК 681.7.03; 667.6; 54.056

Оптические композиты на основе органических полимеров и полупроводниковых пигментов

Валерий Михайлович Волынкин[1], Сергей Константинович Евстропьев[2]✉, Дмитрий Владимирович Булыга[3], Артем Владимирович Морковский[4], Станислав Сергеевич Пашин[5], Константин Владимирович Дукельский[6], Антон Владимирович Бурдин[7], Игорь Борисович Бондаренко[8]

[1,2,6,7] АО «Научно-производственное объединение Государственный оптический институт им. С.И. Вавилова», Санкт-Петербург, 192171, Российская Федерация
[3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[4,5] Поволжский государственный университет телекоммуникаций и информатики, Самара, 443010, Российская Федерация
[8] Санкт-Петербургский государственный университет телекоммуникаций им. профессора М.А. Бонч-Бруевича, Санкт-Петербург, 193232, Российская Федерация

[1] Vvolynkin@yandex.ru, https://orcid.org/0000-0002-6325-1507
[2] evstropiev@bk.ru✉, https://orcid.org/0000-0002-0160-8443
[3] dmbulyga@yandex.ru, https://orcid.org/0000-0001-8861-2527
[4] artuom.igrun@gmail.com, https://orcid.org/0000-0003-3461-3350
[5] pashinstanislav@outlook.com, https://orcid.org/0000-0002-7898-974X
[6] kdukel@mail.ru, https://orcid.org/0000-0002-1627-7499
[7] bourdine-av@psuti.ru, https://orcid.org/0000-0001-8737-5486
[8] igorlitmo@rambler.ru, https://orcid.org/0000-0001-9857-7200

Аннотация

Предмет исследования. Представлены результаты разработки оптических органо-неорганических композиционных материалов, которые обладают высоким поглощением света в видимой части спектра и высоким отражением в ближней инфракрасной области спектра. Такие оптические материалы используются в качестве покрытий в промышленности и строительстве. **Метод.** Для создания оптических композитов использованы эпоксидные и эпоксидно-полиуретановые полимерные матрицы, содержащие неорганические полупроводниковые частицы (CuS, PbS, Fe_3O_4). При получении однородных композиционных материалов применены высокодисперсные порошки неорганических пигментов. Синтез дисперсных порошков сульфидов меди (CuS) и свинца (PbS) осуществлен методом химического осаждения из растворов в присутствии органического стабилизатора. Исследования кристаллической структуры и морфологии порошков неорганических пигментов выполнены методами оптической микроскопии и рентгенофазового анализа. Микротвердость по Виккерсу полученных композитов определена с помощью прибора ПМТ-3. **Основные результаты.** На основании данных рентгенофазового анализа выполнен расчет среднего размера кристаллитов по формуле Шеррера. Установлено, что свежеосажденные порошки CuS и PbS состоят из нанокристаллов, имеющих размер 11–20 нм. Данные оптической микроскопии свидетельствуют о формировании в порошках агрегатов полупроводниковых нанокристаллов. Эксперименты продемонстрировали, что все синтезированные композиты обладают низким коэффициентом отражения (менее 0,06) в видимой части спектра и повышенным коэффициентом отражения в ближней инфракрасной области спектра (0,13–0,15 и более). Результаты исследования показали, что применение эпоксидно-полиуретановых полимерных матриц обеспечивает большую микротвердость композиционных материалов, по сравнению с композитами на основе эпоксидных полимеров. Наибольшие значения микротвердости наблюдались в композиционных материалах на основе эпоксидно-полиуретановых полимеров, содержащих высокодисперсные частицы Fe_3O_4. **Практическая значимость.** Полученные в работе органо-неорганические композиты могут быть использованы в промышленности и строительстве в качестве материала черных светопоглощающих покрытий, которые характеризуются повышенным коэффициентом отражения света в ближней инфракрасной области спектра и высокой микротвердостью.

© Волынкин В.М., Евстропьев С.К., Булыга Д.В., Морковский А.В., Пашин С.С., Дукельский К.В., Бурдин А.В., Бондаренко И.Б., 2022

Ключевые слова
эпоксидный полимер, полиуретан, PbS, CuS, магнетит

Ссылка для цитирования: Волынкин В.М., Евстропьев С.К., Булыга Д.В., Морковский А.В., Пашин С.С., Дукельский К.В., Бурдин А.В., Бондаренко И.Б. Оптические композиты на основе органических полимеров и полупроводниковых пигментов // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 10–17. doi: 10.17586/2226-1494-2022-22-1-10-17

Optical composites based on organic polymers and semiconductor pigments

Valery M. Volynkin[1], Sergey K. Evstropiev[2]✉, Dmitry V. Bulyga[3], Artyom V. Morkovsky[4], Stanislav S. Pashin[5], Konstantin V. Dukelsky[6], Anton V. Burdin[7], Igor B. Bondarenko[8]

[1,2,6,7] JSC S.I. Vavilov State Optical Institute, Saint Petersburg, 192171, Russian Federation
[3] ITMO University, Saint Petersburg, 197101, Russian Federation
[4,5] Povolzhskiy State University of Telecommunications & Informatics, Samara, 443010, Russian Federation
[8] The Bonch-Bruevich Saint Petersburg State University of Telecommunications, Saint Petersburg, 193232, Russian Federation

[1] Vvolynkin@yandex.ru, https://orcid.org/0000-0002-6325-1507
[2] evstropiev@bk.ru✉, https://orcid.org/0000-0002-0160-8443
[3] dmbulyga@yandex.ru, https://orcid.org/0000-0001-8861-2527
[4] artuom.igrun@gmail.com, https://orcid.org/0000-0003-3461-3350
[5] pashinstanislav@outlook.com, https://orcid.org/0000-0002-7898-974X
[6] kdukel@mail.ru, https://orcid.org/0000-0002-1627-7499
[7] bourdine-av@psuti.ru, https://orcid.org/0000-0001-8737-5486
[8] igorlitmo@rambler.ru, https://orcid.org/0000-0001-9857-7200

Abstract
The aim of the work was the development of optical organic-inorganic composite materials with high absorption of light the visible part of the spectrum and high reflection in the near infrared region of the spectrum. Such materials are used in industry and construction as coatings. To create these optical composites, epoxy and epoxy-polyurethane polymer matrices containing inorganic semiconducting particles (CuS, PbS, Fe_3O_4) were used. Highly dispersive powders of inorganic pigments were used for the preparation of homogeneous composite materials. The wet precipitation method with the application of organic stabilizing additions was applied for the preparation of dispersive CuS and PbS powders. Optical microscopy and X-ray diffraction analysis helped to study the crystal structure and morphology of the obtained semiconductor pigments. PMT-3 device was applied for microhardness measurements of the prepared composite materials. Based on the data of X-ray diffraction analysis, the average crystallite size was calculated using the Scherrer formula. It was found that freshly precipitated CuS and PbS powders consist of nanocrystals with a size of 11–20 nm. Optical microscopy data indicate the formation of aggregates of semiconductor nanocrystals in powders. Experiments have shown that all synthesized composites have low light reflection coefficient (less than 0.06) in the visible part of the spectrum and an increased light reflection coefficient in the near infrared region of the spectrum (0.13–0.15 and more). The results of the study showed that the use of epoxy-polyurethane polymer matrices provides greater microhardness of composite materials, compared to the composites based on epoxy polymers. The highest microhardness values were observed in composite materials based on epoxy-polyurethane polymers containing highly dispersed Fe_3O_4 particles. Obtained organic-inorganic composites could be used as materials for light-absorbing coatings in different industrial applications.

Keywords
epoxy polymer, polyurethane, PbS, CuS, magnetite

For citation: Volynkin V.M., Evstropiev S.K., Bulyga D.V., Morkovsky A.V., Pashin S.S., Dukelsky K.V., Burdin A.V., Bondarenko I.B. Optical composites based on organic polymers and semiconductor pigments. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 10–17 (in Russian). doi: 10.17586/2226-1494-2022-22-1-10-17

Введение

Композиционные материалы черного цвета широко применяются в промышленности и строительстве. Такие материалы сильно поглощают свет всего видимого спектрального диапазона [1–3]. Композиты часто содержат пигменты черного цвета различного химического состава: технический углерод [1–3], хромит меди [2], смешанный хромит железа и никеля [2], различные смешанные оксиды [1, 4, 5], сульфид свинца (PbS) [6–8] и др.

Для некоторых применений композитов низкое отражение света в видимой части спектра должно сочетаться с относительно высоким отражением в инфракрасной (ИК) области спектра [2]. В работах [6–10] рассмотрены светопоглощающие материалы, обладающие такими спектральными свойствами. Эти композиционные материалы имеют сложную структуру и их практическое применение во многих случаях нецелесообразно.

Для получения красящих композиций, обладающих повышенным отражением света в ближней ИК области спектра в работах [7, 8] в качестве неорганического компонента использован порошок свежеосажденного PbS. Этот метод основан на химической реакции между ионами Pb^{2+} и сульфид-анионами в растворе,

сопровождающейся образованием малорастворимого PbS.

Метод осаждения из растворов — технологически простой и распространенный метод получения неорганических пигментов [5]. Методика осаждения из растворов, содержащих органические стабилизирующие добавки, широко используется для получения различных наночастиц, в том числе PbS [11–19]. Для ограничения роста нанокристаллов PbS и их стабилизации применяется широкий круг органических стабилизаторов: олеиновая кислота [18], поливинилпирролидон [13, 16], тиофенол [11], этилендиаминтетрауксусная кислота [19] и др. [12, 14, 15, 17].

Для многих практических приложений существенным является высокая механическая прочность и твердость композитов. Воздействие различных внешних факторов (ветер, пыль), атмосферных осадков может привести к микродеформациям, повреждению и разрушению композиционных материалов. В связи с этим проблема сочетания требуемых спектральных свойств материалов с их высокой механической прочностью и твердостью — актуальна.

Использование PbS, являющегося токсичным материалом, во многих практических приложениях ограничено, и задача создания светопоглощающего органо-неорганического композита, содержащего менее токсичные неорганические пигменты, является актуальной. Цель настоящей работы — разработка и исследование свойств композиционных материалов на основе эпоксидных полимеров или их модифицированных аналогов, содержащих различные полупроводниковые пигменты, как уже использовавшийся ранее PbS, так и более экологически безопасные сульфид меди (CuS) и магнетит (Fe_3O_4).

Материалы и методы

В качестве полимерной основы светопоглощающих органо-неорганических композитов выбрана эпоксидная полимерная матрица. Эпоксидные полимерные композиции широко применяются в оптике в качестве оптических клеев [20, 21] и различных нелинейно-оптических композиционных материалов [22, 23].

В качестве неорганических модифицирующих компонентов в настоящей работе использованы полупроводниковые соединения, имеющие небольшую ширину запрещенной зоны и, соответственно, сильно поглощающие свет видимой части спектра. В состав полимерной матрицы авторами были введены PbS и CuS, а также Fe_3O_4.

Для формирования однородного материала использованы высокодисперсные неорганические модификаторы. С помощью высокодисперсного коммерческого микропорошка Fe_3O_4 был получен композит, содержащий оксидный модификатор. Для создания высокодисперсных порошков сульфидов металлов применен метод осаждения из водных растворов в присутствии органических стабилизаторов, ограничивающих рост осаждаемых частиц сульфидов. Синтез высокодисперсного PbS данным методом описан в работах [11–13, 16].

Для синтеза частиц сульфидов металлов использованы водные растворы нитрата свинца ($Pb(NO_3)_2$), сульфата меди ($CuSO_4$) и сульфида натрия (Na_2S), в качестве стабилизаторов — поливинилпирролидон (M_w = 1 300 000; Sigma Aldrich). В водные растворы $Pb(NO_3)_2$ и $CuSO_4$ был добавлен раствор поливинилпирролидона, а затем медленно, при интенсивном перемешивании, в эту смесь был добавлен раствор Na_2S. Полученный осадок отделялся фильтрованием и после тщательной промывки дистиллированной водой подвергался сушке при 70 °C. В табл. 1 приведены химические составы растворов, использовавшихся для получения порошков неорганических модификаторов.

Высушенные неорганические модификаторы небольшими порциями при перемешивании вводились в жидкую полимерную композицию. Основу полимерной композиции составила эпоксидная смола YD-128 (Kukdo Chemical Co., Ltd; Республика Корея). В качестве отвердителя эпоксидной смолы использован джеффамин Д-230 (JEFFAMINE® D-230, Huntsman Corp., Belgium).

Выбор данной марки отвердителя обусловлен необходимостью формирования композита с повышенной связностью полимерного каркаса и большей твердостью. Предварительные эксперименты показали, что применение более высокомолекулярных аналогов выбранного отвердителя (джеффамин Д-400 или джеффамин Д-2000) снижает характеристики композита. В некоторые композиции в качестве отвердителя, играющего также роль пластификатора, вводился нетоксичный отвердитель аминного типа (торговая марка «ЭТАЛ-45»).

Полимеризация композиционных материалов выполнена при комнатной температуре в течение 24 ч. Образцы синтезированных материалов имели форму пластин прямоугольной формы толщиной 20–30 мм.

Для оценки возможности повышения механических свойств материалов в состав некоторых композиций были введены добавки полиуретана. Химический состав полученных композитов приведен в табл. 2.

Таблица 1. Химические составы растворов, использованных для синтеза порошков различных неорганических модификаторов

Table 1. Chemical compositions of the solutions used for the synthesis of the powders of different inorganic modificators

Номер образца	Исходные компоненты, масс.%					
	$Pb(NO_3)_2$	Fe_3O_4	$Cu(SO_4)_2 \cdot 5H_2O$	$Na_2S \cdot 9H_2O$	H_2O	Поливинилпирролидон
1	4,6	—	—	1,8	92,2	1,4
2	—	—	4,5	2,1	92,1	1,3
3	—	4,7	—	—	93,9	1,4

Таблица 2. Химические составы полимерных и композиционных материалов

Table 2. Chemical compositions of polymeric and composite materials

Номер образца	Химический состав, масс.%						
	PbS	Fe_3O_4	CuS	Эпоксидная смола 128	Джеффамин Д-230	Отвердитель аминного типа ЭТАЛ-45М	Полиуретан
4	—	—	—	75	25	—	—
5	—	—	—	72	21	—	7
6	16	—	—	63	21	—	—
7	—	16	—	63	21	—	—
8	11	—	—	62	21	—	6
9	—	13	—	61	20	—	6
10	40	—	—	40	—	20	—
11	—	—	20	54	—	26	—
12	7,5	7,5	—	57	—	28	—
13	30	—	—	47	—	23	—

Выполнено исследование методом рентгенофазового анализа на приборе Rigaku Ultima IV структуры порошков PbS и CuS, полученных авторами настоящей работы. На основании полученных данных осуществлен расчет среднего размера кристаллитов по формуле Шеррера:

$$d = k\lambda/(\beta\cos\theta), \qquad (1)$$

где d — размер кристаллита; k = 0,9; λ – длина волны рентгеновского излучения для линии CuKα (λ = 0,15406 нм); β — FWHM (ширина пика на дифрактограмме на половине его максимума). Для расчетов использованы пики максимальной интенсивности.

Микроидентирование поверхности твердых материалов – широко распространенный метод оценки прочности поверхностных слоев материалов [24, 25]. Измерение микротвердости в настоящей работе проведено на приборе ПМТ-3 по стандартной методике, описанной в работе [25]. При измерениях продолжительность приложения нагрузки составляла 10 с, а измерение размеров отпечатка индентора осуществлено сразу же после снятия нагрузки. Микротвердость рассчитана по формуле:

$$H_V = 1{,}854 \cdot P/d^2,$$

где P — нагрузка на индентор; d — длина диагонали остаточного отпечатка индентора.

Измерение спектров отражения материалов произведено на спектрофотометре Perkin Elmer 900 с приставкой для измерения отражения света.

Экспериментальные результаты

Рассмотрим экспериментальные результаты синтезированных авторами настоящей работы порошков PbS, CuS и Fe_3O_4, использованных при получении композиционных материалов в образцах № 6, 11 и 7 соответственно.

На рис. 1, *a* приведена рентгенограмма синтезированного порошка PbS, на которой видны пики, характерные для кубических кристаллов PbS. Наиболее интенсивный пик связан с отражением от плоскости (200) кристаллов PbS. Рассчитанные параметры кристаллической ячейки синтезированных кристаллов PbS следующие: a = 0,59414(18) нм, a — период решетки (расстояние между атомами по ребру элементарной кубической решетки); V = 0,210 нм^3, где V — объем элементарной кубической ячейки. Средний размер кристаллов составил 11 нм, что свидетельствует о высокой эффективности применения поливинилпирролидона для ограничения роста формирующихся наночастиц PbS. При использовании в качестве стабилизатора трилона-Б в работе [23] синтезированы более крупные нанокристаллы PbS, имеющие средний размер 64 нм.

На рис. 1, *b* приведена рентгенограмма синтезированного порошка CuS, полученного при применении в качестве стабилизатора поливинилпирролидона. Заметим, что видны пики, характерные для гексагональных кристаллов CuS (JCPDS № 06-0464).

Рентгенограмма порошка Fe_3O_4, использованного при получении композиций образцов № 7, 9 и 12, показана на рис. 1, *c*, где хорошо видны пики, характерные для Fe_3O_4. Параметры элементарной ячейки кристаллов составили a = 0,59461(46) нм, c = 1,6779(32) нм (c — параметр элементарной ячейки кристалла), V = 0,593(1) нм^3. Расчеты по формуле Шеррера (1) показали, что средний размер кристаллитов в порошке составляет 20 нм.

Данные рентгенофазового анализа свидетельствуют о том, что как синтезированные порошки PbS и CuS, так и коммерческий порошок Fe_3O_4, состоят из нанокристаллов, размер которых не превышает 30 нм. Малый размер частиц неорганических модификаторов необходим для получения однородных по структуре композитов.

Все синтезированные материалы имели черный цвет и обеспечивали интенсивное поглощение света в видимой части спектра.

Приведем спектры отражения образцов композитов в ближней ИК области спектра (рис. 2). Видно, что с увеличением длины волны для всех образцов

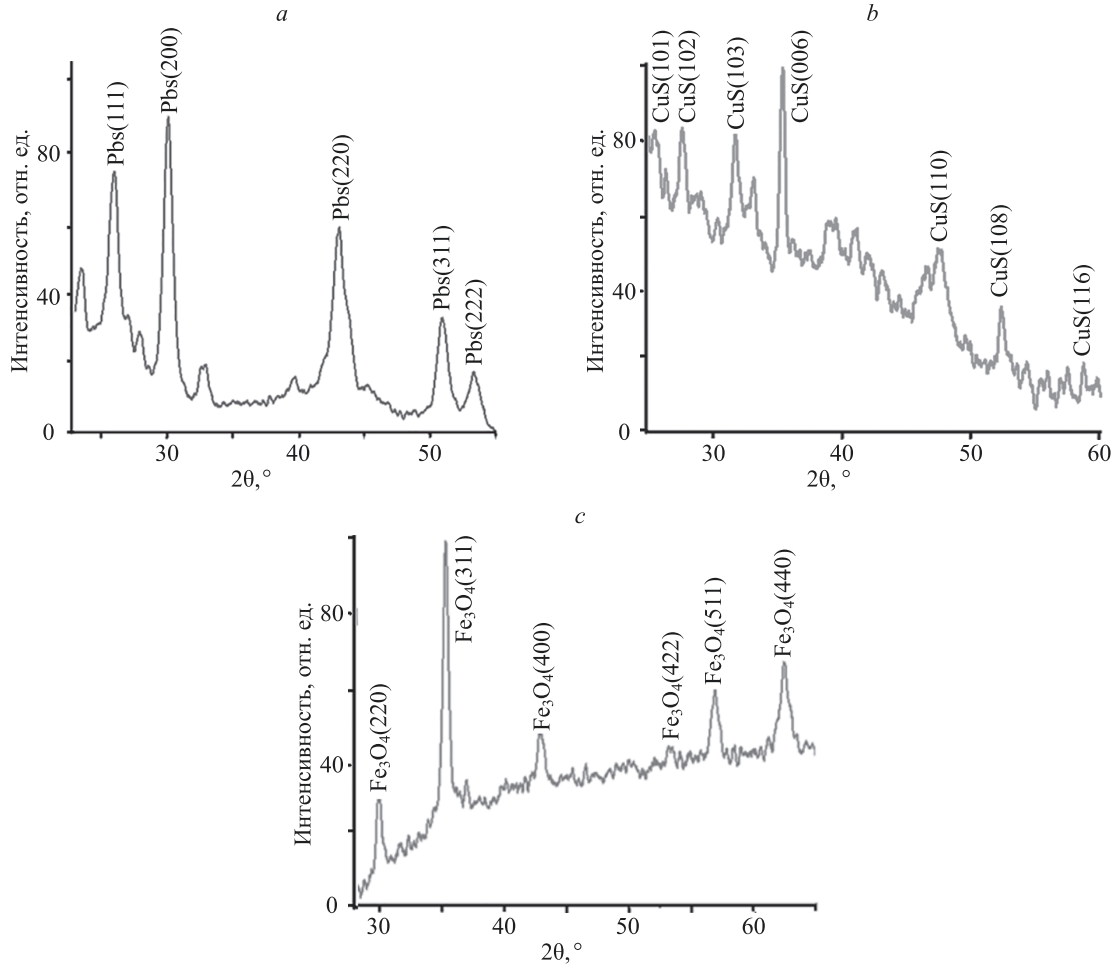


Рис. 1. Рентгенограммы синтезированных порошков: PbS (*a*), CuS (*b*) и Fe_3O_4 (*c*), использованных при получении композиций образцов № 6, 11 и 7 соответственно

Fig. 1. XRD patterns of prepared powders PbS (*a*), CuS (*b*) and Fe_3O_4 (*c*) used for the fabrication of the composites 6 (*a*), 11 (*b*) and 7 (*c*)

происходит увеличение отражения света в ближней ИК области спектра. Наиболее заметное увеличение в области 1000–2000 нм наблюдается для образца № 11 (CuS/эпоксидная смола), но его отличие от других образцов невелико. Термообработка материалов несколько увеличивает отражение света от их поверхности. В целом вид полученных спектров аналогичен виду спектров отражения многослойных PbS-содержащих покрытий, описанных в [8].

Результаты измерения микротвердости полученных покрытий приведены в табл. 3. Эксперименты показали, что с увеличением содержания неорганических пигментов микротвердость покрытий возрастает. Из данных табл. 3 видно, что введение полиуретана в состав полимерной матрицы также привело к некоторому увеличению их микротвердости.

Среди использованных неорганических модификаторов наиболее твердым является Fe_3O_4 и его добавки в составе композита привели к сильному увеличению микротвердости материалов. При введении PbS в эпоксидную полимерную матрицу микротвердость покрытия возрастает на 14 %, а при использовании смешанных добавок PbS и Fe_3O_4 увеличение микротвердости составило более 70 %. CuS является довольно мягким материалом, поэтому его добавление в состав композита не привело к возрастанию микротвердости.

Таблица 3. Микротвердость полимерных и композиционных материалов

Table 3. Microhardness of polymeric and composite materials

Номер образца	Микротвердость, МПа
4	110
5	116
6	125
7	132
8	125
9	190
11	100
12	187
13 (после термообработки)	194

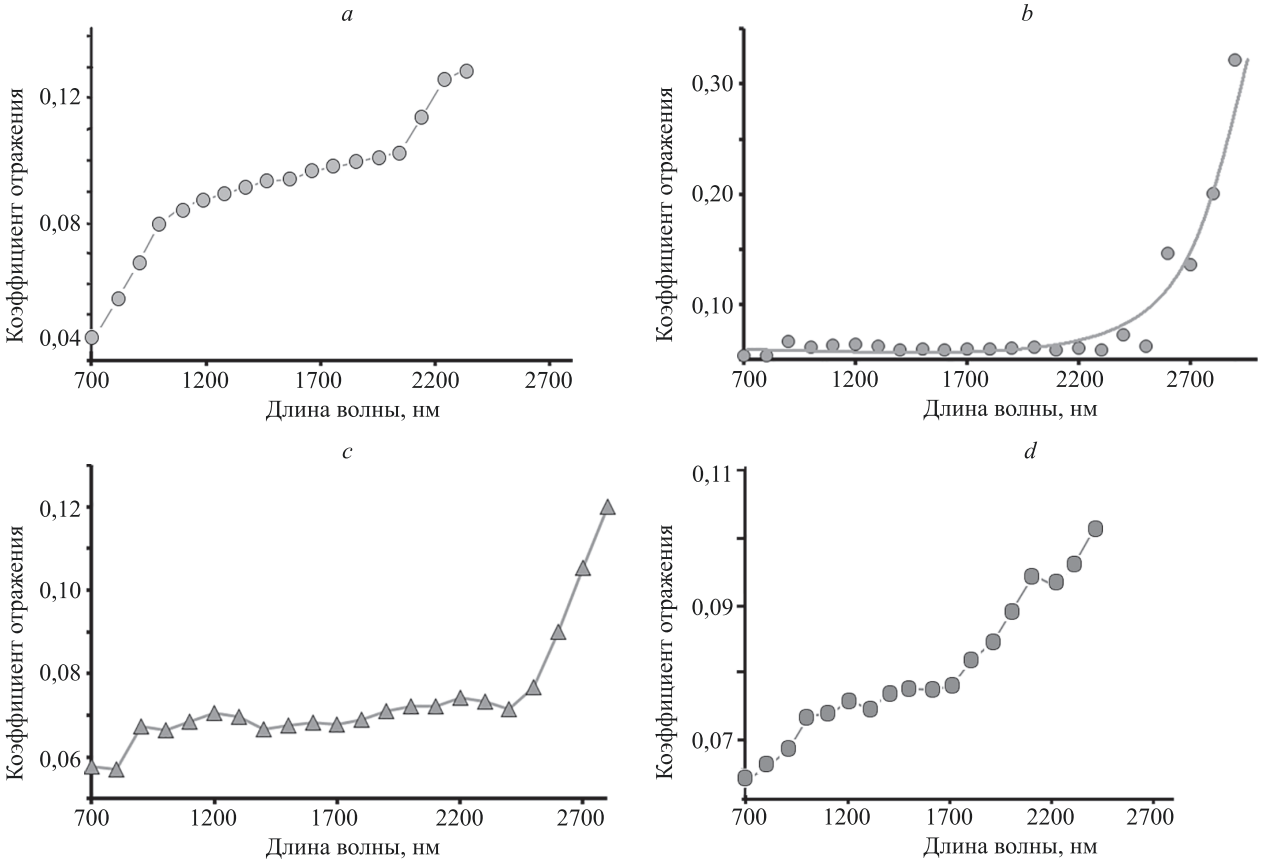


Рис. 2. Спектры отражения композиционных материалов в ближней инфракрасной области спектра для образцов: № 11 (*a*); № 12 (*b*); № 13 (*c*) и № 13 после дополнительной термообработки (120 °С) в течение двух часов (*d*)

Fig. 2. Reflection spectra of composite materials 11 (*a*); 12 (*b*); 13 (*c*); 13 after additional thermal treatment (120 °C) for two hours (*d*)

Заключение

Введение полупроводниковых неорганических модификаторов (PbS, CuS, Fe_3O_4) в состав эпоксидных и эпоксидно-полиуретановых полимерных матриц позволяет формировать композиты, обладающие высоким поглощением света в видимой части спектра и повышенным (до 0,15) коэффициентом отражения света в ближней инфракрасной области спектра. Добавление полиуретана в состав эпоксидных полимерных матриц увеличивает микротвердость композитов, не уменьшая светоотражения в ближней инфракрасной области спектра. Композиты, содержащие сульфид меди, характеризуются высоким светоотражением в ближней инфракрасной области спектра, но небольшой микротвердостью.

Литература

1. Takeda H., Yabuki K. Visible light absorbing film, structural member having this visible light absorbing film and visible light absorbing ink which forms visible light absorbing film. Patent US 7927696 B2. 2011.
2. Qin J., Qu J., Song J.R., Song Z.N., Zhang W.D., Shi Y.X., Zhang T., Xue X., Zhang R.P., Zhang H.Q., Zhang Z.Y., Wu X. The optical properties of black coatings and their estimated cooling effect and cooling energy savings potential // Journal of Power and Energy Engineering. 2014. V. 2. P. 68–75. http://dx.doi.org/10.4236/jpee.2014.24011
3. Гершбейн А.М. Способ получения изображения и печатная краска для осуществления этого способа. Патент RU 2244631 C2. Бюл. 2005. № 2.
4. Sanada K. Infrared reflecting black pigment, coating material and resin composition. Patent EP 1847512 A1. 2006.
5. Мартинкевич А.А., Прокопчук Н.Р. Пигменты для современных лакокрасочных материалов: учеб.-метод. пособие. Минск: БГТУ, 2014. 130 с.
6. Fang V., Kenedy J., Futter J., Manning J. A review of infrared reflectance properties of metal oxide nanostructures: GNS Science Report 2013/39. 23 p.

References

1. Takeda H., Yabuki K. Visible light absorbing film, structural member having this visible light absorbing film and visible light absorbing ink which forms visible light absorbing film. *Patent US7927696B2*, 2011.
2. Qin J., Qu J., Song J.R., Song Z.N., Zhang W.D., Shi Y.X., Zhang T., Xue X., Zhang R.P., Zhang H.Q., Zhang Z.Y., Wu X. The optical properties of black coatings and their estimated cooling effect and cooling energy savings potential. *Journal of Power and Energy Engineering*, 2014, vol. 2, pp. 68–75. http://dx.doi.org/10.4236/jpee.2014.24011
3. Gershbejn A.M. Method for production of image and printing ink for realization of this method (modifications). *Patent RU 2244631 C2*, 2005. (in Russian)
4. Sanada K. Infrared reflecting black pigment, coating material and resin composition. *Patent EP 1847512 A1*, 2006.
5. Martinkevich A.A., Prokopchuk N.R. *Pigments for Modern Paints and Varnishes*. Minsk, BSTU Publ., 2014, 130 p. (in Russian)
6. Fang V., Kenedy J., Futter J., Manning J. A review of infrared reflectance properties of metal oxide nanostructures. *GNS Science Report 2013/39*, 23 p.

7. Волынкин В.М., Киселев В.М., Евстропьев С.К., Бурчинов А.Н., Матвеенцев А.В. Светопоглощающее покрытие. Патент RU 2626838 C2. Бюл. 2017. № 22.
8. Киселев В.М., Бурчинов А.А., Волынкин В.М., Евстропьев С.К., Матвеенцев А.В. Светопоглощающие композиционные покрытия на основе оксидных полых микросфер и сульфида свинца // Оптический журнал. 2015. Т. 82. № 12. С. 60–64.
9. Graighead H.G., Howard R.E. Light-absorbing materials. Patent US 4284689 A. 1981.
10. Самсонов К.Н. Светопоглощающее покрытие. Полезная модель RU 126149 U1. Бюл. 2013. № 8.
11. Nabiyouni G., Moghimi E., Hedayati K., Jalajerdi R. Room temperature synthesis of lead sulfide nanoparticles // Main Group Metal Chemistry. 2012. V. 35. N 5-6. P. 173–178. https://doi.org/10.1515/mgmc-2012-0036
12. Chongad L.S., Sharma A., Banerjee M., Jain A. Synthesis of lead sulfide nanoparticles by chemical precipitation method // Journal of Physics: Conference Series. 2016. V. 755. P. 012032. https://doi.org/10.1088/1742-6596/755/1/012032
13. Evstropiev S.K., Kislyakov I.M., Bagrov I.V., Belousova I.M. Stabilization of PbS quantum dots by high molecular polyvinylpyrrolidone // Polymers for Advanced Technology. 2016. V. 27. N 3. P. 314–317. https://doi.org/10.1002/pat.3642
14. Lu S., Sohling U., Krajewski T., Mennig M., Schmidt H.K. Synthesis and characterization of PbS nanoparticles in ethanolic solution stabilized by hydroxypropyl cellulose // Journal of Materials Science Letters. 1998. V. 17. N 24. P. 2071–2073. https://doi.org/10.1023/A:1006671814597
15. Huang Z., Zhai G., Zhang Z., Zhang C., Xia Y., Lian L., Fu X., Zhang D., Zhang J. Low cost and large scale synthesis of PbS quantum dots with hybrid surface passivation // CrystEngComm. 2017. V. 19. N 6. P. 946–951. https://doi.org/10.1039/C6CE02471H
16. Багров И.В., Данилов В.В., Евстропьев С.К., Киселев В.М., Кисляков И.М., Панфутова А.С., Хребтов А.И. Фотоиндуцированное изменение люминесцентных свойств суспензий наночастиц PbS, стабилизированных поливинилпирролидоном // Письма в Журнал технической физики. 2015. Т. 41. № 2. С. 25–33.
17. Jiao Y., Gao X., Lu J., Chen Y., Zhou J., Li X. A novel method for PbS quantum dot synthesis // Materials Letters. 2012. V. 72. P. 116–118. https://doi.org/10.1016/j.matlet.2011.12.068
18. Fu H., Tsang S.-W., Zhang Y., Ouyang J., Lu J., Yu K., Tao Y. Impact of the growth conditions of colloidal PbS nanocrystals on photovoltaic device performance // Chemistry of Materials. 2011. V. 23. N 7. P. 1805–1810. https://doi.org/10.1021/cm200051j
19. Shyju T.S., Anandhi S., Sivakumar R., Gopalakrishnan R. Studies on lead sulfide (PbS) semiconducting thin films deposited from nanoparticles and its NLO application // International Journal of Nanoscience. 2014. V. 13. N 1. P. 1450001. https://doi.org/10.1142/S0219581X1450001X
20. Чурсова Л.В., Панина Н.Н., Гребенева Т.А., Кутергина И.Ю. Эпоксидные смолы, отвердители, модификаторы и связующие на их основе. СПб.: ЦОП «Профессия», 2020. 576 с.
21. Sang J.J., Kohli D.K., Shah K.G. Structural adhesive and bonding application thereof. Patent USA 8974905 B2. 2015.
22. Kulagina A.S., Sandulenko A.V., Volynkin V.M., Evstropiev S.K. Synthesis and nonlinear optical properties of vanadium-doped plasticized epoxy polymer composites // Advanced Composites and Hybrid Materials. 2021. V. 4. N 2. P. 324–331. https://doi.org/10.1007/s42114-021-00227-y
23. Belousova I.M., Videnichev D.A., Volynkin V.M., Evstropiev S.K., Kislyakov I.M., Murav'ova T.D., Rakov E.G. Nonlinear optical limiters of pulsed laser radiation based on carbon-containing nanostructures in viscous and solid matrices // Polymers for Advanced Technologies. 2014. V. 25. N 9. P. 1008–1013. https://doi.org/10.1002/pat.3343
24. Нацик В.Д., Фоменко Л.С., Лубенец С.В. Исследование ползучести и стеклования эластомеров методом микроиндентирования: эпоксидная смола и нанокомпозиты на ее основе // Физика твердого тела. 2013. Т. 55. № 5. С. 940–952.
25. Low I.M., Shi C. Vickers indentation responses of epoxy polymers // Journal of Materials Science Letters. 1998. V. 17. N 14. P. 1181–1183. https://doi.org/10.1023/A:1006517005082

7. Volynkin V.M., Kiselev V.M., Evstropev S.K., Burchinov A.N., Matveentsev A.V. Light-absorbing coating. *Patent RU 2626838 C2*, 2017. (in Russian)
8. Kiselev V.M., Burchinov A.A., Volynkin V.M., Evstropev S.K., Matveentsev A.V., Composite light-absorbent coatings based on hollow oxide microspheres and lead sulfide. *Journal of Optical Technology*, 2015, vol. 82, no. 12, pp. 824–826. https://doi.org/10.1364/JOT.82.000824
9. Graighead H.G., Howard R.E. Light-absorbing materials. *Patent US 4284689 A*, 1981.
10. Samsonov K.N. Light-absorbing coating. *Patent RU 126149 U1*, 2013. (in Russian)
11. Nabiyouni G., Moghimi E., Hedayati K., Jalajerdi R. Room temperature synthesis of lead sulfide nanoparticles. *Main Group Metal Chemistry*, 2012, vol. 35, no. 5-6, pp. 173–178. https://doi.org/10.1515/mgmc-2012-0036
12. Chongad L.S., Sharma A., Banerjee M., Jain A. Synthesis of lead sulfide nanoparticles by chemical precipitation method. *Journal of Physics: Conference Series*, 2016, vol. 755, pp. 012032. https://doi.org/10.1088/1742-6596/755/1/012032
13. Evstropiev S.K., Kislyakov I.M., Bagrov I.V., Belousova I.M. Stabilization of PbS quantum dots by high molecular polyvinylpyrrolidone. *Polymers for Advanced Technology*, 2016, vol. 27, no. 3, pp. 314–317. https://doi.org/10.1002/pat.3642
14. Lu S., Sohling U., Krajewski T., Mennig M., Schmidt H.K. Synthesis and characterization of PbS nanoparticles in ethanolic solution stabilized by hydroxypropyl cellulose. *Journal of Materials Science Letters*, 1998, vol. 17, no. 24, pp. 2071–2073. https://doi.org/10.1023/A:1006671814597
15. Huang Z., Zhai G., Zhang Z., Zhang C., Xia Y., Lian L., Fu X., Zhang D., Zhang J. Low cost and large scale synthesis of PbS quantum dots with hybrid surface passivation. *CrystEngComm*, 2017, vol. 19, no. 6, pp. 946–951. https://doi.org/10.1039/C6CE02471H
16. Bagrov I.A., Danilov V.V., Evstrop'ev S.K., Kiselev V.M., Kislyakov I.M., Panfutova A.S., Khrebtov A.I. Photoinduced variation of the luminescent properties of PbS nanoparticle suspensions stabilized by polyvinylpyrrolidone. *Technical Physics Letters*, 2015, vol. 41, no. 1, pp. 65–68. https://doi.org/10.1134/S1063785015010198
17. Jiao Y., Gao X., Lu J., Chen Y., Zhou J., Li X. A novel method for PbS quantum dot synthesis. *Materials Letters*, 2012, vol. 72, pp. 116–118. https://doi.org/10.1016/j.matlet.2011.12.068
18. Fu H., Tsang S.-W., Zhang Y., Ouyang J., Lu J., Yu K., Tao Y. Impact of the growth conditions of colloidal PbS nanocrystals on photovoltaic device performance. *Chemistry of Materials*, 2011, vol. 23, no. 7, pp. 1805–1810. https://doi.org/10.1021/cm200051j
19. Shyju T.S., Anandhi S., Sivakumar R., Gopalakrishnan R. Studies on lead sulfide (PbS) semiconducting thin films deposited from nanoparticles and its NLO application. *International Journal of Nanoscience*, 2014, vol. 13, no. 1, pp. 1450001. https://doi.org/10.1142/S0219581X1450001X
20. Chursova L.V., Panina N.N., Grebeneva T.A., Kutergina I.Iu. *Epoxy Resins, Hardeners, Densifiers and Binders Based on Them*. St. Petersburg, Professija Publ., 2020, 576 p. (in Russian)
21. Sang J.J., Kohli D.K., Shah K.G. Structural adhesive and bonding application thereof. *Patent USA 8974905 B2*, 2015.
22. Kulagina A.S., Sandulenko A.V., Volynkin V.M., Evstropiev S.K. Synthesis and nonlinear optical properties of vanadium-doped plasticized epoxy polymer composites. *Advanced Composites and Hybrid Materials*, 2021, vol. 4, no. 2, pp. 324–331. https://doi.org/10.1007/s42114-021-00227-y
23. Belousova I.M., Videnichev D.A., Volynkin V.M., Evstropiev S.K., Kislyakov I.M., Murav'ova T.D., Rakov E.G. Nonlinear optical limiters of pulsed laser radiation based on carbon-containing nanostructures in viscous and solid matrices. *Polymers for Advanced Technologies*, 2014, vol. 25, no. 9, pp. 1008–1013. https://doi.org/10.1002/pat.3343
24. Natsik V.D., Fomenko L.S., Lubenets S.V. Investigation of the creep and glass transition of elastomers by the microindentation method: epoxy resin and related nanocomposites. *Physics of the Solid State*, 2013, vol. 55, no. 5, pp. 1020–1033. https://doi.org/10.1134/S1063783413050260
25. Low I.M., Shi C. Vickers indentation responses of epoxy polymers. *Journal of Materials Science Letters*, 1998, vol. 17, no. 14, pp. 1181–1183. https://doi.org/10.1023/A:1006517005082

Авторы

Волынкин Валерий Михайлович — кандидат химических наук, ведущий научный сотрудник, АО «Научно-производственное объединение Государственный оптический институт им. С.И. Вавилова», Санкт-Петербург, 192171, Российская Федерация, SC 6601999426, https://orcid.org/0000-0002-6325-1507, Vvolynkin@yandex.ru

Евстропьев Сергей Константинович — доктор химических наук, доцент, начальник отдела, АО «Научно-производственное объединение Государственный оптический институт им. С.И. Вавилова», Санкт-Петербург, 192171, Российская Федерация, SC 6507317768, https://orcid.org/0000-0002-0160-8443, evstropiev@bk.ru

Булыга Дмитрий Владимирович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57217158694, https://orcid.org/0000-0001-8861-2527, dmbulyga@yandex.ru

Морковский Артем Владимирович — студент, Поволжский государственный университет телекоммуникаций и информатики, Самара, 443010, Российская Федерация, SC 57217158694, https://orcid.org/0000-0003-3461-3350, artuom.igrun@gmail.com

Пашин Станислав Сергеевич — заместитель директора, Поволжский государственный университет телекоммуникаций и информатики, Самара, 443010, Российская Федерация, SC 57190843183, https://orcid.org/0000-0002-7898-974X, pashinstanislav@outlook.com

Дукельский Константин Владимирович — кандидат технических наук, доцент, генеральный директор, АО «Научно-производственное объединение Государственный оптический институт им. С.И. Вавилова», Санкт-Петербург, 192171, Российская Федерация, SC 6602633236, https://orcid.org/0000-0002-1627-7499, kdukel@mail.ru

Бурдин Антон Владимирович — доктор технических наук, доцент, заместитель генерального директора, АО «Научно-производственное объединение Государственный оптический институт им. С.И. Вавилова», Санкт-Петербург, 192171, Российская Федерация, SC 6701518987, https://orcid.org/0000-0001-8737-5486, bourdine-av@psuti.ru

Бондаренко Игорь Борисович — кандидат технических наук, доцент, доцент, Санкт-Петербургский государственный университет телекоммуникаций им. профессора М.А. Бонч-Бруевича, Санкт-Петербург, 193232, Российская Федерация, SC 57191340641, https://orcid.org/0000-0001-9857-7200, igorlitmo@rambler.ru

Authors

Valery M. Volynkin — PhD, Leading Researcher, JSC S.I. Vavilov State Optical Institute, Saint Petersburg, 192171, Russian Federation, SC 6601999426, https://orcid.org/0000-0002-6325-1507, Vvolynkin@yandex.ru

Sergey K. Evstropiev — D.Sc., Associate Professor, Head of Department, JSC S.I. Vavilov State Optical Institute, Saint Petersburg, 192171, Russian Federation, SC 6507317768, https://orcid.org/0000-0002-0160-8443, evstropiev@bk.ru

Dmitry V. Bulyga — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57217158694, https://orcid.org/0000-0001-8861-2527, dmbulyga@yandex.ru

Artyom V. Morkovsky — Student, Povolzhskiy State University of Telecommunications & Informatics, Samara, 443010, Russian Federation, SC 57217158694, https://orcid.org/0000-0003-3461-3350, artuom.igrun@gmail.com

Stanislav S. Pashin — Deputy Director, Povolzhskiy State University of Telecommunications & Informatics, Samara, 443010, Russian Federation, SC 57190843183, https://orcid.org/0000-0002-7898-974X, pashinstanislav@outlook.com

Konstantin V. Dukelsky — PhD, Associate Professor, CEO, JSC S.I. Vavilov State Optical Institute, Saint Petersburg, 192171, Russian Federation, SC 6602633236, https://orcid.org/0000-0002-1627-7499, kdukel@mail.ru

Anton V. Burdin — D.Sc., Associate Professor, Deputy CEO, JSC S.I. Vavilov State Optical Institute, Saint Petersburg, 192171, Russian Federation, SC 6701518987, https://orcid.org/0000-0001-8737-5486, bourdine-av@psuti.ru

Igor B. Bondarenko — PhD, Associate Professor, Associate Professor, The Bonch-Bruevich Saint Petersburg State University of Telecommunications, Saint Petersburg, 193232, Russian Federation, SC 57191340641, https://orcid.org/0000-0001-9857-7200, igorlitmo@rambler.ru

Статья поступила в редакцию 01.12.2021
Одобрена после рецензирования 13.01.2021
Принята к печати 29.01.2022

Received 01.12.2021
Approved after reviewing 13.01.2021
Accepted 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-18-24
УДК 681.51

Новый алгоритм идентификации частоты синусоидального сигнала с постоянными параметрами

Хак Тунг Нгуен[1], Сергей Михайлович Власов[2], Антон Александрович Пыркин[3], Иван Владимирович Попков[4]✉

[1,2,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] nguyenkhactunghvhq1994@gmail.com, https://orcid.org/0000-0001-6430-1927
[2] smvlasov@itmo.ru, https://orcid.org/0000-0002-8345-7553
[3] a.pyrkin@gmail.com, https://orcid.org/0000-0001-8806-4057
[4] vanta.jesus.2015@gmail.com✉, https://orcid.org/0000-0003-4565-5652

Аннотация

Предмет исследования. Рассмотрено решение проблемы идентификации частоты синусоидального сигнала с постоянными паратметрами. Данная задача актуальна при компенсации возмущений, управлении динамическими объектами и в ряде других задач. Предложен метод улучшения качества оценивания частоты синусоидального сигнала и обеспечения экспоненциальной сходимости к нулю ошибок оценивания. **Метод.** На первом этапе синусоидальный сигнал представлен как выходной сигнал линейного генератора конечной размерности. Параметры сигнала (амплитуда, фаза и частота) неизвестны. На втором этапе применены Жорданова форма матрицы и оператор запаздывания для параметризации синусоидального сигнала. После ряда специальных преобразований получено простейшее уравнение, содержащее произведение одного зависящего от частоты неизвестного параметра и известной функции времени. Для поиска неизвестного параметра использованы методы градиентного спуска и наименьших квадратов. **Основные результаты.** Представлен новый алгоритм параметризации синусоидального сигнала. Решение задачи основано на преобразовании модели сигнала к линейному регрессионному уравнению. Задача решена с использованием методов градиентного спуска и наименьших квадратов настройки на базе линейного регрессионного уравнения, полученного путем параметризации синусоидального сигнала. Выполнен анализ возможностей предложенного метода оценивания с использованием компьютерного моделирования в среде MATLAB (Simulink). Результаты моделирования подтвердили сходимость ошибок оценивания частоты к истинным значениям. **Практическая значимость.** Разработанный метод может быть эффективно применен для широкого класса прикладных задач, связанных с компенсацией или подавлением возмущений, описываемых синусоидальным или мультисинусоидальным сигналами, в частности в задачах управления надводным судном с компенсацией синусоидальных возмущений.

Ключевые слова

синусоидальные сигналы, идентификация, Жорданова форма матрицы, линейная регрессионная модель

Ссылка для цитирования: Нгуен Х.Т., Власов С.М., Пыркин А.А., Попков И.В. Новый алгоритм идентификации частоты синусоидального сигнала с постоянными параметрами // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 18–24. doi: 10.17586/2226-1494-2022-22-1-18-24

A new algorithm for the identification of sinusoidal signal frequency with constant parameters

Khac Tung Nguyen[1], Sergey M. Vlasov[2], Anton A. Pyrkin[3], Ivan V. Popkov[4]✉

[1,2,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] nguyenkhactunghvhq1994@gmail.com, https://orcid.org/0000-0001-6430-1927
[2] smvlasov@itmo.ru, https://orcid.org/0000-0002-8345-7553
[3] a.pyrkin@gmail.com, https://orcid.org/0000-0001-8806-4057
[4] vanta.jesus.2015@gmail.com✉, https://orcid.org/0000-0003-4565-5652

© Нгуен Х.Т., Власов С.М., Пыркин А.А., Попков И.В., 2022

Abstract

The paper presents a solution for identifying the frequency of a sinusoidal signal with constant parameters. The issue can be relevant for compensation of disturbances, control of dynamic objects, and other tasks. The authors propose a method to improve the quality of the estimation of the sinusoidal signal frequency and to ensure exponential convergence to zero of the estimation errors. At the first stage, the sinusoidal signal is presented as an output signal of a linear generator of finite dimension. The signal parameters (amplitude, phase, and frequency) are unknown. At the second stage, the Jordan form of the matrix and the delay operator are applied to parameterize the sinusoidal signal. After a series of special transformations, the simplest equation is obtained containing product of one frequency-dependent unknown parameter and a known function of time. To find the unknown parameter, the authors used the methods of gradient descent and least squares. A new algorithm for the parametrization of a sinusoidal signal is presented. The solution is based on transforming the signal model to a linear regression equation. The problem is solved using gradient descent and least squares tuning methods based on a linear regression equation obtained by parametrizing a sinusoidal signal. The results involve the analysis of the capabilities of the proposed estimation method using computer modeling in the Matlab environment (Simulink). The results confirmed the convergence of the frequency estimation errors to the true values. The developed method can be effectively applied to a wide class of tasks related to compensating or suppressing disturbances described by sinusoidal or multisinusoidal signals, for example, to control a surface vessel with compensation of sinusoidal disturbances.

Keywords

sinusoidal signal, identification, Jordan form of the matrix, linear regression model

For citation: Nguyen Kh.T., Vlasov S.M., Pyrkin A.A., Popkov I.V. A new algorithm for the identification of sinusoidal signal frequency with constant parameters. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 18–24 (in Russian). doi: 10.17586/2226-1494-2022-22-1-18-24

Введение

В работе рассмотрена задача идентификации синусоидального сигнала с неизвестными стационарными амплитудой и смещением. Проблема идентификации частоты синусоидальных сигналов с постоянными параметрами изучена в работах [1–7]. Задача идентификации параметров синусоидальных сигналов является очень важной и актуальной. Ее решение может найти применение при компенсации возмущений, в управлении динамическими объектами и др. [8–12].

В работе [13] предложен глобально сходящийся алгоритм оценки частоты на основе адаптивного фильтра для гармонического сигнала. Используя реализацию синусоидального сигнала в пространстве состояний, задача оценки частоты может быть преобразована в комбинированную задачу оценки состояния и параметров. В [4, 14] разработана глобальная оценка частоты на основе адаптивных наблюдателей. Принимая во внимание белый шум и изменяющиеся во времени частоты в работе [15] разработан модифицированный фильтр Калмана. Основной недостаток перечисленных подходов – невозможность определения параметров за конечное время.

В [16] представлен альтернативный метод частотно-временного анализа сигналов, который сравнивается с традиционными методами дискретного преобразования Фурье. Предлагаемый алгоритм имеет нелинейную структуру, обеспечивающую частотную адаптивность. Алгоритм может быть реализован как в аналоговой, так и в цифровой форме и особенно подходит для приложений реального времени, где важна вычислительная эффективность. В работе [17] предложен адаптивный фильтр, основанный на обобщенном интеграторе третьего порядка, для оценки всех параметров смещенной синусоиды. Теория усреднения использована для доказательства, что фильтр определяет неизвестную частоту сигнала в случае только смещенной синусоиды на входе. Основной недостаток данных подходов — локальная сходимость ошибки оценивания параметров к нулю, что сказывается на качестве управления при применении оценок в контуре обратной связи.

В работе [18] для оценки частоты синусоидального сигнала с шумом, где разрешены мгновенные изменения частоты, использована адаптированная модифицированная ньютоновская оптимизация. В [19] предложено решение проблемы обеспечения глобально сходящейся оценки параметров синусоидального сигнала. В частности, рассмотрен новый адаптивный режекторный фильтр, динамические уравнения которого демонстрируют исключительные особенности, при которых все сигналы глобально ограничены, и оцененная частота является асимптотически правильной для всех начальных условий и значений частоты. Общий недостаток упомянутых подходов к оцениванию параметров синусоидального сигнала состоит в отсутствии оценки частоты гармоник напрямую.

В настоящей работе рассмотрено решение проблемы идентификации частот синусоидального сигнала и предложен новый подход к параметризации. Модель линейной регрессии первого порядка простроена с использованием операторов задержки. Исходным условием является то, что частота, амплитуда и фаза синусоидального сигнала — постоянные, параметры неизвестны. Предлагаемый алгоритм идентификации должен обеспечивать сходимость настроенной оценки частоты к истинному значению.

Постановка задачи

Рассмотрим измеряемый сигнал

$$y(t) = A\sin(\omega t + \varphi),$$

где $A \in \mathbb{R}$ — амплитуда, $\omega \in \mathbb{R}$ — частота и $\varphi \in \mathbb{R}$ — фаза. A, ω, φ — неизвестные постоянные параметры.

Требуется синтезировать алгоритм оценивания частоты $\hat{\omega}(t)$, обеспечивающей выполнение условия

$$\lim_{t\to\infty} (\omega(t) - \hat{\omega}(t)) = 0.$$

Введем допущение.

Допущение. Известны минимальная частота $\underline{\omega}$ и максимальная частота $\overline{\omega}$, причем $0 < \underline{\omega} \le \omega \le \overline{\omega}$.

Параметризация синусоидального сигнала

Рассмотрим задачу построения модели линейной регрессии с измеряемыми переменными и вектором постоянных параметров, зависящих от неизвестной частоты ω.

Сигнал $y(t)$ может быть представлен как выходы линейных генераторов [20]

$$y(t) = \mathbf{H}^T\boldsymbol{\xi}(t), \tag{1}$$

$$\dot{\boldsymbol{\xi}}(t) = \boldsymbol{\Gamma}\boldsymbol{\xi}(t), \tag{2}$$

где $\boldsymbol{\xi} \in \mathbb{R}^q$ — матрица состояния генератора с начальным значением $\boldsymbol{\xi}(0)$; $\boldsymbol{\Gamma} \in \mathbb{R}^{q\times q}$ — матрица постоянных коэффициентов; $\mathbf{H} \in \mathbb{R}^q$ — вектор соответствующей размерности.

Построим генератор сигнала $y(t)$

Выберем в качестве первой координаты вектора состояния генератора сам сигнал $\xi_1 = y$. Продифференцировав ξ_1, получим

$$\xi_1 = A\sin(\omega t + \varphi),$$
$$\dot{\xi}_1 = \dot{y}(t) = A\omega\cos(\omega t + \varphi).$$

Примем в качестве первой координаты вектора состояния генератора производную синусоидального сигнала $\xi_2 = \dot{y}$. Продифференцируем ξ_2, получим

$$\dot{\xi}_2 = \ddot{y}(t) = -A\omega^2\sin(\omega t + \varphi) = -\omega^2\xi_1.$$

Для векторно-матричной формы выражения (1)–(2) имеют вид:

$$\boldsymbol{\xi} = \begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}, \boldsymbol{\Gamma} = \begin{bmatrix}0 & 1\\ -\omega^2 & 0\end{bmatrix}, \mathbf{H} = \begin{bmatrix}1\\ 0\end{bmatrix}.$$

Преобразуем уравнение (2) к виду:

$$\dot{\boldsymbol{\xi}}(t) = \boldsymbol{\Gamma}\boldsymbol{\xi}(t) \Rightarrow \boldsymbol{\xi}(t) = e^{\boldsymbol{\Gamma}t}\boldsymbol{\xi}(0). \tag{3}$$

Подставляя уравнение (3) в уравнение (1), получим:

$$y(t) = \mathbf{H}^T e^{\boldsymbol{\Gamma}t}\boldsymbol{\xi}(0). \tag{4}$$

Наряду с измеряемым сигналом $y(t)$ рассмотрим запаздывающий сигнал

$$y(t-d) = \mathbf{H}^T e^{\boldsymbol{\Gamma}(t-d)}\boldsymbol{\xi}(0) = \mathbf{H}^T e^{\boldsymbol{\Gamma}t}e^{-\boldsymbol{\Gamma}d}\boldsymbol{\xi}(0), \tag{5}$$

где $d \in \mathbb{R}_+$ — постоянная величина запаздывания.

Замечание 1. Введем ограничения на выбранную величину запаздывания d из уравнения (5).

$$d < \frac{\pi}{\overline{\omega}}.$$

На основе уравнений (3) и (5), получим:

$$y(t-d) = \mathbf{H}^T e^{-\boldsymbol{\Gamma}d}\boldsymbol{\xi}(t). \tag{6}$$

Утверждение 1. *Сигнал* (4) *связан соотношением*:

$$y(t-d) = \mathbf{H}^T e^{-\boldsymbol{\Gamma}d}\boldsymbol{\xi}(t) = \mathbf{H}^T\begin{bmatrix}\cos\omega d & -\omega^{-1}\sin\omega d\\ \omega\sin\omega d & \cos\omega d\end{bmatrix}\boldsymbol{\xi}(t).$$

Доказательство утверждения 1.

Применим Жорданову форму матрицы для преобразователя $e^{-\boldsymbol{\Gamma}d}$ уравнения (6).

Сначала вычислим собственные значения матрицы $\boldsymbol{\Gamma}$

$$\det(\boldsymbol{\Gamma} - \lambda\mathbf{I}) = \begin{vmatrix}0-\lambda & 1\\ -\omega^2 & 0-\lambda\end{vmatrix} = 0, \Rightarrow \lambda^2 + \omega^2 = 0,$$
$$\Rightarrow \lambda_1 = i\omega, \lambda_2 = -i\omega.$$

Для каждого из собственных значений λ_1, λ_2 найдем собственные векторы.

Для числа $\lambda_1 = i\omega$, имеем

$$\begin{bmatrix}-i\omega & 1\\ -\omega^2 & -i\omega\end{bmatrix}\begin{bmatrix}x_1\\ y_1\end{bmatrix} = 0, \Rightarrow \begin{cases}-i\omega x_1 + y_1 = 0\\ -\omega^2 x_1 - i\omega y_1 = 0\end{cases} \Rightarrow y_1 = i\omega x_1.$$

Полагая $x_1 = 1$, получим собственный вектор $\mathbf{V_1} = (x_1, y_1)^T$

$$x_1 = 1, \Rightarrow y_1 = i\omega, \Rightarrow \mathbf{V_1} = \begin{bmatrix}1\\ i\omega\end{bmatrix}.$$

Аналогично найдем вектор $\mathbf{V_2} = (x_2, y_2)^T$, ассоциированный с собственным значением $\lambda_2 = -i\omega$, получим

$$\begin{bmatrix}-i\omega & 1\\ -\omega^2 & -i\omega\end{bmatrix}\begin{bmatrix}x_2\\ y_2\end{bmatrix} = 0, \Rightarrow \begin{cases}-i\omega x_2 + y_2 = 0\\ -\omega^2 x_2 + i\omega y_2 = 0\end{cases} \Rightarrow y_2 = -i\omega x_2.$$

Полагая $x_2 = 1$, получим собственный вектор $\mathbf{V_2} = (x_2, y_2)^T$

$$x_2 = 1, \Rightarrow y_2 = i\omega, \Rightarrow \mathbf{V_2} = \begin{bmatrix}1\\ -i\omega\end{bmatrix}.$$

Составим матрицу $\mathbf{V}$ из найденных собственных векторов $\mathbf{V_1}$ и $\mathbf{V_2}$ в виде:

$$\mathbf{V} = \begin{bmatrix}1 & 1\\ i\omega & -i\omega\end{bmatrix}.$$

Запишем Жорданову форму $\mathbf{J}_\Gamma$ для заданной матрицы $\boldsymbol{\Gamma}$, используя форму

$$\mathbf{J}_\Gamma = \mathbf{V}^{-1}\boldsymbol{\Gamma}\mathbf{V} = \frac{1}{-2i\omega}\begin{bmatrix}-i\omega & -1\\ -i\omega & 1\end{bmatrix}\begin{bmatrix}0 & 1\\ -\omega^2 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ i\omega & -i\omega\end{bmatrix} =$$
$$= \frac{1}{-2i\omega}\begin{bmatrix}\omega^2 & -i\omega\\ -\omega^2 & -i\omega\end{bmatrix}\begin{bmatrix}1 & 1\\ i\omega & -i\omega\end{bmatrix} = \frac{1}{-2i\omega}\begin{bmatrix}2\omega^2 & 0\\ 0 & -2\omega^2\end{bmatrix} =$$
$$= \begin{bmatrix}\dfrac{-\omega}{i} & 0\\ 0 & \dfrac{\omega}{i}\end{bmatrix} = \begin{bmatrix}i\omega & 0\\ 0 & -i\omega\end{bmatrix}.$$

Составим матрицу $e^{\mathbf{J}_\Gamma d}$ в виде [21]

$$e^{\mathbf{J}_\Gamma d} = \begin{bmatrix}e^{i\omega d} & 0\\ 0 & e^{-i\omega d}\end{bmatrix}.$$

Вычислим матричную экспоненту $e^{\Gamma d}$ по формуле:

$$e^{\Gamma d} = \mathbf{V}e^{\mathbf{J}_\Gamma d}\mathbf{V}^{-1} = \begin{bmatrix} 1 & 1 \\ i\omega & -i\omega \end{bmatrix}\begin{bmatrix} e^{i\omega d} & 0 \\ 0 & e^{-i\omega d} \end{bmatrix}\frac{1}{-2i\omega}\begin{bmatrix} -i\omega & -1 \\ -i\omega & 1 \end{bmatrix} =$$

$$= \frac{1}{-2i\omega}\begin{bmatrix} e^{i\omega d} & e^{-i\omega d} \\ i\omega e^{i\omega d} & -i\omega e^{-i\omega d} \end{bmatrix}\begin{bmatrix} -i\omega & -1 \\ -i\omega & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} \dfrac{e^{i\omega d} + e^{-i\omega d}}{2} & \omega^{-1}\dfrac{e^{i\omega d} - e^{-i\omega d}}{2i} \\ -\omega^{-1}\dfrac{e^{i\omega d} - e^{-i\omega d}}{2i} & \dfrac{e^{i\omega d} + e^{-i\omega d}}{2} \end{bmatrix}.$$

Разложим экспоненциальные функции $e^{i\omega d}$, $e^{-i\omega d}$ по формулам Эйлера:

$$\frac{e^{i\omega d} + e^{-i\omega d}}{2} = \cos\omega d, \frac{e^{i\omega d} - e^{-2i\omega d}}{2i} = \sin\omega d.$$

Получаем соотношения:

$$e^{\Gamma d} = \begin{bmatrix} \cos\omega d & \omega^{-1}\sin\omega d \\ -\omega\sin\omega d & \cos\omega d \end{bmatrix}.$$

$$e^{-\Gamma d} = \begin{bmatrix} \cos\omega d & -\omega^{-1}\sin\omega d \\ \omega\sin\omega d & \cos\omega d \end{bmatrix}. \quad (7)$$

Из уравнений (5) и (7) получим соотношение вида

$$y(t-d) = \mathbf{H}^T\begin{bmatrix} \cos\omega d & -\omega^{-1}\sin\omega d \\ \omega\sin\omega d & \cos\omega d \end{bmatrix}\xi(t).$$

$$y(t-d) = [\cos\omega d \quad -\omega^{-1}\sin\omega d]\xi(t), \quad (8)$$

что и требовалось доказать.

Аналогично для запаздывающего сигнала $2d$.

$$y(t-2d) = [\cos 2\omega d \quad -\omega^{-1}\sin 2\omega d]\xi(t). \quad (9)$$

Представим в матричном виде уравнения (8)–(9)

$$\begin{bmatrix} y(t-d) \\ y(t-2d) \end{bmatrix} = \begin{bmatrix} \cos\omega d & -\omega^{-1}\sin\omega d \\ \cos 2\omega d & -\omega^{-1}\sin 2\omega d \end{bmatrix}\xi(t).$$

Рассмотрим следующее выражение:

$$\mathbf{Y} = \mathbf{\Phi}\xi, \quad (10)$$

где $\mathbf{Y} = \begin{bmatrix} y(t-d) \\ y(t-2d) \end{bmatrix}$, $\mathbf{\Phi} = \begin{bmatrix} \cos\omega d & -\omega^{-1}\sin\omega d \\ \cos 2\omega d & -\omega^{-1}\sin 2\omega d \end{bmatrix}$.

Из выражения (10) получим

$$\mathbf{\Phi}^{-1}\mathbf{Y} = \xi = \frac{1}{\det(\mathbf{\Phi})}\mathrm{adj}(\mathbf{\Phi})\mathbf{Y} \Rightarrow \mathrm{adj}(\mathbf{\Phi})\mathbf{Y} = \det(\mathbf{\Phi})\xi. \quad (11)$$

Умножим уравнение (11) на $\mathbf{H}^T$, тогда

$$\mathbf{H}^T\mathrm{adj}(\mathbf{\Phi})\mathbf{Y} = \det(\mathbf{\Phi})y(t), \quad (12)$$

$$\mathrm{adj}\mathbf{\Phi} = \begin{bmatrix} -\omega^{-1}\sin 2\omega d & \omega^{-1}\sin\omega d \\ \cos 2\omega d & \cos\omega d \end{bmatrix}, \quad (13)$$

$$\det\mathbf{\Phi} = -\frac{1}{\omega}\cos\omega d\sin 2\omega d + \frac{1}{\omega}\cos 2\omega d\sin\omega d =$$

$$= -\frac{1}{\omega}\sin\omega d. \quad (14)$$

Из выражений (12)–(14) получим уравнение

$$[1 \quad 0]\begin{bmatrix} -\omega^{-1}\sin 2\omega d & \omega^{-1}\sin\omega d \\ -\cos 2\omega d & \cos\omega d \end{bmatrix}\begin{bmatrix} y(t-d) \\ y(t-2d) \end{bmatrix} =$$

$$= -\omega^{-1}\sin(\omega d)y(t). \quad (15)$$

Запишем уравнение (15) в виде

$$-\omega^{-1}2\sin(\omega d)\cos(\omega d)y(t-d) + \omega^{-1}\sin(\omega d)y(t-2d) =$$

$$= -\omega^{-1}\sin(\omega d)y(t). \quad (16)$$

Разделив обе части выражения (16) на $-\omega^{-1}\sin(\omega d)$, получим

$$2\cos(\omega d)y(t-d) - y(t-2d) = y(t),$$

$$\cos(\omega d)y(t-d) = \frac{1}{2}[y(t-2d) + y(t)]. \quad (17)$$

Построим регрессионную модель из уравнения (17) в виде

$$\chi(t) = \varphi(t)\theta, \quad (18)$$

где $\chi(t) = \frac{1}{2}[y(t-2d) + y(t)]$, $\varphi(t) = y(t-d)$, $\theta = \cos(\omega d)$.

Алгоритм оценивания параметров θ модели (18)

Метод градиентного спуска **[22].** Рассмотрим задачу оценивания параметра θ модели (18) с помощью стандартного метода градиентного спуска:

$$\dot{\hat{\theta}}(t) = \gamma\varphi(t)(\chi(t) - \varphi(t)\hat{\theta}(t)),$$

где γ — любое постоянное число.

Метод наименьших квадратов **[22].** Оценка определяется по формуле

$$\dot{\hat{\theta}}(t) = KP(t)\varphi(t)(\chi(t) - \varphi(t)\hat{\theta}(t)),$$

где K — положительный коэффициент, $P(t) > 0$ рассчитаем по формуле

$$\dot{P}(t) = K(\eta K - P(t)\varphi^2(t)P(t)),$$

где η — любое постоянное число.

Оценивание частоты

Для оценивания частоты воспользуемся функцией arccos(.) на основе параметра θ из (17)

$$\hat{\omega}(t) = 1/2\arccos(\hat{\theta}(t)).$$

Оценка $\hat{\theta}$ в модели (18) и принятое допущение должны иметь ограничение

$$\cos\overline{\omega}d \le \hat{\theta} \le \cos\underline{\omega}d.$$

Математическое моделирование

Рассмотрим результаты численного моделирования, иллюстрирующие эффективность предложенного алгоритма оценивания частоты несмещенного гармонического сигнала с постоянными параметрами. Моделирование выполнено с использованием программной среды MATLAB (Simulink).

В качестве примера рассмотрим сигнал вида

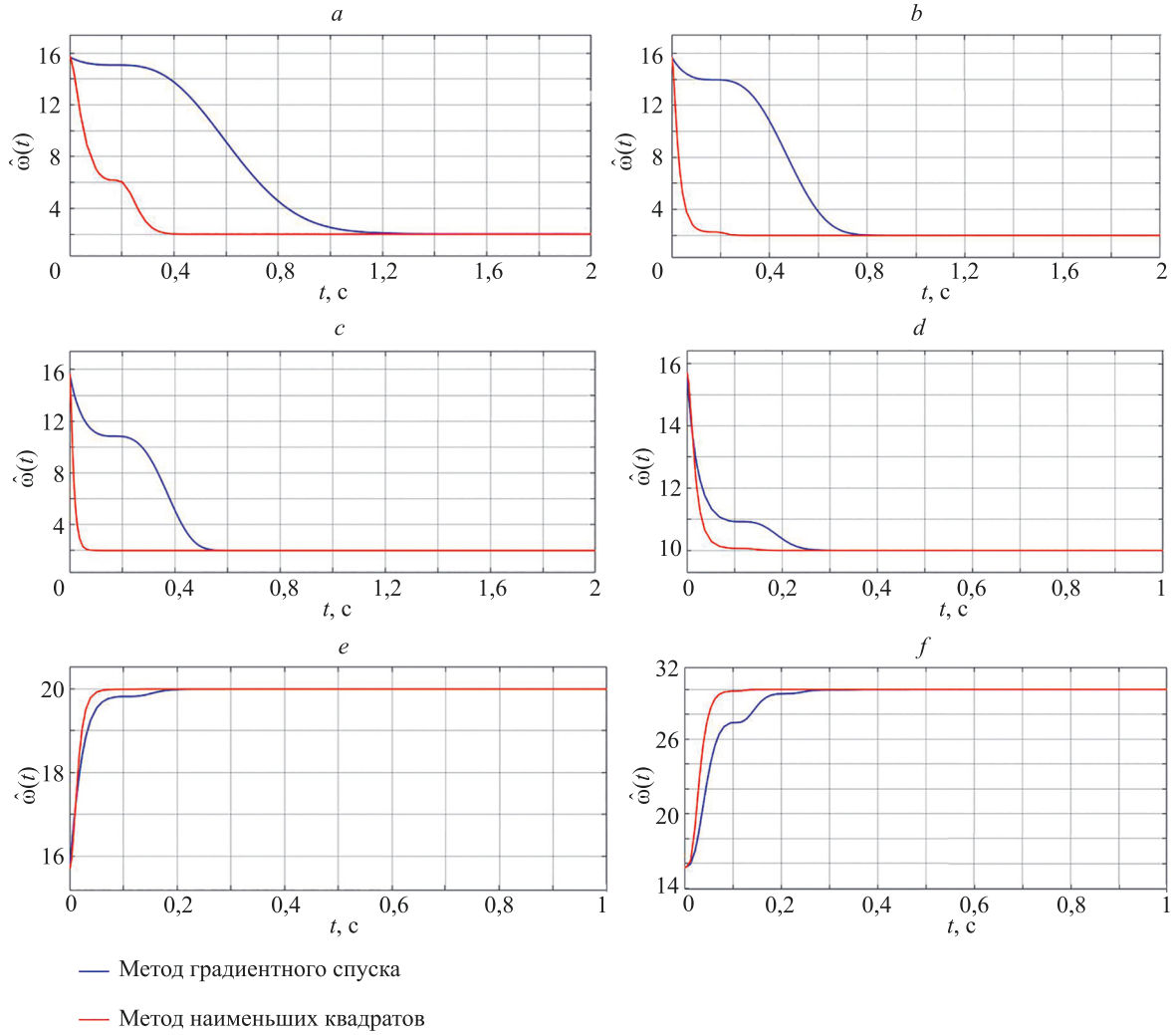


Рис. 1. Оценка $\hat{\omega}(t)$ для значений параметров: $\gamma = 1$, $K = 1$, $\eta = 1$ (*a*); $\gamma = 5$, $K = 5$, $\eta = 5$ (*b*); $\gamma = 10$, $K = 10$, $\eta = 10$ (*c*) и значений частоты ω: 10 (*d*); 20 (*e*); 30 (*f*)

Fig. 1. Estimate $\hat{\omega}(t)$ for parameter values $\gamma = 1$, $K = 1$, $\eta = 1$ (*a*); $\gamma = 5$, $K = 5$, $\eta = 5$ (*b*); $\gamma = 10$, $K = 10$, $\eta = 10$ (*c*) and frequency values ω: 10 (*d*); 20 (*e*); 30 (*f*)

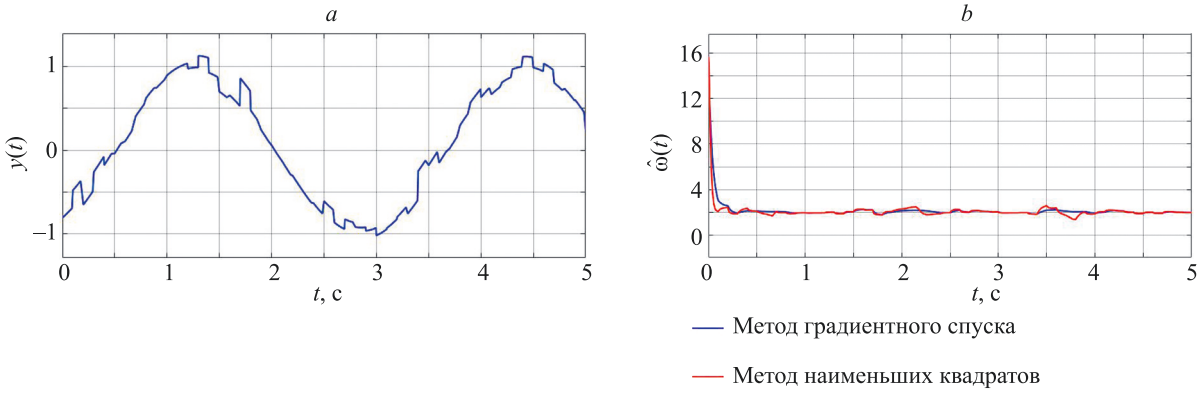


Рис. 2. Синусоидальный сигнал $y(t)$ с аддитивным шумом (*a*); график оценки частоты $\hat{\omega}(t)$ при $\omega = 2$ (*b*)

Fig. 2. Sinusoidal signal $y(t)$ with additive noise (*a*); plot estimation $\hat{\omega}(t)$ frequency $\omega = 2$ (*b*)

$$y(t) = \sin(2t - 1). \quad (19)$$

Приведем значения параметров для предложенного метода: $d = 0{,}1$ с, $\overline{\omega} = 10$ рад/с, $\underline{\omega} = 0{,}01$ рад/с.

На рис. 1 показаны результаты оценивания частоты ω сигнала (19) методами градиентного спуска и наименьших квадратов с различными коэффициентами и частотами.

Как видно из графиков, предложенный алгоритм оценивания обеспечивает экспоненциальную сходимость с истинными значениями оценивания частоты сигнала $y(t)$.

Чтобы продемонстрировать робастность предложенного алгоритма, рассмотрим зашумленный измеренный сигнал

$$y(t) = 5\sin(2t - 2) + \delta(t),$$

где $\delta(t)$ — аддитивный шум, который моделируется как равномерно распределенный процесс в пределах [–0,1, 0,1]. Это ограничение делает синусоиду шумной и неровной, как показано на рис. 2, *a*.

Результат моделирования оценки шумового сценария исследования методами градиентного спуска и наименьших квадратов с параметрами $\gamma = 1$, $K = 10$, $\eta = 10$, при частоте $\omega = 2$ показан на рис. 2, *b*.

Заключение

В работе предложен алгоритм идентификации частоты синусоидального сигнала. Предложен новый подход для параметризации синусоидального сигнала. Выполнена оценка вектора параметров регрессионных моделей с использованием методов стандартного градиентного спуска и наименьших квадратов. Результаты моделирования подтверждают сходимость ошибок оценивания частоты к истинным значениям. В дальнейшем планируется расширить область применения алгоритма на случай мультисинусоидальных сигналов.

Литература

1. Pyrkin A.A., Bobtsov A.A., Efimov D., Zolghadri A. Frequency estimation for periodical signal with noise in finite time // Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC). 2011. P. 3646–3651. https://doi.org/10.1109/CDC.2011.6160655
2. Aranovskiy S., Bobtsov A., Ortega R., Pyrkin A. Improved transients in multiple frequencies estimation via dynamic regressor extension and mixing // IFAC-PapersOnLine. 2016. V. 49. N 13. P. 99–104. https://doi.org/10.1016/j.ifacol.2016.07.934
3. Bobtsov A., Lyamin A., Romasheva D. Algorithm of parameters' identification of polyharmonic function // IFAC Proceedings Volumes. 2002. V. 35. N 1. P. 439–443. https://doi.org/10.3182/20020721-6-ES-1901.01059
4. Marino R., Tomei R. Global estimation of n unknown frequencies // IEEE Transactions on Automatic Control. 2002. V. 47. N 8. P. 1324–1328. https://doi.org/10.1109/TAC.2002.800761
5. Bodson M., Douglas S.C. Adaptive algorithms for the rejection of sinusoidal disturbances with unknown frequency // Automatica. 1997. V. 33. N 12. P. 2213–2221. https://doi.org/10.1016/S0005-1098(97)00149-0
6. Khac T., Vlasov S.M., Iureva R.A. Estimating the Frequency of the Sinusoidal Signal using the Parameterization based on the Delay Operators // Proc. of the 18th International Conference on Informatics in Control, Automation and Robotics (ICINCO). 2021. P. 656–660. https://doi.org/10.5220/0010536506560660
7. Севастеева Е.С., Чернов В.А., Бобцов А.А. Алгоритм увеличения скорости идентификации частоты синусоидального сигнала // Известия вузов. Приборостроение. 2019. Т. 62. № 9. С. 767–771. https://doi.org/10.17586/0021-3454-2019-62-9-767-771
8. Пыркин А.А., Бобцов А.А., Никифоров В.О., Колюбин С.А., Ведяков А.А., Борисов О.И., Громов В.С. Компенсация полигармонического возмущения, действующего на состояние и выход линейного объекта с запаздыванием в канале управления // Автоматика и телемеханика. 2015. № 12. С. 43–64.
9. Бобцов А.А., Пыркин А.А. Компенсация неизвестного синусоидального возмущения для линейного объекта любой относительной степени // Автоматика и телемеханика. 2009. № 3. С. 114–122.
10. Бобцов А.А., Колюбин С.А., Пыркин А.А. Компенсация неизвестного мультигармонического возмущения для нелинейного объекта с запаздыванием по управлению // Автоматика и телемеханика. 2010. № 11. С. 136–148.
11. Vlasov S., Margun A., Kirsanova A., Vakhvianova P. Adaptive controller for uncertain multi-agent system under disturbances // Proc. of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO). 2019. V. 2. P. 198–205. https://doi.org/10.5220/0007827701980205
12. Власов С.М., Борисов О.И., Громов В.С., Пыркин А.А., Бобцов А.А. Алгоритмы адаптивного и робастного управления

References

1. Pyrkin A.A., Bobtsov A.A., Efimov D., Zolghadri A. Frequency estimation for periodical signal with noise in finite time. *Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 3646–3651. https://doi.org/10.1109/CDC.2011.6160655
2. Aranovskiy S., Bobtsov A., Ortega R., Pyrkin A. Improved transients in multiple frequencies estimation via dynamic regressor extension and mixing. *IFAC-PapersOnLine*, 2016, vol. 49, no. 13, pp. 99–104. https://doi.org/10.1016/j.ifacol.2016.07.934
3. Bobtsov A., Lyamin A., Romasheva D. Algorithm of parameters' identification of polyharmonic function. *IFAC Proceedings Volumes*, 2002, vol. 35, no. 1, pp. 439–443. https://doi.org/10.3182/20020721-6-ES-1901.01059
4. Marino R., Tomei R. Global estimation of n unknown frequencies. *IEEE Transactions on Automatic Control*, 2002, vol. 47, no. 8, pp. 1324–1328. https://doi.org/10.1109/TAC.2002.800761
5. Bodson M., Douglas S.C. Adaptive algorithms for the rejection of sinusoidal disturbances with unknown frequency. *Automatica*, 1997, vol. 33, no. 12, pp. 2213–2221. https://doi.org/10.1016/S0005-1098(97)00149-0
6. Khac T., Vlasov S.M., Iureva R.A. Estimating the Frequency of the Sinusoidal Signal using the Parameterization based on the Delay Operators. *Proc. of the 18th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2021, pp. 656–660. https://doi.org/10.5220/0010536506560660
7. Sevasteeva E.S., Chernov V.A., Bobtsov A.A. Algorithm for increasing the speed of sinusoidal signal frequency identification. *Journal of Instrument Engineering*, 2019, vol. 62, no. 9, pp. 767–771. (in Russian). https://doi.org/10.17586/0021-3454-2019-62-9-767-771
8. Pyrkin A.A., Bobtsov A.A., Nikiforov V.O., Kolyubin S.A., Vedyakov A.A., Borisov O.I., Gromov V.S. Compensation of polyharmonic disturbance of state and output of a linear plant with delay in the control channel. *Automation and Remote Control*, 2015, vol. 76, no. 12, pp. 2124–2142. https://doi.org/10.1134/S0005117915120036
9. Bobtsov A.A., Pyrkin A.A. Compensation of unknown sinusoidal disturbances in linear plants of arbitrary relative degree. *Automation and Remote Control*, 2009, vol. 70, no. 3, pp. 449–456. https://doi.org/10.1134/S0005117909030102
10. Bobtsov A.A., Kolyubin S.A., Pyrkin A.A. Compensation of unknown multi-harmonic disturbances in nonlinear plants with delayed control. *Automation and Remote Control*, 2010, vol. 71, no. 11, pp. 2383–2394. https://doi.org/10.1134/S000511791011010X
11. Vlasov S., Margun A., Kirsanova A., Vakhvianova P. Adaptive controller for uncertain multi-agent system under disturbances. *Proc. of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2019, vol. 2, pp. 198–205. https://doi.org/10.5220/0007827701980205

по выходу роботизированным макетом надводного судна // Мехатроника, автоматизация, управление. 2016. Т. 17. № 1. С. 18–25. https://doi.org/10.17587/mau.17.18-25
13. Власов С.М., Борисов О.И., Громов В.С., Пыркин А.А., Бобцов А.А. Робастная система динамического позиционирования для роботизированного макета надводного судна // Известия вузов. Приборостроение. 2015. Т. 58. № 9. С. 713–719. https://doi.org/10.17586/0021-3454-2015-58-9-713-719
14. Hsu L., Ortega R., Damm G. A globally convergent frequency estimator // IEEE Transactions on Automatic Control. 1999. V. 44. N 4. P. 698–713. https://doi.org/10.1109/9.754808
15. Hou M. Amplitude and frequency estimator of a sinusoid // IEEE Transactions on Automatic Control. 2005. V. 50. N 6. P. 855–858. https://doi.org/10.1109/TAC.2005.849244
16. Lee S.W., Lim J.S., Baek S., Sung K.M. Time-varying signal frequency estimation by VFF Kalman filtering // Signal Processing. 1999. V. 77. N 3. P. 343–347. https://doi.org/10.1016/S0165-1684(99)00085-7
17. Karimi-Ghartemani M., Ziarani A.K. A nonlinear time-frequency analysis method // IEEE Transactions on Signal Processing. 2004. V. 52. N 6. P. 1585–1595. https://doi.org/10.1109/TSP.2004.827155
18. Fedele G., Ferrise A. A frequency-locked-loop filter for biased multi-sinusoidal estimation // IEEE Transactions on Signal Processing. 2014. V. 62. N 5. P. 1125–1134. https://doi.org/10.1109/TSP.2014.2300057
19. Hall S., Wereley N. Performance of higher harmonic control algorithms for helicopter vibration reduction // Journal of Guidance, Control, and Dynamics. 1993. V. 16. N 4. P. 793–797. https://doi.org/10.2514/3.21085
20. Nikiforov V.O. Adaptive servomechanism controller with an implicit reference model // International Journal of Control. 1997. V. 68. N 2. P. 277–286. https://doi.org/10.1080/002071797223604
21. Umari A.M.J., Gorelick S.M. Evaluation of the matrix exponential for use in ground-water-flow and solute-transport simulations: theoretical framework: Water-Resources Investigations Report 86-4096. U.S. Geological Survey, 1986. P. 12. https://doi.org/10.3133/wri864096
22. Ljung L. System Identification: Theory for the User. NJ: Prentice-Hall, 1987.

12. Vlasov S.M., Borisov O.I., Gromov V.S., Pyrkin A.A., Bobtsov A.A. Algorithms of adaptive and robust output control for a robotic prototype of a surface vessel. *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2016, vol. 17, no. 1, pp. 18–25. (in Russian). https://doi.org/10.17587/mau.17.18-25
13. Vlasov S.M., Borisov O.I., Gromov V.S., Pyrkin A.A., Bobtsov A.A. Robust system of dynamic positioning for robotized model of surface craft. *Journal of Instrument Engineering*, 2015, vol. 58, no. 9, pp. 713–719. (in Russian). https://doi.org/10.17586/0021-3454-2015-58-9-713-719
14. Hsu L., Ortega R., Damm G. A globally convergent frequency estimator. *IEEE Transactions on Automatic Control*, 1999, vol. 44, no. 4, pp. 698–713. https://doi.org/10.1109/9.754808
15. Hou M. Amplitude and frequency estimator of a sinusoid. *IEEE Transactions on Automatic Control*, 2005, vol. 50, no. 6, pp. 855–858. https://doi.org/10.1109/TAC.2005.849244
16. Lee S.W., Lim J.S., Baek S., Sung K.M. Time-varying signal frequency estimation by VFF Kalman filtering. *Signal Processing*, 1999, vol. 77, no. 3, pp. 343–347. https://doi.org/10.1016/S0165-1684(99)00085-7
17. Karimi-Ghartemani M., Ziarani A.K. A nonlinear time-frequency analysis method. *IEEE Transactions on Signal Processing*, 2004, vol. 52, no. 6, pp. 1585–1595. https://doi.org/10.1109/TSP.2004.827155
18. Fedele G., Ferrise A. A frequency-locked-loop filter for biased multi-sinusoidal estimation. *IEEE Transactions on Signal Processing*, 2014, vol. 62, no. 5, pp. 1125–1134. https://doi.org/10.1109/TSP.2014.2300057
19. Hall S., Wereley N. Performance of higher harmonic control algorithms for helicopter vibration reduction. *Journal of Guidance, Control, and Dynamics*, 1993, vol. 16, no. 4, pp. 793–797. https://doi.org/10.2514/3.21085
20. Nikiforov V.O. Adaptive servomechanism controller with an implicit reference model. *International Journal of Control*, 1997, vol. 68, no. 2, pp. 277–286. https://doi.org/10.1080/002071797223604
21. Umari A.M.J., Gorelick S.M. *Evaluation of the matrix exponential for use in ground-water-flow and solute-transport simulations: theoretical framework*. Water-Resources Investigations Report 86-4096. U.S. Geological Survey, 1986, pp. 12. https://doi.org/10.3133/wri864096
22. Ljung L. *System Identification: Theory for the User*. NJ, Prentice-Hall, 1987.

Авторы

Нгуен Хак Тунг — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-6430-1927, nguyenkhactunghvhq1994@gmail.com
Власов Сергей Михайлович — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 55355689600, https://orcid.org/0000-0002-8345-7553, smvlasov@itmo.ru
Пыркин Антон Александрович — доктор технических наук, профессор, профессор, декан факультета, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 26656070700, https://orcid.org/0000-0001-8806-4057, a.pyrkin@gmail.com
Попков Иван Владимирович — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-4565-5652, vanta.jesus.2015@gmail.com

Authors

Khac Tung Nguyen — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-6430-1927, nguyenkhactunghvhq1994@gmail.com
Sergey M. Vlasov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 55355689600, https://orcid.org/0000-0002-8345-7553, smvlasov@itmo.ru

Anton A. Pyrkin — D.Sc., Full Professor, Dean, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 26656070700, https://orcid.org/0000-0001-8806-4057, a.pyrkin@gmail.com

Ivan V. Popkov — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-4565-5652, vanta.jesus.2015@gmail.com

Статья поступила в редакцию 15.10.2021
Одобрена после рецензирования 12.12.2021
Принята к печати 20.01.2022

Received 15.10.2021
Approved after reviewing 12.12.2021
Accepted 20.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

НОВЫЕ МАТЕРИАЛЫ И НАНОТЕХНОЛОГИИ

MATERIAL SCIENCE AND NANOTECHNOLOGIES

doi: 10.17586/2226-1494-2022-22-1-25-32

A study of silicon *p-n* structures with mono- and multifacial photosensitive surfaces

Avazbek Mirzaalimov[1], Jasurbek Gulomov[2]✉, Rayimjon Aliev[3], Navruzbek Mirzaalimov[4], Suhrob Aliev[5]

[1,2,3,4] Andijan State University, Andijan, 170100, Uzbekistan
[5] Andijan Machine-Building Institute, Andijan, 170100, Uzbekistan

[1] avazbek.mirzaalimov@mail.ru, https://orcid.org/0000-0003-2846-1901
[2] jasurbekgulomov@yahoo.com✉, https://orcid.org/0000-0001-7516-987X
[3] alievuz@yahoo.com, https://orcid.org/0000-0003-1986-2199
[4] mirzaalimov90@mail.ru, https://orcid.org/0000-0002-9264-3710
[5] suhrob_asr89@mail.ru, https://orcid.org/0000-0002-4494-8261

Abstract
Increase in the efficiency and reduction of silicon consumption in production of solar cells are relevant problems. Designing two and three facial solar cells can be seen as a solution for such tasks. Compared to usual SC, the output power of two and three facial solar cells exceeds by 1.72 times by 2.81 times, respectively. Illumination of solar cells with high intensity light makes the temperature of its heating an important characteristic. Therefore, the paper investigates the influence of temperature on properties of multifacial solar cells. We defined the nature of change of temperature coefficients for the main photovoltaic parameters that are inherent to silicon solar cells under various (one, two and three facial) conditions of lighting. Temperature coefficients of three facial solar cells are $2.52 \cdot 10^{-3}$ V/K for open circuit voltage and $1.8 \cdot 10^{-3}$ K^{-1} for fill factor of I-V. At temperature change of SC from 300 K to 350 K, the density of short circuit current decreases only by 4 %.

Keywords
silicon, solar cell, simulation, three facial sensitivity, p-n junction, efficiency, lighting, temperature

Acknowledgements
The authors are grateful to the staff of the Renewable Energy Sources Laboratory at Andijan State University for their invaluable assistance in preparing this article.

For citation: Mirzaalimov A., Gulomov J., Aliev R., Mirzaalimov N., Aliev S. A study of silicon *p-n* structures with mono- and multifacial photosensitive surfaces. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 25–32. doi: 10.17586/2226-1494-2022-22-1-25-32

УДК 621.382

Исследование кремниевых *p-n* структур с моно- и мультифоточувствительными поверхностями

Авазбек Алишерович Мирзаалимов[1], Жасурбек Гуломов[2]✉, Райимжон Алиев[3], Наврузбек Алишер угли Мирзаалимов[4], Сухроб Алиев[5]

[1,2,3,4] Андижанский государственный университет, Андижан, 170100, Узбекистан
[5] Андижанский машиностроительный институт, Андижан, 170100, Узбекистан

[1] avazbek.mirzaalimov@mail.ru, https://orcid.org/0000-0003-2846-1901
[2] jasurbekgulomov@yahoo.com✉, https://orcid.org/0000-0001-7516-987X
[3] alievuz@yahoo.com, https://orcid.org/0000-0003-1986-2199
[4] mirzaalimov90@mail.ru, https://orcid.org/0000-0002-9264-3710
[5] suhrob_asr89@mail.ru, https://orcid.org/0000-0002-4494-8261

© Mirzaalimov A., Gulomov J., Aliev R., Mirzaalimov N., Aliev S., 2022

Аннотация
Повышение эффективности и сокращение расхода кремния при производстве солнечных элементов являются актуальными проблемами. Один из путей их решения — создание двух- и трехлицевых солнечных элементов. Выходная мощность солнечного элемента при двухстороннем освещении в 1,72 раза и при трехстороннем освещении в 2,81 раза превышает мощность односторонне освещенного (обычного) солнечного элемента. При освещении солнечного элемента светом высокой интенсивности актуальным становится учет температуры его нагрева. В работе исследовано влияние температуры на свойства многосторонне освещаемого солнечного элемента. Определен характер изменения температурных коэффициентов основных фотоэлектрических параметров кремниевых солнечных элементов при различных условиях (одно-, двух- и трехстороннего) освещения. Температурный коэффициент трехсторонне освещаемых солнечных элементов для напряжения холостого хода составил величину $2{,}52 \cdot 10^{-3}$ В/К, коэффициент заполнения вольтамперной характеристики $1{,}8 \cdot 10^{-3}$ $К^{-1}$. При изменении температуры солнечного элемента от 300 до 350 К плотность тока короткого замыкания уменьшается всего на 4 %.

Ключевые слова
кремний, солнечный элемент, моделирование, трехсторонняя чувствительность, *p-n* переход, эффективность, освещение, температура

Благодарности
Авторы выражают благодарность сотрудникам Лаборатории возобновляемых источников энергии Андижанского государственного университета за активную помощь в подготовке статьи.

Ссылка для цитирования: Мирзаалимов А.А., Гуломов Ж., Алиев Р., Мирзаалимов Н.А., Алиев С. Исследование кремниевых *p-n* структур с моно- и мультифоточувствительными поверхностями // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 25–32 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-1-25-32

Introduction

Continuous growth of the needs in energy that humanity has stimulates faster development of renewable energy sources. They are the most primary solutions that ensure growing energy consumption without negative impact on the environment [1]. According to official information from “International Energy Agency”, the contribution of renewable energy sources (RES) in 2021 reaches up to 276 GW. The widest circulation among different types of RES has evolved from sunlight to thermal and electric energy [2]. For transformation of sunlight to electric energy, the semiconductor solar cells (SC) are generally used. More than 85 % of the produced by the industry SC are made on a silicon basis [3]. For this reason, we chose silicon SC as the main object of research. The efficiency of silicon SC that agree with theoretical expectations does not exceed 29 % [4]. But both industrial and laboratory silicon SC did not reach this value yet. Therefore, there is a need in researches on the increase in efficiency of SC and decrease in prime cost of silicon. Numerous theoretical and experimental works characterized by different scientific and technical approaches are carried out in this direction. For example, for the purpose of passivation of recombination activity and improvement of optical properties, a frontal surface of silicon SC is covered with a layer of SiN_x by 75 nanometers thick or SiO_2 by 100 nanometers thick [5]. Authors in [6] introduced nanoparticles of metals to increase the efficiency of silicon SC. For the increase in the photovoltaic transformed energy, silicon bifacial SC were designed [7]. However, in experiments, it was established that lighting the front side of the silicon bifacial SC showed the efficiency by 19.4 %, while lighting the back side of the silicon bifacial SC reached the efficiency equal to 16.5 %. In experiment, front and rear side of bifacial SC were 18.4 % and 18.1 %, respectively. So, the values for the efficiency of both sides reached symmetry [8]. In our work, the three facial silicon SC (Fig. 1), which is offered in [9], is studied for the first-time. In addition to physical mechanism of energy transformation in multifacial solar cells, influence of illumination and temperature on them has not been studied yet. For widespread introduction of any new design of SC, it is important to estimate the influence of illumination conditions and temperature changes on their photovoltaic parameters. For example, the influence of sunlight incident angle changes and temperature changes on the main photovoltaic parameters of SC on silicon base with volume metals nanoparticles was investigated in [10, 11]. Assessment of silicon SC temperature and impact of temperature changes on its photovoltaic parameters is crucial and of practical interest at three facial lighting conditions. Our work deals with the solution of this task.

If one facial solar cell is called a solar cell with monofacial photosensitive surfaces, hence two or three facial solar cells can be called solar cells with multifacial photosensitive surfaces. In the feature, solar cells with two,

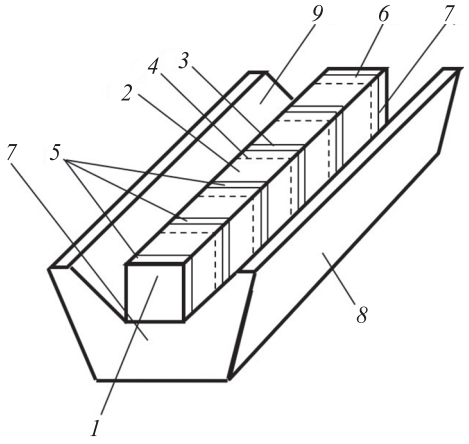


Fig. 1. Construction of a three facial solar cell: *1* — silicon in prism form; *2* — *p-n* junction; *3* — frontal contact; *4* — back contact; *5* — frontal surface; *6* — lateral face; *7* — reflector; *8* — the substrate in a tray form; *9* — silver mirror

three or four photosensitive surfaces will represent one family of multifacial solar cells.

Method

There are various research methods, for example, experimental, theoretical and modeling ones. Theoretical methods are divided into the following two groups:

1) fundamental theory of physical processes;
2) analytical solutions of the empirical equations received on experimental data.

The method of modeling is considered more difficult, as both fundamental theories and empirical formulas are used. A basis of modeling of semiconductor structures work out the private differential equations. Calculation of private differential equations in the analytical way is very laborious, therefore numerical methods are used. For the sequence of calculations by numerical methods, it is possible to create appropriate algorithms. For the solution of private differential equations, it is important to accept entry boundary conditions. Entry boundary conditions can take certain values or a form of function. When modeling physical processes as entry boundary conditions, one should use empirical formulas.

For modeling a design of SC, the simplified geometrical model is created. The geometrical model of SC with various illumination systems is given in Fig. 2. One, two and three facial solar cells depend on illumination conditions in our research. So, if only (L_B) or (L_B, L_A) or (L_B, L_A, L_C) faces are illuminated, they can be taken into account respectively as a one, two or three facial solar cells. On the front side of SC, the ray of sunlight falls perpendicularly. Flat reflectors from silver are placed on two sides of solar cells with an angle of 45 degree to ensure uniform light on all three surfaces of the solar cell. SC has a design in the form of a long prism, which width and height are identical: l = 175 μm. The base of SC represents conductivity p-type silicon with concentration of acceptor impurity $p = 10^{15}$ cm^{-3} and 171.5 μm thick, and an emitter layer of n-type of conductivity with concentration of phosphorus impurity $n = 10^{17}$ cm^{-3} and 3.5 μm thick. In the industry, silicon solar cells are made with thickness of 175 μm. If industrial silicon solar is divided into many pieces with width of 175 μm, a solar cell, which is like our geometric model, can be formed for the experiment. So, in this paper, the geometrical parameters of the cell are taken as parameters of an industrial silicon solar cell. Illumination of SC corresponds to AM1.5G range condition.

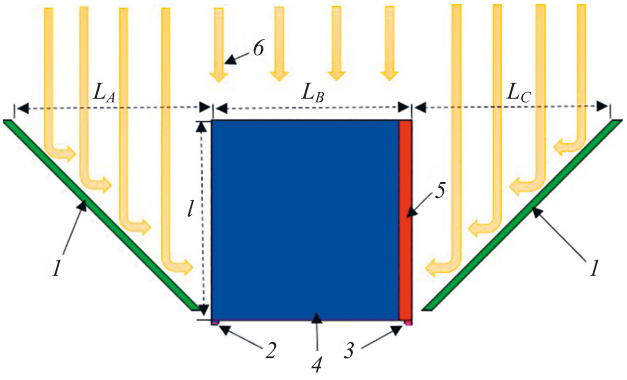


Fig. 2. Geometrical model of a three facial solar cell: *1* — silver reflector; *2* and *3* — electrodes; *4* — p-type silicon base; *5* — n-type emitter layer; *6* — sunlight beam; l — width and height (the prism basis) of a solar cell. $L_A = L_B = L_C = l$

As noted above, during the modulation of semiconductor structures, private differential equations play the main role. For example, for determination of electric parameters of semiconductor p-n structure in equilibrium conditions it is enough to use a formula for Fermi statistics:

$$n = N_c F_{1/2}\left(\frac{E_{F,n} - E_c}{kT}\right) \text{ and } p = N_V F_{1/2}\left(\frac{E_V - E_{F,p}}{kT}\right), \quad (1)$$

and a formula for Poisson's equation:

$$\Delta\varphi = -\frac{q}{\varepsilon}(p - n + N_D + N_A), \quad (2)$$

where $F_{1/2}$ is Fermi half integral; E_c is conduction band energy; E_V is valence band energy; $E_{F,n}$ is quasi fermi energy for electrons; $E_{F,p}$ is quasi fermi energy for holes; T is absolute temperature; N_c is density of states in the conduction band; N_V is density of states in valence band; k is Boltzmann constant; ε is permittivity; n and p are concentration of electrons and holes; N_D and N_A are donor and acceptor concentration; q is electron charge; φ is electric potential; Δ is Laplace operator.

Based on formula (1), it is possible to determine distribution of charge carriers concentration by semiconductor structure. Analytical function evaluation of Fermi statistics is impossible because its part is half integral Fermi. Therefore, Fermi statistics can be calculated by using Boltzmann approximation. But, Boltzmann distribution yields more exact results only at certain intervals of energy.

In this work, numerical calculation on the basis of "Sentaurus TCAD" is executed and used directly a function of Fermi statistics. Inserting the calculated concentration of electrons and holes into Poisson's equation, it is possible to define the distribution of tension or potential of the electric field in a semiconductor structure.

Modeling a semiconductor SC is considered among various instrument structures to be more difficult. It is connected with the fact that in the SC models except electric and thermal properties, it is still necessary to consider optical properties. For determination of optical properties by means of the "Sentaurus Device" tool, the transfer matrix method (TMM), ray tracing method and beam propagation method are also used. Each of the mentioned methods has its irreplaceable advantages. For example, TMM allows one to consider the interferential phenomena when calculating optical properties of the thin layer of SC. For determination of optical properties, the ray tracing method is used. When using the ray tracing method, there is an opportunity to consider the beams that fall, refract and are reflected as independent beams. Population of rays is stopped when ray energy is below the critical intensity that is defined in simulation command field. Therefore, in the modeling mode at the end of calculation, light beams are divided into parts: absorbed, free and stopped. One beam can be divided into a thousand parts.

A part of beams can be reflected from a surface, and a part can pass through, and they are considered to be free. The beam absorbed in SC loses energy and its further growth stops. The general energy falling on SC is expressed by a formula:

$$P_{total} = P_{abs} + P_{escape} + P_{stopped},$$

where P_{total} is total energy of falling beam; P_{abs} is energy of absorbed light; P_{escape} is energy of escaped light; $P_{stopped}$ is energy of stopped light.

For determination of optical properties of devices, it is necessary to know values of refraction and absorption indexes and for the materials which are contained in the structure. Optical parameters of any material are expressed as a complex index of refraction by the formula, which consists of a real and imaginary parts:

$$n_{tot}(\lambda) = n(\lambda) + ik(\lambda),$$

where n_{tot} is complex refractive index; n is real part of complex refractive index; k is imaginary part of complex refractive index; λ is wavelength.

Imaginary part of a refraction index expresses excitement coefficient. The size of both coefficients depends on light wavelength. The exact function of dependence of these coefficients for various materials is not defined. Therefore, in the modeling, table of dependencies of complex refractive index on light wavelength is used in simulation. So, accuracy and reliability of the optical properties of devices determined by way of modeling is provided.

The absorption coefficient of material is expressed through dependencies of the coefficient of excitement on wavelength in the following form:

$$\alpha(\lambda) = \frac{4\pi k}{\lambda},$$

where α is absorption coefficient.

When calculating optical properties using the Ray tracing method, certain boundary conditions are entered. The relationship between the angle of the incident, reflected and refracted beams are expressed by Snell's the law in the form:

$$\frac{n_1}{n_2} = \frac{\sin(\gamma)}{\sin(\theta)}, \beta = \theta,$$

where n_1 and n_2 are refractive indices of media; β is angle of incident light; γ is angle of refracted light; θ is angle of reflected light.

The relationship between energy of the falling, reflected and refracted beams are expressed by Fresnel equations:

$$\begin{cases} r_t = \dfrac{n_1\cos\beta - n_2\cos\gamma}{n_1\cos\beta + n_2\cos\gamma} \\ t_t = \dfrac{2n_1\cos\beta}{n_1\cos\beta + n_2\cos\gamma} \end{cases} \text{and} \begin{cases} r_p = \dfrac{n_1\cos\gamma - n_2\cos\beta}{n_1\cos\gamma + n_2\cos\beta} \\ t_p = \dfrac{2n_1\cos\beta}{n_2\cos\beta + n_1\cos\gamma} \end{cases},$$

where r_p, t_p, r_t and t_t are Fresnel coefficients.

When Snell's law and Fresnel coefficients are used as boundary conditions, distributions of light on the interface between two media will also depend on light wavelength. Due to the opportunities of "Sentaurus Device", the fact that values of coefficient of reflection and passing of light on border of two environments can be set by constants is important. In the real work, Snell's and Fresnel laws are used as boundary conditions for zones of contact of different environments in SC. As boundary conditions for an interval of the silver reflector and air (Fig. 2) Snell's law and constant value of coefficient of reflection $R = 1$ are adopted. It provides falling of the reflected beam perpendicularly on two lateral faces of SC without loss of energy.

The quantity of the generated electrons and holes in SC layers during the absorption of sunlight is defined by the expression for optical generation:

$$G^{opt}(x, y, z, t) = I(x, y, z, t)[1 - e^{-\alpha L}],$$

where G^{opt} is optical generation; I is light intensity; L is medium thickness; x, y, z are Cartesian coordinates; t is time; α is absorption coefficient of medium.

In addition to optical generation of charge carriers, our simulation also considers thermal generation. A recombination of carriers is considered to be the basis for SC accepted silicon and when modeling processes of "Auger" and "Shokley-Read-Hall". Then were calculated distribution of electrons and holes in the structure of SC by using of Fermi functions of distribution (1). Inserting a certain distribution of electrons and holes into the Poisson (2) equation, a numerical calculation is executed. For implementation of a numerical calculation, it is necessary to form plane (or volume) mesh in the structure of SC. Steps for mesh on an axis "x" from the minimum 0.04 μm to the maximum 0.08 μm, and on an axis "y" from the minimum 0.01 μm to the maximum 0.02 μm are accepted for this purpose. For the calculation of both optical and SC electric properties, we used an identical technique of formation of mesh with identical step values. During simulation, optical parameters such as absorption coefficient or absorbed photon density are calculated for a point of each optical mesh. Therefore, the ray tracing method needs optical mesh.

As a result of absorption of a photon in SC, electron-hole couples are generated which are divided by *p-n* junction and move towards contact electrodes. Change of concentration of charge carriers in a certain volume of SC on time causes the emergence of an electric current. The relationship between changes of concentration of electrons and holes over time and also the created electric current is expressed by the equation of continuity:

$$\nabla\vec{J}_n = qR_{net,n} + q\frac{\partial n}{\partial t} \text{ and } -\nabla\vec{J}_p = qR_{net,p} + q\frac{\partial p}{\partial t},$$

where J_n, J_p are the current densities of electrons and holes; $R_{net,p}$, $R_{net,n}$ are the net recombination of electrons and holes; t is time; ∇ is Nabla operator.

In the "Sentaurus Device" system there are 4 models for calculating the movement of charge carriers: drift-iffusion, thermodynamic, hydrodynamic and Monte Carlo. The drift-diffusion model is a default one, and it starts automatically. In this work, for calculating the density of the current created by the directed movement of electrons

and holes, we used the thermodynamic model according to the following formula:

$$J_n = -nq\mu_n(\nabla F_n + P_n\nabla T) \text{ and } J_p = -pq\mu_p(\nabla F_p + P_p\nabla T),$$

where μ_n, μ_p are the mobility of electrons and holes; F_n, F_p are the electron and hole quasi-Fermi potentials; P_n, P_p are the thermoelectric power of electrons and holes; T is absolute temperature.

The choice of this model is connected with the fact that it is necessary to consider the influence of temperature on charge transfer and formation of phonons in processes of a recombination of charge carrier. But at the same time, it is necessary to use additional calculation by the temperature equation:

$$\frac{\partial}{\partial t}(c_L T) - \nabla(k\nabla T) = -\nabla[(P_n T + F_n)\vec{J}_n + (P_p T + F_p)\vec{J}_p] - \\ -\frac{1}{q}\left(E_c + \frac{3}{2}kT\right)(\nabla\vec{J}_n - qR_{net,n}) - \\ -\frac{1}{q}\left(E_v + \frac{3}{2}kT\right)(-\nabla\vec{J}_p - qR_{net,n}) + \hbar\omega G^{opt}, \quad (3)$$

where k is heat conductance; c_L is heat capacity; E_c is minimum energy of conduction band; E_v is maximum energy of valence band; G^{opt} is optical generation; $R_{net,n}$ and $R_{net,p}$ are net recombination, J_n and J_p are current densities of electrons and holes; t is time; F_m is metal Fermi state; J_m is current density in metal; ω is frequency of photons; $\hbar$ is Planck constant.

In equation (3), the temperature equation is given as a non-stationary equation. But, in this paper, the stationary form of temperature equation has been used to calculate temperature of the lattice. If it is taken that temperature does not depend on the time, the time derivative of temperature, which is located the left side of equation (3), will be equal to zero. So, the equation (3) will be changed from non-stationary to its stationary form.

For the solution of equation (3), thermal boundary conditions should be applied. Thermopowers in the semiconductor and metal are various. Therefore, the thermal resistive boundary condition is calculated:

$$k\hat{n}\nabla T = \frac{T_{ext} - T}{R_{th}}, \quad (4)$$

where R_{th} is external thermal resistance; T_{ext} is external temperature; $\hat{n}$ is a unit vector in the direction of the outer normal.

If the right side of equation (4) is equal to zero, it can be used for thermally insulated interfaces. Assuming an interface between materials 1 and 2 such that $\Delta T = T_2 - T_1 > 0$, $\vec{H}_1$ is the heat flux density entering material 1 and $\vec{H}_2$ is the heat flux density leaving material 2. Hence, the thermal boundary condition at the interface between materials 1 and 2 can written as:

$$\vec{H}_1 = \vec{H}_2,$$
$$\vec{H}_1 = \frac{T_2 - T_1}{R_{d,th}}$$

where $R_{d,th}$ is the interface-distributed resistance.

In other words, it is not enough to choose the "Thermodynamic" command for the correct use of the thermodynamic model in the command file of the "Sentaurus Device" tool in the section "Physics". In addition, in the command file in the section "Solve", it is necessary to start calculation of the equation of "Thermodynamic".

Thus, according to the specified sequence of operations, the calculation of SC properties in the conditions of three facial lighting is executed.

Results and discussion

The main photovoltaic parameters of silicon SC with thickness of 175 μm are determined under 4 various conditions of lighting. The influence of temperature on the main photovoltaic parameters of SC is studied. I-V characteristic of silicon SC calculated for various conditions of lighting is given in Fig. 1. Note that the volume of a short circuit current of SC for the lighting conditions in parallel (the front side with vertical *p-n* junction) or perpendicularly (lateral face with horizontal *p-n* junction) to the front of *p-n* junction are identical. This fact means that the mechanics of photo generation and collecting of charge carriers by *p-n* junction are identical for the accepted SC sizes. Short circuit current and output power of SC have exceeded usual monofacial SC parameters by 1.72 times in bifacial lighting and by 2.81 times in three facial lighting. Open circuit voltage of SC has been greater 1.023 times in bifacial lighting and 1.047 times in three facial lighting than that of custom monofacial SC. The analysis of the known literary data demonstrated that at bifacial lighting, silicon SC can exceed short circuit current from 20.1 % to 68.1 % of parameter at its monofacial lighting [12]. It is possible to consider that the received estimated results confirmed by the known experimental data Fig. 3 shows I-V characteristics of one, two and three facial silicon solar cells at temperature of 300 K.

For determination of temperature dependencies of the main photovoltaic parameters of the silicon SC model, such I-V characteristics as shown in Fig. 3 were calculated for the temperature range from 300 K to 350 K. Each data of I-V characteristics was reworked by using new software created for calculating the photoelectric parameters from data of

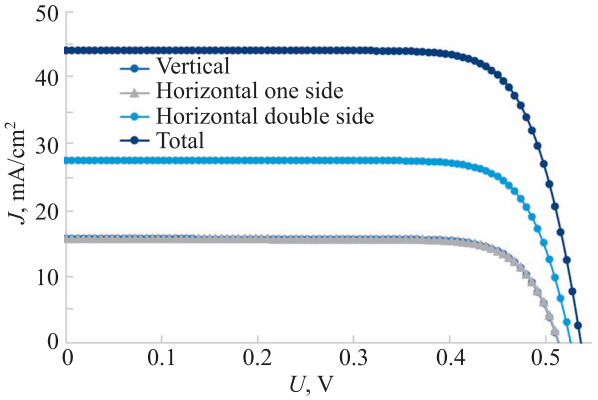


Fig. 3. I-V characteristics of silicon solar cells at various illumination conditions

I-V characteristics. So, photoelectric parameters of a solar cell were identified for the temperature range from 300 K to 350 K. The photovoltaic parameter of silicon SC that is the most sensitive to the change of temperature is open circuit voltage. For this reason, reduction of silicon SC efficiency with temperature increase can be interpreted through the change of the open circuit voltage. The contribution of processes of "Auger" recombination during the change of open circuit voltage of SC with temperature is rather high in comparison with other types of a recombination. The dependence of open circuit voltage of a silicon SC on temperature under various conditions of its illumination is given in Fig. 4, *a*. We note that for all conditions of illumination of a silicon SC, the temperature coefficient of open circuit voltage accepts identical value and makes $2.52 \cdot 10^{-3}$ V/K. Means, change of the area of lighting silicon SC has no impact on temperature coefficients of the open circuit voltage. At the same time, Tiedje experimentally determined the size for the temperature coefficient of open circuit voltage $1.36 \cdot 10^{-3}$ V/K [13]. The difference in values of certain settlement and experiments can be determined by methods, apparently, with construction features of SC and contact electrodes.

Temperature change of SC will change parameters and concentration of phonons and also the concentration of electrons and holes due to thermogeneration of charge carriers. Therefore, short circuit current has to change. Curve dependencies of short circuit current of a silicon SC on temperature are given in Fig. 4, *b* under various conditions of its lighting. Similar to data in Fig. 3, the volume of short circuit current of silicon SC when lighting it from two and three sides significantly exceeds the parameter for it during monoside lighting. Therefore, in Fig. 4, *b*, the changes from temperature of short circuit current of a silicon SC are given in relative units. At the increase in temperature of SC from 300 K to 350 K, the volume of a short circuit current of a silicon SC at horizontal monofacial lighting changes for 3.1 %, at horizontal bifacial lighting for 4.4 %, vertically monofacial lighting for 3.8 % and at three facial lighting for 4 %.

In the experimental work by He Wang, it is defined that the temperature coefficient of short circuit current of silicon SC at the conditions of bifacial lighting with an intensity 600 W/m^2 reaches $0.7 \cdot 10^{-3}$ K^{-1} [14].

Other important photovoltaic parameter of silicon SC is fill factor. Usually, consecutive resistance of silicon SC has significant effect on its fill factor volume. Dependencies on the temperature of short circuit current of a silicon SC under various conditions of its lighting are given in Fig. 4, *c*. The results demonstrated that value of temperature coefficients of fill factor of SE at three facial lighting is less by 2 %, than in the case of monofacial lighting. Some increase of the fill factor at the temperature of 330 K for SC with horizontal and vertical lighting is revealed. However, the function of temperature coefficient change for fill factor is uniform for all conditions of lighting. The value of temperature coefficient of the fill factor is $1.8 \cdot 10^{-3}$ K^{-1}. The reduction of temperature coefficient of fill factor of silicon SC with the increase of temperature is caused by the increase of a surface recombination of charge carriers.

Development of bifacial and three facial SC attracts economic interest. It is connected with the fact that output power of bifacial SC by 1.72 times and three facial SC by

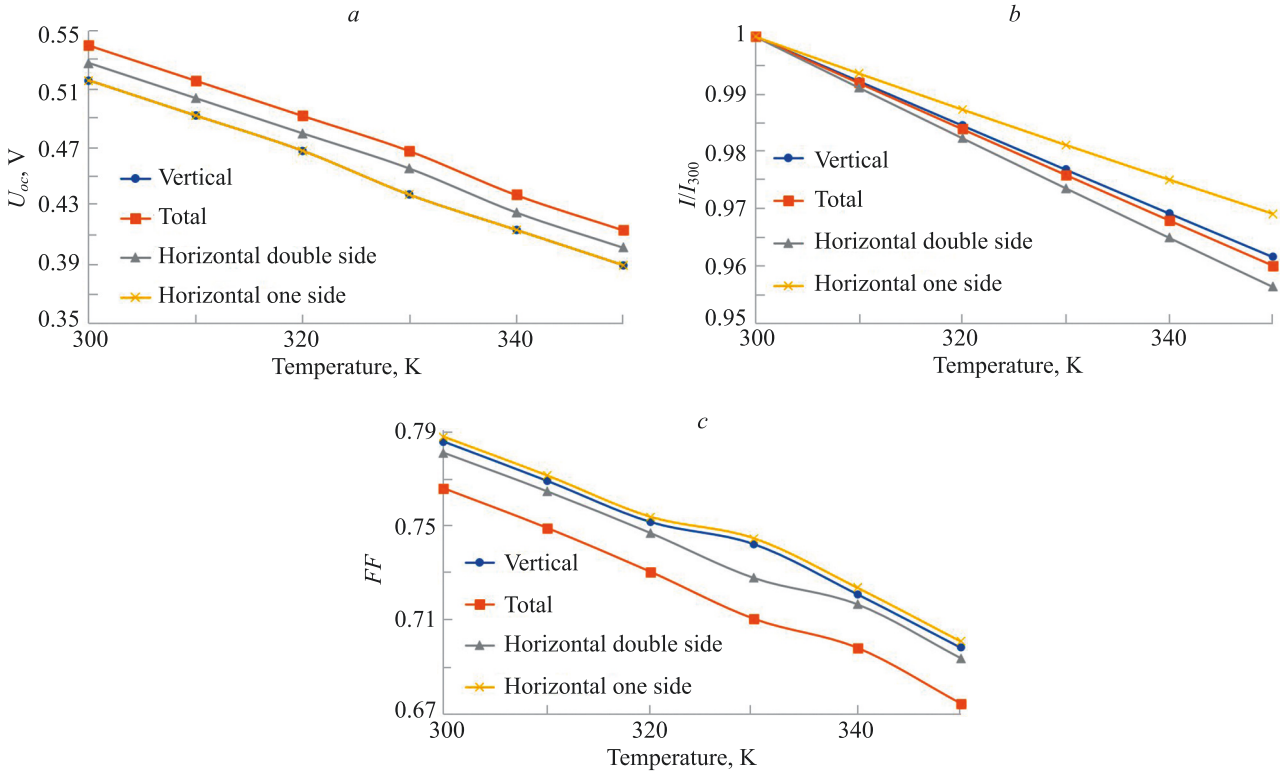


Fig. 4. Dependence on temperature of: open circuit voltage (*a*), short circuit current (*b*), fill factor (*FF*) (*c*) of a silicon solar cell under various conditions of its lighting

2.81 times exceeds the power of usual monofacial SC. By spending the single volume of silicon for the creation of three facial SC, it is possible to receive by 2.81 times more energy than in case of usual monofacial SC.

Conclusion

The quantity of the photovoltaic energy transformed by silicon SC depends on the size of its area. The increase of the area of SC causes the increase of the photons falling on its surface and, therefore, the number of the photogenerated charge carriers. Therefore, the short circuit current increases with the increase in the lighting area. Usually, the increase of open circuit voltage is insignificant under such conditions. A considerable part of photons which energy is insufficient for photogeneration of charge carriers is absorbed by silicon and other SC construction elements. It leads to heating of a body of SC, causing reduction of its efficiency. Illumination of SC with high intensity light makes the temperature of its heating an important characteristic. The nature of temperature coefficients change of the main photovoltaic parameters of silicon SC has been defined under various conditions.

Temperature coefficients for three facial SC reached $2.52 \cdot 10^{-3}$ V/K for open circuit voltage and $1.8 \cdot 10^{-3}$ K^{-1} for fill factor. Output power of SC at bifacial lighting by 1.72 times and three facial lighting by 2.81 times exceeds power usual monofacial lighting. At temperature change of SC from 300 K to 350 K, the density of short circuit current decreases by only 4 %. Comparison of material consumption for production of bifacial and three facial SC demonstrates the corresponding economy of silicon for obtaining significantly bigger photoelectric energy. Thus, it is possible to note that production and introduction of silicon SC with bifacial, three facial and multifacial lighting is expedient.

References

1. Mardani A., Jusoh A., Zavadskas E.K., Cavallaro F., Khalifah Z. Sustainable and renewable energy: An overview of the application of multiple criteria decision making techniques and approaches. *Sustainability*, 2015, vol. 7, no. 10, pp. 13947–13984. https://doi.org/10.3390/SU71013947
2. Guangul F.M., Chala G.T. Solar energy as renewable energy source: SWOT analysis. *Proc. 4th MEC International Conference on Big Data and Cmart City (ICBDSC)*, 2019, pp. 8645580. https://doi.org/10.1109/ICBDSC.2019.8645580
3. Fath P., Keller S., Winter P., Jooß W., Herbst W. Status and perspective of crystalline silicon solar cell production. *Proc. of the 34th IEEE Conference Photovoltaic Specialists*, 2009, pp. 002471–002476. https://doi.org/10.1109/PVSC.2009.5411274
4. Shockley W., Queisser H.J. Detailed balance limit of efficiency of *p-n* junction solar cells. *Journal of Applied Physics*, 1961, vol. 32, no. 3, pp. 510–519. https://doi.org/10.1063/1.1736034
5. Schmidt J., Werner F., Veith B., Zielke D., Steingrube S., Altermatt P.P., Gatz S., Dullweber T., Brendel R. Advances in the surface passivation of silicon solar cells. *Energy Procedia*, 2012, vol. 15, pp. 30–39. https://doi.org/10.1016/J.EGYPRO.2012.02.004
6. Gulomov J., Aliev R. The way of the increasing two times the efficiency of silicon solar cell. *Physics and Chemistry of Solid State*, 2021, vol. 22, no. 4, pp. 756–760. https://doi.org/10.15330/pcss.22.4.756-760
7. Hiroshi M. Radiation energy transducing device. *Patent US3278811A*. Available at: https://patents.google.com/patent/US3278811A/en (accessed: 14.09.2021).
8. Hübner A., Aberle A.G., Hezel R. Novel cost-effective bifacial silicon solar cells with 19.4% front and 18.1% rear efficiency. *Applied Physics Letters*, 1997, vol. 70, no. 8, pp. 1008–1010. https://doi.org/10.1063/1.118466
9. Aliev R., Muxtarov E., Mirzaalimov A. Quyosh generator. *Patent UZ.* № FAP00623, 2011.
10. Gulomov J., Aliev R. Numerical analysis of the effect of illumination intensity on photoelectric parameters of the silicon solar cell with various metal nanoparticles. *Nanosystems: Physics, Chemistry, Mathematics*, 2021, vol. 12, no. 5, pp. 569–574. https://doi.org/10.17586/2220-8054-2021-12-5-569-574
11. Gulomov J., Aliev R. Study of the temperature coefficient of the main photoelectric parameters of silicon solar cells with various nanoparticles. *Journal of Nano- and Electronic Physics*, 2021, vol. 13, no. 4, pp. 04033. https://doi.org/10.21272/JNEP.13(4).04033
12. Pal S., Reinders A., Saive R. Simulation of bifacial and monofacial silicon solar cell short-circuit current density under measured spectro-angular solar irradiance. *IEEE Journal of Photovoltaics*, 2020, vol. 10, no. 6, pp. 1803–1815. https://doi.org/10.1109/JPHOTOV.2020.3026141

Литература

1. Mardani A., Jusoh A., Zavadskas E.K., Cavallaro F., Khalifah Z. Sustainable and renewable energy: An overview of the application of multiple criteria decision making techniques and approaches // Sustainability. 2015. V. 7. N 10. P. 13947–13984. https://doi.org/10.3390/SU71013947
2. Guangul F.M., Chala G.T. Solar energy as renewable energy source: SWOT analysis // Proc. of the 4th MEC International Conference on Big Data and Cmart City (ICBDSC). 2019. P. 8645580 https://doi.org/10.1109/ICBDSC.2019.8645580
3. Fath P., Keller S., Winter P., Jooß W., Herbst W. Status and perspective of crystalline silicon solar cell production // Proc. of the 34th IEEE Conference Photovoltaic Specialists. 2009. P. 002471–002476. https://doi.org/10.1109/PVSC.2009.5411274
4. Shockley W., Queisser H.J. Detailed balance limit of efficiency of *p-n* junction solar cells // Journal of Applied Physics. 1961. V. 32. N 3. P. 510–519. https://doi.org/10.1063/1.1736034
5. Schmidt J., Werner F., Veith B., Zielke D., Steingrube S., Altermatt P.P., Gatz S., Dullweber T., Brendel R. Advances in the surface passivation of silicon solar cells // Energy Procedia. 2012. V. 15. P. 30–39. https://doi.org/10.1016/J.EGYPRO.2012.02.004
6. Gulomov J., Aliev R. The way of the increasing two times the efficiency of silicon solar cell // Physics and Chemistry of Solid State. 2021. V. 22. N 4. P. 756–760. https://doi.org/10.15330/pcss.22.4.756-760
7. Hiroshi M. Radiation energy transducing device. Patent US3278811A [Электронный ресурс]. URL: https://patents.google.com/patent/US3278811A/en (дата обращения: 14.09.2021).
8. Hübner A., Aberle A.G., Hezel R. Novel cost-effective bifacial silicon solar cells with 19.4% front and 18.1% rear efficiency // Applied Physics Letters. 1997. V. 70. N 8. P. 1008–1010. https://doi.org/10.1063/1.118466
9. Aliev R., Muxtarov E., Mirzaalimov A. Quyosh generator. Patent UZ. № FAP00623. 2011.
10. Gulomov J., Aliev R. Numerical analysis of the effect of illumination intensity on photoelectric parameters of the silicon solar cell with various metal nanoparticles // Наносистемы: физика, химия, математика. 2021. Т. 12. № 5. С. 569–574. https://doi.org/10.17586/2220-8054-2021-12-5-569-574
11. Gulomov J., Aliev R. Study of the temperature coefficient of the main photoelectric parameters of silicon solar cells with various nanoparticles // Journal of Nano- and Electronic Physics. 2021. V. 13. N 4. P. 04033. https://doi.org/10.21272/JNEP.13(4).04033
12. Pal S., Reinders A., Saive R. Simulation of bifacial and monofacial silicon solar cell short-circuit current density under measured spectro-angular solar irradiance // IEEE Journal of Photovoltaics. 2020. V. 10. N 6. P. 1803–1815. https://doi.org/10.1109/JPHOTOV.2020.3026141
13. Tiedje T., Engelbrecht D.A. Temperature dependence of the limiting efficiency of bifacial silicon solar cells // Proc. of the 47th IEEE

13. Tiedje T., Engelbrecht D.A. Temperature dependence of the limiting efficiency of bifacial silicon solar cells. *Proc. of the 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 2020, pp. 1789–1791. https://doi.org/10.1109/PVSC45281.2020.9300921
14. Wang H., Cheng X., Yang H. Temperature coefficients and operating temperature verification for passivated emitter and rear cell bifacial silicon solar module. *IEEE Journal of Photovoltaics*, 2020, vol. 10, no. 3, pp. 729–739. https://doi.org/10.1109/JPHOTOV.2020.2974289

Photovoltaic Specialists Conference (PVSC). 2020. P. 1789–1791. https://doi.org/10.1109/PVSC45281.2020.9300921
14. Wang H., Cheng X., Yang H. Temperature coefficients and operating temperature verification for passivated emitter and rear cell bifacial silicon solar module // IEEE Journal of Photovoltaics. 2020. V. 10. N 3. P. 729–739. https://doi.org/10.1109/JPHOTOV.2020.2974289

Authors

Avazbek Mirzaalimov — PhD Student, Andijan State University, Andijan, 170100, Uzbekistan, SC 55807268500, https://orcid.org/0000-0003-2846-1901, avazbek.mirzaalimov@mail.ru

Jasurbek Gulomov — Student, Andijan State University, Andijan, 170100, Uzbekistan, SC 57221531752, https://orcid.org/0000-0001-7516-987X, jasurbekgulomov@yahoo.com
Rayimjon Aliev — D.Sc., Full Professor, Andijan State University, Andijan, 170100, Uzbekistan, SC 7102561277, https://orcid.org/0000-0003-1986-2199, alievuz@yahoo.com

Navruzbek Mirzaalimov — PhD Student, Andijan State University, Andijan, 170100, Uzbekistan, SC 57223835161, https://orcid.org/0000-0002-9264-3710, mirzaalimov90@mail.ru

Suhrob Aliev — PhD, Vice-rector, Andijan Machine-Building Institute, Andijan, 170100, Uzbekistan, SC 7006390145, https://orcid.org/0000-0002-4494-8261, suhrob_asr89@mail.ru

Авторы

Мирзаалимов Авазбек Алишерович — аспирант, Андижанский государственный университет, Андижан, 170100, Узбекистан, SC 55807268500, https://orcid.org/0000-0003-2846-1901, avazbek.mirzaalimov@mail.ru
Гуломов Жасурбек — студент, Андижанский государственный университет, Андижан, 170100, Узбекистан, SC 57221531752, https://orcid.org/0000-0001-7516-987X, jasurbekgulomov@yahoo.com
Алиев Райимжон — доктор технических наук, профессор, профессор, Андижанский государственный университет, Андижан, 170100, Узбекистан, SC 7102561277, https://orcid.org/0000-0003-1986-2199, alievuz@yahoo.com
Мирзаалимов Навруzбек Алишер угли — аспирант, Андижанский государственный университет, Андижан, 170100, Узбекистан, SC 57223835161, https://orcid.org/0000-0002-9264-3710, mirzaalimov90@mail.ru
Алиев Сухроб — PhD, проректор, Андижанский машиностроительный институт, Андижан, 170100, Узбекистан, SC 7006390145, https://orcid.org/0000-0002-4494-8261, suhrob_asr89@mail.ru

Received 30.10.2021
Approved after reviewing 10.12.2021
Accepted 31.01.2022

Статья поступила в редакцию 30.10.2021
Одобрена после рецензирования 10.12.2021
Принята к печати 31.01.2022



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2022-22-1-33-46
УДК 004.04

Детектирование состояния зевоты у водителя транспортного средства при помощи модели сверточной нейронной сети

Игорь Борисович Лашков

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербург, 199178, Российская Федерация

igor-lashkov@ya.ru✉, https://orcid.org/0000-0001-6418-4660

Аннотация

Предмет исследования. Среди факторов, потенциально являющихся причиной возникновения дорожно-транспортных происшествий, является усталость водителя, накапливающаяся во время поездки или присутствующая еще до ее начала. Распространенный признак усталости или утомления у водителя транспортного средства — зевота. Выявление признаков зевоты у человека способно охарактеризовать состояние усталости. Методы компьютерной обработки изображений активно применяются в задачах детектирования открытости рта и определения зевоты у человека. Однако такой подход обладает перечнем недостатков, к которым относятся различные окружающие условия работы и множество ситуационных вариантов зевоты для разных людей. Представлена схема детектора выявления признаков зевоты, заключающаяся в обработке изображений лица водителя с использованием методов анализа данных, компьютерной обработки изображений и сверточной модели нейронной сети. **Метод.** Сущность предложенного метода заключается в детектировании признаков состояния зевоты в поведении водителя в кабине транспортного средства на основе анализа последовательности изображений, полученных с видеокамеры. Показано, что возникновение состояния зевоты водителя сопровождается широким и продолжительным открытием рта. Длительная открытость рта сигнализирует о возникновении признаков зевоты у человека. Представлена концептуальная модель обнаружения открытости рта у человека, и разработана схема обработки и полуавтоматической разметки наборов данных YawDD и Kaggle Drowsiness Dataset. Разработанная модель сверточной нейронной сети показала точность равную 0,992 и полноту — 0,871 на тестовом 10 % наборе данных. **Основные результаты.** Предложенная схема детектора выявления призраков зевоты апробирована на тестовой выборке видео, сформированной набором данных YawDD: Yawning Detection Dataset. Данная схема детектирования успешно обнаружила 124 зевка среди всех видеофайлов из тестового набора данных. Доля правильно классифицированных объектов составила 98,2 %, точность — 96,1 %, полнота — 98,4 %, а F-мера — 97,3 % при определении признаков зевоты у водителя транспортного средства. **Практическая значимость.** Детектирование признаков зевоты в поведении водителя позволяет уточнить о нем информацию, и тем самым повысить эффективность существующих систем мониторинга водителя в кабине транспортного средства, ориентированных на предотвращение и снижение риска возникновения дорожно-транспортных происшествий. Предлагаемый подход может быть использован совместно с другими технологиями мониторинга поведения водителя при построении интеллектуальной системы поддержки водителя.

Ключевые слова

транспортное средство, водитель, зевота, камера, мониторинг, обработка информации

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-06099 (разработка методов поиска уязвимостей интерфейсов взаимодействия человека с искусственным интеллектом транспортной среды «умного города»). Концептуальная модель мониторинга поведения водителя была поддержана бюджетной темой СПб ФИЦ РАН № FFZF-2022-0005.

Ссылка для цитирования: Лашков И.Б. Детектирование состояния зевоты у водителя транспортного средства при помощи модели сверточной нейронной сети // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 33–46. doi: 10.17586/2226-1494-2022-22-1-33-46

© Лашков И.Б., 2022

Detection of yawning in driver behavior based on a convolutional neural network

Igor B. Lashkov

St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

igor-lashkov@ya.ru✉, https://orcid.org/0000-0001-6418-4660

Abstract

Among the factors that usually cause road accidents in the world is driver fatigue, which accumulates during the trip or is present even before it begins. One of the most common signs of fatigue or tiredness of a vehicle driver is yawning. The detection of signs of yawning in human behavior is potentially able to further characterize its state of fatigue. Computer image processing methods are actively used to detect the openness of the mouth and yawning for a person. However, this approach has many disadvantages, which include different environmental conditions and a variety of situational yawning options for different people. The paper presents a scheme of a detector for determining signs of yawning, which is focused on processing images of the driver's face using data analysis methods, computer image processing, and a convolutional neural network model. The essence of the proposed method is to detect yawning in the driver's behavior in the cabin of a vehicle based on the analysis of a sequence of images obtained from a video camera. It is shown that the driver's yawning state is accompanied by a wide and prolonged openness of the mouth. Prolonged openness of the mouth signals the appearance of signs of yawning. A conceptual model for detecting the openness of the mouth for a vehicle driver is presented and a scheme for processing and labeling the YawDD and Kaggle Drowsiness Dataset datasets is developed. The developed convolutional neural network model showed an accuracy of 0.992 and recall of 0.871 on a test 10 % data set. The proposed scheme for detecting the yawning state has been validated on a test video subset extracted from the YawDD: Yawning Detection Dataset. This detection scheme successfully detected 124 yawns among all video files from the test dataset. The proportion of correctly classified objects is 98.2 % accuracy, precision is equal to 96.1 %, recall is 98.4 %, and F score is 97.3 % while detecting signs of yawning in driver behavior. Detecting signs of yawning in the driver's behavior allows one to clarify information about the driver and thereby to increase the effectiveness of existing driver monitoring systems in the vehicle cabin, aimed at preventing and reducing the risk of road accidents. The proposed approach can be combined with other technologies for monitoring driver behavior when building an intelligent driver support system.

Keywords

vehicle, driver, yawning, camera, monitoring, information processing

Acknowledgements

The research was carried out with the financial support of the RFBR within the scientific project No. 19-29-06099 (development of methods for finding vulnerabilities in human interaction interfaces with artificial intelligence of the "smart city" transport environment). The conceptual model of driver behavior monitoring was supported by the budget theme of the SPC RAS No. FFZF-2022-0005.

For citation: Lashkov I.B. Detection of yawning in driver behavior based a convolutional neural network. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 33–46 (in Russian). doi: 10.17586/2226-1494-2022-22-1-33-46

Введение

Дорожно-транспортные происшествия (ДТП) по-прежнему остаются одной из серьезных проблем в мире. Сложно найти страну, которую не затронули дорожные инциденты. Тысячи людей ежегодно погибают или получают серьезные ранения по причине того, что водители засыпают за рулем. Согласно аналитическим отчетам Всемирной организации здравоохранения, увеличение средней скорости движения на 1 % приводит к увеличению числа аварий со смертельным исходом на 4 % и числа аварий с серьезными травмами — на 3 %[1]. Согласно публично открытым отчетам, в Европе за 2019 год в результате ДТП погибло почти 23 000 чел.[2]. В соответствии с отчетом национального управления безопасности дорожного движения (БДД) США (NHTSA), более 91 000 зарегистрированных полицией аварий в 2017 г. были связаны с засыпанием водителей за рулем транспортного средства (ТС)[3]. Результаты расследования, проведенного Национальным управлением БДД [1] в США, показали, что каждый третий респондент испытывал усталость во время вождения автомобиля. Среди тех, у кого были инциденты с вождением в состоянии сонливости, 10 % признались, что у них были такие инциденты в течение последнего месяца вождения. Установлено, что усталость водителя приводит к возникновению 22 % ДТП. В отсутствие какого-либо предупреждения усталость водителя повышает вероятность столкновения в шесть раз, чем при вождении в нормальном состоянии. Таким образом, исследования в области распознавания усталости водителя, несомненно, являются необходимыми для повышения БДД. За последние десятилетия предложено множество методов определения усталости водителей [2, 3], чтобы обратить внимание водителей на безопасное управление автомобилем и, в целом, повысить БДД. Поведенческие

[1] WHO, Road traffic injuries [Электронный ресурс]. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (дата обращения: 09.08.2021).

[2] Road safety: Europe's roads are getting safer but progress remains too slow | Mobility and Transport [Электронный ресурс]. URL: https://ec.europa.eu/transport/media/news/2020-06-11-road-safety-statistics-2019_en (дата обращения: 09.08.2021).

[3] N. Administration, «Drowsy Driving», NHTSA, 2019 [Электронный ресурс]. URL: https://www.nhtsa.gov/risky-driving/drowsy-driving/ (дата обращения: 10.08.2021).

характеристики водителей при утомительном вождении включают моргание, кивание, закрытие глаз и зевоту. Среди водителей автотранспорта зевота – одна из ключевых форм проявления усталости при вождении [4].

Усталость обычно воспринимается как чувство сонливости. В контексте, связанном с работой человека, усталость — состояние умственного и/или физического истощения, которое снижает способность безопасно и эффективно выполнять работу [5]. Усталость может быть вызвана рядом факторов, включая недостаточный отдых, чрезмерную физическую или умственную активность, нарушения режима сна или чрезмерный стресс. Существует множество последствий, которые могут повлиять на отдельных людей и которые считаются непосредственно связанными с усталостью, такими как нарушение зрения, снижение зрительно-моторной координации, низкая мотивация, сниженное внимание, медленные рефлексы, недостаточная реакция или неспособность сосредоточиться. Усталость считается основной причиной ДТП, приводящих к человеческим жертвам. Данные за 2010 год от Министерства транспорта Великобритании[1] указывают на 1850 погибших и 22 660 серьезно раненых, при этом 20 % от общего числа несчастных случаев [6] произошли по причине усталости водителя за рулем автомобиля. Федеральное управление автомобильных дорог США утверждает, что около 100 000 аварий в год происходят по причине усталости водителя за рулем ТС [7].

Предмет исследования

Интеллектуальные технологии в транспорте, выполняющие мониторинг и оказывающие помощь водителю в управлении ТС [8], в последние годы стали активно исследуемой тематикой в области повышения БДД. В связи с этим автопроизводители стремятся оборудовать свои автомобили различными видами передовых вспомогательных систем для водителя, которые способствуют безопасной эксплуатации и помогают водителю в различных небезопасных ситуациях. Контроль безопасности включает: съезд и удержание в полосе движения автомобиля; предупреждение о возникающих препятствиях, ТС и пешеходах; об экстренном торможении в случае, если водитель не предпринимает действий, чтобы избежать столкновения; визуализацию дорожных знаков на приборной панели автомобиля и др. Эффективный мониторинг поведения водителей позволяет повысить БДД не только для самих водителей, но и дает преимущества страховым и логистическим компаниям с точки зрения сбережений и потребностей общества.

Сонливость является немаловажным физиологическим состоянием человека, по причине которого происходит около 24 % всех ДТП в мире [9]. Вождение автомобиля в состоянии сонливости повышает риск возникновения дорожной аварии в 4–6 раз в сравнении с состоянием бдительности. Одним из предупреждающих сигналов, по возникновению которого можно судить о состоянии усталости у водителя, является зевота. Зевота — активно исследуемая тематика в научной сфере, предоставляющая неинвазивный механизм определения физиологического состояния человека.

Исследователи предложили различные методы для обнаружения состояния зевоты. Существующие методы обнаружения зевоты сосредоточены на измерении открывания рта при анализе одного статического изображения в единицу времени. При исследованиях применяются методы компьютерной обработки изображений для поиска и локализации области рта человека, необходимых для выявления признаков зевоты, основываясь на различиях в характеристиках между открытым и закрытым ртом. Нейронные модели глубокого обучения начинают получать распространение в области систем интеллектуальных транспортных средств [10] и, в том числе, находят применение при обнаружении состояния открытости и закрытости рта. Подходы, основанные на работе со статическим изображением, сталкиваются со значительными трудностями и не всегда могут обеспечить приемлемое качество и скорость распознавания признаков небезопасного поведения по одному изображению. Для решения перечисленных трудностей в настоящей работе предложен подход к выявлению зевоты, основанный на распознавании набора статических изображений лица. Временной промежуток от возникновения признаков усталости у водителя за рулем с признаками зевоты до серьезных травм, включая летальный исход, может быть достаточно коротким. Таким образом, эффективность алгоритма обнаружения зевоты у водителя при управлении ТС во времени имеет решающее значение. Эффективность алгоритма — свойство алгоритма, которое связано с численной оценкой его качества, которая, как правило, определяется при помощи метрик полноты, точности, правильно классифицированных объектов, F-меры.

Цель работы — разработка подхода для детектирования признаков зевоты в поведении водителя в кабине ТС на основе обработки и анализа изображений с видеокамеры. Предложенный подход основан на мониторинге поведения водителя в кабине ТС и позволяет заблаговременно выявить признаки небезопасного поведения человека при вождении.

Научная значимость работы заключается в разработке подхода, ориентированного на выявление признаков состояния зевоты у водителя ТС с использованием алгоритмов анализа и обработки данных, позволивших сформировать единый датасет данных. Датасет включает в себя: наборы данных, представленные открытым и доступным для загрузки источником видео YawDD, а также аннотированным набором данных на платформе Kaggle, с разделением на классы «зевающий» и «не зевающий» водитель; модели нейронной сети, обученной с использованием данного сформированного датасета, цель которого — детектирование открытости рта на изображении; схемы выявления признаков зевоты у водителя ТС на временном промежутке. Одновременное использование данных о визуальном (внешнем) состоянии водителя из разных источников данных способно повысить точность и надежность работы системы под-

[1] D. f. T. UK, Reported Road Casualties in Great Britain: 2010 Annual Report, Department for Transport UK, United Kingdom, Statistical Release – 29 September 2011.

держки водителя. Данное исследование представляет продолжение существующих исследований авторов настоящей работы в области детектирования усталости водителя за рулем ТС [11]. Полученные результаты могут быть использованы при построении алгоритма выявления признаков состояния зевоты для системы мониторинга поведения водителя в кабине ТС [12], например Drive Safely [13].

Исследования в области систем мониторинга состояния зевоты водителя

Существующие методы обнаружения зевоты можно условно разделить на две категории: методы компьютерной обработки изображений и методы, использующие в своей работе алгоритмы машинного обучения и математические модели нейронных сетей.

Методы компьютерной обработки изображений. Такие методы основаны на извлечении геометрических и цветовых признаков. Исследователи [14] предложили метод, который обнаруживает признаки зевоты через изменения в соотношении сторон области границ рта. Область границ рта измеряется на основе самой темной области между губами. Такая область сегментируется с использованием адаптивного порога. Для выявления ситуации, когда при зевоте рот водителя закрыт рукой, использован оператор локальных бинарных шаблонов (Local Binary Patterns) для классификации и извлечения признаков области рта, с целью дальнейшей оценки такого состояния при помощи классификатора нейронной сети. Чтобы убедиться, что закрытие рта рукой вызвано именно зевотой человека, а не каким-то другим его действием или состоянием, производится анализ появления морщин на лице. Была проведена серия экспериментов, чтобы найти предельное значение соотношений сторон. Когда соотношение сторон рта меньше фиксированного порогового значения, состояние рта классифицируется как зевота. Состояние зевоты детектируется, когда происходят значимые искажения в рассмотренных областях лица. Предложенный алгоритм обнаруживает зевоту при открытии рта длительностью от 3 до 5 с. Если в рассматриваемый момент открытость рта меньше порогового значения, рот считается закрытым. Тогда на следующем шаге происходит выявление состояния закрытия рта рукой человеком и наличие искажения в верхней области лица. Если оба условия выполняются, алгоритмом фиксируется состояние зевоты у человека, а в ином случае — ее отсутствие в рассматриваемый момент.

Достаточно распространенным подходом к определению состояния зевоты у человека — наблюдение за величиной открытости рта на протяжении некоторого промежутка времени, приблизительно равного нескольким секундам. При решении задачи детектирования признаков зевоты большая часть существующих исследований концентрируется на детектировании области лица человека, и затем остальных его частей, таких как рот, нос и глаза, с последующим обнаружением состояния зевоты. Один из популярных подходов к распознаванию лиц — применение метода Виола–Джонса (Viola–Jones) [15, 16], представленного в библиотеке компьютерного зрения OpenCV [17], и обученного детектора Хаара. Существенным недостатком использования такого детектора лиц является то, что идентификация лиц среди множества кандидатов, вырезанных с кадров при различном масштабе с помощью скользящего окна и других методов, сопровождается значительной вычислительной нагрузкой на целевые устройства. Также стоит отметить, что данный каскадный детектор лиц способен находить область лиц только при относительно небольшом определенном угле обзора камеры. Эффективность работы такого детектора лиц непосредственно зависит от внешних окружающих условий, к которым относятся окружающая освещенность, наличие бликов, теней и др. При недостаточной освещенности или наличии каких-либо объектов, частично или полностью перекрывающих видимость лица человека в кадре, точность и другие показатели работы данного алгоритма могут быть значительно снижены.

Работы [18, 19] основаны на детектировании открытости рта по лицевым характеристикам и множеству 68 опорных точек на лице. Отметим работу [20], в которой предложен подход, заключающийся в локализации и сегментации области интереса по кадрам входного видеофайла. Алгоритм Виолы–Джонса использован для определения местоположения области лица и рта человека на изображении. Чтобы провести границу между примерами положительного и отрицательного класса, при которых лицо человека на изображении присутствует и там, где его нет, анализ и классификация объектов выполнены при помощи алгоритма машинного обучения адаптивного бустинга Adaboost [21]. С целью обнаружения состояния зевоты область губ была локализована и выбрана из изображения. Для определения краев губ использован метод активных контуров [22]. Принцип активных контуров был впервые предложен в 1998 г. [23]. Активный контур — параметрическая кривая, которая стремится изменить свое положение таким образом, чтобы ее энергия оказалась минимальной. Далее к полученному сигналу применяется функция пересечения нуля, чтобы определить состояние зевоты у водителя.

Методы, основанные на применении математических моделей нейронных сетей. В других исследованиях активно прорабатываются подходы, основанные на использовании не только методов компьютерной обработки изображений, но и глубоких нейронных сетей. В [24] описан алгоритм для определения усталости и сонливости водителя по признакам зевоты и состоянию закрытости глаз. Обнаружение открытости рта и зевоты водителя становится возможным благодаря нахождению кончика носа путем получения информации о глубине лица на изображении. Зная расположение кончика носа, нижняя часть лица может быть отделена (отрезана) от всего изображения. На полученном изображении с помощью адаптивной маски, которая действует как детектор границ, область рта обнаружена с точностью 91 %. Два метода, один из которых основан на максимальной глубине, а другой — на модели активного контура, были применены для задачи поиска области рта, и их результаты были объединены для выявления открытости рта, продемонстрировав

точность в 95 %. Несмотря на то, что среднее количество последовательных кадров в видео, на которых обнаружена зевота, составляет 85, в качестве порогового значения видеокадров для определения состояния зевоты было выбрано 45, обозначающих число таких кадров, на которых рот человека открыт, описывающих проявление признаков зевоты. Далее, как только найдена верхняя область лица, происходит обнаружение области глаз на изображении при помощи алгоритма Виолы–Джонса. Затем, используя предварительно обученную сверточную нейронную сеть для выделения признаков VGG (Visual Geometry Group), открытость/закрытость глаз была классифицирована, показав точность в 90 %. В такой нейронной модели были обучены только последние два полно связанных слоя и слой softmax для распознавания открытости/закрытости глаз человека. Входное изображение для нейронной сети имеет размер 24 × 24 пикселов. Исследователи выбрали значение для показателя PERCLOS (PERcentage of eye CLOSure), равным доле времени, в течение которого глаза закрыты, на уровне в 0,5 (т. е. глаза закрыты 15 кадров из доступных 30). Также, с целью снижения количества ошибок второго рода (false negative), было предложено игнорировать 20 % кадров (что соответствует трем кадрам), на которых было обнаружено, что глаза водителя закрыты, среди всех 15 кадров, на которых отмечено, что глаза закрыты. Путем объединения результатов алгоритмов обнаружения зевоты и детектирования закрытости глаз принимается решение о предупреждении водителя. Использование информации о глубине, которая не чувствительна к изменениям освещенности, является основным из преимуществ предложенного решения. Разработанный алгоритм не сработает в тех случаях, когда водитель прикрывает рот рукой или он носит очки.

Большинство современных методов обнаружения зевоты основаны на одном статическом изображении, и могут вызывать ложные срабатывания, поскольку зевота представляет собой действие, а не статическое состояние. Так, для решения этой проблемы в работе [25] представлен метод детектирования состояния усталости, осуществляющий мониторинг уровня сонливости, и способный отображать выходные данные сверточной нейронной сети (CNN) на уровне кадров в вероятность (уверенность) нахождения человека в таком состоянии. Задача в данном исследовании сформулирована как бинарная классификация, заключающаяся в определении класса для каждого кадра видео: «зевающий» или «не зевающий». В качестве классификатора выступает глубокая нейронная сеть, построенная на основе архитектуры сети ResNet-50. Обучение нейронной модели выполнено следующим способом. Параметры данной нейронной модели инициализированы с использованием параметров, взятых от модели, обученной на наборе данных ImageNet. Обучение модели произведено с использованием алгоритма стохастического градиентного спуска для оптимизации функции потерь перекрестной энтропии. На этапе тестирования нейронной модели генерируется 10 случайных примеров тестовых изображений, и окончательный результат прогнозирования записан путем голосования большинством голосов из 10 прогнозов для этих примеров. При проведении начальных экспериментов использован открытый набор данных YawDD [26], содержащий 322 видеофрагмента, записанных внутри кабины ТС. Результаты работы [26] подтверждают достоверность гипотезы о том, что обнаружение зевоты следует проводить в течение длительного непрерывного захвата видео, а не на отдельно взятых изображениях. Среди недостатков данного метода стоит выделить нестабильное качество его работы в условиях низкого уровня освещенности.

В [27] предложен подход к обнаружению зевоты на основе видеофайлов. Такой подходит состоит из трех основных компонентов. Первый компонент представляет собой модуль предварительной обработки входных данных. Этот этап включает распознавание лиц и сегментацию, нормализацию размера изображения и шумоподавление. Алгоритм Виолы–Джонса используется для обнаружения и сегментации области лица, чтобы уменьшить влияние информации о внешнем виде фона на изображении. Размер последовательных кадров должен быть одинаковым после сегментации. Затем однородные последовательные кадры исключаются с помощью алгоритма быстрой медианной фильтрации для уменьшения шума изображения. Обработанные кадры подаются на вход второго компонента — модуля выбора ключевых кадров. Ключевые кадры выбираются посредством вычисления гистограммы изображения и обнаружения выбросов для устранения избыточных кадров. Далее ключевые кадры поступают на вход в нейронную сеть распознавания лиц 3D-LTS (3D deep learning network with a low time sampling characteristic). В качестве третьего компонента выступает модуль классификации движений лица. В этом модуле 3D-сеть глубокого обучения характеристикой LTS построена с использованием ключевых видеокадров для выявления различных незначительных движений лица. LTS — это характеристика изменения скорости дискретизации временных измерений для улучшения распознавания едва заметных лицевых действий. 3D-LTS можно разделить на две части. Кратковременные пространственно-временные характеристики лица извлекаются с помощью 3D-свертки. Двунаправленная долговременная кратковременная память (LSTM, Bidirectional long short-term memory) используется для создания долговременных последовательностей с помощью извлеченных кратковременных пространственно-временных функций для улучшения эффекта распознавания. Распознавание лицевых действий способно обнаруживать признаки зевоты у человека и уменьшать количество ложных срабатываний.

Существующие исследования показали, что усталость водителя — одна из основных причин возникновения ДТП. Исследования, основанные на применении методов компьютерного зрения для выявления признаков физической усталости водителя, таких как зевота, продемонстрировали хороший потенциал при разработке систем технологий выявления таких состояний. Методы, основанные на компьютерном зрении, популярны, хорошо изучены в научном сообществе и зачастую не требовательны к вычислительным ресурсам и применимы к изображению, полученному с

видеокамеры, установленной перед водителем. Такие методы не лишены недостатков, к которым относятся условия окружающей освещенности и индивидуальные различия в цвете и форме лица человека.

В настоящей работе предложено разработать новый подход к обнаружению зевоты, основанный на распознавании признаков проявления состояния зевоты с использованием специализированных подходов к анализу и обработке данных, и сверточной нейронной сети, обученной на разных наборах данных, позволяющей выделять опорные признаки и паттерны на изображениях. Особое внимание сосредоточено на подготовке набора данных, используемого для обучения модели, и архитектуре самой модели.

Подготовка данных для обучения глубокой нейронной сети определения открытости рта человека

Входными источниками данных для задачи детектирования величины открытости рта человека послужили следующие открытые информационные ресурсы.

1) Набор данных YawDD: Yawning Detection Dataset:
   - 322 видео в формате RGB, записанные при частоте в 30 кадров/с; по 3–4 видео соотносится с каждым водителем (4,8 ГБ); видеокамера установлена под зеркалом заднего вида;
   - 29 видеофайла в формате RGB, записанные при частоте в 30 кадров/с; каждое видео ассоциировано с каждым водителем; камера установлена напротив водителя, на бортовой панели автомобиля;
   - три вида вождения: нормальное вождение (нет разговоров); разговор или пение; зевота;
   - мужчины и женщины за рулем автомобиля, с очками/солнцезащитными очками и без, различные этнические группы;
   - рамки лица разного размера и смещения;
   - различные условия освещения и разные погодные условия (облачно, солнечно, и др.);
   - переменный угол поворота головы 0–30°;
   - неразмеченный набор данных.
2) Набор данных Kaggle: Drowsiness Dataset[1]:
   - 1450 изображений;
   - две категории: «зевающий»/«не зевающий»;
   - мужчины и женщины за рулем автомобиля, с очками/солнцезащитными очками и без, различные этнические группы;
   - переменный угол поворота головы 0–30°;
   - размеченный набор данных.

Поскольку набор данных YawDD состоит только из видеофайлов, неаннотированных покадрово, т. е. является неразмеченным, требуется классифицировать каждое изображение из видео на предмет присутствия одного из двух состояний у водителя: рот открыт и рот закрыт. Схема подготовки YawDD датасета и формирования собственного набора данных представлена на рис. 1.

На вход данному алгоритму, изображенному на схеме (рис. 1), поступает список RGB видеофайлов в разрешении 640 × 480 пикселов с частотой 30 кадров/с, на которых изображен водитель за рулем ТС. Данные видеофайлы записаны в двух положениях видеокамеры: со стороны заднего зеркала и в анфас (фронтально). С целью уменьшения шума на изображении и сужении области интереса изображения, конечным результатом в сформированном наборе данных должно стать изображение самого лица человека без каких-либо посторонних объектов. Чтобы выполнить поиск и обнаружение положения лица на изображении требуется воспользоваться методами компьютерной обработки изображений или возможностями глубоких нейронных сетей. Среди алгоритмов поиска объекта лица человека на изображении выбор происходит между Dlib [28], Faceboxes [29], SSD (Single Shot MultiBox Detector) [30],

[1] Drowsiness_dataset | Kaggle [Электронный ресурс]. URL: https://www.kaggle.com/dheerajperumandla/drowsiness-dataset (дата обращения: 06.08.2021).

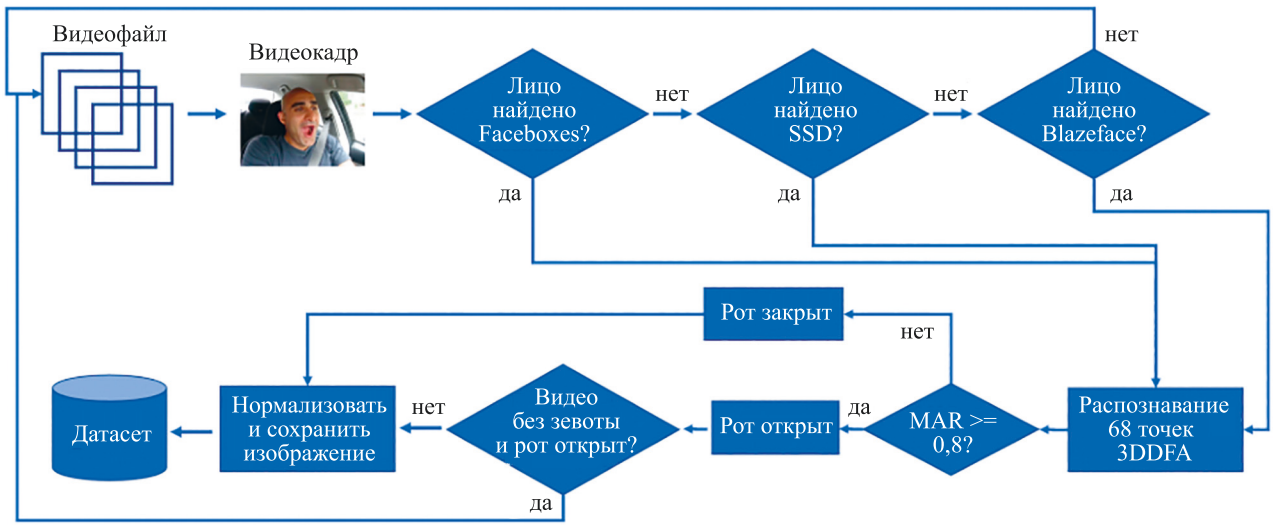


Рис. 1. Схема подготовки набора данных из YawDD

Fig. 1. Scheme of processing of the YawDD dataset

Blazeface [31]. Выполнено сравнение эффективности и точности работы вышеперечисленных алгоритмов поиска лиц на подмножестве тестовых образцов из набора данных YawDD. Алгоритм поиска Dlib показал на порядок хуже результаты по сравнению с остальными решениями. Среди недостатков фреймворка Dlib можно выделить неспособность найти лица при переменном и недостаточном освещении, при повороте головы влево или вправо, а также некорректно найденные объекты на изображении, не относящиеся к лицу. Остальные решения для поиска лица на изображении показали достаточно стабильные результаты по точности работы. Для задачи поиска лица на изображении использованы перечисленные алгоритмы, поскольку разные решения способны независимо друг от друга находить рамку лица одного и того же человека с незначительным смещением и различиями по высоте и ширине. Данные о рамке лица человека, полученные в результате работы разных алгоритмов, способны разнообразить выборку изображений и улучшить качество всего набора данных, используемого в дальнейшем для обучения нейронной сети и для предсказания открытости рта на целевом изображении. Отметим, что алгоритмы поиска лиц на изображениях исполняются поочередно для каждого последующего изображения, тем самым позволяя получить различные рамки лица. Как только лицо найдено одним из трех алгоритмов, оно передается на вход алгоритму 3D Face alignment [32], отвечающему за поиск 68 лицевых опорных точек. Такой набор точек позволяет определить положение и размеры рта на изображении и вычислить долю открытости рта MAR (mouth aspect ratio):

$$\mathrm{MAR} = \frac{\|p_2 - p_8\| + \|p_3 - p_7\|\,\|p_4 - p_6\|}{2\|p_1 - p_5\|},$$

где p_1–p_8 — лицевые точки, соответствующие расположению, показанному на рис. 2.

По результатам экспериментов на наборе данных YawDD получено пороговое значение показателя MAR = 0,8 для определения события открытости рта. На рис. 3 приведен пример возникновения события зевоты, характеризующейся продолжительным открытием рта на временном промежутке, состоящим из последовательности видеокадров с 55 по 110, на которой показатель MAR выше установленного порога.

Отметим, что открытость рта человека при возникновении состояния зевоты характеризуется большей величиной и иной продолжительностью непрерывной

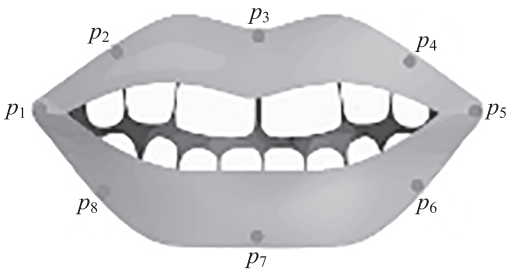


Рис. 2. Положение лицевых точек для области рта (MAR)

Fig. 2. Location of facial points related to the mouth (MAR)

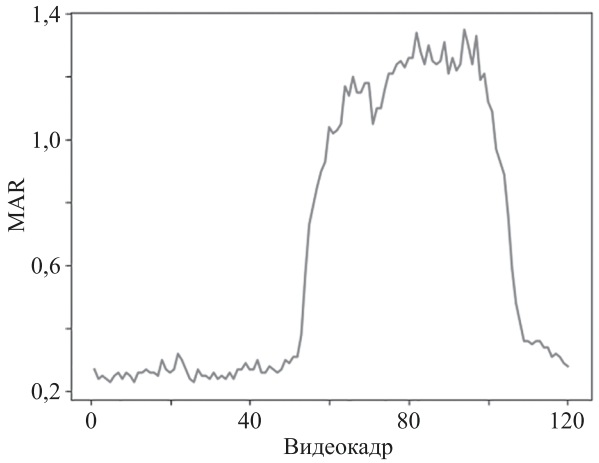


Рис. 3. Пример проявления состояния зевоты на временном промежутке

Fig. 3. Example of yawning occurrence within time period

открытости рта (рис. 3), чем при других видах активностях человека, таких как «разговор» (рис. 4) или «пение». На рис. 4 приведен пример возникновения состояния «разговор» на временном промежутке, характеризующимся непродолжительным открытием рта на данном промежутке, состоящим из последовательности видеокадров от 100 до 200, на котором показатель MAR изменяется незначительно и варьируется в интервале от 0,35 до 0,65.

Видео, полученные из набора данных YawDD, описываются очень кратко и могут только подсказать, зевает ли человек на записанном видеофрагменте или нет. Если по имеющемуся описанию видео становится понятно, что человек не зевает за рулем автомобиля, и при этом алгоритм фиксирует открытость рта у водителя, которая может определяться при разговоре с пассажиром или по телефону, такое изображение пропускается из обработки, и алгоритм переходит к следующему кадру из видео. Если же такое условие не выполняется, изображение лица человека проходит предфинальную нормализацию — переводится в формат полутонового

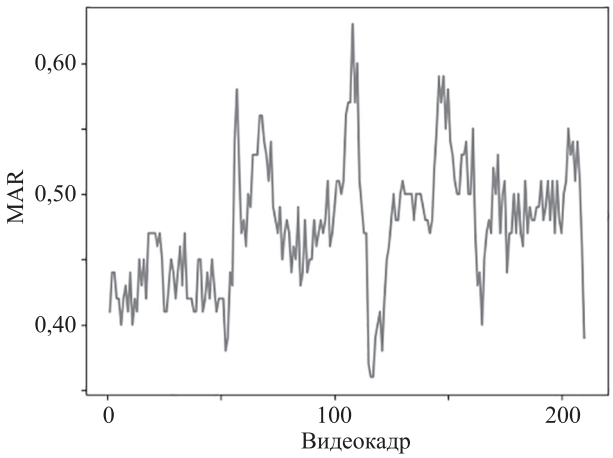


Рис. 4. Пример разговора человека на временном промежутке

Fig. 4. Example of a person talking within time period

изображения (grayscale), трансформируется в размер ширины и высоты 100 × 100 пикселов и сохраняется в новый полученный набор данных, запоминая данные о показателе MAR.

В результате проведенной обработки набора данных YawDD Dataset количество изображений лиц водителя с закрытым ртом составило приблизительно 49 000 изображений, а с открытым ртом — 11 000 изображений. Также в результирующий набор данных (на схеме компонент «Датасет») вошли изображения лиц водителей, представленные существующим открытым набором данных «Kaggle: Drowsiness Dataset»: 1448 изображений лица водителя, из которых 723 — с открытым ртом, и 725 изображений — с закрытым.

Концептуальная модель определения открытости рта водителя

Разработанная концептуальная модель определения открытости рта представлена на рис. 5. Подготовленные наборы примеров изображений лиц должны быть перемешены, чтобы исключить неявный порядок следования изображений и равномерно распределить их для следующего этапа обработки. Как только наборы данных для обучения и тестирования нейронной сети готовы, на следующем этапе происходит разделение данных на три разных подмножества выборок: обучающая, проверочная и тестовая (training-validation-test). На обучающей выборке происходит первичная настройка параметров модели, ее весов и смещений посредством применения алгоритма градиентного спуска. Проверочная выборка — это та выборка, которая не принимала участие в процессе обучения модели и на которой оценивалось качество созданной модели и оптимизация ее гиперпараметров. Наборы данных, составляющих тестовую выборку, использовались для независимой оценки получившейся нейронной модели, получения конечных результатов качества модели и выбора наилучшей модели из некоторого множества. В качестве обучающей, проверочной и тестовой выборок для нейронной модели использован единый набор данных, полученный из подготовленного на предыдущем этапе набора данных «YawDD Dataset», размеченного в полуавтоматизированном режиме, а также открытого размеченного набора данных «Kaggle Drowsiness Dataset». С учетом особенностей подготовки исходных наборов данных, а также на основе общепринятых практик в машинном обучении, наборы обучающей, проверочной и тестовой выборок были определены в соотношениях 80, 10 и 10 % соответственно.

Разработанная модель имеет структуру сверточной глубокой нейронной сети (DNN, Deep Neural Network). При разработке данной модели за основу взята архитектура модели MNIST, специализирующейся на распознавании рукописных цифр [33]. Модификация архитектуры MNIST, по большей мере, основана на корректировке весов слоев Flatten, Dense, Dropout, представляющих собой эффективный и достаточно простой способ уменьшения эффекта переобучения модели и улучшения обобщения рассматриваемых при обучении образцов изображений. Так, например, наилучший показатель для слоя Dropout составил 60 %.

Количество параметров, участвующих при обучении нейронной сети, примерно составляет 440 000. Обучение финальной версии модели выполнено на 80 эпохах (epochs) при количестве обучающих примеров за одну итерацию (batch size) — 64, а в качестве алгоритма оптимизации использована реализация Adam с шагом 0,001. Построение нейронной сети произведено при помощи фреймворка Keras совместно с TensorFlow. Функция $f(p)$ определяет выходной результат нейронной сети, описываемый вещественным числом в диапазоне от 0 до 1, где 0 — модель уверена, что рот человека закрыт; а 1 — рот открыт.



Рис. 5. Концептуальная модель обнаружения открытости/закрытости рта

Fig. 5. Reference model for the detection of mouth openness/closeness

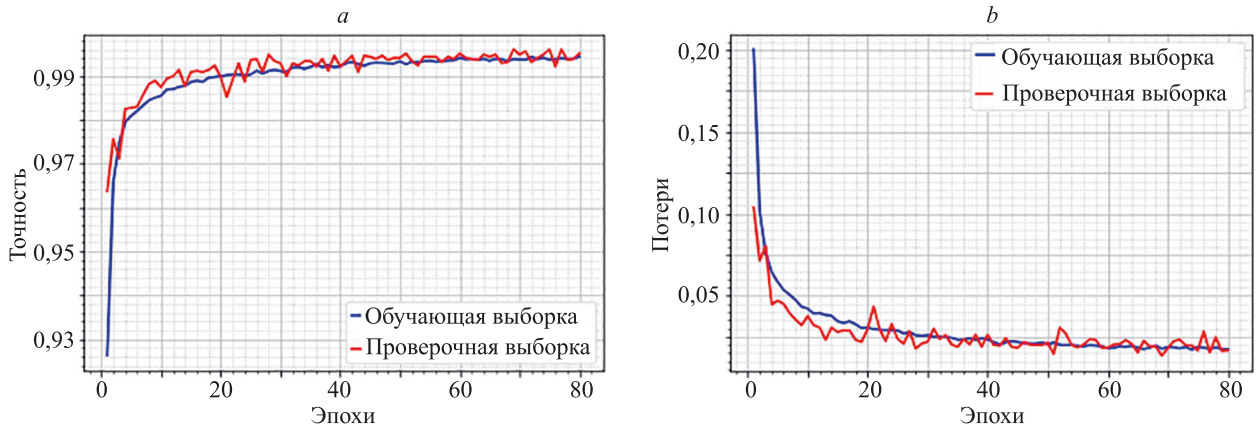


Рис. 6. Графики точности нейронной сети (*a*) и потерь (*b*) на этапе обучения

Fig. 6. Plots for accuracy of neural network (*a*) and loss (*b*) on the training stage

Эффективность обучения нейронной сети приведена на рис. 6. Можно заметить, что эффект переобучения модели сведен к минимуму, а точность (доля правильных ответов, accuracy) обучающей и проверочной выборок составили более 0,99 (рис. 6, *a*), а потери (loss) (рис. 6, *b*) около 0,02 единиц.

Оценка работы модели нейронной сети по детектированию открытости рта человека

Оценка полученной модели сверточной нейронной сети произведена на тестовой 10 % выборке, образованной набором данных YawDD и размеченным датасетом Kaggle Dataset. В табл. 1 при сравнении результатов работы модели нейронной сети на разных наборах данных использованы метрики доли правильно классифицированных объектов (accuracy), точности (precision), полноты (recall) и F-меры. Отметим, что на сегодняшний день в открытом доступе отсутствуют размеченные наборы данных, одновременно включающие как состояние зевоты, так и другие действия человека, представляющие собой «разговор», «улыбку» и другие проявления эмоций. Таким образом, оценка получившейся модели глубокого обучения на существующих наборах данных является не совсем корректной. Несмотря на это, для получения более полной оценки разработанной нейронной модели, модель была апробирована на открытом наборе данных[1], содержащем изображения лиц людей в состоянии «разговор» и «улыбка». Результаты проведенного эксперимента показали, что модель нейронной сети безошибочно детектирует отсутствие событий открытости рта, которые бы характеризовали проявление состояния зевоты для всех изображений из данного набора данных. Другой набор данных, использованный при оценке нейронной модели — Emotion Detection Dataset [34], опубликован в открытом доступе и содержит более 32 000 изображений, классифицированных по семи типам эмоций, проявляемых человеком: злость, раздражение, страх, радость, нейтральное состояние, грусть и удивление. При проверке всех изображений на отсутствие продолжительного открытия рта, характерного для состояния зевоты, из данного набора точность бинарного классификатора составила 0,806. Полученный результат объясняется в том числе тем, что все изображения в наборе представлены в недостаточном разрешении 48 × 48 пикселов, а модель нейронной сети обучена на изображениях лиц людей в разрешении 100 × 100 пикселов.

Из табл. 2 видно, что полученная модель нейронной сети превосходит по метрикам оценки качества модель, опубликованную в работе [11]. Традиционный подход, основанный на алгоритмах компьютерного зрения, к распознаванию доли открытости рта по 68 лицевым координатным точкам [35] показывает значительно хуже результат по всем используемым статистическим метрикам — точность, полнота, F-мера. Проведено сравнение с подходами, базирующимися на методах машинного обучения — на методе опорных векторов (Support vector machines, SVM) [36, 37]. Результаты оценки метода [36] показывают, что метод опорных векторов значительно превосходит по качеству подход [35], но не достигает результатов, полученных новой разработанной нейронной моделью.

Размер нейронной сети на файловом диске составил 1,7 МБ в формате ONNX. Исходный код нейронной модели, а также данные и остальные результаты сравнения моделей опубликованы в открытом репозитории Github[2].

Оценка скорости работы разработанной нейронной модели проведена на персональном ноутбуке, и на смартфонах различной мощности. Сравнение характеристик быстродействия работы полученной модели в различных режимах и настройках, включающих конфигурацию целевого устройства, используемый акселератор, формат представления нейронной модели, среднее время работы модели на устройстве и версию фреймворка машинного обучения для исполнения модели, приведено в табл. 3. Заметим, что время работы

[1] Talking Face Video [Электронный ресурс]. URL: https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html (дата обращения: 11.10.2021).

[2] HippoYD | Github [Электронный ресурс]. URL: https://github.com/iglaweb/HippoYD (дата обращения: 09.10.2021).

Таблица 1. Сравнение результатов работы модели нейронной сети на разных наборах данных
Table 1. Comparison of the results obtained from the neural network model on different datasets

Набор данных	Отрицательный класс («рот закрыт»)	Положительный класс («рот открыт»)	Метрика			
			Доля правильно классифицированных объектов	Точность	Полнота	F-мера
Тестовая выборка на основе YawDD и Kaggle Drowsiness Dataset	4900	1100	0,932	0,992	0,871	0,927
Talking Face Video	5 000	—	1,000	—	—	—
Emotion Detection Dataset [34]	32 000	—	0,806	—	—	—

Таблица 2. Сравнение работы предложенной модели нейронной сети и других подходов к детектированию открытости рта у человека
Table 2. Comparison of the neural network model and other approaches aimed at detecting mouth openness of a person

Модель	Метрика			
	Доля правильно классифицированных объектов	Точность	Полнота	F-мера
Предложенная в настоящей работе	0,932	0,992	0,871	0,927
Предыдущее исследование авторов настоящей работы [11]	0,740	0,910	0,540	0,680
Традиционный подход [35]	0,280	0,317	0,384	0,347
Метод опорных векторов [36]	0,803	0,921	0,662	0,770

Таблица 3. Сравнение быстродействия модели нейронной сети
Table 3. Comparison of model performance of a neural network model

Конфигурация	Акселератор	Формат модели	Среднее время, мс	TensorFlow/TFLite версия
Macbook Pro, CPU 2 GHz Quad-Core Intel Core i5	CPU	TFLite (Floating)	5	2.3
	CPU	TFLite (Quantized)	8	2.3
	CPU	Keras H5 (Floating)	30	2.3
	CPU	ONNX	2	2.3
	CPU	Frozen pb	4	2.3
	Wasm (Safari 14.0, Firefox 84)	TensorFlowJS	30	2.3
Смартфон Xiaomi MI8	GPU/CPU 3 потока	TFLite (Floating)	4	2.4
	CPU 3 потока	TFLite (Quantized)	10	2.4
Смартфон Xiaomi Redmi 9	GPU/CPU 3 потока	TFLite (Floating)	11	2.4
	CPU 3 потока	TFLite (Quantized)	9	2.4
Смартфон Samsung Galaxy S20+	GPU/CPU 3 потока	TFLite (Floating)	2	2.4
	CPU 3 потока	TFLite (Quantized)	2	2.4

нейронной модели на смартфоне редко составляет более 10 мс, что, в свою очередь, не превышает лимит количества обрабатываемых кадров в 30 кадров/с.

Схема детектора выявления признаков состояния зевоты по видеоряду с камеры

Определение состояния зевоты целесообразно рассматривать не на одном кадре, а с учетом последовательности изображений, полученных с видеокамеры. В качестве следующего шага после построения нейронной модели детектирования открытости рта, предложена схема детектора выявления признаков состояния зевоты у водителя ТС, показанная на рис. 7. Данная схема представляет собой демонстрационный пример детектирования состояния зевоты, при условии, что характеристики зевоты (вероятность открытости рта $p_1, p_2, \ldots, p_k$) могут быть распределены для каждого водителя и в каждый момент времени ($t_{e1}, t_{e2}, \ldots, t_{ei}$) по-разному. Состояние зевоты человека принято рас-

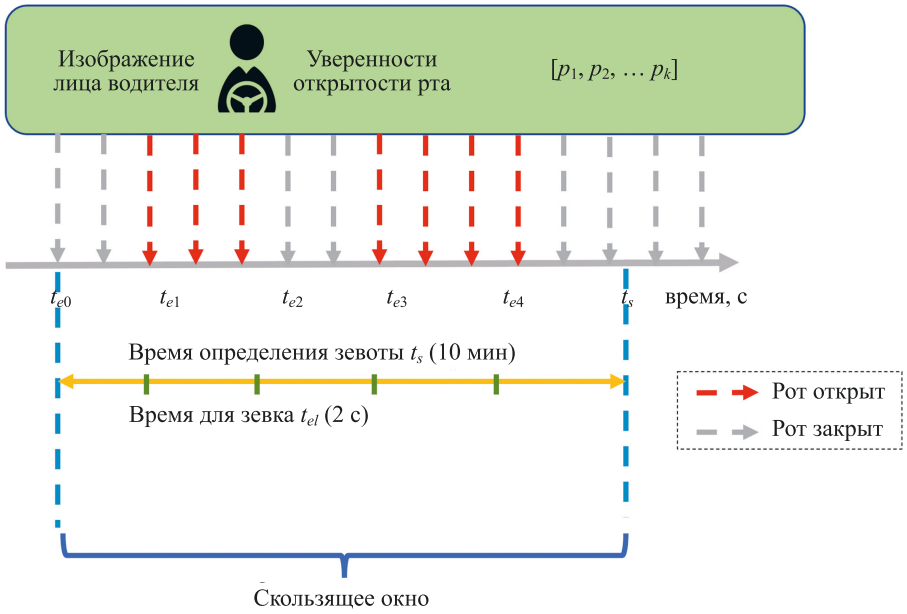


Рис. 7. Схема детектора выявления признаков состояния зевоты

Fig. 7. Scheme for the detection of yawning

сматривать на промежутке t_s равном 10 мин. В течение этого промежутка времени с использованием алгоритма скользящего окна [38] определяется число зевков, совершенных водителем. Если водитель совершает три зевка в течение данного интервала, фиксируется состояние зевоты, и водитель может быть оповещен о таком опасном состоянии. На основе существующих научных исследований время определения одного зевка t_{el} выбрано в промежутке до 2,5 с. При условии, что рот водителя был открыт более 80 % времени в рамках данного промежутка времени, принимается решение, что водитель совершил один зевок. Открытость рта водителя определяется на основе ранее представленной нейронной модели глубокого обучения, выходным значением для которой является уверенность открытости рта.

В качестве эксперимента по анализу эффективности работы предложенной схемы, она была апробирована на тестовой выборке видео, сформированной из набора данных YawDD. Общее количество видеофайлов, на которых отмечен хотя бы один зевок, составляет 90 видеофайлов. Размер тестовой выборки равен 35 видеофайлам, что составляет 10 % от количества всех видеофайлов в исходном наборе. Общее количество зевков, которые совершили водители, составило 128 раза. Данная схема детектирования успешно обнаружила 124 зевка среди всех видеофайлов из тестового набора данных. Доля правильно классифицированных объектов составила 98,2 %, точность — 96,1 %, полнота — 98,4 %, а F-мера — 97,3 % при определении признаков состояния зевоты у водителя ТС.

С целью сглаживания результатов работы нейронной сети, другим вариантом выявления признаков состояния зевоты на промежутке времени является анализ последовательности видеокадров с использованием алгоритма скользящего среднего. Суть подхода заключается в усреднении вероятностей принадлежности изображений к одному из классов, «рот открыт» и «рот закрыт» работы модели нейронной сети, на промежутке времени. Размер скользящего окна, включающего набор усредняемых значений, определен в 2 с и включает все видеокадры, вошедшие в данный период времени. Такой размер окна для детектирования состояния зевоты выбран на основе проанализированных ранее исследований. В результате работы данного алгоритма изображение относится к классу «рот открыт» или «рот закрыт», усредненное значение вероятностей которого является максимальным в рассматриваемый момент времени. Данный алгоритм также был протестирован на наборе данных YawDD: Yawning Detection Dataset. Доля правильно классифицированных объектов достигла 94,7 %, точность — 86 %, полнота — 97,4 %, а F-мера — 91,4 % при определении признаков состояния зевоты.

Таблица 4. Сравнение подходов к детектированию состояния зевоты, %

Table 4. Comparison of approaches of recognizing yawning states

Подход	Метрика			
	Правильно классифицированные объекты	Точность	Полнота	F-мера
Схема детектирования зевоты	98,2	96,1	98,4	97,3
Алгоритм скользящего среднего	94,7	86,0	97,4	91,4

Исходя из оценки результатов работы обоих алгоритмов детектирования состояния зевоты (табл. 4) видно, что первоначальная схема детектирования зевоты, описанная на рис. 7, превосходит по всем перечисленным статистическим показателям алгоритм скользящего среднего.

Заключение

В работе предложена схема эффективного выявления признаков зевоты у водителя в кабине транспортного средства с использованием методов анализа и обработки данных, компьютерной обработки изображений и сверточной модели нейронной сети. В рамках данного подхода представлена концептуальная модель обнаружения открытости/закрытости рта у человека за рулем транспортного средства, и разработана схема обработки и разметки наборов данных YawDD и Kaggle Drowsiness Dataset. Разработанная модель сверточной нейронной сети показала точность (precision) равную 0,992 и полноту (recall) — 0,871 на тестовом 10 % наборе данных, что является приемлемым результатом для дальнейшей интеграции и апробации в интеллектуальной системе поддержки водителю в реальных условиях вождения. Представленные результаты работы по совмещению данных из разных наборов способны найти применение при построении технологии детектирования признаков зевоты для систем поддержки водителя и улучшить эффективность их работы. В качестве дальнейших перспектив по улучшению полученных результатов можно рассмотреть возможность разработки гибридного подхода, основанного на одновременном использовании анализа открытости/закрытости рта, состояния глаз человека, и оценки его позы в пространстве при выявлении зевоты и состояния сонливости в целом.

Физиологическое состояние водителя позволяет проанализировать и оценить текущую обстановку в кабине транспортного средства. Присутствие признаков небезопасного поведения человека за рулем может потенциально снизить безопасность для водителя и других участников дорожного движения. В целях повышения безопасности и экономических факторов информация об управлении транспортным средством способна найти применение в работе страховых и логистических компаний, специализирующихся на грузоперевозках, а также в области пассажирских перевозок.

Литература

1. Rau P.S. Drowsy driver detection and warning system for commercial vehicle drivers: Field operational test design, data analyses, and progress // Proc. Nat. Highway Traffic Saf. Admin. 2005. P. 05–0192.
2. Dua M., Singla R., Singla R., Raj S., Jangra A. Deep CNN models-based ensemble approach to driver drowsiness detection // Neural Computing and Applications. 2021. V. 33. N 8. P. 3155–3168. https://doi.org/10.1007/s00521-020-05209-7
3. Bakheet S., Al-Hamadi A. A framework for instantaneous driver drowsiness detection based on improved HOG features and naïve bayesian classification // Brain Sciences. 2021. V. 11. N 2. P. 240. https://doi.org/10.3390/brainsci11020240
4. Li L., Chen Y., Li Z. Yawning detection for monitoring driver fatigue based on two cameras // Proc. 12th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2009. P. 12–17. https://doi.org/10.1109/ITSC.2009.5309841
5. Daquin G., Micallef J., Blin O. Yawning // Sleep Medicine Reviews. 2001. V. 5. N 4. P. 299–312. https://doi.org/10.1053/smrv.2001.0175
6. Jackson P., Hilditch C., Holmes A., Reed N., Merat N., Smith L. Fatigue and Road Safety: A Critical Analysis of Recent Evidence. Department for Transport, London, 2011. 88 p.
7. Resendes R., Martin K.H. Saving Lives Through Advanced Safety Technology. Faderal Highway Administration, US, Washington, DC, 2003.
8. Wang L., Sun P., Xie M., Ma S., Li B., Shi Y., Su Q. Advanced driver-assistance system (ADAS) for intelligent transportation based on the recognition of traffic cones // Advances in Civil Engineering. 2020. P. 8883639. https://doi.org/10.1155/2020/8883639
9. Zhang W., Murphey Y.L., Wang T., Xu Q. Driver yawning detection based on deep convolutional neural learning and robust nose tracking // Proc. of the International Joint Conference on Neural Networks (IJCNN). 2015. P. 1–8. https://doi.org/10.1109/IJCNN.2015.7280566
10. Chevalier Y., Fenzl F., Kolomeets M., Rieke R., Chechulin A., Krauss C. Cyberattack detection in vehicles using characteristic functions, artificial neural networks, and visual analysis // Информатика и автоматизация. 2021. Т. 20. № 4. С. 845–868. https://doi.org/10.15622/ia.20.4.4
11. Hasan F., Kashevnik A. State-of-the-Art Analysis of modern drowsiness detection algorithms based on computer vision // Proc. of the 29th Conference of Open Innovations Association (FRUCT). 2021. P. 141–149. https://doi.org/10.23919/FRUCT52173.2021.9435480
12. Kashevnik A., Lashkov I., Parfenov V., Mustafin N., Baraniuc O. Context-based driver support system development: Methodology and

References

1. Rau P.S. Drowsy driver detection and warning system for commercial vehicle drivers: Field operational test design, data analyses, and progress. *Proc. Nat. Highway Traffic Saf. Admin.*, 2005, pp. 05–0192.
2. Dua M., Singla R., Singla R., Raj S., Jangra A. Deep CNN models-based ensemble approach to driver drowsiness detection. *Neural Computing and Applications*, 2021, vol. 33, no. 8, pp. 3155–3168. https://doi.org/10.1007/s00521-020-05209-7
3. Bakheet S., Al-Hamadi A. A framework for instantaneous driver drowsiness detection based on improved HOG features and naïve bayesian classification. *Brain Sciences*, 2021, vol. 11, no. 2, pp. 240. https://doi.org/10.3390/brainsci11020240
4. Li L., Chen Y., Li Z. Yawning detection for monitoring driver fatigue based on two cameras. *Proc. 12th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2009, pp. 12–17. https://doi.org/10.1109/ITSC.2009.5309841
5. Daquin G., Micallef J., Blin O. Yawning. *Sleep Medicine Reviews*, 2001, vol. 5, no. 4, pp. 299–312. https://doi.org/10.1053/smrv.2001.0175
6. Jackson P., Hilditch C., Holmes A., Reed N., Merat N., Smith L. *Fatigue and Road Safety:* A *Critical Analysis of Recent Evidence*. Department for Transport, London, 2011, 88 p.
7. Resendes R., Martin K.H. *Saving Lives Through Advanced Safety Technology*. Faderal Highway Administration, US, Washington, DC, 2003.
8. Wang L., Sun P., Xie M., Ma S., Li B., Shi Y., Su Q. Advanced driver-assistance system (ADAS) for intelligent transportation based on the recognition of traffic cones. *Advances in Civil Engineering*, 2020, pp. 8883639. https://doi.org/10.1155/2020/8883639
9. Zhang W., Murphey Y.L., Wang T., Xu Q. Driver yawning detection based on deep convolutional neural learning and robust nose tracking. *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8. https://doi.org/10.1109/IJCNN.2015.7280566
10. Chevalier Y., Fenzl F., Kolomeets M., Rieke R., Chechulin A., Krauss C. Cyberattack detection in vehicles using characteristic functions, artificial neural networks, and visual analysis. *Informatics and Automation*, 2021, vol. 20, no. 4, pp. 845–868. https://doi.org/10.15622/ia.20.4.4
11. Hasan F., Kashevnik A. State-of-the-Art Analysis of modern drowsiness detection algorithms based on computer vision. *Proc. of the 29th Conference of Open Innovations Association (FRUCT)*, 2021, pp. 141–149. https://doi.org/10.23919/FRUCT52173.2021.9435480

case study // Proc. of the 21st Conference of Open Innovations Association (FRUCT). 2017. P. 162–171. https://doi.org/10.23919/FRUCT.2017.8250179
13. Kashevnik A., Lashkov I., Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention // IEEE Transactions on Intelligent Transportation Systems. 2020. V. 21. N 6. P. 2427–2436. https://doi.org/10.1109/TITS.2019.2918328
14. Ibrahim M.M., Soraghan J.J., Petropoulakis L., Di Caterina G. Yawn analysis with mouth occlusion detection // Biomedical Signal Processing and Control. 2015. V. 18. P. 360–369. https://doi.org/10.1016/j.bspc.2015.02.006
15. Ji Y., Wang S., Lu Y., Wei J., Zhao Y. Eye and mouth state detection algorithm based on contour feature extraction // Journal of Electronic Imaging. 2018. V. 27. N 5. P. 051205. https://doi.org/10.1117/1.JEI.27.5.051205
16. Omidyeganeh M., Shirmohammadi S., Abtahi S., Khurshid A., Farhan M., Scharcanski J., Hariri B., Laroche D., Martel L. Yawning detection using embedded smart cameras // IEEE Transactions on Instrumentation and Measurement. 2016. V. 65. N 3. P. 570–582. https://doi.org/10.1109/TIM.2015.2507378
17. Khan S., Akram A., Usman N. Real time automatic attendance system for face recognition using face API and OpenCV // Wireless Personal Communications. 2020. V. 113. N 1. P. 469–480. https://doi.org/10.1007/s11277-020-07224-2
18. Jose J., Vimali J.S., Ajitha P., Gowri S., Sivasangari A., Jinila B. Drowsiness detection system for drivers using image processing technique // Proc. of the 5th International Conference on Trends in Electronics and Informatics (ICOEI). 2021. P. 1527–1530. https://doi.org/10.1109/ICOEI51242.2021.9452864
19. Al-Madani A.M., Gaikwad A.T., Mahale V., Ahmed Z.A., Shareef A.A.A. Real-time driver drowsiness detection based on eye movement and yawning using facial landmark // Proc. of the International Conference on Computer Communication and Informatics (ICCCI). 2021. P. 1–4. https://doi.org/10.1109/ICCCI50826.2021.9457005
20. Akrout B., Mahdi W. Yawning detection by the analysis of variational descriptor for monitoring driver drowsiness // Proc. of the 2nd International Image Processing, Applications and Systems Conference (IPAS). 2016. P. 1–5. https://doi.org/10.1109/IPAS.2016.7880127
21. Zharmagambetov A., Gabidolla M., Carreira-Perpinán M.A. Improved multiclass AdaBoost for image classification: The role of tree optimization // Proc. of the IEEE International Conference on Image Processing (ICIP). 2021. P. 19–22. https://doi.org/10.1109/ICIP42928.2021.9506569
22. Ковшов Е.Е., Завистовская Т.А. Разработка автоматизированной системы контроля и управления доступом на основе анализа динамики носогубных мышц лица человека // Современные проблемы науки и образования. 2013. № 2. С. 185.
23. Kass M., Witkin A., Terzopoulos D. Snakes: Active contour models // International Journal of Computer Vision. 1988. V. 1. N 4. P. 321–331. https://doi.org/10.1007/BF00133570
24. Yazdi Z.M., Soryani M. Driver drowsiness detection by identification of yawning and eye closure // Automotive Science and Engineering. 2019. V. 9. N 3. P. 3033–3044. https://doi.org/10.22068/ijae.9.3.3033
25. Kassem H.A., Chowdhury M.U., Abawajy J., Al-Sudani A.R. Yawn based driver fatigue level prediction // EPiC Series in Computing. 2020. V. 69. P. 372–382. https://doi.org/10.29007/67kk
26. Abtahi S., Omidyeganeh M., Shirmohammadi S., Hariri B. YawDD: A yawning detection dataset // Proc. of the 5th ACM Multimedia Systems Conference. 2014. P. 24–28. https://doi.org/10.1145/2557642.2563678
27. Yang H., Liu L., Min W., Yang X., Xiong X. Driver yawning detection based on subtle facial action recognition // IEEE Transactions on Multimedia. 2021. V. 23. P. 572–583. https://doi.org/10.1109/TMM.2020.2985536
28. Suwarno S., Kevin K. Analysis of face recognition algorithm: Dlib and openCV // Journal of Informatics and Telecommunication Engineering. 2020. V. 4. N 1. P. 173–184. https://doi.org/10.31289/jite.v4i1.3865
29. Zhang S., Zhu X., Lei Z., Shi H., Wang X., Li S.Z. FaceBoxes: A CPU real-time face detector with high accuracy // Proc. of the IEEE International Joint Conference on Biometrics (IJCB). 2017. P. 1–9. https://doi.org/10.1109/BTAS.2017.8272675
30. Zhang K., Zhang Z., Li Z., Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks // IEEE Signal

12. Kashevnik A., Lashkov I., Parfenov V., Mustafin N., Baraniuc O. Context-based driver support system development: Methodology and case study. *Proc. of the 21st Conference of Open Innovations Association (FRUCT)*, 2017, pp. 162–171. https://doi.org/10.23919/FRUCT.2017.8250179
13. Kashevnik A., Lashkov I., Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention. *IEEE Transactions on Intelligent Transportation Systems*, 2020, vol. 21, no. 6, pp. 2427–2436. https://doi.org/10.1109/TITS.2019.2918328
14. Ibrahim M.M., Soraghan J.J., Petropoulakis L., Di Caterina G. Yawn analysis with mouth occlusion detection. *Biomedical Signal Processing and Control*, 2015, vol. 18, pp. 360–369. https://doi.org/10.1016/j.bspc.2015.02.006
15. Ji Y., Wang S., Lu Y., Wei J., Zhao Y. Eye and mouth state detection algorithm based on contour feature extraction. *Journal of Electronic Imaging*, 2018, vol. 27, no. 5, pp. 051205. https://doi.org/10.1117/1.JEI.27.5.051205
16. Omidyeganeh M., Shirmohammadi S., Abtahi S., Khurshid A., Farhan M., Scharcanski J., Hariri B., Laroche D., Martel L. Yawning detection using embedded smart cameras. *IEEE Transactions on Instrumentation and Measurement*, 2016, vol. 65, no. 3, pp. 570–582. https://doi.org/10.1109/TIM.2015.2507378
17. Khan S., Akram A., Usman N. Real time automatic attendance system for face recognition using face API and OpenCV. *Wireless Personal Communications*, 2020, vol. 113, no. 1, pp. 469–480. https://doi.org/10.1007/s11277-020-07224-2
18. Jose J., Vimali J.S., Ajitha P., Gowri S., Sivasangari A., Jinila B. Drowsiness detection system for drivers using image processing technique. *Proc. of the 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1527–1530. https://doi.org/10.1109/ICOEI51242.2021.9452864
19. Al-Madani A.M., Gaikwad A.T., Mahale V., Ahmed Z.A., Shareef A.A.A. Real-time driver drowsiness detection based on eye movement and yawning using facial landmark. *Proc. of the International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–4. https://doi.org/10.1109/ICCCI50826.2021.9457005
20. Akrout B., Mahdi W. Yawning detection by the analysis of variational descriptor for monitoring driver drowsiness. *Proc. of the 2nd International Image Processing, Applications and Systems Conference (IPAS)*, 2016, pp. 1–5. https://doi.org/10.1109/IPAS.2016.7880127
21. Zharmagambetov A., Gabidolla M., Carreira-Perpinán M.A. Improved multiclass AdaBoost for image classification: The role of tree optimization. *Proc. of the IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 19–22. https://doi.org/10.1109/ICIP42928.2021.9506569
22. Kovshov E.E., Zavistovskaya T.A. Development of automated system of an access control based on analysis of the nasolabial muscles dynamics of the human face. *Modern problems of science and education*, 2013, no. 2, pp. 185–185. (in Russian)
23. Kass M., Witkin A., Terzopoulos D. Snakes: Active contour models. *International Journal of Computer Vision*, 1988, vol. 1, no. 4, pp. 321–331. https://doi.org/10.1007/BF00133570
24. Yazdi Z.M., Soryani M. Driver drowsiness detection by identification of yawning and eye closure. *Automotive Science and Engineering*, 2019, vol. 9, no. 3, pp. 3033–3044. https://doi.org/10.22068/ijae.9.3.3033
25. Kassem H.A., Chowdhury M.U., Abawajy J., Al-Sudani A.R. Yawn based driver fatigue level prediction. *EPiC Series in Computing*, 2020, vol. 69, pp. 372–382. https://doi.org/10.29007/67kk
26. Abtahi S., Omidyeganeh M., Shirmohammadi S., Hariri B. YawDD: A yawning detection dataset. *Proc. of the 5th ACM Multimedia Systems Conference*, 2014, pp. 24–28. https://doi.org/10.1145/2557642.2563678
27. Yang H., Liu L., Min W., Yang X., Xiong X. Driver yawning detection based on subtle facial action recognition. *IEEE Transactions on Multimedia*, 2021, vol. 23, pp. 572–583. https://doi.org/10.1109/TMM.2020.2985536
28. Suwarno S., Kevin K. Analysis of face recognition algorithm: Dlib and openCV. *Journal of Informatics and Telecommunication Engineering*, 2020, vol. 4, no. 1, pp. 173–184. https://doi.org/10.31289/jite.v4i1.3865
29. Zhang S., Zhu X., Lei Z., Shi H., Wang X., Li S.Z. FaceBoxes: A CPU real-time face detector with high accuracy. *Proc. of the IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 1–9. https://doi.org/10.1109/BTAS.2017.8272675

Processing Letters. 2016. V. 23. N 10. P. 1499–1503. https://doi.org/10.1109/LSP.2016.2603342
31. Bazarevsky V., Kartynnik Y., Vakunov A., Raveendran K., Grundmann M. Blazeface: Sub-millisecond neural face detection on mobile gpus // arXiv.org. 2019. arXiv:1907.05047.
32. Bulat A., Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks) // Proc. of the 16th IEEE International Conference on Computer Vision (ICCV). 2017. P. 1021–1030. https://doi.org/10.1109/ICCV.2017.116
33. Cohen G., Afshar S., Tapson J., Van Schaik A. EMNIST: Extending MNIST to handwritten letters // Proc. of the International Joint Conference on Neural Networks (IJCNN). 2017. P. 2921–2926. https://doi.org/10.1109/IJCNN.2017.7966217
34. Goodfellow I.J., Erhan D., Carrier P.L., Courville A., Mirza M., Hamner B., Cukierski W., Tang Y., Thaler D., Lee D., Zhou Y., Ramaiah C., Feng F., Li R., Wang X., Athanasakis D., Shawe-Taylor J., Milakov M., Park J., Ionescu R., Popescu M., Grozea C., Bergstra J., Xie J., Romaszko L., Xu B., Chuang Z., Bengio Y. Challenges in representation learning: A report on three machine learning contests // Lecture Notes in Computer Science. 2013. V. 8228. P. 117–124. https://doi.org/10.1007/978-3-642-42051-1_16
35. Ramos A.L., Erandio J.C., Enteria E.M., Del Carmen N., Enriquez L.J., Mangilaya D.H. Driver drowsiness detection based on eye movement and yawning using facial landmark analysis // International Journal of Simulation: Systems, Science & Technology. 2019. V. 20. N S2. P. 37. https://doi.org/10.5013/IJSSST.a.20.S2.37
36. Savaş B.K., Becerikli Y. Real time driver fatigue detection based on SVM algorithm // Proc. of the 6th International Conference on Control Engineering and Information Technology (CEIT). 2018. P. 1–4. https://doi.org/10.1109/CEIT.2018.8751886
37. Saradadevi M., Bajaj P. Driver fatigue detection using mouth and yawning analysis // International Journal of Computer Science and Network Security. 2008. V. 8. N 6. P. 183–188.
38. Castella F.R. Sliding window detection probabilities // IEEE Transactions on Aerospace and Electronic Systems. 1976. V. AES-12. N 6. P. 815–819. https://doi.org/10.1109/TAES.1976.308363

30. Zhang K., Zhang Z., Li Z., Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016, vol. 23, no. 10, pp. 1499–1503. https://doi.org/10.1109/LSP.2016.2603342
31. Bazarevsky V., Kartynnik Y., Vakunov A., Raveendran K., Grundmann M. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv.org*, 2019, arXiv:1907.05047.
32. Bulat A., Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks). *Proc. of the 16th IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030. https://doi.org/10.1109/ICCV.2017.116
33. Cohen G., Afshar S., Tapson J., Van Schaik A. EMNIST: Extending MNIST to handwritten letters. *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2921–2926. https://doi.org/10.1109/IJCNN.2017.7966217
34. Goodfellow I.J., Erhan D., Carrier P.L., Courville A., Mirza M., Hamner B., Cukierski W., Tang Y., Thaler D., Lee D., Zhou Y., Ramaiah C., Feng F., Li R., Wang X., Athanasakis D., Shawe-Taylor J., Milakov M., Park J., Ionescu R., Popescu M., Grozea C., Bergstra J., Xie J., Romaszko L., Xu B., Chuang Z., Bengio Y. Challenges in representation learning: A report on three machine learning contests. *Lecture Notes in Computer Science*, 2013, vol. 8228, pp. 117–124. https://doi.org/10.1007/978-3-642-42051-1_16
35. Ramos A.L., Erandio J.C., Enteria E.M., Del Carmen N., Enriquez L.J., Mangilaya D.H. Driver drowsiness detection based on eye movement and yawning using facial landmark analysis. *International Journal of Simulation: Systems, Science & Technology*, 2019, vol. 20, no. S2, pp. 37. https://doi.org/10.5013/IJSSST.a.20.S2.37
36. Savaş B.K., Becerikli Y. Real time driver fatigue detection based on SVM algorithm. *Proc. of the 6th International Conference on Control Engineering and Information Technology (CEIT)*, 2018, pp. 1–4. https://doi.org/10.1109/CEIT.2018.8751886
37. Saradadevi M., Bajaj P. Driver fatigue detection using mouth and yawning analysis. *International Journal of Computer Science and Network Security*, 2008, vol. 8, no. 6, pp. 183–188.
38. Castella F.R. Sliding window detection probabilities. *IEEE Transactions on Aerospace and Electronic Systems*, 1976, vol. AES-12, no. 6, pp. 815–819. https://doi.org/10.1109/TAES.1976.308363

Автор

Лашков Игорь Борисович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербург, 199178, Российская Федерация, SC 56719631000, https://orcid.org/0000-0001-6418-4660, igor-lashkov@ya.ru

Author

Igor B. Lashkov — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 56719631000, https://orcid.org/0000-0001-6418-4660, igor-lashkov@ya.ru

Статья поступила в редакцию 09.08.2021
Одобрена после рецензирования 30.11.2021
Принята к печати 21.01.2022

Received 09.08.2021
Approved after reviewing 30.11.2021
Accepted 21.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-47-59

A Game Theory approach for communication security and safety assurance in cyber-physical systems with Reputation and Trust-based mechanisms

Ilia I. Viksnin[1]✉, Egor D. Marinenkov[2], Sergey S. Chuprov[3]

[1] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation
[2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] wixnin@mail.ru✉, https://orcid.org/0000-0001-6240-0390
[2] egormarinenkov@gmail.com, https://orcid.org/0000-0001-9895-239X
[3] drmyscull@gmail.com, https://orcid.org/0000-0001-7081-8797

Abstract

Cyber-physical systems' security and safety assurance is a challenging research problem for Smart City concept development. Technical faults or malicious attacks over communication between its elements can jeopardize the whole system and its users. Reputation systems implementation is an effective measure to detect such malicious agents. Each agent in the group has its indicator, which reflects how trustworthy it is to the other agents. However, in the scenario when it is not possible to calculate the Reputation indicator based on objective characteristics, malicious or defective agents can negatively affect the system's performance. In this paper, we propose an approach based on Game Theory to address the Reputation and Trust initial values calculation challenge. We introduced a mixed strategies game concept and a probability indicator. The possible outcomes of using different strategies by the system agents are represented with a payoff matrix. To evaluate the approach effectiveness, an empirical study using a software simulation environment was conducted. As a Cyber-physical system implementation scenario, we considered an intersection management system with a group of unmanned autonomous vehicles, the aim of which is to perform conflict-free optimal intersection traversal. To simulate the attack scenario, some vehicles were able to transmit incorrect data to other traffic participants. The obtained results showed that the Game Theory approach allowed us to increase the number of detected intruders compared to the conventional Reputation and Trust model.

Keywords

Game Theory, reputation, trust; security, safety, cyber-physical systems

Acknowledgements

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Project 'Goszadanie' No. 075-01024-21-02 from 29.09.2021).

For citation: Viksnin I.I., Marinenkov E.D., Chuprov S.S. A Game Theory approach for communication security and safety assurance in cyber-physical systems with Reputation and Trust-based mechanisms. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 47–59. doi: 10.17586/2226-1494-2022-22-1-47-59

УДК 004.896

Применение теории игр для обеспечения безопасности коммуникации киберфизической системы с использованием механизмов репутации и доверия

Илья Игоревич Викснин[1]✉, Егор Денисович Мариненков[2], Сергей Сергеевич Чупров[3]

[1] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова, Санкт-Петербург, 197022, Российская Федерация
[2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] wixnin@mail.ru✉, https://orcid.org/0000-0001-6240-0390
[2] egormarinenkov@gmail.com, https://orcid.org/0000-0001-9895-239X
[3] drmyscull@gmail.com, https://orcid.org/0000-0001-7081-8797

© Viksnin I.I., Marinenkov E.D., Chuprov S.S., 2022

Аннотация

Предмет исследования. Обеспечение безопасности и надежности киберфизических систем является сложной исследовательской проблемой для разработки концепции «умного города». Технические неполадки или злонамеренные атаки на коммуникации между элементами системы могут поставить под угрозу всю систему и ее пользователей. Реализация репутационных систем — эффективная мера для обнаружения таких вредоносных агентов. Каждый агент в группе имеет свой показатель, который отражает, насколько он заслуживает доверия других агентов. Вместе с тем в сценарии, когда невозможно рассчитать показатель репутации на основе объективных характеристик, вредоносные или дефектные агенты негативно влияют на работу системы. **Метод.** Предложен подход, основанный на теории игр, для решения проблемы расчета начальных значений репутации и доверия. Введены концепция игры со смешанными стратегиями и индикатор вероятности. Возможные результаты использования различных стратегий агентами системы представлены с помощью матрицы выплат. **Основные результаты.** Для оценки эффективности подхода выполнено эмпирическое исследование с использованием программной среды моделирования. В качестве сценария реализации киберфизической системы рассмотрена система управления перекрестком с группой беспилотных автономных транспортных средств, цель которой бесконфликтное оптимальное прохождение перекрестка. Для имитации сценария атаки часть транспортных средств может передавать неверные данные другим участникам движения. Полученные результаты показали, что подход теории игр позволил увеличить количество обнаруживаемых нарушителей по сравнению с необработанной моделью репутации и доверия.

Ключевые слова

теория игр, репутация, доверие, информационная безопасность, функциональная безопасность, киберфизические системы

Благодарности

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (проект «Госзадание» № 075-01024-21-02 от 29.09.2021).

Ссылка для цитирования: Викснин И.И., Мариненков Е.Д., Чупров С.С. Применение теории игр для обеспечения безопасности коммуникации киберфизической системы с использованием механизмов репутации и доверия // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 47–59 (на агл. яз.). doi: 10.17586/2226-1494-2022-22-1-47-59

Introduction

Striking development of information, communication and automation technologies over the past few decades has had a tremendous impact on various areas of human life. The endeavor to optimize various routine processes and make our life more convenient have led to the emergence of such concepts as Smart Home, Smart City and Smart Manufacturing [1, 2]. These approaches are based on the communication (most often, wireless) between the informational and physical components, the combination of which became known as Cyber-physical systems (CPSs) [3]. The aims of physical elements are to interact with the environment, in which they are located, and collect and/or measure its characteristics. For example, it can be light brightness, humidity, or temperature sensors, which measure the characteristics at predetermined time intervals and transmit the collected data to the informational elements. Informational elements perform computational operations, and, according to predetermined algorithms, generate decisions based on the data received. For instance, if the light brightness level has fallen to a certain threshold, the system needs to turn on the lights.

The implementation of unmanned autonomous vehicles (AVs) is a vital direction for future transportation systems and the Smart City concept development and modernization [4]. Such AVs can be terrestrial, aerial, water or underwater, and can also be described as a set of CPS elements. At the present development stage, AVs are widely available on the market and are actively used in various spheres to perform different work, including those that could previously be performed by highly qualified specialists, e.g. aircraft pilots or train drivers. However, there are tasks that can be performed more effectively by AVs group than using individuals, for instance, territory surveillance or people search during rescue operations. To coordinate group actions, AVs have to use one of the control strategies: centralized or decentralized. They both have their advantages and drawbacks, and the choice depends on such factors as group participants number, task types, or system requirements.

A more detailed review of their properties and an example of practical application can be found in [5, 6]. In the present work, we use a decentralized agent control strategy, as it is more reliable and fault-tolerant from the safety perspective.

CPSs, like any information systems, are exposed to various cybersecurity threats. Conventional security methods, such as authentication, authorization, or cryptography mechanisms are effective to counter or mitigate most information attacks. However, there are so-called “soft” types of attacks that cannot be identified by conventional security mechanisms. These attacks can be aimed at unauthorized changes in the contextual integrity of data transmitted between group members. Moreover, such attacks can be both intentional and unintentional. For instance, in the event when the legitimate agent’s hardware or software components fail, and it starts broadcasting false data about its current location. To combat “soft” attacks, the mechanism based on the agents’ Reputation and Trust was proposed. Group member’s Reputation level is based on their behavior and calculated according to the other group members’ opinions. However, this method has a drawback: since the Reputation is a retrospective indicator, it cannot be calculated at the initial system functioning moment, or at the moment when a new member joins the group.

Earlier in our study [7], we proposed and physically implemented the Reputation and Trust-based approach for AVs security and safety assurance in the intersection management system. To address the initial Reputation value calculation challenge, in [8] we provided the mechanism based on Game Theory fundamentals, which allowed us to calculate this value relying on objective indicators. The major contribution of this paper is twofold. First, we provided an improved and more rigorous approach formalization, with a novel dynamical hybrid decision-making strategy and a probability indicator. In addition, we developed our custom software simulator, that can be found in public access1, and conducted an empirical study with multiple robotic devices able to communicate with each other under "soft" attacks conditions. The results showed that the Game Theory approach implementation allows one to reduce probability of classification intruders as legitimate agents and to increase accuracy of their detection, compared with using classical Reputation and Trust metrics, provided in [7].

The paper contains the definition and description of the terms "Trust" and "Reputation", and a brief discussion of the documents that offer security mechanisms based on this approach. A description of the use of the game theory approach for computer system security is presented, with a brief discussion of the research that followed this concept. The CPS model, information interaction between agents, and group goal optimization problem were formalized. A solution to the problem of transmitting false data by agents and its impact on the work and security of the entire group was proposed. The calculation of trust and reputation was introduced, and the issue of the initial reputation value calculation was discussed for cases where it is impossible to assign this value due to the lack of historical data on the agent behavior. The classification of data transmitted between agents and the formalization of the costs are presented. The concept of the game between two agents, payoff matrix, possible mixed strategies, and their outcomes for the case when the agent does not have enough data to calculate the current reputation value are defined. An approach to the effectiveness evaluation of the proposed model, a modeling scheme and metrics, and a discussion of the obtained result interpretation are considered.

Literature review

Trust and Reputation

In some social networks, online stores, and e-commerce applications, user reputation rating systems have gained popularity. The presence of a reputation indicator implies the existence of certain generally accepted norms and behavior rules on a resource. Violation of such rules and norms by the user leads to a decrease in his reputation indicator, as well as to a decreasing trust to him from other users. For instance, if one of the online store's sellers sells a product with characteristics different from the declared ones, or the delivery time is not corresponding to the expected, it is less likely that buyers want to buy goods from him if there are other more trustworthy sellers.

Depending on the sources, interpretations of Trust and Reputation may vary. The content of these concepts goes deep into antiquity, with the advent of the first people communities and the interaction between them. Those concepts can now be described as Trust and Reputation. Study [9] defines trust as an open and positive relationship between people, containing confidence in decency and goodwill. If we move away from the human relationship and describe the trust between some agents in a computer system, in [10] trust described as a subjective expectation of agent A of certain behavior from agent B based on the history of the interaction. It follows from the definition that trust allows us to assume what kind of expected action or inaction might come from the agent. From the same definition we can trace the subjectivity of trust in relation to one or another object of relationships.

Reputation is defined as an opinion about the intentions and norms of a particular agent, based on his behavior retrospective and interactions with him [10]. Quantification can be calculated based on the opinions or observations of other group members. Unlike subjective trust (relying on one's own experience and other factors), reputation allows reflecting a public measure of the agent's reliability based on group members' observations or assessments.

To use the Trust and Reputation-based approach in information systems, it is necessary to formalize and consider quantitative Reputation and Trust indicators, and data on observations and assessments. This can be especially relevant in decentralized networks, where there is a lack of network infrastructure and the nodes interact directly with each other. Such networks became known as peer-to-peer (P2P) networks [11]. P2P networks have gained widespread popularity with the advent of the Internet of Things (IoT) concept [12], vehicular (VANETs) and mobile (MANETs) ad-hoc networks [13]. P2P allows to transfer and process large amounts of information, at a cost lower than using a centralized infrastructure network [14]. However, due to the decentralized architecture, presence of heterogeneous elements, and specific features, such networks are subject to "soft" attacks aimed at the contextual integrity of the transmitted data. "Conventional" cybersecurity methods, such as authentication or cryptography, are ineffective against such attacks.

In the AVs case, VANETs allow transmitting data from one vehicle to another and to the transport infrastructure objects. Such data transfer can be used by the Intelligent Transport System (ITS) to build optimal routes, generate informational and emergency messages warning of bad weather conditions, construction and maintenance road works, and etc. Papers studying Reputation-based data security techniques may offer different approaches to calculate these metrics. In [15], the authors suggest calculating the trust indicator in the range from –1 to 1, as in [7], we proposed to calculate the Reputation and Trust indicators in the range from 0 to 1. In the present paper, we use the calculus described in [7] and improve it with the Game Theory-based approach.

In [16], Starub et al. proposed a multi-level intrusion detection system (IDS) to protect self-driving vehicles from malicious attacks. The system is based on the method of determining nodes' reputation value. The system contains shared knowledge generated by all communication participants. The reputation level depends on the nodes'

retrospective behavior. Despite the interesting system's architecture proposed by the authors, it is difficult to evaluate the effectiveness of their solution. The work lacks both reputation level calculus and the solution effectiveness validation and comparison with other existing Trust and Reputation-based mechanisms.

In [17], Kim and Viksnin proposed a method for calculating Trust and Reputation indicators, which is based on the loan theory to ensure communication of security flying drones. The main idea of the approach is that it would be unprofitable for intruders to perform a destructive informational impact on the group. In case when the agent transmits incorrect information, its indebtedness increases. The experiment results showed that the intruder transmitting incorrect data was blocked in 90.2 % cases.

To verify the data reliability, two approaches are proposed [18]: objective and subjective. In the second case, the nodes rely on the opinion of other nodes to calculate the trust indicator. The authors addressed the data privacy problem when calculating nodes' trust indicators, and proposed a framework that allowed them to find a balance between trust and privacy in the system. Experiments conducted using the ONE network simulator showed that the application of the proposed linkability protocol allowed increasing transmitted data privacy by using pseudonyms for nodes and offered more flexibility than the standard secure broadcast authentication protocol utilized in the ONE network simulator.

One of the main challenges in existing Reputation and Trust-based approaches is generating the initial value for the system agents, based on their retrospective behavior. Moreover, this issue is actual when a new agent joins the group, and other participants need to decide, how trustworthy it is. In this paper, we proposed a novel dynamic approach for initializing the Reputation value, which depends on particular situations and considers current agents' conditions. In Table 1, we summarize and compare the main characteristics of the related literature in the context of this challenge.

As can be seen from Table 1, we outlined a number of important characteristics that are necessary for evaluating further potential and implementation of the proposed mechanisms. We compared the approach, proposed in this paper, with our previous research [7] and other four studies, related to Reputation and Trust-based security methods. According to the presented Table, only our approach provides dynamic Reputation value initialization for self-driving vehicles, which is vital to reduce the negative influence of malicious or defective agents on the system, when their Reputation value cannot be determined on the basis of retrospective behavior. Moreover, our plans include implementing of this Game-Theory approach on the physical testing ground, demonstrated in [19] and conducting real-world performance evaluation.

Game Theory

Game theory is a branch of mathematical economics that studies the resolution of conflicts between players and the optimality of their strategies. It is widely used in various fields of human activity, such as economics and management, industry and agriculture, military and construction, trade and transport, communications, etc [20].

One of the Game Theory implementation tasks in the cybersecurity area is to optimize security administrators' actions in network systems. In the Game Theory context, this task can be formalized as follows: there are two coalitions: defenders (administrators) and attackers; the goal of administrators is to minimize the damage to the system by optimal tasks distribution among themselves, and the goal of the attackers is to compromise the system. Considering different attackers' behaviors, it is possible to identify such strategies for the administrators' behavior (both for a coalition and for each administrator), in which the system's damage is minimized, regardless of the attackers' strategy. One of the approaches is described in [21]. The authors proposed a strategy, in which Nash equilibrium can be achieved, which guarantees an optimal solution to the defending side regardless of the attackers' decisions. The authors conducted a comparative approach analysis to ensure Game Theory-based safety circuit and common sense decision algorithms. To verify the developed model, real statistics were used.

Table 1. Reputation and Trust-based approaches characteristics comparison

Characteristics	Reviewed Studies					Our approach
	[7]	[15]	[16]	[17]	[18]	
Implementation scenario	Self-driving vehicles	Cloud computing	VANETs	Unmanned aerial vehicles (UAV)	VANETs	Self-driving vehicles
Reputation (or Trust) initial value	Constant (0.5)	Constant (0)	Constant (0)	Constant (0)	Constant (not specified)	Dynamic (depends on situation)
Behavior evaluation	Collective	Individual and collective	Individual and collective	Collective	Collective	Collective
Calculus	Provided	Provided	Not provided	Provided	Provided	Provided
Soft attacks	Addressed	Not addressed	Addressed (lack of details)	Addressed	Addressed	Addressed
Empirical study	Software, physical	No	No	Software	Software	Software

In [22], Roy et al. provided an overview of the Game-Theoretic models' application for network security assurance. The authors reviewed static games and divided them into complete imperfect information and incomplete imperfect information games. In the former game type, the authors cited the example of an information war and a quantitative risk assessment for effective investment decisions in the cybersecurity area. The latter gave examples of games in the framework to counter DDoS and intrusions in ad-hoc networks. Moreover, the authors analyzed dynamic games and subdivided them into 4 types: complete perfect information, complete imperfect information, incomplete perfect information, and incomplete imperfect information games. The first game type is used for risk analysis in computer networks, where, as a rule, there are only two participants: a network administrator and an attacker. Implementation of Game Theory allows determining the optimal strategy for several iterations, which helps to optimally distribute resources for long time periods. For the second type, an IDS and several scenarios, based on the attackers' knowledge completeness on the system were considered. This approach determines the optimal players' strategies, which can subsequently be applied as a deciding rule when implementing or modifying the system. The third type described a game, in which network participants reduce worm-attack propagation speed, which allows scanning a system for important and valuable information. In the fourth type, games like admin-attacker are also considered.

In [23], Game Theory is used for security assurance in e-commerce applications. The authors described the security game model using the penalty parameter, calculated replicator dynamics and analyzed the evolutionarily stable strategy of the game model. As a result, the authors concluded that investment cost reduction leads to the stimulation of investment in cybersecurity. With an increase in investment costs, the penalty parameter saves the incentive for investments. The described papers on Game Theory approaches show the expediency of applying such approaches in the areas related to distributed networks and automated systems. However, there still is a challenge of initial value calculation for Trust and Reputation indicators, and our Game Theory-based approach allows estimating the behavior of elements within a distributed system. Thus, we propose our Game Theory-based approach to address this challenge.

Cyber-physical system model formalization

As mentioned above, we consider the CPS with a decentralized group control strategy. In addition, we assume that all group participants are homogeneous. Then, CPS can be formalized as a set of homogeneous agents with the cardinality of n: $CPS = \{e_i | i = \overline{1, n}\}$. Let us assume that agent e_i is a dynamic object and is able to move. Moreover, CPS agents possess the following characteristics:

— agent's current location;
— maximum possible distance to perform informational interaction (II) with other group members;
— on-board sensors' maximum possible distance to perform surroundings monitoring.

The agents are able to perform the tasks assigned to them. Tasks are distributed between group participants via collective task-allocation auction. All tasks are aimed to reach the common CPS's group goal. Generally, this goal can be interpreted as an optimization problem: the group needs to complete maximum tasks with the minimum costs, where costs can be understood as time, energy or other characteristics. In the task execution process, every action performed by agents increases the group costs for goal reaching, therefore, these actions need to be optimal. To perform optimal action, agents are necessary to analyze the data circulating inside the CPS and decide which action to perform on the basis of these data. The data circulating in CPS at the discrete time t can be classified in the following way:

— data on e_i current technical state TS_{ei}^t which include hardware and software components condition, current location and velocity and other agent's characteristics;
— data on e_i current status S_{ei}^t, which can be interpreted as "occupied" or "unoccupied" with a task at a current moment;
— data on e_i current surroundings condition E_{ei}^t, which is obtained by agent's on-board sensors;
— other agent's e_i useful data O_{ei}^t, which are relevant for reaching the CPS goal;
— data on other group participant $I_{ei}^t = \{I_{e_ie_j} | i \neq j, j = \overline{1, m}\}$, that e_i possesses, where $I_{e_ie_j} = \{TS_{ej}, S_{ej}, E_{ej}, O_{ej}\}$ is a data on e_j obtained in t time or earlier, $m | m \leq n - 1$ is a number of elements, on which e_i has knowledge.

To calculate the task's completion costs, it is proposed to calculate actions' costs that need to be performed when completing this task. To perform this, we introduce a cost calculation function based on the selected action:

$$c_{ei}^t = costs(act(TS_{ei}^t, S_{ei}^t, E_{ei}^t, I_{ei}^t, O_{ei}^t)), \quad (1)$$

where c_{ei}^t is the amount of resources spent to execute an action at the t time; *costs* is the function for calculating the costs amount; *act* is the function for e_i optimal action determining at the t time. The set of completed tasks T^d can be represented as a subset of all available tasks T with a cardinality of k: $\exists T^d \subseteq T$: $T = \{tsk_l | l = \overline{1, k}\}$, where tsk_l is a l'th task need to be performed by the group.

According to (1) and to the introduced task's subset, the CPS goal can be formalized as:

$$\begin{cases} \sum_{s=0}^{t} \sum_{i=0}^{n} costs(act(TS_{ei}^t, S_{ei}^t, E_{ei}^t, I_{ei}^t, O_{ei}^t)) \to 0 \\ |T^d| \to |T| \end{cases}.$$

Problem statement

The data transmitted by agents can be either correct or false. In the first case, the data reflects the actual (real) location and environment characteristics of the agent e_i at the time of transmission t_j. In the second case, the data is incorrect and does not reflect the real characteristics of the agent e_i at the time of the data transfer t_j. The data may be incorrect due to malfunction, sensors failure, or malicious interference with the software or hardware agent's e_i components.

To identify agents that transmit false data, earlier, we proposed the procedure based on Reputation and Trust

indicators evaluation [7]. Each of the group agents has a Reputation indicator. The assessment is based on the transmitted data verification at each time t by group agents, which are able to perform this evaluation. To describe our approach, we introduced three indicators: Truth, Trust, and Reputation. A brief description of these indicators is provided below, and a more detailed explanation can be found in [7].

In [8], we applied pure strategies and obtained better results than using raw Reputation and Trust metrics. However, the pure strategies application did not show a considerable gain in effectiveness.

Therefore, in this study, we formulate the hypothesis that the *Truth* indicator calculation in the incomplete data conditions, based on the information impact on the CPS's aim assessment, allows us to improve false data providers detection accuracy compared to setting initial Reputation value as 0.5.

Reputation and Trust approach formalization

To perform the data correctness evaluation, we need to introduce three indicators: *Truth*, *Reputation* (*R*), and *Trust*.

Truth is an indicator that displays a subjective correctness assessment of the transferred data by other agents. Correctness is determined using the sensors of agents and can be described as:

$$Truth_t = f_{tr_t}(data),$$

where $Truth_t$ is the evaluation of data at the time t; *data* is the data to be evaluated; f_{tr_t} is the evaluation function of *Truth* at the t time.

Reputation (*R*) is an indicator based on a retrospective of the *Truth* indicator assessed by each group agent. It can be described as:

$$R_t = f_{r_t}(Truth_t) = f_{r_t}(f_{tr_t}(data)),$$

where R_t is the R value at the t time; f_{r_t} is the R evaluation function at the t time.

Trust is an indicator characterizing a subjective assessment of agent's behavior by other group members. It is calculated based on a *Truth* and *R* combination, and can be represented as:

$$Trust_t = f_{trust_t}(R_{t-1}, Truth_t) = f_{trust_t}(f_{r_{t-1}}(f_{tr_{t-1}}(data), f_{tr_t}(data)),$$

where $Trust_t$ is the indicator of *Trust* at the t time; f_{trust_t} is the function of evaluating *Trust* at the t time.

Reputation initial value calculating challenge

Existing Reputation-based models use indicators of Reputation and Trust to detect intruders on the basis of their behavior and the content of informational messages, transmitted by them [24–27]. However, during the system operation, situations may occur when none of the agents has the opportunity to assess the correctness of the data transmitted to them. For example, such a situation may arise at the t_0 time (initialization of the system), when agents are distributed over the area and do not have a retrospective assessment, or when a new agent joins the group. As a limitation, each of the above indicators is in the range of [0, 1]. In general, the initial *R* value is defined as 0.5 (as average value). Such an approach does not allow characterizing transmitted data as either correct or not, which leads to a further unpredictable agent's behavior assessment.

To address this issue, in [8], we provided an improved *Truth* calculating mechanism for the case of data incompleteness, based on the transmitted data impact evaluation on the task performance process. Considering the data incompleteness case in the Game Theory context, we formalized, implemented, and evaluated our approach via software simulations. The proposed model implementation allowed us to slightly increase malicious agents' detection accuracy and to decline false-negative errors by almost 8 times. However, false-positive errors increased by almost 12 times. These results were obtained using pure game strategies, which led to *Truth* = 0 assigning for both correct and incorrect data. Such obstacles encouraged us to evolve the approach's accuracy and reliability.

In this study, we suggest that in the data incompleteness case — when the agent is unable to assess data transmitted from another agent — a probabilistic data correctness assessment increases the malicious agents' detection accuracy. Such an approach can be implemented using Game Theory, namely a mixed game extension, in which an equilibrium situation always exists [28]. This mechanism allows calculating *Truth* indicator even in the data incompleteness cases. Moreover, the probabilistic nature of *Truth* indicator formation assumes obtaining a dynamic solution, using which it is possible to assess the optimal *Truth* value for various system's conditions.

False data identification model

False data impact on the group's performance

To verify our mechanism, we propose a calculus for assessing the false data impact on the group's performance during the goal achievement process. The information in the system can be divided by its relevance into the following categories: actual, less actual, and disinformation. The information relevance is substantiated by the combination of the information receiving time, and the time at which this information is used to determine the agent's further actions and is characterized by a linear costs increase. In this case, the costs are calculated according to:

$$c_{ei}^{t} = k \times (t - (t' - 1)) + c_{ei}^{t'-1}, \quad (2)$$

where k is a static coefficient that determines costs increasing rate using actual information; t' is a moment of information reception.

In the less actual information case, the costs grow exponentially since the information becomes outdated in time. This can lead to various scenarios that maliciously affect agent's or whole CPS's operation. Let us introduce the information block *Inf* relevance indicator $a_{Inf} \in (0; 1)$, which characterizes the information obsolescence rate and the growth of costs, estimated by the agent e_i. The costs of using less relevant information are calculated according to:

$$c_{ei}^{t} = k \times a_{Inf}^{t'-1} \times (t - (t' - 1)) + c_{ei}^{t'-1}. \quad (3)$$

In the disinformation case, when the data is incorrect, costs grow faster than using actual and less actual information. Therefore, it is necessary to introduce the

disinformation impact coefficient $a'_{Inf} \in (0; 1)$, which characterizes the damage caused by the false data. The costs are then calculated according to:

$$c_{ei}^{t} = k \times a_{Inf}^{t'-1-t} \times (t - (t' - 1)) + c_{ei}^{t'-1}. \quad (4)$$

Game Theory approach formalization for data incompleteness case

To solve the problem stated in section "Problem Statement" and test the hypothesis, an approach to the formation of the *Truth* indicator based on Game Theory is proposed. Herein under the "game", we mean the process of assessing the information that an evaluating agent receives from another agent in the case when it is impossible to evaluate received data by his sensor devices or to rely on other agents' opinions. Therefore, two players have two strategies in this game. In the case of an evaluating agent, the strategy is the definition of the received data as correct, and its further processing or determination of these data as incorrect. In the case of the transmitting agent, there are strategies to transmit correct or incorrect data. Thus, the solution to the game is to find an equilibrium in a given situation. That is, for the evaluating agent, this is a strategy that gives it the maximum gain regardless of the transmitting agent. The payoff of the evaluating agent is the difference in the cost' growth rate for actions performance based on the evaluated data, received from the transmitting agent. In other words, the evaluating agent decides which information will lead to a smaller deviation in the cost' growth rate: less relevant information or disinformation.

In the previous study [8], we proposed to solve this game in pure strategies, which required to initialize the *Truth* indicator as a constant (0 in our case). The simulation results showed insufficient effectiveness (the *Accuracy* increased by only 1 %) since, in the case of *Truth* = 0, the effectiveness grows in proportion to the malicious agents in the system. To address this issue, we decided to solve the game in a mixed form, the outcome of which directly depends on the data obsolescence indicators, the cost' growth for disinformation, and the time of using these data. The solution of such a game gives the probability of choosing a particular strategy, which allows finding an equilibrium for games with different conditions and thereby gives a general solution to the problem.

To calculate the *Truth* indicator in the data incompleteness case, the process of information receiving is considered as a game with two players in normal form, where each agent has a finite number of possible strategies. The game can be characterized as [29]:

— discrete — the strategies set is discrete;
— finite — the strategies set is finite;
— strategic — the uncertainty comes from another player;
— in normal form — the payment matrix exists;
— antagonistic — the loss of one player is equal to the gain of the other.

Thus, let us define the game G according to the antagonistic game in normal form [28]:

$$G = (X, Y, K),$$

where X and Y are player's 1 and 2 strategies sets respectively; $K : X \times Y \to \mathbb{R}$ is player's 1 gain function. In this case, under player 1 e_T trusted agent that receives information is considered. Under player 2 e_U potential intruder agent that transmits information is meant. Table 2 represents agents' e_T and e_U strategies $x_i \in X$, $i \geq 1$ and $y_j \in Y$, $j \geq 1$, respectively.

Information correctness evaluation for data incompleteness case

To form a payoff matrix, we introduce the agent's e_T payoff function $K(x_i, y_j)$. Let there be a function for cost' growth rate calculating, which depends on the $H(x_i, y_j)$. Strategies, selected by the agents. According to the strategies outcomes, defined in Table 2, e_T's information is considered as: actual if agent e_T chooses 1st strategy and e_U — 2nd strategy $(x_1; y_2)$; less actual if e_T chooses 2nd strategy and e_U chooses 1st strategy $(x_2; y_1)$ or e_T chooses 2nd strategy and e_U also chooses 2nd strategy $(x_2; y_2)$; and disinformation if both e_T and e_U choose 1st strategy $(x_1; y_1)$. In the actual information and disinformation cases $t' = t$. Thus, determining the functions' (2), (3), and (4) first degree derivative, H is determined according to:

$$H(x_i, y_i) = \begin{cases} \begin{cases} k \times (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1), j = 1, i = 1 \\ k, j = 2 \end{cases} \\ k \times (a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t' - 1)) + 1), i = 2 \end{cases}. \quad (5)$$

Let us introduce the function for determining agent's e_T optimal strategy, which depends on the strategy chosen by e_U agent: $opt(y_j) = x_{jmod2+1}$. Then the payoff function $K(x_i, y_j)$ can be defined as the difference between the cost growth rate in the case when e_T knows e_U's strategy and the cost growth rate, which depends on the strategies selected by the agents:

$$K(x_i, y_j) = H(opt(y_j), y_j) - H(x_i, y_j). \quad (6)$$

According to the (5) and (6) equations, the generated payoff matrix is presented in Table 3.

As one can see from Table 3, situations (x_1, y_1) and (x_2, y_2) defined for the general case, and maximin cannot be defined. However, as the calculation of these functions results strictly less than zero, then $maximin \neq minimax$, $minimax = 0$. Therefore, it is not possible to solve the game in pure strategies and mixed strategies should be used. According to the book [28], $\exists \chi_i$: $\sum_{i=1}^{|X|} \chi_i = 1$, $1 \leq i \leq |X|$ is the

Table 2. Agents' strategies

Strategy counter l	x_l	y_l
1	To estimate information as a correct	To send incorrect information
2	To estimate information as an incorrect	To send correct information

Table 3. Payoff matrix

		e_U	
		y_1	y_2
e_T	x_1	$k \times \left((a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right) \right)$	0
	x_2	0	$k \times \left(1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} \right)$

pure strategy χ_i selection probability, and $\exists \gamma_j: \sum_{j=1}^{|Y|} \gamma_j = 1$, $1 \le j \le |Y|$ be the pure strategy y_j selection probability.

Then $\overline{X} = (\chi_1, \ldots, \chi_{|X|})$ and $\overline{Y} = (\gamma_1, \ldots, \gamma_{|Y|})$ are agents' e_T and e_U mixed strategies, respectively. Therefore, it is possible to define mixed strategies using the following equation systems:

$$\begin{cases} K(x_1, y_1) \times \chi_1 + K(x_2, y_1) \times \chi_2 = v_\Gamma \\ K(x_1, y_2) \times \chi_1 + K(x_2, y_2) \times \chi_2 = v_\Gamma \\ \chi_1 + \chi_2 = 1 \end{cases}$$

$$\Rightarrow \begin{cases} \left((a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right) \right) \times \chi_1 = \\ = \left(1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} \right) \times \chi_2, \\ \chi_2 = 1 - \chi_1 \end{cases}$$

$$\begin{cases} K(x_1, y_1) \times \gamma_1 + K(x_1, y_2) \times \gamma_2 = v_\Gamma \\ K(x_2, y_1) \times \gamma_1 + K(x_2, y_2) \times \gamma_2 = v_\Gamma \\ \gamma_1 + \gamma_2 = 1 \end{cases}$$

$$\Rightarrow \begin{cases} \left((a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right) \right) \times \gamma_1 = \\ = \left(1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} \right) \times \gamma_2, \\ \gamma_2 = 1 - \gamma_1 \end{cases}$$

where v_Γ is game value. As one could see these equation systems looking similar, so, we will solve the equation system only for the $\overline{X}$:

$$\begin{cases} \left((a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right) \right) \times \chi_1 = \\ = \left(1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} \right) \times \chi_2, \\ \chi_2 = 1 - \chi_1 \end{cases} \Rightarrow$$

$$\Rightarrow \left((a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right) \right) \times \chi_1 =$$

$$= \left(1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} \right) \times (1 - \chi_1) \Rightarrow$$

$$\Rightarrow \chi_1 = \frac{1 - (a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}}}{1 - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right)} \Rightarrow$$

$$\Rightarrow \chi_2 = \frac{(a_{Inf})^{t'-t} \times \ln \frac{e}{(a_{Inf})^{t'-t+1}} - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right)}{1 - (a'_{Inf})^{-1} \times \ln\left(\frac{e}{a'_{Inf}} \right)}.$$

The equation system for the $\overline{Y}$ are solving the same, therefore, $\chi_1 = \gamma_1$ and $\chi_2 = \gamma_2$.

As a result of solving the game, the obtained mixed strategies can be formalized according to:

$$\overline{X} = \left(\frac{1 - (a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t'-1)) + 1)}{1 - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)}, \right.$$

$$\left. \frac{(a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t'-1)) + 1) - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)}{1 - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)} \right)$$

$$\overline{Y} = \left(\frac{(1 - (a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t'-1)) + 1))}{1 - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)}, \right.$$

$$\left. \frac{(a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t'-1)) + 1) - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)}{1 - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)} \right).$$

On the basis of agent's e_T strategies, defined in Table 2, the *Truth* indicator directly depends on the probability of evaluating the information as correct. Therefore, $Truth = \chi_1$, and in the data incompleteness case the *Truth* indicator is calculated according to:

$$Truth = \frac{(1 - (a_{Inf})^{t'-t} \times (-\ln a_{Inf} \times (t - (t'-1)) + 1))}{1 - (a'_{Inf})^{-1} \times (-\ln a'_{Inf} + 1)}.$$

Since the *Truth* indicator can be greater than 1 in this case, we assume that all values greater than 1 is equated to 1: $Truth > 1 \rightarrow Truth = 1$.

Empirical study

Simulation setup

To evaluate the effectiveness of the proposed model, we conducted an empirical study using a custom software simulator. As one of the CPS's possible implementations, we considered the simulation of the intersection management system with multiple unmanned autonomous vehicles, which need to perform conflict-free optimal intersection traversal with minimal costs [7]. In the present study, costs are represented as a number of sectors that agent overcomes to reach its path's finish point.

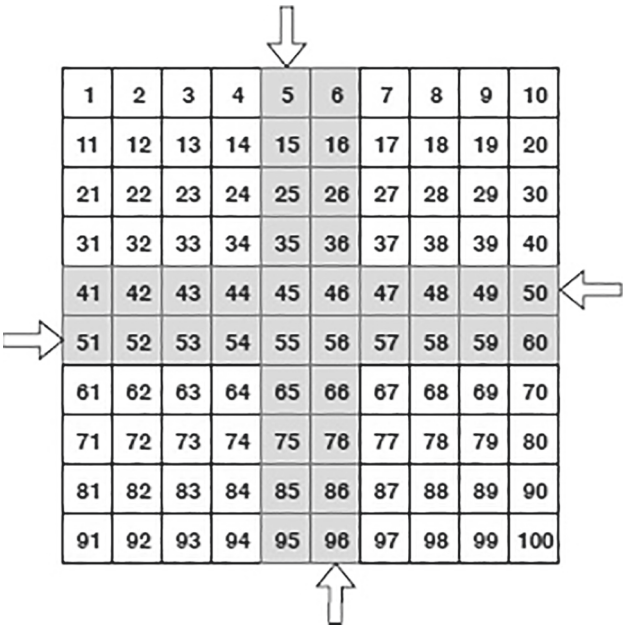


Fig. 1. Model of intersection and schematic representation of the vehicles' driving direction

The intersection scheme is represented in Fig. 1. It has the following properties:

— software testing ground is divided into equal sectors, and each sector has its unique number;
— software testing ground size: 10 × 10 sectors;
— software testing ground has 4 roads: two vertical (oncoming and passing) and two horizontal (oncoming and passing) ones.

As an assumption, we initially set the number of agents, which provide false data (intruders). During the system operation process, intruders implement on-off attack [30]. The purpose of this attack is to compromise Reputation mechanism and decrease system's effectiveness via the alternating transmission of correct and false data. In our experiments, on-off attack cycle is 2-on, and 1-off, i.e. intruders transmit false data during 2 discrete time moments and correct data during the next 1 discrete time moment.

To detect intruders, agents transmit and evaluate information about their current location. For such an assessment, they use their sensor devices, which can obtain data from the surroundings within a radius of 1 sector, that is, in 8 sectors around the agent. Moreover, the agent can request the assessment of other agents in case when it is not able to evaluate the data received. The radius of information interaction between agents is 9 sectors. All agents located on the software testing ground have the ability to interact.

Experiments were performed using raw Reputation and Trust indicators and using the proposed Game Theory approach. In each group of the experiment series, 1000 simulations were conducted with the various intruders percentage: 10, 20, 30, and 40 % from 1000 agents in a group. To evaluate the approach effectiveness, we introduced the following metrics:

— False Positive (FP) — the ratio of legitimate agents, that were incorrectly identified as intruders (relatively to all agents in a group);
— False Negative (FN) — the ratio of intruders, that were incorrectly identified as legitimate agents;
— Accuracy — the ratio of agents, that were correctly identified as legitimate agents or saboteurs;
— Precision — the ratio of intruders, that were correctly identified as false data providers relative to all agents identified as intruders;
— Recall — the ratio of intruders, that were correctly identified as false data provides relative to all agents in the group;
— F0.5 — weighted harmonic mean of Precision and Recall metrics, when $\beta = 0.5$;
— F1 — harmonic mean of Precision and Recall metrics, when $\beta = 1$;
— F2 — harmonic mean of Precision and Recall metrics, when $\beta = 2$.

During the simulation process, Accuracy, Recall, FP and FN error metrics were defined and employed in the following way:

— FN error occurred when the agent provides false data and is perceived by the rest of the group as legitimate. The likelihood of collision increases in this case;
— FP error occurred when the agent provides correct data and is perceived by the rest of the group as an intruder, which results in a system's effectiveness decrease;
— True Negative (TN) case is occurred when the incorrect information transmitted by the intruder is perceived by the rest of the group as incorrect;
— True Positive (TP) case is occurred when the correct information transmitted by the legitimate agent is perceived by the rest of the group as correct;
— $\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}$;
— $\text{Recall} = \dfrac{TP}{TP + FN}$.

Simulation results

Fig. 2 demonstrates the obtained results for 10, 20, 30, and 40 % intruders in the group. Averaged indicators' values are presented in Fig. 3. TP and TN values are not presented in the figures, although they were used for Accuracy and Recall calculation.

Fig. 2 demonstrates that the proposed Game Theory-based approach seems to be more sensible in relation to intruders detection, i.e., more elements are likely to be identified as intruders if their behavior deviates from "normal". Thus, the basic approach is more characterized by "skipping" intruders to increase the number of elements involved in the system. In addition, Game Theory-based approach did not show a significant change in the efficiency of legitimate agents' identification. This is evidenced by the values of $F(\beta = 0.5)$, $F(\beta = 1)$ and $F(\beta = 2)$.

Comparing the results presented in Fig. 2, one can observe that the values of Accuracy, $F(\beta = 0.5)$, and $F(\beta = 1)$ tend to the values obtained with $R = 0.5$ as the intruders proportion in the group increases. According to Fig. 3, the Accuracy of intruder identification increased by 15 % on average. Moreover, FN errors decreased by an average of 3 times, and FP errors increased by 1.7 times, which also decreased the average value of the Precision metric. As

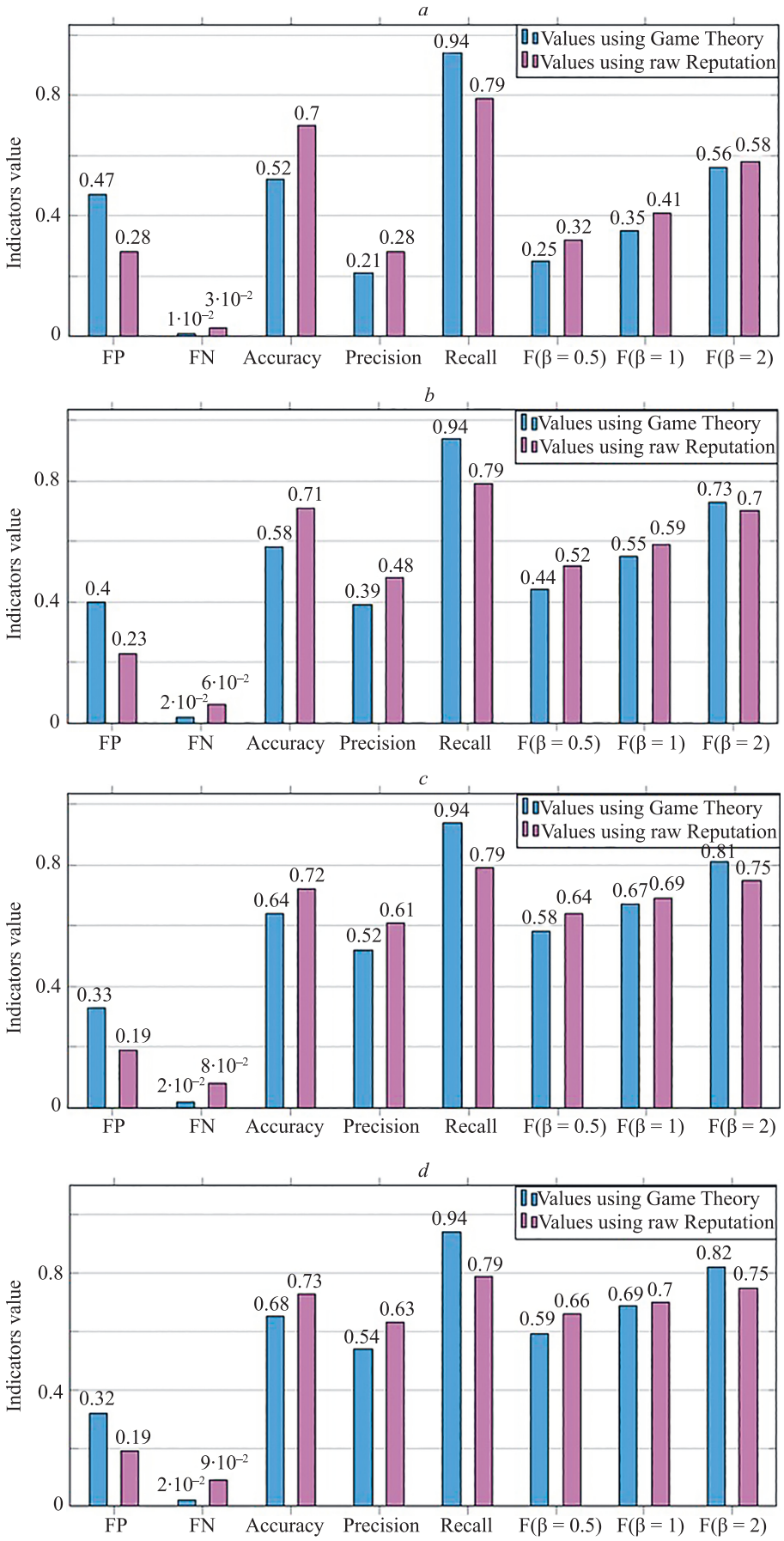


Fig. 2. FP, FN, Accuracy, Precision, Recall, and F-measure values, averaged on 1000 agents classification experiments, with and without Game Theory approach, for the case with: 10 % (*a*); 20 % (*b*); 30 % (*c*); 40 % (*d*) of intruders in the group

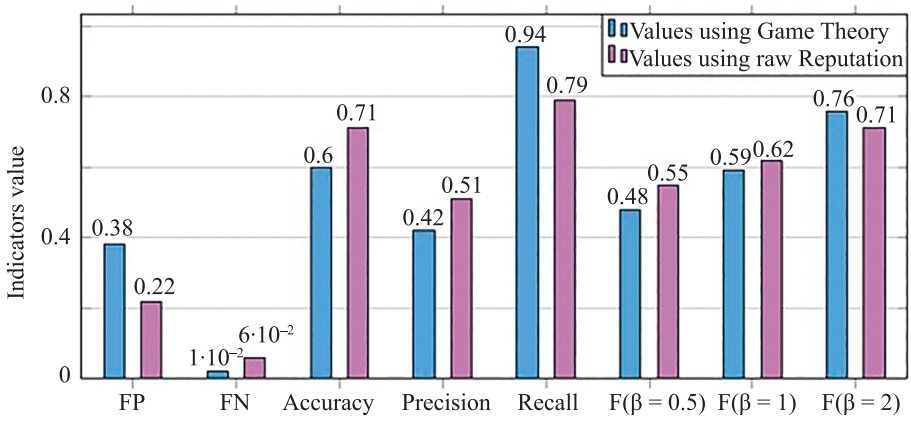


Fig. 3. Averaged FP, FN, Accuracy, Precision, Recall, and F-measure values for all experiment series (10–40 % of intruders), with and without Game Theory approach

a result of modeling the developed approach, the Recall metric had been increased. As the Recall shows the ratio of detected intruders, the developed model demonstrates a better result than proposed in previous studies. Given the larger number of elements functioning in "normal" mode, we can say that despite the use of the proposed approach may reduce the performance of the system (speed of task execution, cost of task execution, etc.), it can also increase the probability of successful tasks' execution. The implementation of the proposed approach can be practically appropriate in case of cyber-physical systems supposed to work in an aggressive environment, for instance, in the group of UAV designed for environmental monitoring tasks [31]. The use of the proposed approach allows one to organize the verification of sensitive information and, as an example, can increase the chance of human rescue in case of emergency rescue operations.

Further work will be aimed at improving the results on other indicators. Compared to previous work [8], the results obtained in this study are more reliable, as during the experiments, the number of agents in the group was increased, and various intruders ratio were simulated. The advantage of the presented improved approach is the dynamic calculation of the Truth value. In the earlier work, such an indicator was constant in cases when it was not possible to obtain the data on the agent's preceding behavior. The advancement of the presented approach allows us to make the Truth indicator more flexible and to adjust it to the conditions of the system. In addition, further research will focus on the implementation of the proposed approach in real UAV groups designed for ground objects detection purposes.

Conclusion

In this paper, we proposed the enhanced Reputation, Trust, and Game Theory-based model to improve cyber-physical system elements' security and safety. To address the Reputation initial value calculation challenge, we described the intruders identification procedure in terms of Game Theory, applied the game concept between intruders and legitimate agents, and formalized group members strategies. The possible outcomes of using different strategies are represented with a payoff matrix. To verify our enhanced approach, we conducted an empirical study using a custom software simulator. Multiple experiments were performed with a group of agents able to interact with each other. Cases with a 10–40 % of intruders from the whole agents' group were simulated. Despite the fact that the probability to incorrectly classify a legitimate agent as an intruder increased, which also reduced the Precision metric, results analysis showed that our model implementation allowed us to significantly increase intruders detection Accuracy and to reduce the intruders incorrect classification probability compared with raw Reputation and Trust model. This specific characteristic can be vital in systems, which are not tolerant of the high risk of damage acceptance.

Our further research plans include implementation and assessing the proposed model on a developed intersection management physical testing ground, with models of autonomous vehicles, presented by us in [7]. As the previous study has shown, Reputation and Trust approach's practical implementation allows one to effectively detect "soft" attacks in the intersection management system, organized by the agents that transmit incorrect data. We assume that implementation of the proposed Game Theory mechanisms on real physical models will allow increasing "soft" attacks detection accuracy, including the cases when agents do not have retrospective data, on the basis of which they can calculate the Reputation value.

References

1. Bastos D., Shackleton M., El-Moussa F. Internet of Things: A survey of technologies and security risks in smart home and city environments. *IET Conference Publications*, 2018, vol. 2018, no. CP740. https://doi.org/10.1049/cp.2018.0030
2. Kang H.S., Lee J.Y., Choi S., Kim H., Park J.H., Son J.Y., Kim B.H., Do Noh S. Smart manufacturing: Past research, present findings, and future directions. *International Journal of Precision Engineering and Manufacturing — Green Technology*, 2016, vol. 3, no. 1, pp. 111–128. https://doi.org/10.1007/s40684-016-0015-5
3. Wolf W. Cyber-physical systems. *Computer*, 2009, vol. 42, no. 3, pp. 88–89. https://doi.org/10.1109/MC.2009.81
4. Tokody D., Albini A., Ady L., Rajnai Z., Pongrácz F. Safety and security through the design of autonomous intelligent vehicle systems and intelligent infrastructure in the smart city. *Interdisciplinary Description of Complex Systems*, 2018, vol. 16, no. 3, pp. 384–396. https://doi.org/10.7906/indecs.16.3.11
5. Kalyaev I., Gaiduk A., Kapustyan S. *Models and Algorithms of the Collective Control of Robots Group*. Moscow, Fizmatlit Publ., 2009, 278 p. (in Russian)
6. Furno L., Nielsen M.C., Blanke M. Centralised versus decentralised control reconfiguration for collaborating underwater robots. *IFAC-PapersOnLine*, 2015, vol. 48, no. 21, pp. 732–739. https://doi.org/10.1016/j.ifacol.2015.09.614
7. Chuprov S., Viksnin I., Kim I., Marinenkov E., Usova M., Lazarev E., Melnikov T., Zakoldaev D. Reputation and trust approach for security and safety assurance in intersection management system. *Energies*, 2019, vol. 12, no. 23, pp. 4527. https://doi.org/10.3390/en12234527
8. Marinenkov E., Chuprov S., Viksnin I., Kim I. Empirical study on trust, reputation, and game theory approach to secure communication in a group of unmanned vehicles. *CEUR Workshop Proceedings*, 2019, vol. 2590.
9. Gibb J.R. Trust: *A New View of Personal and Organizational Development*. Guild of Tutors Press, 1978. 320 p.
10. Mui L., Mohtashemi M., Halberstadt A. A computational model of trust and reputation. *Proc. of the 35th Annual Hawaii International Conference on System Sciences*, 2002, pp. 2431–2439. https://doi.org/10.1109/HICSS.2002.994181
11. Schollmeier R. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. *Proc. of the 1st International Conference on Peer-to-Peer Computing (P2P)*, 2001, pp. 101–102. https://doi.org/10.1109/P2P.2001.990434
12. *Sector S. Series Y: Global information infrastructure, internet protocol aspects and next-generation networks. Next generation networks – frameworks and functional architecture models*: Recommendation ITU-T Y 2012 / International Telecommunication Union. Switzerland, Geneva, P. 2060.
13. Singh A., Kumar M., Rishi R., Madan D. A relative study of MANET and VANET: Its applications, broadcasting approaches and challenging issues. *Communications in Computer and Information Science*, 2011, vol. 132 CCIS, part 2, pp. 627–632. https://doi.org/10.1007/978-3-642-17878-8_63
14. Chmaj G., Walkowiak K. A P2P computing system for overlay networks. *Future Generation Computer Systems*, 2013, vol. 29, no. 1, pp. 242–249. https://doi.org/10.1016/j.future.2010.11.009
15. Nojoumian M., Stinson D.R. Social secret sharing in cloud computing using a new trust function. *Proc. of the 10th Annual International Conference on Privacy, Security and Trust (PST)*, 2012, pp. 161–167. https://doi.org/10.1109/PST.2012.6297936
16. Straub J., McMillan J., Yaniero B., Schumacher M., Almosalami A., Boatey K., Hartman J. CyberSecurity considerations for an interconnected self-driving car system of systems. *Proc. of the 12th System of Systems Engineering Conference (SoSE)*, 2017, pp. 7994973. https://doi.org/10.1109/SYSOSE.2017.7994973
17. Kim I., Viksnin I. Secure information interaction within a group of unmanned aerial vehicles based on economic approach. *Advances in Intelligent Systems and Computing*, 2019, vol. 997, pp. 59–72. https://doi.org/10.1007/978-3-030-22871-2_5
18. Pham T.N.D., Yeo C.K. Adaptive trust and privacy management framework for vehicular networks. *Vehicular Communications*, 2018, vol. 13, pp. 1–12. https://doi.org/10.1016/j.vehcom.2018.04.006
19. Chuprov S., Viksnin I., Kim I., Nedosekin G. Optimization of autonomous vehicles movement in urban intersection management system. *Proc. of the 24th Conference of Open Innovations Association* (FRUCT), 2019, pp. 60–66. https://doi.org/10.23919/FRUCT.2019.8711967

Литература

1. Bastos D., Shackleton M., El-Moussa F. Internet of Things: A survey of technologies and security risks in smart home and city environments // IET Conference Publications. 2018. V. 2018. N CP740. https://doi.org/10.1049/cp.2018.0030
2. Kang H.S., Lee J.Y., Choi S., Kim H., Park J.H., Son J.Y., Kim B.H., Do Noh S. Smart manufacturing: Past research, present findings, and future directions // International Journal of Precision Engineering and Manufacturing – Green Technology. 2016. V. 3. N 1. P. 111–128. https://doi.org/10.1007/s40684-016-0015-5
3. Wolf W. Cyber-physical systems // Computer. 2009. V. 42. N 3. P. 88–89. https://doi.org/10.1109/MC.2009.81
4. Tokody D., Albini A., Ady L., Rajnai Z., Pongrácz F. Safety and security through the design of autonomous intelligent vehicle systems and intelligent infrastructure in the smart city // Interdisciplinary Description of Complex Systems. 2018. V. 16. N 3. P. 384–396. https://doi.org/10.7906/indecs.16.3.11
5. Каляев И.А., Гайдук А.Р., Капустян С.Г. Модели и алгоритмы коллективного управления в группах роботов. М.: Физматлит, 2009. 278 с.
6. Furno L., Nielsen M.C., Blanke M. Centralised versus decentralised control reconfiguration for collaborating underwater robots // IFAC-PapersOnLine. 2015. V. 48. N 21. P. 732–739. https://doi.org/10.1016/j.ifacol.2015.09.614
7. Chuprov S., Viksnin I., Kim I., Marinenkov E., Usova M., Lazarev E., Melnikov T., Zakoldaev D. Reputation and trust approach for security and safety assurance in intersection management system // Energies. 2019. V. 12. N 23. P. 4527. https://doi.org/10.3390/en12234527
8. Marinenkov E., Chuprov S., Viksnin I., Kim I. Empirical study on trust, reputation, and game theory approach to secure communication in a group of unmanned vehicles // CEUR Workshop Proceedings. 2019. V. 2590.
9. Gibb J.R. Trust: A New View of Personal and Organizational Development. Guild of Tutors Press, 1978. 320 p.
10. Mui L., Mohtashemi M., Halberstadt A. A computational model of trust and reputation // Proc. of the 35th Annual Hawaii International Conference on System Sciences. 2002. P. 2431–2439. https://doi.org/10.1109/HICSS.2002.994181
11. Schollmeier R. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications // Proc. of the 1st International Conference on Peer-to-Peer Computing (P2P). 2001. P. 101–102. https://doi.org/10.1109/P2P.2001.990434
12. Sector S. Series Y: Global information infrastructure, internet protocol aspects and next-generation networks. Next generation networks – frameworks and functional architecture models: Recommendation ITU-T Y 2012 / International Telecommunication Union. Switzerland, Geneva. P. 2060.
13. Singh A., Kumar M., Rishi R., Madan D. A relative study of MANET and VANET: Its applications, broadcasting approaches and challenging issues // Communications in Computer and Information Science. 2011. V. 132 CCIS. Part 2. P. 627–632. https://doi.org/10.1007/978-3-642-17878-8_63
14. Chmaj G., Walkowiak K. A P2P computing system for overlay networks // Future Generation Computer Systems. 2013. V. 29. N 1. P. 242–249. https://doi.org/10.1016/j.future.2010.11.009
15. Nojoumian M., Stinson D.R. Social secret sharing in cloud computing using a new trust function // Proc. of the 10th Annual International Conference on Privacy, Security and Trust (PST). 2012. P. 161–167. https://doi.org/10.1109/PST.2012.6297936
16. Straub J., McMillan J., Yaniero B., Schumacher M., Almosalami A., Boatey K., Hartman J. CyberSecurity considerations for an interconnected self-driving car system of systems // Proc. of the 12th System of Systems Engineering Conference (SoSE). 2017. P. 7994973. https://doi.org/10.1109/SYSOSE.2017.7994973
17. Kim I., Viksnin I. Secure information interaction within a group of unmanned aerial vehicles based on economic approach // Advances in Intelligent Systems and Computing. 2019. V. 997. P. 59–72. https://doi.org/10.1007/978-3-030-22871-2_5
18. Pham T.N.D., Yeo C.K. Adaptive trust and privacy management framework for vehicular networks // Vehicular Communications. 2018. V. 13. P. 1–12. https://doi.org/10.1016/j.vehcom.2018.04.006
19. Chuprov S., Viksnin I., Kim I., Nedosekin G. Optimization of autonomous vehicles movement in urban intersection management system // Proc. of the 24th Conference of Open Innovations Association (FRUCT). 2019. P. 60–66. https://doi.org/10.23919/FRUCT.2019.8711967

20. Basar T., Zaccour G. *Handbook of Dynamic Game Theory*. Springer, 2018.
21. Fielder A., Panaousis E., Malacaria P., Hankin C., Smeraldi F. Game theory meets information security management. *IFIP Advances in Information and Communication Technology*, 2014, vol. 428, pp. 15–29. https://doi.org/10.1007/978-3-642-55415-5_2
22. Roy S., Ellis C., Shiva S., Dasgupta D., Shandilya V., Wu Q. A survey of game theory as applied to network security. *Proc. of the 43rd Hawaii International Conference on System Sciences (HICSS-43)*, 2010, pp. 5428673. https://doi.org/10.1109/HICSS.2010.35
23. Sun W., Kong X., He D., You X. Information security problem research based on game theory. *Proc. of the International Symposium on Electronic Commerce and Security (ISECS)*, 2008, pp. 554–557. https://doi.org/10.1109/ISECS.2008.147
24. Guo J., Chen R. A classification of trust computation models for service-oriented internet of things systems. *Proc. of the IEEE International Conference on Services Computing*, 2015, pp. 324–331. https://doi.org/10.1109/SCC.2015.52
25. Bankovic Z., Vallejo J.C., Fraga D., Moya J.M. Detecting false testimonies in reputation systems using self-organizing maps. *Logic Journal of the IGPL*, 2013, vol. 21, no. 4, pp. 549–559. https://doi.org/10.1093/jigpal/jzs028
26. Li W., Song H., Zeng F. Policy-based secure and trustworthy sensing for internet of things in smart cities. *IEEE Internet of Things Journal*, 2018, vol. 5, no. 2, pp. 716–723. https://doi.org/10.1109/JIOT.2017.2720635
27. Rawat D.B., Yan G., Bista B.B., Weigle M.C. Trust on the security of wireless vehicular Ad-hoc Networking. *Ad-Hoc and Sensor Wireless Networks*, 2015, vol. 24, no. 3-4, pp. 283–305.
28. Petrosian L.A., Zenkevich N.A., Shevkoplias E.V. *Game Theory*. St. Petersburg, BHV-Peterburg, 2012, 432 p. (in Russian)
29. Kuzin L. *Foundations of Cybernetics*. In 2 vol. Moscow, Energia Publ., 1979. (in Russian)
30. Perrone L.F., Nelson S.C. A study of on-off attack models for wireless ad hoc networks. *Proc. of the 1st Workshop on Operator-Assisted (Wireless Mesh) Community Networks (OpComm)*, 2006, pp. 4138221. https://doi.org/10.1109/WOACN.2006.337180
31. Baker C.A.B., Ramchurn S., Teacy W.T.L., Jennings N.R. Planning search and rescue missions for UAV teams. *Frontiers in Artificial Intelligence and Applications*, 2016, vol. 285, pp. 1777–1778. https://doi.org/10.3233/978-1-61499-672-9-1777

20. Basar T., Zaccour G. Handbook of Dynamic Game Theory. Springer, 2018.
21. Fielder A., Panaousis E., Malacaria P., Hankin C., Smeraldi F. Game theory meets information security management // IFIP Advances in Information and Communication Technology. 2014. V. 428. P. 15–29. https://doi.org/10.1007/978-3-642-55415-5_2
22. Roy S., Ellis C., Shiva S., Dasgupta D., Shandilya V., Wu Q. A survey of game theory as applied to network security // Proc. of the 43rd Hawaii International Conference on System Sciences (HICSS-43). 2010. P. 5428673. https://doi.org/10.1109/HICSS.2010.35
23. Sun W., Kong X., He D., You X. Information security problem research based on game theory // Proc. of the International Symposium on Electronic Commerce and Security (ISECS). 2008. P. 554–557. https://doi.org/10.1109/ISECS.2008.147
24. Guo J., Chen R. A classification of trust computation models for service-oriented internet of things systems // Proc. of the IEEE International Conference on Services Computing. 2015. P. 324–331. https://doi.org/10.1109/SCC.2015.52
25. Bankovic Z., Vallejo J.C., Fraga D., Moya J.M. Detecting false testimonies in reputation systems using self-organizing maps // Logic Journal of the IGPL. 2013. V. 21. N 4. P. 549–559. https://doi.org/10.1093/jigpal/jzs028
26. Li W., Song H., Zeng F. Policy-based secure and trustworthy sensing for internet of things in smart cities // IEEE Internet of Things Journal. 2018. V. 5. N 2. P. 716–723. https://doi.org/10.1109/JIOT.2017.2720635
27. Rawat D.B., Yan G., Bista B.B., Weigle M.C. Trust on the security of wireless vehicular Ad-hoc Networking // Ad-Hoc and Sensor Wireless Networks. 2015. V. 24. N 3-4. P. 283–305.
28. Петросян Л.А., Зенкевич Н.А., Шевкопляс Е.В. Теория игр. СПб.: БХВ-Петербург, 2012. 432 с.
29. Кузин Л.Т. Основы кибернетики. В 2 т. М.: Энергия, 1979.
30. Perrone L.F., Nelson S.C. A study of on-off attack models for wireless ad hoc networks // Proc. of the 1st Workshop on Operator-Assisted (Wireless Mesh) Community Networks (OpComm). 2006. P. 4138221. https://doi.org/10.1109/WOACN.2006.337180
31. Baker C.A.B., Ramchurn S., Teacy W.T.L., Jennings N.R. Planning search and rescue missions for UAV teams // Frontiers in Artificial Intelligence and Applications. 2016. V. 285. P. 1777–1778. https://doi.org/10.3233/978-1-61499-672-9-1777

Authors

Ilia I. Viksnin — PhD, Associate Professor, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation, SC 57191359693, https://orcid.org/0000-0001-6240-0390, wixnin@mail.ru

Egor D. Marinenkov — Scientific Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57212198065, https://orcid.org/0000-0001-9895-239X, egormarinenkov@gmail.com

Sergey S. Chuprov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57208318420, https://orcid.org/0000-0001-7081-8797, drmyscull@gmail.com

Авторы

Викснин Илья Игоревич — кандидат технических наук, доцент, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова, Санкт-Петербург, 197022, Российская Федерация, SC 57191359693, https://orcid.org/0000-0001-6240-0390, wixnin@mail.ru

Мариненков Егор Денисович — научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57212198065, https://orcid.org/0000-0001-9895-239X, egormarinenkov@gmail.com

Чупров Сергей Сергеевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57208318420, https://orcid.org/0000-0001-7081-8797, drmyscull@gmail.com

Received 10.11.2021
Approved after reviewing 21.12.2021
Accepted 29.01.2022

Статья поступила в редакцию 10.11.2021
Одобрена после рецензирования 21.12.2021
Принята к печати 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-60-66
УДК 004.942

Исследование влияния человеческих факторов на скорость движения рельсового городского транспорта

Юлия Александровна Ляховенко[1], Илья Юрьевич Попов[2]✉

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] lyakhovenko.kam@gmail.com, https://orcid.org/0000-0001-7396-2831
[2] ilyapopov27@gmail.com✉, https://orcid.org/0000-0002-6407-7934

Аннотация

Предмет исследования. Представлена модель оценки состояния транспортных узлов движения общественного рельсового транспорта. Исследована зависимость скорости движения рельсового городского транспорта от влияния внешних случайных факторов, зависящих от человека (human factors). Учтены такие факторы, как движение других транспортных средств и пешеходов, ремонтные работы на участках дорог, а также плотность потока транспортных средств и пешеходов. Предложенная модель оценки транспортного узла содержит множество перекрестков, линий движения рельсового транспорта и полос смешенного движения в рамках правил дорожного движения. **Метод.** Решение задачи построено на методологии мультиагентных систем. Основа предложенного подхода — определение архитектуры отдельных агентов, входных параметров предполагаемых откликов системы. Для построения модели движения транспортных потоков с различными видами транспортных средств использована программная платформа PTV Vissim. **Основные результаты.** В ходе симуляции мультиагентной системы выявлена значительная зависимость скорости движения городского рельсового транспорта от плотности потока движения и наличия ремонтных работ. Отличительная особенность предложенного подхода — учет влияния человеческого фактора. **Практическая значимость.** Полученный подход может быть использован при проектировании транспортных узлов беспрепятственного движения беспилотного рельсового городского транспорта.

Ключевые слова

многоагентные системы, моделирование, железнодорожный транспорт, human factors, моделирование транспортных потоков

Благодарности

Работа выполнена в Университете ИТМО при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках проекта 2019-0898 «Многоуровневое управление сложными техническими системами».

Ссылка для цитирования: Ляховенко Ю.А., Попов И.Ю. Исследование влияния человеческих факторов на скорость движения рельсового городского транспорта // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 60–66. doi: 10.17586/2226-1494-2022-22-1-60-66

A study of the influence of human factors on the speed of urban rail transport

Julia A. Lyakhovenko[1], Ilya Yu. Popov[2]✉

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] lyakhovenko.kam@gmail.com, https://orcid.org/0000-0001-7396-2831
[2] ilyapopov27@gmail.com✉, https://orcid.org/0000-0002-6407-7934

Abstract

The paper presents a model for assessing the state of transport hubs of public rail transport and investigates the dependence of the movement speed of urban rail transport on the influence of external random human-based factors. The study considers the following factors: the movement of other vehicles and pedestrians, repair on road sections, and the density of traffic of vehicles and pedestrians. The proposed model of the transport hub contains many intersections,

© Ляховенко Ю.А., Попов И.Ю., 2022

rail traffic lines, and mixed traffic lanes within the framework of the traffic rules. The solution to the problem is based on the methodology of multi-agent systems. The basis of the proposed approach is the definition of the architecture of individual agents and the input parameters of the expected system responses. The software platform PTV Vissim, which allows building models of traffic flows with various types of vehicles, is used. During the simulation of the multi-agent system, a significant dependence of the speed of urban rail transport on the traffic density and the presence of repair work was revealed. A distinctive feature of the proposed approach is that it considers the influence of the human factor. The obtained approach can be used to design transport hubs for the unimpeded movement of unmanned urban rail transport.

Keywords
multi-agent systems, modeling, railway transport, human factors, modeling of traffic flows

Acknowledgements
This work is partially supported by the Ministry of Science and Higher Education of Russian Federation, state assignment No. 2019-0898.

For citation: Lyakhovenko Ju.A., Popov I.Yu. A study of the influence of human factors on the speed of urban rail transport. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 60–66 (in Russian). doi: 10.17586/2226-1494-2022-22-1-60-66

Введение

Влияние человека способствует большому количеству инцидентов, которые в том числе происходят в железнодорожной системе [1]. Оперативное выявление и устранение человеческих факторов (human factors), а также сдерживание последствий, возникающих в результате их возникновения, позволяют сделать железнодорожный транспорт безопаснее [2]. Первый шаг в процессе обеспечения безопасности — выявление факторов, которые часто приводят к возникновению инцидентов, что позволяет разработать соответствующие стратегии предотвращения и/или смягчения последствий.

На сегодняшний день в России нет исследований, которые бы систематически изучали влияние человеческого фактора на систему железнодорожного движения.

В настоящей работе выполнен обзор существующих подходов к выявлению зависимости железнодорожного движения от человеческого фактора. Представлен анализ влияния внешних случайных факторов на скорость движения рельсового городского транспорта — трамвая, на основе которого составлена мультиагентная система зависимости скорости движения трамвая от этих факторов. Выполнено моделирование представленной системы и обсуждение его результатов.

Обзор предметной области

Существует широкая область, охватывающая бдительность и восприятие машинистов поездов, их распознавание и действия в соответствии со знаками и сигналами. Это включает также исследования дизайна систем указателей и сигнализации. Из всех вопросов человеческого фактора на железнодорожном транспорте эта область была наиболее изучена на протяжении многих лет [3, 4]. Одна из причин данного изучения — связь с сильными исследовательскими программами, охватывающими те же темы, что и вождение автомобиля.

В работах [5, 6] изучено влияние человеческого фактора на работу планировщиков, например, в составлении расписания, организации владения (пути для технического обслуживания) и для работы в чрезвычайных ситуациях. В [7] рассмотрены модели системного уровня, включая модели распределенного познания всей железнодорожной системы, а также взаимодействия и репрезентации человек-человек, и человек-артефакт. В [8] начат анализ руководств и стандартов безопасности, связанных с человеческим фактором для получения соответствия требованиям железнодорожных приложений и руководства ресурсами, планирующими и проектирующими сети завтрашнего дня [8]. В значительной степени связанный с подходом к общей системной эргономике, человеческий фактор на железнодорожном транспорте также дал дальнейшее развитие понятию и практике интеграции человеческого фактора при разработке систем. Достижения в этой области часто скрываются в документации по коммерческим проектам [9].

Анализ факторов

Скорость движения трамвая в контексте представленной работы зависит от следующих факторов: от других участников дорожного движения (транспортных средств (ТС) и пешеходов (пассажиров), нуждающихся в посадке/высадке в/из трамвая); ремонтных работ и уровня загруженности участка дорожного движения.

Трамвай описывается следующим набором атрибутов параметров движения: координаты и направление; средняя скорость на участке; наличие остановки на участке.

Разрабатываемая для исследования модель имитирует процессы дорожного движения трамвая с влиянием на него всех перечисленных выше факторов.

Трамваи могут двигаться только по рельсам (на дороге или выделенных полосах), останавливаясь на перекрестках при запрещающем сигнале светофора, а также при высадке и посадке пассажиров.

Автомобили могут двигаться по любым участкам дороги, кроме тротуаров и выделенных полос, уступая дорогу трамваям.

На рис. 1 представлена блок-схема алгоритма движения трамвая в условиях влияния human factors.

Исследование зависимости скорости движения трамвая от human factors

Система дорожного движения с участием трамваев, ТС и пешеходов (далее — система) описана с помо-

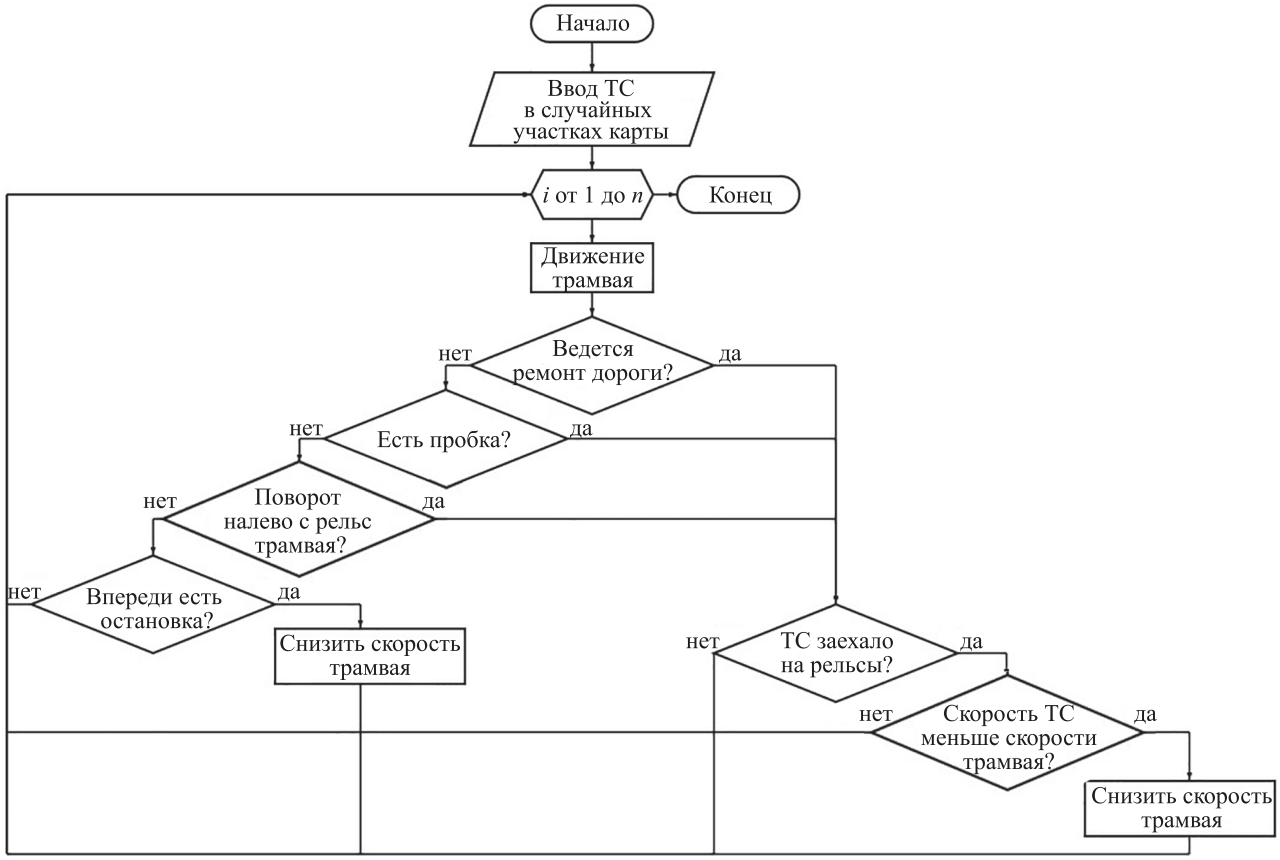


Рис. 1. Блок-схема алгоритма движения трамвая в условиях влияния human factors, где *i* — индекс итерации цикла движения

Fig. 1. The block diagram of the tram movement algorithm under the influence of external random factors, where *i* is the motion loop iteration index

щью мультиагентного подхода. Предлагаемая система рассматривает процесс дорожного движения как набор временных интервалов $T = \{t_0, t_1, \ldots, t_m\}$, т. е. можно предположить о существовании системы в момент времени t_i и в предыдущий момент времени t_{i-1}.

Система имеет следующие виды агентов: трамваи; ТС; пешеходы.

Пусть группа трамваев состоит из n элементов: $A = \{a_1, \ldots, a_n\}$. Каждый агент трамвай в момент времени t_j описывается следующим набором атрибутов:

$$a_{i,t_j} = \{(x, y)_{i,t_j}, v_{i,t_j}, stop_{i,t_{j+1}}\},$$

где $(x, y)_{i,t_j}$ и v_{i,t_j} — местоположение и скорость агента a_i в момент времени t_j; $stop_{t_{j+1}}$ — наличие остановки в момент времени t_{j+1} для агента a_i.

Отметим, что обозначение *fun* представляет некоторую зависимость параметров, т. е. $b = fun(c)$ — зависимость параметра b от параметра c.

Определим состояние участка дороги в момент времени t_j:

$$S_{x,y,t_j} = \{traffic\_jam_{roal\_line_{x,y,t_j}}, repair_{x,y,t_j}\},$$

где $traffic\_jam_{roal\_line_{x,y,t_j}}$ — плотность потока на дороге в момент времени t_j в баллах на дороге $roal\_line_{x,y,t_j}$; $repair_{x,y,t_j}$ — коэффициент, показывающий наличие ремонтных работ на участке (x, y): при 1 — ведется ремонт дороги, 0 — в обратном случае.

Зависимость скорости от плотности потока многополосной и однополосной дорог приведена в работе [10]. Авторы предложили переход к усредненному представлению скорости и плотности потока дороги, по аналогии с многокомпонентной односкоростной моделью баротропного газа и решения численного интегрирования системы из $S + 1$ уравнений для многоскоростной или односкоростной систем, где S — количество полос на дороге. Так, перенос концентрации ТС между разными полосами (в том числе заезд на полосы движения трамвая) описаны уравнениями конвективно-диффузного переноса.

В рамках настоящей работы под состоянием дороги понимается величина, сильно влияющая на скорость движения трамвая. Тогда скорость движения трамвая a_i в момент времени t_j на участке (x, y) определяется состоянием трамвая и внешними факторами (состоянием дороги S_{x,y,t_j}):

$$v_{i,x,y,t_j} = fun(a_{i,t_j}, S_{x,y,t_j}).$$

Получим величину скорости, на которую влияет S_{x,y,t_j}:

$$v_{i,x,y,t_j} = fun(S_{x,y,t_j}) \times fun(a_{i,t_j}) =$$
$$= Z_{x,y,t_j} \times fun(repair_{x,y,t_j}) \times fun(a_{i,t_j}),$$

где $Z_{x,y,t} = fun(traffic\_jam_{roal\_line_{x,y,t_j}})$ — коэффициент линейной зависимости скорости движения трамвая от вероятности наезда ТС на трамвайные рельсы.

Определим коэффициент Z_{x,y,t_j} на основании исследования Яндекса изменения скорости движения транспорта от уровня загруженности дороги[1]:

$$Z_{x,y,t_j} = \begin{cases} traffic_{jam_{t_j}} \in [1, 3]: [0{,}7, 1] \\ traffic_{jam_{t_j}} \in [4, 6]: [0{,}3, 0{,}7] \\ traffic_{jam_{t_j}} \in [7, 10]: [0, 0{,}3] \end{cases}.$$

Аналогично, в контексте представляемой системы можно сказать, что $k\_repair_{x,y,t_j} = fun(repair_{x,y,t_j})$ — является коэффициентом зависимости наличия ремонтных работ на дороге от времени t_j и местоположения участка дороги (x, y):

$$v_{i,x,y,t_j} = k\_repair_{x,y,t_j} \times fun(a_i) \times Z_{x,y,t_j}.$$

Так как $a_{i,t_j} = \{(x, y)_{i,t_j}, v_{i,x,y,t_j}, stop_{i,t_{j+1}}\}$, тогда:

$$v_{i,x,y,t_j} = Z_{x,y,t_j} \times k\_repair_{x,y,t_j} \times v_{x,y} \times k_{stop_{t_{j+1}}},$$

где $v_{x,y}$ — средняя скорость агента a_i на участке (x, y) в условиях отсутствия влияния human factors; $k_{stop_{t_{j+1}}}$ — коэффициент скорости трамвая при наличии остановки трамвая в момент времени t_{j+1}:

— $stop_{t_{j+1}} = False$: $k_{stop_{t_{j+1}}} = 1$,

— $stop_{t_{j+1}} = True$: $k_{stop_{t_{j+1}}} \in [0, 1)$.

Описание симулятора

Для построения моделей транспортных потоков на микроуровне применяются средства: PTV Vissim, Quadstone Paramics и TTS Aimsun [11].

В качестве программной платформы для реализации рассматриваемой модели зависимости влияния внешних human factors на скорость движения трамвая выбран инструмент имитационного моделирования PTV Vissim Demo, позволяющий строить модели движения транспортных потоков. Система PTV Vissim позволяет имитировать и визуализировать взаимодействие любых видов транспорта в рамках единой модели с использованием правил дорожного движения, что говорит о приближении проведенного эксперимента к реальным условиям.

Система PTV Vissim обеспечивает реалистичное и всестороннее представление транспортных потоков, а также всех аспектов управления этими потоками, которые динамически адаптируются к дорожно-транспортной обстановке, определяя или воспроизводя индивидуальные параметры моделируемого участка.

Количество данных, используемых в симуляторе: трамваев — 5; ТС — 80; пешеходов — 80; участков ремонтных работ — 2.

На рис. 2 представлен план карты дороги с участками *1–4*. Участок *1* — дорога только для ТС, участки *2*, *3* — трамвайные рельсы, на которые не могут заехать ТС. Участок дороги *4* является круговым по две полосы в каждую сторону: по обеим полосам могут двигаться ТС, а по внутренним полосам — трамваи.

Симуляция начинается с равномерного распределения агентов в помещении.

На рис. 3 показана работа симулятора: длинными овалами синего, зеленого и черного цвета обозначены трамваи, короткими овалами черного, белого, серого и синего — ТС, желтыми линиями обозначены стоп-линии на второстепенных участках, черными линиями обозначены места появления агентов симуляции.

На первом этапе выполнено 10 запусков системы с 5 агентами трамваями для получения значений средних скоростей на участках дороги.

На втором этапе проведено 10 запусков системы с 5 агентами трамваями, 20 агентами ТС, 80 агентами пешеходами для проверки влияния на скорость трамваев human factors без пробок и ремонтных работ, но с моделью поведения агентов ТС с заездом на трамвайные рельсы.

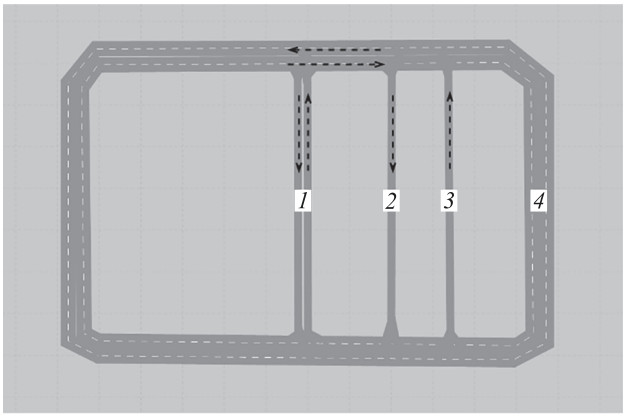


Рис. 2. План карты дороги. Движение правостороннее (направления движения вправо и влево обозначены пунктирными стрелками)

Fig. 2. Road map plan. Right-hand traffic (arrows stand for the direction of traffic)

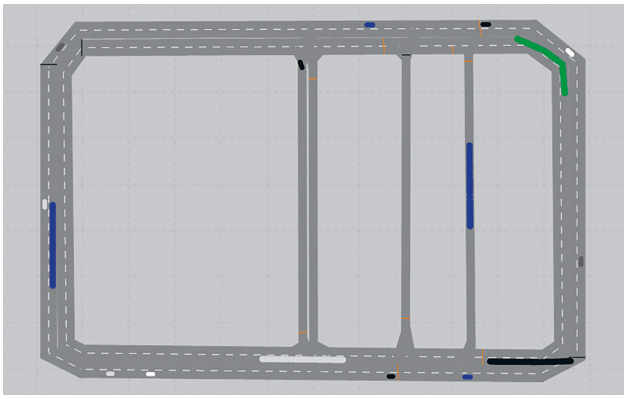


Рис. 3. Схема движения трамваев и транспортных средств по дороге

Fig. 3. The scheme of the movement of trams and cars on the road

[1] Исследование Яндекса – Пробки в Москве 2013–2017 [Электронный ресурс]. URL: https://yandex.ru/company/researches/2017/moscow_traffic_2017 (дата обращения: 30.11.2021).

На третьем этапе проведено 10 запусков системы с 5 агентами трамваями, 40 агентами ТС, 80 агентами пешеходами, а также с ремонтными работами на одном участке дороги для проверки влияния на скорость human factors с небольшими пробками (не более 5 баллов), а также влияния ремонтных работ.

На четвертом этапе выполнено 10 запусков системы с 5 агентами трамваями, 80 агентами ТС, 80 агентами пешеходами, а также с ремонтными работами на двух участках дороги для проверки влияния на скорость human factors на загруженной дороге (пробки более 5 баллов), а также влияния ремонтных работ на нескольких участках дороги.

На пятом этапе проведена агрегация полученных результатов, а также их оценка качества в результате эксперимента и формирование выводов.

Вычислительный эксперимент

В ходе проведения эксперимента произведено по 10 запусков модели для каждого этапа. На третьем этапе на участок *4* был добавлен один участок ремонтных работ (направление вправо (рис. 2)). На четвертом этапе — добавлены два участка ремонтных работ (обе стороны направления (рис. 2)).

В таблице представлены средние скорости для каждого участка дороги в зависимости от этапа эксперимента.

Из результатов эксперимента видно, что при добавлении небольшого количества машин (второй этап) скорость на участке *4* изменилась в среднем на 15 %. Скорость на выделенных для трамвая полосах (участки *2*, *3*) не изменилась — эти показатели можно выбрать как эталонные.

При появлении ремонтных работ (участок *4*, направление вправо) средняя скорость движения трамвая значительно снизилась: примерно на 59 %. Это связано с общим снижением скорости потока, а также с заездом ТС на рельсы трамвая, при этом скорость на соседней полосе (участок *4*, направление влево) снизилась на 20 %.

При добавлении ремонтных работ (участок *4*, оба направления) скорость на участке *4* снизилась в среднем на 59 % (направление вправо) и на 69 % (направление влево). При этом скорость на выделенных полосах изменилась 21 % и 16 % для участков *2* и *3* соответственно.

Таблица. Средняя скорость движения трамвая, км/ч

Table. Average speed of the tram, km/h

Номер участка дороги (рис. 2)	Результаты моделирования			
	Первый этап	Второй этап	Третий этап	Четвертый этап
2	25,32	25,32	23,41	20,16
3	22,99	20,16	23,32	19,53
4 (направление вправо)	24,32	17,26	11,20	10,04
4 (направление влево)	24,90	24,46	19,80	7,88

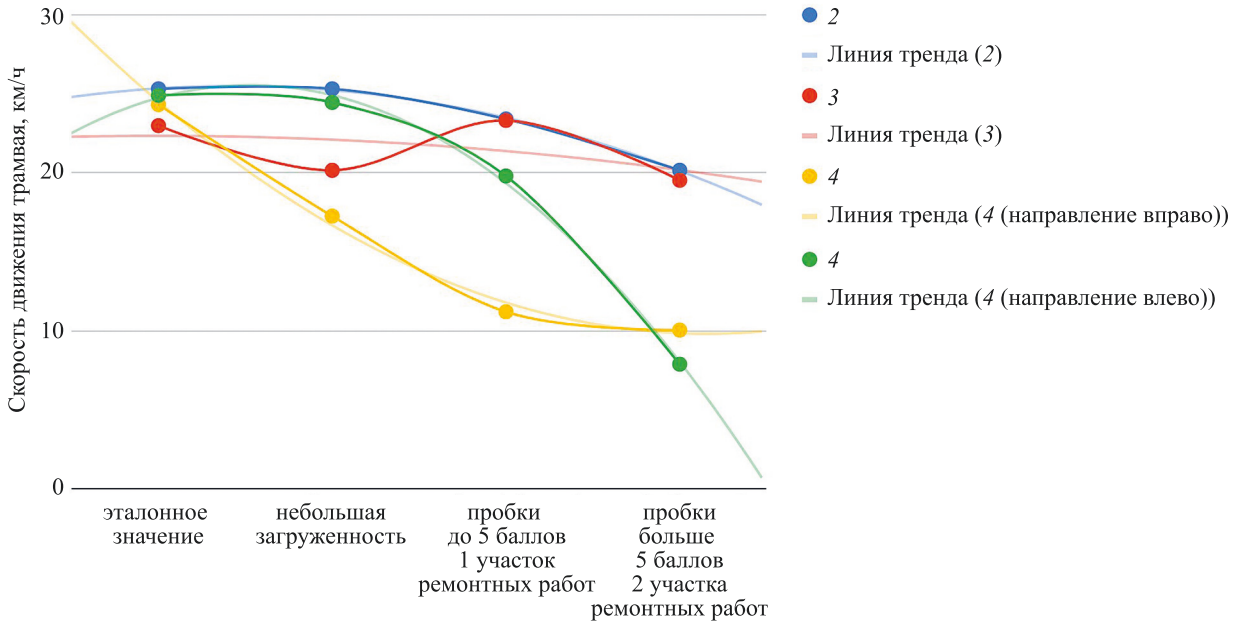


Рис. 4. Графики зависимости скорости на каждом участке дороги в зависимости от факторов, влияющих на скорость движения трамвая

Fig. 4. Graphs of the dependence of the speed on each section of the road, depending on the factors affecting the speed of the tram

На рис. 4 представлены графики изменения скорости на каждом участке дороги в зависимости от факторов, влияющих на скорость движения трамвая, а также линии тренда, используемые для выявления тенденций изменения скорости трамвая. Как видно из рис. 4 зависимость скорости от human factors является полиномиальной.

После анализа полученных результатов можно сделать вывод, что образование пробок на дорогах в значительной степени влияет на скорость движения трамвая. Это связано с большим количеством заездов ТС на трамвайные рельсы. В результате работы видно, что на втором этапе скорость движения на участке дороги *4* снизилась при поворотах налево или обгона. Но скорость на втором этапе снижена не так сильно, как на третьем и четвертом этапах.

Заключение

Предложен и протестирован анализ зависимости скорости движения пассажирского трамвая от человеческих факторов (human factors). Анализ учитывает воздействие внешних human factors, таких как выезд на трамвайные рельсы, проведение ремонтных работ на дороге.

На основе методологии исследования на основе мультиагентных систем выполнен анализ зависимости скорости движения пассажирского трамвая от human factors. Определены множества входных параметров: агентов модели и их атрибутов, а также вида предполагаемого отклика скорости. Проведен итеративный процесс разработки структуры системы посредствам визуального моделирования и проверка гипотезы зависимости скорости движения трамвая от human factors на тестовых данных в среде, приближенной к целевой.

Выполнен анализ полученных результатов и сравнение с данными, полученными при тестировании модели без влияния human factors. В результате тестирования видно, что при добавлении транспортных средств и одного участка ремонтных работ скорость трамвая снижается на 16 %, а при увеличении числа ремонтных работ и значительных пробках – 69 %.

Также отмечено, что скорость трамвая почти не изменяется при движении по специальным выделенным полосам.

Исходя из перечисленных этапов, можно говорить о том, что исследование проведено качественно, выдвинута гипотеза, формализован объект исследования, гипотеза подтверждена экспериментом. Работа имеет практическую значимость в области проектирования дорожного движения.

Представленное исследование может быть использовано при проведении анализа проектируемых транспортных узлов для беспрепятственного движения беспилотного городского рельсового транспорта.

Литература

1. Krokos K.J., Baker D.P. Preface to the special section on classifying and understanding human error // Human Factors. 2007. V. 49. N 2. P. 175–177. https://doi.org/10.1518/001872007X312414
2. Cacciabue P.C. Human error risk management methodology for safety audit of a large railway systems // Rail human factors: supporting the integrated railway / ed. by J.R. Wilson, B. Norris, T. Clarke, A. Mills. London: Ashgate Publishing, 2005. P. 353–365. https://doi.org/10.4324/9781315089201-32
3. Branton R. Investigations into the skills of train driving // Ergonomics. 1979. V. 22. N 2. P. 155–164. https://doi.org/10.1080/00140137908924600
4. Buck L. Errors in the perception of railway signals // Ergonomics. 1963. V. 6. N 2. P. 181–192. https://doi.org/10.1080/00140136308930688
5. Rosmuller N., Beroggi G.E.G. Group decision making in infrastructure safety planning // Safety Science. 2004. V. 42. N 4. P. 325–349. https://doi.org/10.1016/S0925-7535(03)00046-8
6. Slamen A., Schock A., Ryan B., Wilson J.R. Human factors analysis of the work of the engineering supervisor: restricted report of Network Rail. London, 2004.
7. Hale A.R., Heijer T., Koornneef F. Management of safety rules: the case of railways // Proc. of the Third International Symposium on Safety and Hygiene. Porto, Portugal, March 2003.
8. Slamen A., Coleman N. The application of ergonomics to standards development for VDU based signalling control systems // Rail human factors: supporting the integrated railway / ed. by J.R. Wilson, B. Norris, T. Clarke, A. Mills. London: Ashgate Publishing, 2005. P. 239–250. https://doi.org/10.4324/9781315089201-22
9. Bourne A., Carey M. Integrating human factors into the development of railway systems // People in Control: The Second International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres. Manchester, 19–21 June 2001. P. 25–30. https://doi.org/10.1049/cp:20010427
10. Морозов И.И., Гасников А.В., Тарасов В.Н., Холодов Я.А., Холодов А.С. Численное исследование транспортных потоков на основе гидродинамической модели // Компьютерные исследования и моделирование. 2011. Т. 3. № 4. С. 389–412.

References

1. Krokos K.J., Baker D.P. Preface to the special section on classifying and understanding human error. *Human Factors*, 2007, vol. 49, no. 2, pp. 175–177. https://doi.org/10.1518/001872007X312414
2. Cacciabue P.C. Human error risk management methodology for safety audit of a large railway systems. *Rail human factors: supporting the integrated railway* / Ed. by J.R. Wilson, B. Norris, T. Clarke, A. Mills. London: Ashgate Publishing, 2005, pp. 353–365. https://doi.org/10.4324/9781315089201-32
3. Branton R. Investigations into the skills of train driving. *Ergonomics*, 1979, vol. 22, no. 2, pp. 155–164. https://doi.org/10.1080/00140137908924600
4. Buck L. Errors in the perception of railway signals. *Ergonomics*, 1963, vol. 6, no. 2, pp. 181–192. https://doi.org/10.1080/00140136308930688
5. Rosmuller N., Beroggi G.E.G. Group decision making in infrastructure safety planning. *Safety Science*, 2004, vol. 42, no. 4, pp. 325–349. https://doi.org/10.1016/S0925-7535(03)00046-8
6. Slamen A., Schock A., Ryan B., Wilson J.R. *Human factors analysis of the work of the engineering supervisor:* Restricted report of Network Rail. London, 2004.
7. Hale A.R., Heijer T., Koornneef F. Management of safety rules: the case of railways. *Proc. of the Third International Symposium on Safety and Hygiene*, Porto, Portugal, March 2003.
8. Slamen A., Coleman N. The application of ergonomics to standards development for VDU based signalling control systems. *Rail human factors: supporting the integrated railway* / Ed. by J.R. Wilson, B. Norris, T. Clarke, A. Mills. London: Ashgate Publishing, 2005, pp. 239–250. https://doi.org/10.4324/9781315089201-22
9. Bourne A., Carey M. Integrating human factors into the development of railway systems. *People in Control: The Second International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, Manchester, 19–21 June 2001, pp. 25–30. https://doi.org/10.1049/cp:20010427
10. Morozov I.I., Gasnikov A.V., Tarasov V.N., Kholodov Y.A., Kholodov A.S. Numerical study of traffic flows by the hydrodynamic models. *Computer Research and Modeling*, 2011, vol. 3, no. 4, pp. 389–412. (in Russian)

11. Захаров Ю.И., Карнаух Е.С. Основные современные инструменты имитационного моделирования транспортных потоков // Вісник Придніпровської державної академії будівництва та архітектури. 2014. № 1(190). С. 46–51 [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/osnovnye-sovremennye-instrumenty-imitatsionnogo-modelirovaniya-transportnyh-potokov (дата обращения: 15.10.2021).

11. Zakharov U., Karnaukh E. The main modern traffic simulation tools. *Bulletin of Prydniprovs'ka State Academy of Civil Engineering and Architecture*, 2014, no. 1(190), pp. 46–51. Available at: https://cyberleninka.ru/article/n/osnovnye-sovremennye-instrumenty-imitatsionnogo-modelirovaniya-transportnyh-potokov (accessed: 15.10.2021). (in Russian)

Авторы

Ляховенко Юлия Александровна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57216353188, https://orcid.org/0000-0001-7396-2831, lyakhovenko.kam@gmail.com

Попов Илья Юрьевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57202195632, https://orcid.org/0000-0002-6407-7934, ilyapopov27@gmail.com

Authors

Julia A. Lyakhovenko — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57216353188, https://orcid.org/0000-0001-7396-2831, lyakhovenko.kam@gmail.com

Ilya Yu. Popov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57202195632, https://orcid.org/0000-0002-6407-7934, ilyapopov27@gmail.com

Статья поступила в редакцию 02.11.2021
Одобрена после рецензирования 15.12.2021
Принята к печати 20.01.2022

Received 02.11.2021
Approved after reviewing 15.12.2021
Accepted 20.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-67-73
УДК 004.772

Алгоритм обнаружения RFID-дубликатов

Наталия Викторовна Волошина[1], Александр Андреевич Лавринович[2]✉

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] nataliv@yandex.ru, https://orcid.org/0000-0001-9435-9580
[2] Lavrinovich600@gmail.com✉, https://orcid.org/0000-0002-5058-3473

Аннотация

Предмет исследования. При расширении применения RFID (Radio Frequency Identification)-технологии в качестве способа маркировки импортируемых товаров все более актуальной становится проблема использования злоумышленниками дубликатов RFID-меток. На дубликат может быть записана информация о товаре, которая отличается от его фактических характеристик. В работе предложен алгоритм обнаружения RFID-дубликатов как метод достижения целостности сведений, которые поступают в информационные системы международной транспортировки товаров. Актуальность создания алгоритма определена потребностью снижения риска создания и использования импортерами RFID-дубликатов при трансграничном перемещении маркируемых товаров. Существующие алгоритмы обнаружения дубликатов не подходят для применения в системе RFID-маркировки импортируемых в Россию товаров. Предложенный алгоритм ограничивает возможности злоумышленника считывать с оригинальной RFID-метки сведения, необходимые для создания RFID-дубликата. **Метод.** Алгоритм основан на разделении Electronic Product Code (EPC)-области памяти RFID-метки на части и применении команды самоуничтожения метки (kill) для предотвращения несанкционированных считываний. Рассмотрены сценарии реализации алгоритма и определены риски его использования. Алгоритм представлен в виде графической модели на основе нотации Business Process Model and Notation (BPMN). **Основные результаты.** Оценка эффективности предлагаемого алгоритма проведена с использованием формулы гипергеометрической вероятности. В качестве исходных данных приняты результаты проведения выборочной проверки RFID-меток таможенными органами. Показано, что в сравнении с существующим подходом реализация алгоритма в программно-аппаратном комплексе создает условия повышения вероятности обнаружения RFID-дубликатов при условии проведения контроля только в отношении высокорисковых декларантов. **Практическая значимость.** Применение алгоритма снижает риск поступления искаженной или недостоверной информации в информационные системы международной транспортировки товаров и повышает обоснованность принимаемых юридических и экономических решений в информационных системах таможенных органов.

Ключевые слова

RFID-технология, RFID-метка, RFID-дубликат, RFID-маркировка, информационная система, EPC-память, команда kill, целостность информации, маркировка товаров

Ссылка для цитирования: Волошина Н.В., Лавринович А.А. Алгоритм обнаружения RFID-дубликатов // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 67–73. doi: 10.17586/2226-1494-2022-22-1-67-73

An algorithm for detecting RFID duplicates

Natalia V. Voloshina[1], Aleksandr A. Lavrinovich[2]✉

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] nataliv@yandex.ru, https://orcid.org/0000-0001-9435-9580
[2] Lavrinovich600@gmail.com✉, https://orcid.org/0000-0002-5058-3473

Abstract

The problem of using duplicate RFID tags by attackers is becoming more and more actual with the expansion of RFID technology for marking imported goods. The duplicate may contain information about goods, which differs from their

© Волошина Н.В., Лавринович А.А., 2022

actual characteristics. This paper proposes an algorithm for detecting RFID duplicates as a method for achieving the integrity of information that enters the information systems of international goods transportation. The relevance of creating an algorithm deals with the need to reduce the risk of creating and using RFID duplicates by importers during the cross-border movement of marked goods. Existing duplicate detection algorithms are unsuitable for use in the RFID-marking system of goods imported into Russia. The algorithm hinders an attacker from reading data from the original RFID tag, which is necessary to create an RFID duplicate. The proposed algorithm is based on dividing the EPC memory area of an RFID tag into parts and using the tag self-destruction command (kill) to prevent unauthorized readings. The authors considered the scenarios for implementing the algorithm and identified the risks of using the algorithm. The algorithm is presented as a graphical model based on BPMN notation. The efficiency of the proposed algorithm was evaluated using the hypergeometric probability formula. The results of a selective check of RFID tags by the customs authorities were taken as the initial data. It is shown that, in comparison with the existing approach, the implementation of the algorithm in software and hardware complex increases the probability of detecting RFID duplicates, provided that control is carried out only in relation to high-risk declarants. The use of the algorithm reduces the risk of receiving distorted or inaccurate data in the information systems dealing with international goods transportation and increases the validity of legal and economic decisions taken in the information systems of customs authorities.

Keywords
RFID technology, RFID mark, RFID duplicate, RFID marking, information system, EPC memory, kill command, information integrity, marking of goods

For citation: Voloshina N.V., Lavrinovich A.A. An algorithm for detecting RFID duplicates. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 67–73 (in Russian). doi: 10.17586/2226-1494-2022-22-1-67-73

Введение

Для учета товаров в рамках внешней торговли применяется RFID (Radio Frequency Identification)-технология. RFID-метки — устройство, состоящее из чипа и антенны, закрепленных в этикетке — контрольном (идентификационном) знаке. RFID-метки выступают в качестве источника сведений о товарах. С помощью считывания RFID-меток формируются данные для информационных систем (ИС) международной транспортировки товаров, на основе которых пользователи системы могут принимать решения о товаре.

Работа посвящена формированию целостности информации, которая поступает в ИС с помощью RFID-технологии. При использовании изучаемых ИС актуально требование к отсутствию неправомерных искажений и добавлений к информации, что может быть обеспечено, если злоумышленник применяет RFID-дубликат.

В работах [1, 2] проанализированы два способа борьбы с RFID-дубликатами: профилактика и обнаружение. Преобладает мнение [1–7], что способ обнаружения — более эффективный, но менее затратный и универсальный. В работах [1–4, 8] рассмотрены алгоритмы метода обнаружения. Недостатки данных алгоритмов связаны с возможностью их применения лишь при наличии определенных условий, которые отсутствуют при формировании сведений в ИС международной транспортировки товаров.

С учетом невозможности применения существующих алгоритмов обнаружения RFID-дубликатов для решения проблемы формирования искаженных сведений в изучаемых ИС необходимо создание нового алгоритма для применения в данной сфере.

Новизна предлагаемого алгоритма обнаружения RFID-дубликатов заключается в том, что он основан на разделении Electronic Product Code (EPC)-области памяти RFID-меток на три части и применении команды kill (команда уничтожения RFID-метки, после этого метку уже нельзя использовать). Алгоритм применим в сфере использования государственных ИС международной транспортировки товаров. Он позволяет повысить вероятность обнаружения дубликатов на 14,1 % (без применения алгоритма вероятность обнаружения равна 24,8 %, а с наличием алгоритма — 38,9 %). Повышение вероятности обнаружения RFID-дубликатов связано с обеспечением информационной безопасности для снижения рисков поступления недостоверной, искаженной информации в ИС. Авторами рассмотрена государственная ИС международной транспортировки товаров, которая используется таможенными органами, как органами государственной власти, назначение которой — реализация полномочий таможенными органами в пределах своей компетенции.

Постановка задачи повышения безопасности информационных систем международной транспортировки товаров

Рассмотрим RFID-технологии [9] и обеспечение целостности информации при формировании сведений в ИС международной транспортировки товаров с ее применением. Сведения, поступающие с RFID-меток в изучаемые ИС, должны быть достоверными, иначе таможенными органами могут быть приняты некорректные юридические (например, решение о выпуске товаров таможенными органами) и экономические (например, взимание таможенных платежей) решения в отношении товаров. Пользователи ИС должны работать с информацией, которая не содержит несанкционированных модификаций.

Способом нарушения целостности в данном случае является использование RFID-дубликатов злоумышленниками. Реализация данного способа показана на рис. 1.

Чтобы не допустить формирования сведений в ИС с RFID-дубликатов, необходимы способы проверки оригинальности RFID-меток. Характеристики существующих алгоритмов обнаружения RFID-дубликатов представлены в табл. 1.

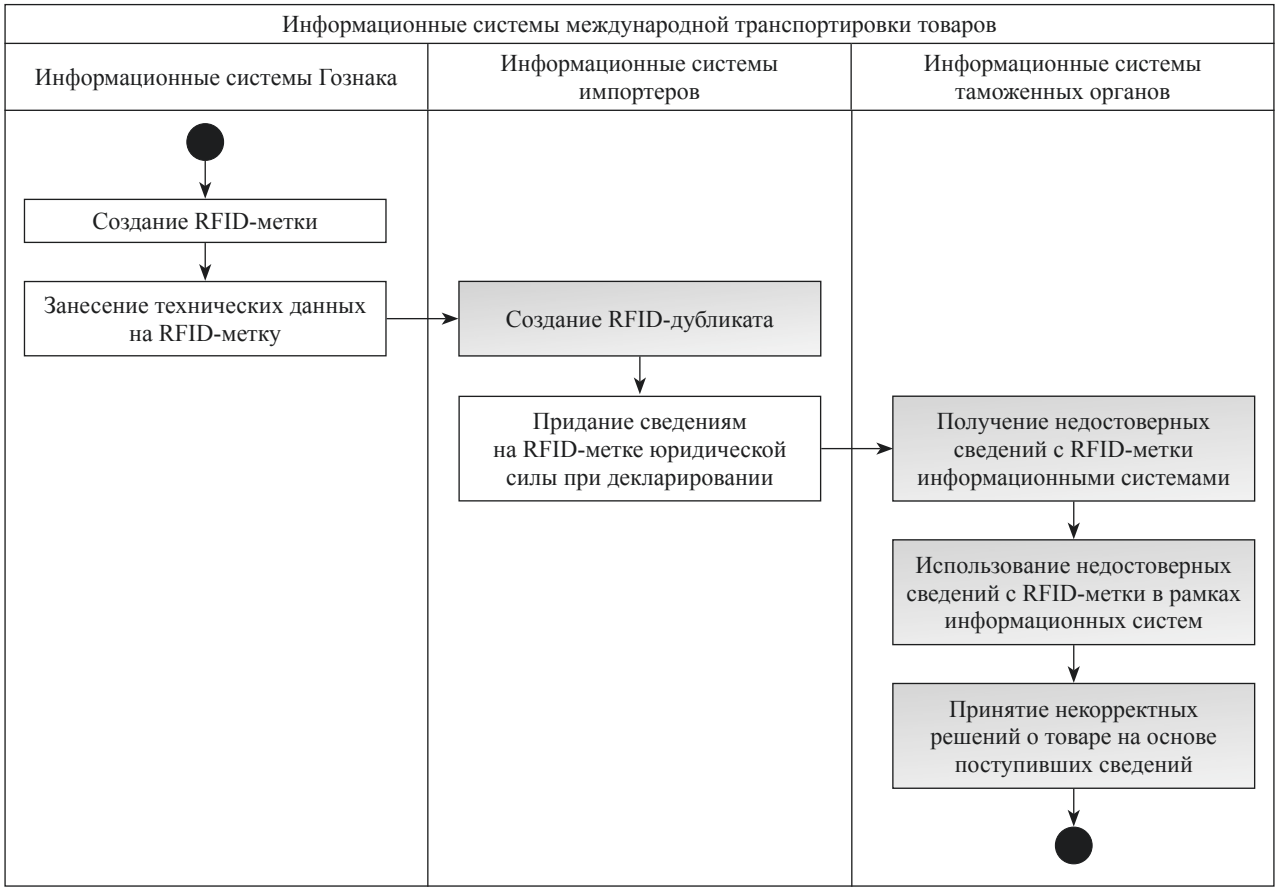


Рис. 1. Модель использования RFID-дубликатов для формирования искаженных сведений в информационных системах международной транспортировки товаров

Fig. 1. The model of using RFID duplicates for the formation of distorted information in information systems for the international transportation of goods

Рассмотренные алгоритмы (табл. 1) действуют по схожему принципу. Для их реализации требуются RFID-метки с перезаписываемой памятью, так как на метки заносится некоторая информация, которая впоследствии сопоставляется с другими сведениями. Применение алгоритмов обусловлено требованием к целостности информации, которую пользователь получает с RFID-метки.

Таблица 1. Критерии алгоритмов обнаружения RFID-дубликатов

Table 1. Characteristics of RFID duplicate detection algorithms

Критерий	Алгоритмы обнаружения RFID-дубликатов, рассмотренные в работах		
	[1]	[2]	[3]
Основной принцип метода	Согласованность двойных хеш-коллизий и модифицированный вектор эскиза count-min	При каждом считывании на метку записываются случайные значения, из них создается «хвост»	Перезаписываемая память меток: при каждом считывании случайное число изменяется
Определение дубликата	Частота чтения дубликата ниже, чем оригинальной метки	Хвосты дубликатов отличаются от хвостов оригинальных меток	Сравнение сведений о случайном числе на метке и случайном числе в базе данных
RFID-считыватели	Минимум три авторизованных считывателя	Минимум четыре авторизованных считывателя	—
Сфера применения	Системы с временными границами для прибытия маркируемых товаров	Цепочка поставок; система с использованием подконтрольных RFID-считывателей	Коммерческие системы с многократным считыванием меток
Команда kill	Не используется	Не используется	Не используется
Базы данных для проверки подлинности метки	Не используются	Используются локальные базы данных	Используются централизованные базы данных

Но недостатки алгоритмов связаны с узкой сферой применения каждого из них. Во всех случаях алгоритмы разработаны для систем, в которых объекты с RFID-метками проходят путь согласно плану, и в рамках этого пути периодически происходит считывание RFID-меток, что позволяет и записать новую информацию, и проверить предыдущие записанные на метку сведения. В случае с изучаемыми ИС международной транспортировки товаров такой возможности нет, так как метка производится уполномоченной организацией, она поставляется импортерам, которые заносят на метку сведения и далее используют метки при таможенном оформлении. В связи с этим ставится задача разработки нового алгоритма по обнаружению RFID-дубликатов, который мог бы быть применен в изучаемых ИС.

Предложенный алгоритм обнаружения RFID-дубликатов

Сфера применения предлагаемого алгоритма — государственные ИС, входящие в систему государственной маркировки импортируемых товаров [10]. Рассмотрим принципы предлагаемого алгоритма.

1. EPC-область памяти делится на три части. В первую часть при создании метки на Гознаке должно быть записано число I. Вторая часть резервируется для результата считывания импортером. Третья часть резервируется для числа T, которое записывается при считывании таможней. После получения сведений считыватель таможни проверяет EPC-область памяти, и делается вывод о статусе метки.
2. Последующие считывания доступны только считывателям таможенного органа путем обновления числа T в третьей части EPC-области памяти. Если это действительно число T, метка направляет информацию этому считывателю, а если нет, то срабатывает команда kill и метка уничтожается. В данном случае команда kill играет защитную функцию [11], она не выступает в качестве угрозы информационной безопасности [12].
3. Данный алгоритм предполагает, что метка содержит два числа, получив которые злоумышленник может изготовить дубликат — Transponder identification number (TID)-номер и число I. Для получения этих чисел требуется два считывания. Но у импортера есть возможность произвести только одно считывание, если им будет произведено второе считывание, то RFID-меткой будет установлено, что оно производится считывателем импортера, а не считывателем таможенного органа, и тогда метка применит команду kill, в результате чего она становится непригодной для использования, и дальнейшие действия по считыванию метки станут невозможными.

Алгоритм применения метода в виде блок-схемы в обобщенном виде представлен на рис. 2.

В алгоритме можно выделить следующие два сценария.

1. «Оригинальная метка». Если импортер планирует легальным образом использовать RFID-метку, то он записывает на нее Global Trade Item Number (GTIN)-номер, который заносится на вторую часть EPC-области памяти RFID-метки. Далее метка представляется к считыванию таможенному органу. Авторизованный RFID-считыватель таможенного органа записывает число T в третью часть EPC-области памяти. Метка проверяет структуру числа T, и если это действительно число T, отвечающее требованиям к этому числу, то метка передает информацию из EPC-области памяти считывателю таможенного органа. Данный считыватель проверяет структуру информации, полученной из EPC-области памяти с помощью сведений, поступивших из централизованной базы данных Гознака. В данном сценарии делается вывод о том, что метка оригинальна, и поступившая с нее информация – достоверна. Следовательно, обеспечивается целостность информации для ИС международной транспортировки товаров, что позволяет принимать корректные решения в отношении товаров.
2. «Дубликат». Если импортер планирует ввести в процесс использования RFID-технологии дубликат, то ему необходимо сначала создать его. При условии применения предлагаемого метода, одним из способов создания дубликата является считывание информации с оригинальной метки [13] (TID-номера и числа I), создание дубликата с этими сведениями и GTIN-номером, который указывает на недостоверные сведения о товаре. Для получения сведений, необходимых для создания дубликата, требуется произвести, во-первых, считывание TID-номера, так как эта информация заносится на метку производителем метки, в данном случае это импортер, а не Гознак. Тогда в EPC-области памяти во вторую часть делается запись TID, а не GTIN, как в первом сценарии. Во-вторых, требуется считывание числа I, которое также заносится производителем метки. Считывание числа I повлечет за собой занесение в третью часть EPC-области памяти числа I, но для третьей части EPC-области памяти установлено требование, что в ней должно находиться число T. Соответственно, меткой при обнаружении записи, отличной от числа T, в третьей части EPC-области памяти, применяется команда kill, что приводит к самоуничтожению метки. Это делает невозможным ее дальнейшее использование, и импортер вынужден будет заново заказывать метки у Гознака. Если импортер решит использовать метку с легальным TID-номером, но с пустой EPC-памятью, то это будет обнаружено RFID-считывателем таможенного органа, так как EPC-область памяти не будет содержать структуру «I-GTIN-T». В таком случае будет сделан вывод о статусе метки – «дубликат».

Риски при использовании алгоритма связаны с тем, что злоумышленнику может быть заранее известно число TID или число I. Контрмерами в таком случае должны выступать меры по обеспечению защиты ИС Гознака, каналов связи, по которым передаются сведения из базы данных Гознака в ИС таможенных органов, обеспечению защиты помещений Гознака.

Предложенный в работе алгоритм отличается от существующих [1, 3] разделением EPC-области памяти на

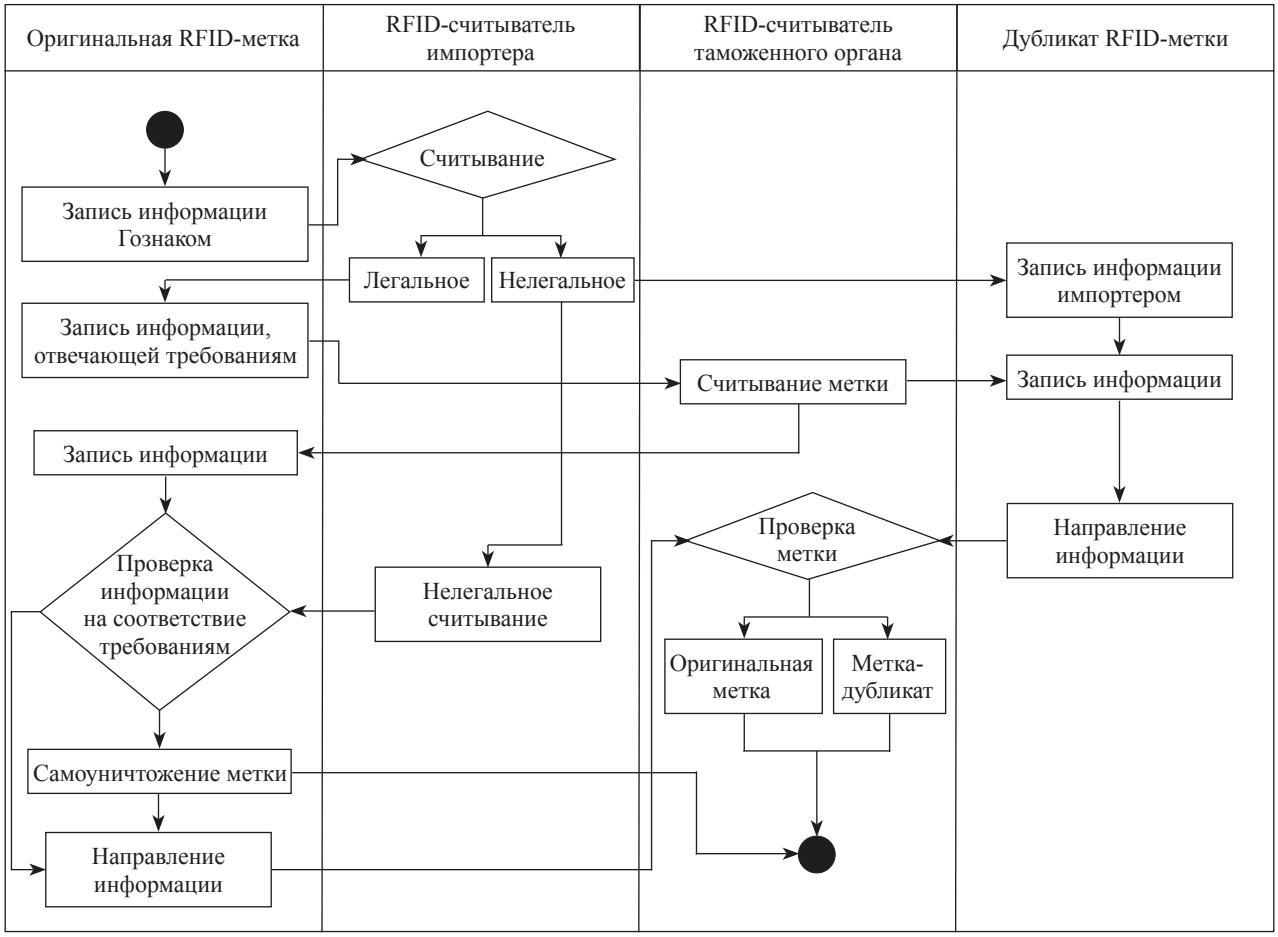


Рис. 2. Предлагаемый алгоритм обнаружения RFID-дубликатов

Fig. 2. The proposed algorithm for detecting RFID duplicates

части, применением команды kill при несанкционированных считываниях, применением централизованных баз данных для проверки подлинности метки, меньшим числом требуемых авторизованных считывателей и меньшим числом операций в протоколе взаимодействия метки и считывателя.

Ситуацию обнаружения дубликатов RFID-меток можно рассмотреть, как обнаружение объекта, обладающего заданными свойствами, среди некоторого количества объектов (выборки). Вероятность обнаружения RFID-дубликата в выборке RFID-меток, нанесенных на товары, можно определить по формуле гипергеометрической вероятности:

$$P = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}, \qquad (1)$$

где P — вероятность обнаружения RFID-дубликата в выборке RFID-меток, нанесенных на товары; k — число обнаруженных RFID-дубликатов среди совокупности RFID-меток; n — число проверенных RFID-меток; N — совокупность RFID-меток; K — число RFID-дубликатов в совокупности RFID-меток; C_K^k — число всех способов обнаружить k RFID-дубликатов из K возможных; C_{N-K}^{n-k} — число всех способов проверить $n - k$ RFID-меток из $N - K$ возможных; C_N^n — число возможных исходов результатов проверки RFID-меток.

Текущая ситуация выявления RFID-дубликатов связана с проведением досмотров должностными лицами таможенных органов (P_1). Повышение безопасности ИС должно быть обеспечено при использовании алгоритма обнаружения, являющегося базой программно-аппаратного комплекса (P_2).

Оценка предлагаемого алгоритма обнаружения RFID-дубликатов для двух ситуаций приведена в табл. 2. Исходные данные для оценки, полученные по результатам таможенного контроля[1,2]: 100 RFID-меток, нанесенных на импортируемые товары; 10 RFID-дубликатов, которые находятся среди RFID-меток.

При проведении оценки учитывалось, что таможенные досмотры проводятся в отношении 3 % товарных партий, поэтому вероятность обнаружить при проверке трех меток один RFID-дубликат $P(1)_1$ составляет 24,8 %. При условии реализации алгоритма обнаруже-

[1] Итоговые доклады о результатах деятельности ФТС России (официальный сайт ФТС России) [Электронный ресурс]. Режим доступа: https://customs.gov.ru/activity/results/itogovye-doklady-o-rezul-tatax-deyatel-nosti, свободный. Яз. рус. (дата обращения: 28.07.2021).

[2] Таможенная статистика внешней торговли РФ [Электронный ресурс]. Режим доступа: https://customsonline.ru/search_ts.html, свободный. Яз. рус. (дата обращения: 28.07.2021).

Таблица 2. Оценка предлагаемого алгоритма обнаружения RFID-дубликатов

Table. 2. Evaluation of the proposed RFID duplicate detection algorithm

Параметр оценки	Ситуация	
	Текущая ситуация обнаружения дубликатов RFID-меток	Ситуация с учетом внедрения алгоритма обнаружения дубликатов RFID-меток
Способ обнаружения	Проведение таможенного досмотра	Проведение автоматической проверки RFID-меток, нанесенных на товары, декларируемые импортерами высокой категории риска по системе категорирования
Охват проверки	Три RFID-метки	Семь RFID-меток
Требуемый результат	Обнаружение одного RFID-дубликата	Обнаружение одного RFID-дубликата
Вероятность получения требуемого результата $P(1)$ по формуле (1)	$P(1)_1 = \frac{C_{10}^1 C_{90}^2}{C_{100}^3} = \frac{10 \times 4005}{161700} = 0{,}248$	$P(1)_2 = \frac{C_{10}^1 C_{90}^6}{C_{100}^7} = \frac{10 \times 622614630}{16007560800} = 0{,}389$

ния RFID-дубликатов в виде программно-аппаратного комплекса в таможенных органах и его применении в отношении декларантов высокой категории риска по системе категорирования (7 % от общего числа декларантов), вероятность обнаружить при проверке семи меток один RFID-дубликат $P(1)_2$ — 38,9 %. Таким образом, реализация предлагаемого алгоритма позволит создать условия, при которых вероятность обнаружения дубликата будет выше на 14,1 %, чем при текущей практике их обнаружения, и при условии, что охват контроля ограничен только участниками внешнеэкономической деятельности (ВЭД) высокой категории риска. При этом не исключается, что по системе управления рисками таможенных органов контроль также может проводиться и в отношении участников ВЭД иных категорий риска, что позволит сделать вероятность получения требуемого результата еще выше.

Рекомендации по реализации алгоритма: применение RFID-меток с типом памяти RW, создание программно-аппаратного комплекса в таможенных органах по считыванию RFID-меток и обработке сведений, позволяющий сделать вывод о статусе метки. С учетом того, что применение RFID-меток с типом памяти RW приведет к росту их стоимости, из экономических соображений целесообразно реализовывать данный проект в отношении категорий товаров, по которым установлен высокий уровень криминогенности при их импорте (уклонение от уплаты таможенных платежей, недостоверное декларирование, занижение таможенной стоимости). Стоит внедрять программно-аппаратный комплекс в рамках проектов по построению интеллектуальных пунктов пропуска, чтобы получение информации с RFID-меток было не только мерой контроля, но и стандартной операцией по перемещению товаров через границу вместо ручного досмотра. За счет этого может увеличиться скорость прохождения таможенного оформления, что приведет к снижению таможенных затрат участников ВЭД.

Заключение

Решение задачи обеспечения целостности при использовании RFID-технологии в информационных системах международной транспортировки товаров достигается путем снижения рисков создания и использования импортерами RFID-дубликатов. Снижение рисков в настоящем исследовании связывается с увеличением вероятности обнаружения RFID-дубликата в выборке RFID-меток, нанесенных на товары, такую вероятность можно определить по формуле гипергеометрической вероятности.

Предложен алгоритм обнаружения RFID-дубликатов, применимый для информационных систем международной транспортировки товаров, входящих в систему государственной маркировки импортируемых товаров. Алгоритм снижает риски создания и использования дубликатов путем ограничения возможностей злоумышленника по получению сведений, необходимых для создания дубликата. При реализации алгоритма в виде программно-аппаратного комплекса могут быть созданы условия, при которых вероятность обнаружения дубликата RFID-метки выше на 14,1 % по сравнению с текущей практикой (без учета алгоритма вероятность обнаружения составляет 24,8%, с учетом алгоритма — 38,9 %).

Таким образом, приведенные сведения позволяют сделать вывод о целесообразности предлагаемого алгоритма, так как он применим к специфической области (в государственных информационных системах в рамках системы маркировки импортируемых товаров), а за счет возможностей по его автоматизации он позволяет добиться увеличения вероятности обнаружения RFID-дубликатов в сравнении с существующим подходом по выборочной проверке RFID-меток в таможенных органах на 14,1 %. Перспективы исследования заключаются в дальнейшей детализации структуры EPC-области памяти RFID-метки для снижения рисков создания и использования злоумышленниками дубликатов RFID-меток.

Литература

1. Kamaludin H., Mahdin H., Abawajy J.H. Clone tag detection in distributed RFID systems // PLoS ONE. 2018. V. 13. N 3. P. e0193951. https://doi.org/10.1371/journal.pone.0193951
2. Zanetti D., Capkun S., Juels A. Tailing RFID tags for clone detection // NDSS Symposium. 2013.
3. Lehtonen M., Ostojic D., Ilic A., Michahelles F. Securing RFID systems by detecting tag cloning // Lecture Notes in Computer Science. 2009. V. 5538. P. 291–308. https://doi.org/10.1007/978-3-642-01516-8_20
4. Jokhio I.A., Jokhio S.H., Baloch J.A. A novel security method to protect RFID cloning attacks. 2012 [Электронный ресурс]. URL: http://oaji.net/articles/2016/2712-1454748209.pdf (дата обращения: 28.07.2021).
5. Shi J., Kywe S.M., Li Y. Batch clone detection in RFID-enabled supply chain // Proc. of the 2014 IEEE International Conference on RFID (IEEE RFID). 2014. P. 118–125. https://doi.org/10.1109/RFID.2014.6810721
6. Chen X., Liu J., Wang X., Zhang X., Wang Y., Chen L. Combating tag cloning with COTS RFID devices // Proc. of the 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). 2018. P. 1–9. https://doi.org/10.1109/SAHCN.2018.8397134
7. Bu K., Xu M., Liu X., Luo J., Zhang S., Weng M. Deterministic detection of cloning attacks for anonymous RFID systems // IEEE Transactions on Industrial Informatics. 2015. V. 11. N 6. P. 1255–1266. https://doi.org/10.1109/TII.2015.2482921
8. Jin B., Jin H. Security analysis of RFID based on multiple readers // Procedia Engineering. 2011. V. 15. P. 2598–2602. https://doi.org/10.1016/j.proeng.2011.08.488
9. Финкенцеллер К. RFID-технологии. Справочное пособие / пер. с нем. М.: Додека XXI век, 2016. 490 с.
10. Волошина Н.В., Лавринович А.А. Рекомендации по автоматическому обнаружению дубликатов RFID-меток таможенными органами // Таможенные чтения — 2020. Стратегия развития 2030: вызовы времени. наука и инновации: Сборник материалов Международной научно-практической конференции. В 2-х т. Т. 1. СПб., 2020. С. 88–93.
11. Okpara O. Detecting Cloning Attack in Low-Cost Passive RFID Tags. 2015 [Электронный ресурс]. URL: https://www.researchgate.net/publication/280077859_Detecting_Cloning_Attack_in_Low-Cost_Passive_RFID_Tags (дата обращения: 28.07.2021). https://doi.org/10.13140/RG.2.1.1709.4240
12. Mitrokotsa A., Rieback M.R., Tanenbaum A.S. Classifying RFID attacks and defenses // Information Systems Frontiers. 2010. V. 12. N 5. P. 491–505. https://doi.org/10.1007/s10796-009-9210-z
13. Huang W., Zhang Y., Feng Y. ACD: An adaptable approach for RFID cloning attack detection // Sensors. 2020. V. 20. N 8. P. 2378. https://doi.org/10.3390/s20082378

References

1. Kamaludin H., Mahdin H., Abawajy J.H. Clone tag detection in distributed RFID systems. *PLoS ONE*, 2018, vol. 13, no. 3, pp. e0193951. https://doi.org/10.1371/journal.pone.0193951
2. Zanetti D., Capkun S., Juels A. Tailing RFID tags for clone detection. *NDSS Symposium*, 2013.
3. Lehtonen M., Ostojic D., Ilic A., Michahelles F. Securing RFID systems by detecting tag cloning. *Lecture Notes in Computer Science*, 2009, vol. 5538, pp. 291–308. https://doi.org/10.1007/978-3-642-01516-8_20
4. Jokhio I., Jokhio S.H., Baloch J.A. *A novel security method to protect RFID cloning attacks*. 2012. Available at: http://oaji.net/articles/2016/2712-1454748209.pdf (accessed: 28.07.2021).
5. Shi J., Kywe S.M., Li Y. Batch clone detection in RFID-enabled supply chain. *Proc. of the 2014 IEEE International Conference on RFID (IEEE RFID)*, Orlando, FL, USA, 2014, pp. 118–125. https://doi.org/10.1109/RFID.2014.6810721
6. Chen X., Liu J., Wang X., Zhang X., Wang Y., Chen L. Combating tag cloning with COTS RFID devices. *Proc. of the 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2018, pp. 1–9. https://doi.org/10.1109/SAHCN.2018.8397134
7. Bu K., Xu M., Liu X., Luo J., Zhang S., Weng M. Deterministic detection of cloning attacks for anonymous RFID systems. *IEEE Transactions on Industrial Informatics*, 2015, vol. 11, no. 6, pp. 1255–1266. https://doi.org/10.1109/TII.2015.2482921
8. Jin B., Jin H. Security analysis of RFID based on multiple readers. *Procedia Engineering*, 2011, vol. 15, pp. 2598–2602. https://doi.org/10.1016/j.proeng.2011.08.488
9. Finkenzeller K. *RFID Handbook Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*. Wiley, 2010, 480 p.
10. Voloshina N.V., Lavrinovich A.A. Recommendations for automatic detection RFID-mark duplicates by customs bodies. *"Customs readings 2020. Strategy 2030: challenges, science and innovations". Proceedings of the International scientific and practical conference.* St. Petersburg, 2020, pp. 88–93. (in Russian)
11. Okpara O. *Detecting Cloning Attack in Low-Cost Passive RFID Tags*. 2015. Available at: https://www.researchgate.net/publication/280077859_Detecting_Cloning_Attack_in_Low-Cost_Passive_RFID_Tags (accessed: 28.07.2021). https://doi.org/10.13140/RG.2.1.1709.4240
12. Mitrokotsa A., Rieback M.R., Tanenbaum A.S. Classifying RFID attacks and defenses. *Information Systems Frontiers*, 2010, vol. 12, no. 5, pp. 491–505. https://doi.org/10.1007/s10796-009-9210-z
13. Huang W., Zhang Y., Feng Y. ACD: An adaptable approach for RFID cloning attack detection. *Sensors*, 2020, vol. 20, no. 8, pp. 2378. https://doi.org/10.3390/s20082378

Авторы

Волошина Наталия Викторовна — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-9435-9580, nataliv@yandex.ru

Лавринович Александр Андреевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-5058-3473, Lavrinovich600@gmail.com

Authors

Natalia V. Voloshina — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-9435-9580, nataliv@yandex.ru

Aleksandr A. Lavrinovich — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-5058-3473, Lavrinovich600@gmail.com

Статья поступила в редакцию 16.09.2021
Одобрена после рецензирования 07.12.2021
Принята к печати 27.01.2022

Received 16.09.2021
Approved after reviewing 07.12.2021
Accepted 27.01.2022



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-74-81
УДК 519.6

Редукция набора детекторов LSB с заданной достоверностью

Роман Александрович Солодуха[1]✉, Геннадий Вадимович Перминов[2], Игорь Викторович Атласов[3]

[1] Воронежский институт МВД России, Воронеж, 394065, Российская Федерация
[2] Российский государственный университет правосудия (центральный филиал), Воронеж, 394006, Российская Федерация
[3] Московский университет МВД России имени В.Я. Кикотя, Москва, 117997, Российская Федерация

[1] standartal@list.ru✉, https://orcid.org/0000-0002-3878-4221
[2] perminovgv@mail.ru, https://orcid.org/0000-0003-4170-2861
[3] mathematic1@rambler.ru, https://orcid.org/0000-0001-6270-6787

Аннотация

Предмет исследования. Предложено решение задачи сокращения набора стеганоаналитических методов определения размера вложения в пространственную область изображения, на примере количественных детекторов Least Significant Bits (LSB) стеганографии. Предположено, что методы могут отслеживать одни и те же закономерности в контейнерах, вследствие чего результаты их работы могут коррелировать. Представлены результаты разработки и тестирования методики редукции набора методов с учетом точности и достоверности для снижения вычислительной сложности стеганоаналитической экспертизы. **Метод.** Теоретическая база предложенного решения — приближение регрессии первого рода линейной регрессией второго рода для многомерных случайных величин. Для верификации результатов выполнен численный эксперимент. В качестве источника контейнеров применена коллекция BOSSbase. Вложения реализованы с шагом 10 % путем автоматизации стеганографических программ freeware-сегмента CryptArkan и The Third Eye с помощью AutoIt. Использованы стеганоаналитические методы Weighted Stego, Sample Pairs, Triples analysis, Asymptotically Uniformly Most Powerful detection, Pair of Values. Датасеты получены в среде MATLAB, программа реализована на языке Python. Для обеспечения воспроизводимости эксперимента датасеты и программный код представлены в Kaggle. **Основные результаты.** На основе экспериментальных данных рассчитаны интервальные оценки коррелированности методов для различных размеров стегановложения. Разработана методика в составе математической модели, алгоритма реализации модели и компьютерной программы. **Практическая значимость.** Предложенную методику целесообразно применять в задачах, где необходимо учитывать точность и достоверность. Такие оценки востребованы при осуществлении экспертно-криминалистической деятельности по вопросам, допускающим вероятностные выводы. С помощью данных оценок аналитик может варьировать количество методов в зависимости от доступных вычислительных мощностей и временных рамок исследования.

Ключевые слова

стеганоанализ, редукция набора методов, достоверность, LSB, стеганоаналитическая экспертиза, регрессия

Ссылка для цитирования: Солодуха Р.А., Перминов Г.В., Атласов И.В. Редукция набора детекторов LSB с заданной достоверностью // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 74–81. doi: 10.17586/2226-1494-2022-22-1-74-81

Reduction of LSB detectors set with definite reliability

Roman A. Solodukha[1]✉, Gennadiy V. Perminov[2], Igor V. Atlasov[3]

[1] Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation
[2] Russian State University of Justice (Central Branch), Voronezh, 394006, Russian Federation
[3] Kikot Moscow University of the Ministry of Internal Affairs of Russia, Moscow, 117997, Russian Federation

[1] standartal@list.ru✉, https://orcid.org/0000-0002-3878-4221
[2] perminovgv@mail.ru, https://orcid.org/0000-0003-4170-2861
[3] mathematic1@rambler.ru, https://orcid.org/0000-0001-6270-6787

© Солодуха Р.А., Перминов Г.В., Атласов И.В., 2022

Abstract
The article focuses on decreasing the set of steganalytical methods that determine the payload value in the image spatial domain using quantitative detectors of Least Significant Bits (LSB) steganography. It is supposed that methods can trace the same image regularities, and hence their results can correlate. The work presents the results of the development and testing of the technique for reducing the set of steganalytical methods taking into account the accuracy and reliability to the diminution of the computational complexity of steganalytical expertise. The theoretical basis of the proposed solution is the approximation of regression of the first kind by linear regression of the second kind for multivariate random variables. To verify the results, the computational experiment was performed. The payloads were implemented in 10 % increments by automating the freeware steganographic programs CryptArkan and The Third Eye with AutoIt. The steganalytical methods, such as Weighted Stego, Sample Pairs, Triples analysis, Asymptotically Uniformly Most Powerful detection, Pair of Values, were used. The datasets were built in the MATLAB environment; the program was implemented in Python. For the experiment's reproducibility, the datasets and program code are provided in Kaggle. Interval estimates of methods correlation are calculated based on experimental data for various payload values. The developed technique includes a mathematical model, an algorithm for implementing the model, and a computer program. The proposed technique can be applied in those tasks where accuracy and reliability are taken into account. One of the subject areas demanding such assessments is computer forensics dealing with expertise with probabilistic conclusions. These estimates allow the analyst to vary the number of methods depending on the available computing resources and the time frame of the research.

Keywords
steganalysis, reduction of methods set, reliability, LSB, steganalytical expertise, regression

For citation: Solodukha R.A., Perminov G.V., Atlasov I.V. Reduction of LSB detectors set with definite reliability. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 74–81 (in Russian). doi: 10.17586/2226-1494-2022-22-1-74-81

Введение

В настоящее время в области анализа цифровой стеганографии изображений разработаны десятки методов, отвечающие на вопрос о размере вложения. Стеганоаналитические методы существуют «с открытым кодом» или в компилированном виде, т. е. представляются аналитику как «черный ящик». С учетом того, что разные методы могут отслеживать одни и те же закономерности в пустых и заполненных контейнерах, поэтому в их результатах может наблюдаться коррелированность, и один метод может включать векторы признаков других методов. С учетом вычислительной сложности стеганоаналитических методов возникает необходимость отбросить коррелированные методы или методы, являющиеся частными случаями других методов.

Данная задача особенно актуальна в рамках компьютерной экспертизы по определению наличия стегановложения при известном стеганографическом алгоритме/программе или в любом исследовании, допускающем вероятностные выводы. Согласно [1] при выполнении экспертизы эксперт не должен применять все методы, так как среди них могут присутствовать дублирующие друг друга. Потому на практике этот вопрос решается исходя из достаточности метода для решения задачи с необходимой надежностью. При этом должны учитываться такие факторы, как экономия средств и времени.

Если в распоряжении эксперта имеется несколько стеганоаналитических методов, то целесообразно предварительно проверить их на коррелированность (схожесть результатов), но с учетом требуемой достоверности, градация которой облегчила бы оценку и использование выводов эксперта следствием и судом [2]. Иначе говоря, отбрасывать какой-либо метод эксперт должен в соответствии с методикой, оценивающей достоверность результатов. При этом принятие решения на основании коэффициента корреляции требует проверки гипотезы о значимости различия между коэффициентами корреляции выборки и генеральной совокупности, а затем проверки гипотез о конкретных, интересующих аналитика, значениях генеральной совокупности [3]. Также для легитимности коэффициента корреляции результаты работы методов должны иметь совместное нормальное распределение, что накладывает дополнительное условие на выборку.

Возникает задача разработки методики редукции множества стеганоаналитических методов с возможностью априорной установки схожести методов с заданным уровнем достоверности при отсутствии требований совместного нормального распределения результатов их работы. Теоретическая база для решения данной задачи — приближение регрессии первого рода линейной регрессией второго рода для многомерных случайных величин. Теоретические вопросы для общего случая рассмотрены в работах [4, 5], а в настоящей работе приведено попарное сравнение методов для частного случая.

Предложенный алгоритм предполагается выполнять перед стеганоанализом, а также по мере появления новых стеганоаналитических методов для выявления и отбрасывания схожих методов в рамках заданных достоверности и точности.

Постановка задачи

Рассмотрим выборку из n групп по g файлов в каждой. Файлы являются пустыми и заполненными контейнерами с шагом вложения 10, т. е. размер вложения $s = 0, 10, \ldots, 100$ в процентах от максимально возможного. В данной работе предполагается, что используется m стеганоаналитических методов.

Обозначим результаты работы второго метода применительно к первой файловой группе $\mathbf{x}_1^2 = \{x_{1,i}^2\}_{i=1}^{g}$, обобщенный вектор результатов $\mathbf{x}_n^m = \{x_{n,i}^m\}_{i=g(n-1)+1}^{gn}$.

Построим n-размерную случайную величину для каждого m-го метода:

$$\zeta_m = (\zeta_{m,1},\zeta_{m,2}, \ldots, \zeta_{m,n-1}, \zeta_{m,n}) = \\ = \left(\frac{1}{\sqrt{g}}\sum_{j=1}^{g} x_{1,j}^m, \ldots, \frac{1}{\sqrt{g}}\sum_{j=(n-1)g+1}^{gn} x_{n,j}^m\right). \quad (1)$$

Согласно центральной предельной теореме для многомерных случайных величин, случайные величины ζ_k, ($k = 1, \ldots, m$) для $\tilde{\mathbf{x}} = (x_1, \ldots, x_m)$ имеют нормальное распределение с неизвестной функцией распределения

$$f_{\eta_k}(\tilde{\mathbf{x}}) = \frac{\sqrt{|\mathbf{A}|}}{(2\pi)^{\frac{m}{2}}}\exp\left(-\frac{1}{2}\langle\mathbf{A}(\tilde{\mathbf{x}} - \mathbf{h})^T, (\tilde{\mathbf{x}} - \mathbf{h})\rangle\right) = \\ = \frac{\sqrt{|\mathbf{A}|}}{(2\pi)^{\frac{m}{2}}}\exp\left(-\frac{1}{2}\sum_{i,j}a_{ij}(x_i - h_i)(x_j - h_j)\right),$$

где вектор $\mathbf{h} = (h_1, \ldots, h_m)$ и $\mathbf{A} = \{a_{ij}\}_{i,j=1}^m$ — симметричная положительно-определенная матрица, $\langle\rangle$ — обозначение скалярного произведения. Значение g выбирается из соображений обеспечения нормальности распределения ζ_m [6].

Применительно к стеганоанализу можно сказать, что есть методы, опирающиеся на одинаковые закономерности естественного контейнера и использующие схожий математический аппарат. Результаты их работы статистически идентичны, поэтому один из таких методов можно отбросить. Заметим, что условие

$$f(\zeta_1) = \zeta_2,$$

где f — измеримая функция, которую можно заменить условием математического ожидания $M(x)$:

$$\min_f M(\zeta_2 - f(\zeta_1))^2 = M(\zeta_2 - f_{21}(\zeta_1))^2 = 0.$$

Также можно приближенно считать, что для некоторых a_{21} и b_{21}

$$f_{21}(x) \approx a_{21} + b_{21}x,$$

поэтому

$$\min_{a,b} M(\zeta_2 - a - b\zeta_1)^2 \approx \min_f M(\zeta_2 - f(\zeta_1))^2 = 0.$$

Окончательно, если для методов 1 и 2 выполнено условие

$$M(\zeta_2 - a_{21} - b_{21}\zeta_1)^2 < \gamma,$$

то случайные величины ζ_1 и ζ_2 можно считать схожими.

Если выполнено условие

$$M(\zeta_2 - a_{21} - b_{21}\zeta_1)^2 > \gamma,$$

то считаем, что случайная величина ζ_2 принципиально отличается от ζ_1 (результаты работы методов 1 и 2 различны). Значение γ представляет собой меру схожести методов и выбирается аналитиком.

Обобщим данное утверждение. Пусть $1 \le k \le m$ произвольное натуральное число. В соответствии с [7], если уравнение

$$\zeta_{k+1} = f(\zeta_1, \ldots, \zeta_k) \quad (2)$$

верно, то случайная величина ζ_{k+1} статистически значимо связана с остальными случайными величинами. Метод $k + 1$ не привносит никакого улучшения в эту группу методов и может быть отброшен. Заметим, что условие (2) можно заменить условием

$$M(\zeta_{k+1} - f_{k+1}(\zeta_1, \ldots, \zeta_k))^2 = \min_f M(\zeta_{k+1} - f(\zeta_1, \ldots, \zeta_k))^2 = 0.$$

Примем, что для некоторых a_i и $b_{k+1,i}$

$$f_{k+1}(\mathbf{x}_1, \ldots, \mathbf{x}_k) \approx a_i + \sum_{i=1}^{k} b_{k+1,i}\mathbf{x}_i.$$

Следовательно

$$M\left(\zeta_{k+1} - a_i - \sum_{i=1}^{k} b_{k+1,i}\zeta_i\right)^2 \approx \min_f M(\zeta_2 - f(\zeta_1))^2 = 0.$$

Окончательно, если для некоторого $\gamma > 0$ (степень схожести) условие

$$M\left(\zeta_{k+1} - a_i - \sum_{i=1}^{k} b_{k+1,i}\zeta_i\right)^2 < \gamma$$

верно, то случайная величина ζ_{k+1} может быть отброшена.

Если условие

$$M\left(\zeta_{k+1} - a_i - \sum_{i=1}^{k} b_{k+1,i}\zeta_i\right)^2 > \gamma$$

верно, то случайная величина ζ_{k+1} фундаментально отличается от $\{\zeta_i\}_{i=1}^m$ и не может быть отброшена.

Математическая модель редукции

В основе предлагаемой методики лежит сравнение каждого метода с другим методом, с парой методов, с тройкой методов и т. д. для выявления оптимальной комбинации методов. Ограничимся рассмотрением самого простого частного случая (1-to-1) — сравнение двух случайных величин. Для каждой случайной величины ζ_i обозначим $m_i = M(\zeta_i)$. Выполним дальнейшие построения на примере двух первых случайных величин:

$$\min_a M(\zeta_2 - a - b_{21}\zeta_1)^2 = M((\zeta_2 - m_2) - \\ - b_{21}(\zeta_1 - m_1) + (m_2 - b_{21}m_1 - a_{12}))^2 = \\ = M(\zeta_2 - m_2)^2 - 2b_{21}M((\zeta_2 - m_2)(\zeta_1 - m_1)) + \\ + b_{21}^2 M(\zeta_1 - m_1)^2 + (m_2 - \beta_{21}m_1 - a_{12})^2 = \\ = M(\zeta_2 - m_2)^2 - 2b_{21}M((\zeta_2 - m_2)(\zeta_1 - m_1)) + \\ + b_{21}^2 M(\zeta_1 - m_1)^2. \quad (3)$$

Равенство (3) выполняется при удовлетворении условия

$$a = a_{21} = m_2 - b_{21}m_1.$$

Минимизируя по b_{21}, имеем

$$b_{21} = \frac{M((\zeta_2 - m_2)(\zeta_1 - m_1))}{M(\zeta_1 - m_1)^2},$$

$$M((\zeta_2 - m_2)(\zeta_1 - m_1)) = M(\zeta_1 - m_1)^2 b_{21}.$$

Согласно (3), имеем

$$\min_a M(\zeta_2 - a - \beta_{21}\zeta_1)^2 = M(\zeta_2 - m_2)^2 - \beta_{21}^2 M(\zeta_1 - m_1)^2. \quad (4)$$

Для оценки дисперсии и ковариации зададим уровень значимости, например, α = 0,95. Найдем коэффициенты β, входящие в выражение (4). Для всех $k = 1, \ldots, n$ рассмотрим m-мерные нормальные независимые случайные векторы $\zeta_k = (\zeta_{1k}, \ldots, \zeta_{mk})$, плотности вероятности которых для всех $k = 1, \ldots, n$ совпадают и равны

$$f_{\eta_k}(\tilde{\mathbf{x}}) = \frac{\sqrt{|\mathbf{A}|}}{(2\pi)^{\frac{m}{2}}} \exp\left(-\frac{1}{2}\langle \mathbf{A}(\tilde{\mathbf{x}} - \mathbf{h})^T, (\tilde{\mathbf{x}} - \mathbf{h})\rangle\right),$$

где $\mathbf{A} = \{a_{ij}\}_{i,j=1}^m$ — симметричная положительно-определенная матрица.

Для $1 \le i,j \le m$ и $1 \le k \le n$ определим случайную величину:

$$\overline{\zeta}_i = \frac{1}{n}\sum_{k=1}^{n} \zeta_{ik}. \quad (5)$$

Обозначим массив коэффициентов ковариации $\mathbf{\Theta} = \{\overline{\theta}_{ij}\}_{i,j=1}^m$, тогда

$$\overline{\theta}_{ij} \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} \zeta_{ik}\zeta_{jk} - \overline{\zeta}_j\overline{\zeta}_i = \frac{1}{n}\sum_{k=1}^{n} (\zeta_{ik} - \overline{\zeta}_i)(\zeta_{jk} - \overline{\zeta}_j). \quad (6)$$

Пользуясь уровнем значимости α и n найдем q, такое, чтобы выполнить условие

$$\int_{\frac{n}{(1+q)^2}}^{\frac{n}{(1-q)^2}} f_{\xi^2}^{n-1}(x)dx = \alpha, \quad (7)$$

где $f_{\xi^2}^{n-1}$ — плотность вероятности распределения ξ-квадрат с $n - 1$ степенями свободы. В этом случае, при $0 < q < 1$ для найденного n, величина $M(\zeta_k - m_k)^2$ с вероятностью α находится в промежутке функции вероятности $P(x)$

$$P(\overline{\theta}_{kk}(1 - q) < M(\zeta_k - m_k)^2 < \overline{\theta}_{kk}(1 + q)) = \int_{\frac{n}{(1+q)^2}}^{\frac{n}{(1-q)^2}} f_{\xi^2}^{n-1}(x)dx = \alpha. \quad (8)$$

При $q \ge 1$, для найденного n, величина $M(\zeta_k - m_k)^2$ с вероятностью α находится в промежутке

$$P(0 < M(\zeta_k - m_k)^2 < \overline{\theta}_{kk}(1 + q)) = \int_{\frac{n}{(1+q)^2}}^{\infty} f_{\xi^2}^{n-1}(x)dx = \alpha. \quad (9)$$

Величину $M((\zeta_i - m_i)(\zeta_j - m_j))$ рассчитаем по формуле

$$M((\zeta_i - m_i)(\zeta_j - m_j)) = \beta_{ij} M(\zeta_j - m_j)^2.$$

Заметим, что индекс коэффициента β_{ij} формируется следующим образом: сначала номер случайной величины, которую хотим приблизить, затем номер величины, на которую устанавливается регрессия. Рассмотрим нахождение β_{21}.

Обозначим

$$w_{21} = \frac{\begin{vmatrix} \overline{\theta}_{11} & \overline{\theta}_{21} \\ \overline{\theta}_{21} & \overline{\theta}_{22} \end{vmatrix}}{\overline{\theta}_{11}} = \frac{\overline{\theta}_{11}\overline{\theta}_{22} - \overline{\theta}_{21}^2}{\overline{\theta}_{11}}, \quad (10)$$

$$u_{21} = \sqrt{\overline{\theta}_{11}}\left(\frac{\overline{\theta}_{21}}{\overline{\theta}_{11}} - \beta_{21}\right) = \sqrt{\overline{\theta}_{11}}(b_{21} - \beta_{21}).$$

В соответствии с [8] случайная величина

$$\sqrt{n-2}\,\frac{u_{21}}{\sqrt{w_{21}}} = \frac{\sqrt{\overline{\theta}_{11}}(b_{21} - \beta_{21})\sqrt{n-2}}{\sqrt{w_{21}}}$$

имеет распределение Стьюдента с $n - 2$ степенями свободы.

Пользуясь уровнем значимости α, найдем t такое, чтобы выполнить условие

$$\int_{-t}^{t} f_{n-2}(x)dx = \alpha, \quad (11)$$

где f_{n-2} — плотность вероятности распределения Стьюдента с $n - 2$ степенями свободы. В этом случае для найденного t величина β_{21} с вероятностью α находится в промежутке

$$P\left(b_{21} - \frac{t\sqrt{w_{21}}}{\sqrt{\overline{\theta}_{11}}\sqrt{n-2}} < \beta_{21} < b_{21} + \frac{t\sqrt{w_{21}}}{\sqrt{\overline{\theta}_{11}}\sqrt{n-2}}\right) = \int_{-t}^{t} f_{n-2}(x)dx = \alpha. \quad (12)$$

Обобщенный алгоритм редукции

Рассмотрим методику работы обобщенного алгоритма редукции.

1. Определение количества групп.

   Исходя из требуемой достоверности α, в соответствии с формулой (9) подберем значения n и q. Выберем степень схожести γ.
2. Формирование исходных данных.
   - — Подбор требуемого количества файлов g.
   - — Реализация стегановложения с заданным шагом.
   - — Формирование группы контейнеров.
   - — Для каждого контейнера расчет стеганоаналитического вектора признаков.
   - — В соответствии с формулой (1) построение n-мерных реализаций случайной величины ζ_m. Проверена нормальность распределения. Если ζ_m распределена не нормально, то выполняется увеличение g и повтор данного пункта с новым файлом g.
3. Попарное сравнение случайных величин ζ_m.
   - — Вычисление элементов матрицы **W**. Элементы матрицы w_{ij} вычисляются по формулам (5), (6) и (10).
   - — Вычисление коэффициента β. В соответствии с выбранным значением определение значения t по формуле (11) и определение коэффициента β_{ij} по формуле (12).
   - — Оценка дисперсии $M(\zeta_k - m_k)^2$. В соответствии со значениями n, q, и α вычисление дисперсии по формулам (8) при $0 < q < 1$, или (9) при $q > 1$.
4. Принятие решения.

   Полученные значения $M(\zeta_k - m_k)^2$ и β_{ij} подставляем в формулу (4). Если максимальное значение $M(\zeta_k - m_i)^2 - \beta_{ij}^2 M(\zeta_j - m_j)^2 < \gamma$, то метод отбрасывается.

Схема программной реализации[1]

Схема программной реализации редукции набора стеганоаналитических методов с заданной достоверностью представлена на рис. 1.

Для программной реализации выбран язык программирования Python. В качестве среды разработки использовано приложение Jupyter Notebook, входящее в состав дистрибутива Anaconda. Все библиотеки, упоминаемые далее, входят в состав данного дистрибутива, если не указано иное.

Рассмотрим шаги программной реализации.

Шаг 1. Ввод исходных данных из файла. Для работы с файлами использованы библиотеки xlrd и xlwt. Исходный файл start_data.xls содержит двумерную таблицу, в которой строки соответствуют методам (m), столбцы — группам файлов (n).

Шаг 2. Вычисление массива коэффициентов ковариации методов с помощью функции cov библиотеки numpy.

Шаг 3. Вычисление массива коэффициентов ширины диапазонов Omega.

Шаг 4. Вычисление матрицы отношений ковариации к дисперсии и помещение ее в массив **b**.

Шаг 5. Определение конечных точек диапазона, содержащего заданный процент распределения Стьюдента. Для этого использована функция interval библиотеки scipy.stats.t для работы с функциями распределения Стьюдента. В качестве исходных данных выступают: заданный процент, степени свободы, среднее значение, среднеквадратическое отклонение.

Шаг 6. Нахождение нижних и верхних границ диапазонов по Стьюденту. Матрицы Beta_from_Stud и Beta_to_Stud содержат границы диапазонов, в которых находится величина Beta для найденных на шаге 5 точек с заданной вероятностью.

Шаг 7. Вычисление Disper_from_Khi и Disper_to_Khi – матриц диапазонов дисперсий методов. Для вычисления использована функция q_fron_Khi с исходными данными: количество степеней свободы, требуемая достоверность, точность определения корней. Данная функция не входит в состав стандартных функций дистрибутива Anaconda, определена непосредственно в программе.

Шаг 8. Определение верхних диапазонов разностей выборочных средних по методам, взятых попарно. Результат помещен в массив Difference_max. Определены все возможные комбинации выборочных средних и выбраны максимальные из них.

Шаг 9. Определение отбрасываемых методов. Выполнено сравнение верхних диапазонов выборочных средних, полученные на шаге 8, с заданной степенью схожести Gamma, таким образом, чтобы разность между методами, взятыми попарно была меньше заданной.

Шаг 10. Коррекция исходных данных с учетом отброшенных методов и запись в файл.

[1] Исходный код доступен в Kaggle: [Электронный ресурс]. Режим доступа: https://www.kaggle.com/romansolodukha/batrlett-1-1 (дата обращения: 10.01.2022).

Программный эксперимент

Проведен эксперимент по верификации предложенной методики. В качестве стеганоаналитических методов для сравнения выбраны количественные детекторы LSB[2] (Least Significant Bits): Weighted Stego (WS) [9], Sample Pairs (SP) [10], Triples analysis (T) [11], Asymptotically Uniformly Most Powerful detection (AUMP) [12], Pair of Values (PoVs) [13]. WS использован в модификации без коррекции (WSn) и с коррекцией (WSy) смещения. Параметры AUMP, рекомендованные в работе [12]: размер блока пикселов — 16, степень полинома — 6. LSB представляет программу, которая вычисляет размер вложения на основании данных изображения. Приведенные детекторы используют один цветовой канал пространственной области изображения с глубиной цвета 8 бит.

В качестве источника контейнеров взята коллекция BOSSbase 1.01[3] (первые 1000 файлов), которая содержит 10 000 8-битных полутоновых изображений размером 512 × 512 пикселов в формате PGM (Portable Grey Map). Для дальнейшей работы файлы были преобразованы в формат BMP 8 бит с помощью XnConvert и в формат BMP 24 бит — пакетного преобразования FastStone Image Viewer. Поскольку все три цветовых плоскости контейнера одинаковые, для анализа использован только канал красного цвета. Контейнеры заполнены с шагом 10 % от максимального размера вложения от 9 до 99 %. Таким образом, выборка составила 11 000 контейнеров[4].

Выбраны следующие стеганографические программы, реализующие LSB пространственной области изображения:

1. CryptArkan 1.0 — не требует установки, позволяет задействовать 1, 2 или 4 LSBs файла, шифровать стегановложение. Вложение реализуется последовательно, что видно из рис. 2, *a–c*. Для эксперимента реализован режим LSB.
2. The Third Eye 1.0 — не требует установки, задействует плоскость младшего бита. Байты для вложения выбраны случайно, что видно из рис. 2, *d–f*.

Изображения на рис. 2 получены с помощью программы WinMerge[5] и иллюстрируют локализацию изменений изображения с размерами вложения 9, 49 и 99 %.

Для достоверности $\alpha = 0{,}95$, устойчивое соответствие нормальному распределению по критерию Шапиро–Уилка на уровне значимости 0,01 было достигнуто при количестве групп $n = 100$. Далее программный модуль редукции запускался с различными

[2] Structural LSB Detectors [Электронный ресурс]. Режим доступа: http://dde.binghamton.edu/download/structural_lsb_detectors/ (дата обращения 10.01.2022).

[3] Image Database BOSSbase 1.01 [Электронный ресурс]. Режим доступа: http://dde.binghamton.edu/download/ImageDB/BOSSbase_1.01.zip (дата обращения 10.01.2022).

[4] Датасет доступен в Kaggle: [Электронный ресурс]. Режим доступа: https://www.kaggle.com/romansolodukha/bartlett-reduction (дата обращения: 10.01.2022).

[5] WinMerge. Open Source инструмент сравнения и слияния для Windows [Электронный ресурс]. Режим доступа: https://winmerge.org/downloads/ (дата обращения: 10.01.2022).

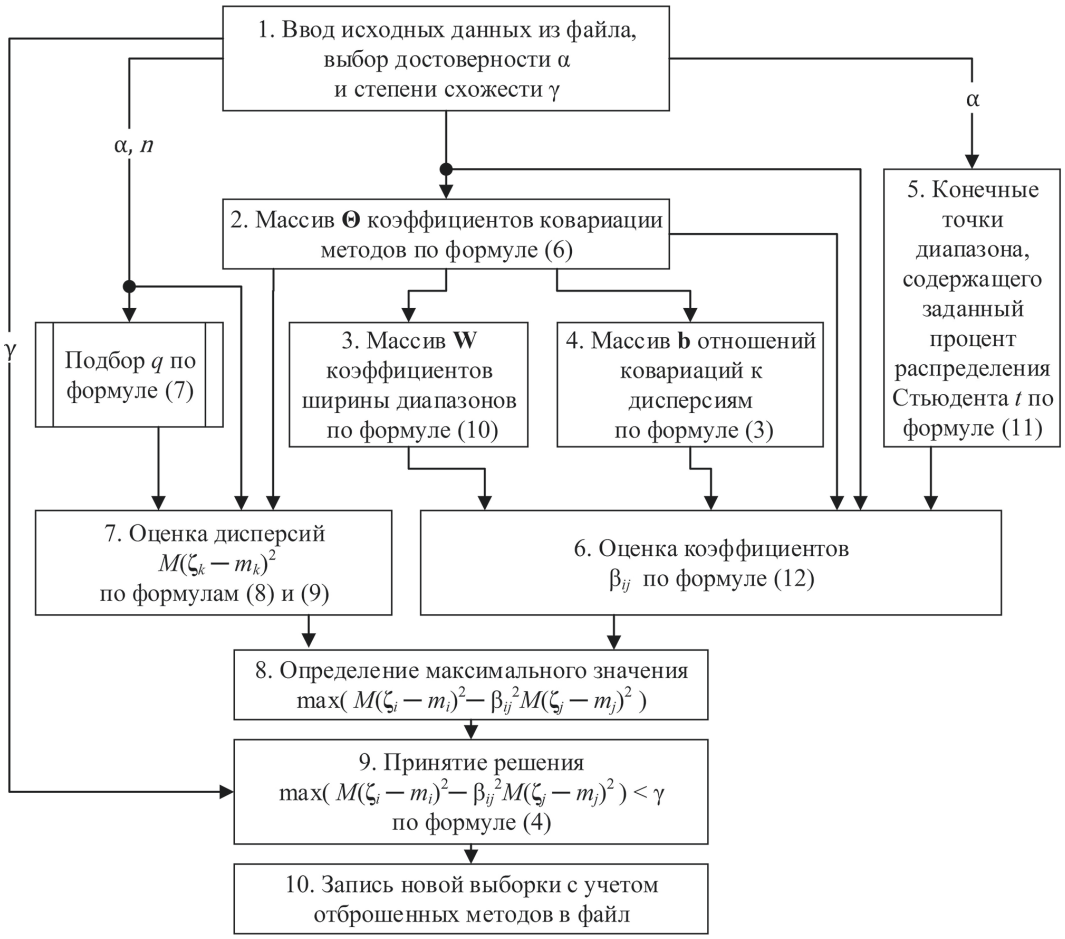


Рис. 1. Схема программной реализации редукции набора методов с заданной достоверностью

Fig. 1. Scheme of software implementation of methods set reduction with definite reliability

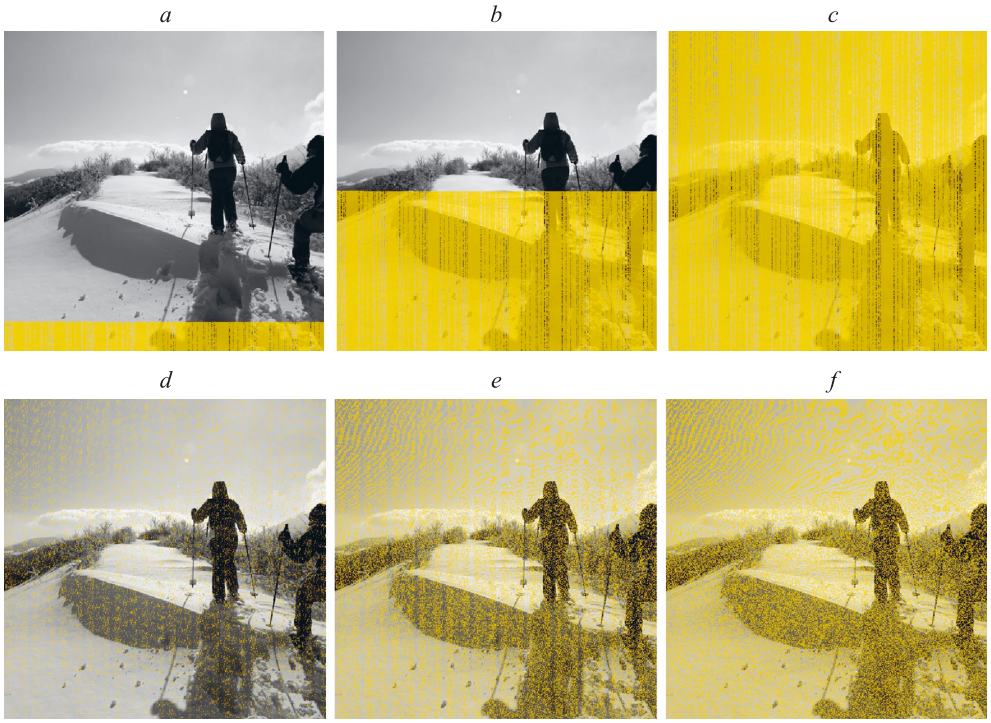


Рис. 2. Модификация пространственной области изображения программ CryptArkan (*a–c*) и The Third Eye (*d–f*)

Fig. 2. Modification of the image spatial domain by the CryptArkan (*a–c*) and The Third Eye (*d–f*)

Таблица. Степени схожести количественных детекторов LSB

Table. Similarity of quantitative LSB detectors

Отброшенные методы	CryptArkan			The Third Eye		
	Размер вложения, %					
	0	49	99	0	49	99
WSn	0,31	0,33	0,36	0,32	0,34	0,39
WSn, AUMP	0,51	0,54	0,59	0,56	0,50	0,58
WSn, AUMP, SP	0,82	0,86	0,91	0,84	0,88	0,91
WSn, AUMP, SP, T	1,00	1,04	1,05	1,01	1,06	1,10
WSn, AUMP, SP, T, PoVs	1,12	1,14	1,18	1,13	1,16	1,21

значениями γ и отбрасывал коррелированные методы. Рассмотрены три уровня вложения 0, 49 и 99 %. Для каждой стеганопрограммы и уровня вложения методы отбрасывались в одинаковом порядке: WSn, AUMP, SP, Triples, PoVs. Результаты в виде значений γ приведены в таблице.

Анализ таблицы позволяет сделать вывод о том, что коррелированность работы методов не зависит от стеганопрограммы и уменьшается с увеличением размера вложения. Интерпретировать полученные результаты можно следующим образом: с вероятностью 0,95 квадрат разности значений отброшенного и какого-либо из оставшихся методов не превышает γ. В плане практического применения видно, что различия между самыми близкими методами WSn и WSy составляют $\gamma = [0{,}31; 0{,}39]$, и для задач с $\gamma < 0{,}4$ WSn может быть отброшен.

Выводы и перспективы

Представлена методика, включающая математическую модель, алгоритм реализации модели и компьютерную программу для редукции набора стеганоаналитических методов. В дальнейших исследованиях по данной тематике предполагается провести программный эксперимент не с методами, а многомерными стеганоаналитическими векторами признаков (SPAM (686D) [14], SRM (34671D) [15], CHEN (486D) [16], etc.), применив предложенный подход для снижения размерности признакового пространства.

Заключение

Экспериментально подтверждена гипотеза о возможности вычисления схожести стеганоаналитических методов с помощью интервальных оценок. Это позволяет использовать предложенную в работе методику в сфере, относящейся к судебной экспертной деятельности. В программном эксперименте в качестве методов использованы количественные LSB детекторы, степени их схожести рассчитаны, применительно к стеганопрограммам The Third Eye и CryptArkan. Распространение предложенного подхода на известные стеганопрограммы/алгоритмы и стеганоаналитические методы позволит уменьшить вычислительные затраты стеганоанализа с заданной достоверностью при проведении исследования на основании известного стеганоалгоритма/программы.

Литература

1. Жижина М.В. Судебно-почерковедческая экспертиза документов / под ред. проф. Е.П. Ищенко. М.: ЮРЛИТИНФОРМ, 2006. С. 110–116.
2. Усов А.И., Градусова О.Б., Кузьмин С.А. Использование вероятностно-статистических методов при оценке значимости результатов экспертного исследования в отечественной и зарубежной судебно-экспертной практике (сравнительный анализ) // Теория и практика судебной экспертизы. 2018. Т. 13. № 4. С. 6–15. https://doi.org/10.30764/1819-2785-2018-13-4-6-15
3. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов / пер. с нем. М.: Финансы и статистика, 1983. 304 с.
4. Атласов И.В., Солодуха Р.А. Стеганоанализ цифровых изображений: автоматизация, оптимизация, достоверность: монография. Электр. дан. и прогр. Воронеж: Воронежский институт МВД России, 2020. 171 с. [Электронный ресурс]. URL: www.kaggle.com/dataset/c736afc689f328127816c59961677b3106468ce9f8e4399f3d18a12745c9e94c (дата обращения: 10.11.2021).
5. Atlasov I., Solodukha R. Reduction of steganalytical methods set with determined reliability // Proc. 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA). 2019. P. 126–131. https://doi.org/10.1109/SUMMA48161.2019.8947470

References

1. Zhizhina M.V. *Forensic Handwriting Document Expertise*. Moscow, JURLITINFORM Publ., 2006, pp. 110–116. (in Russian)
2. Usov A.I., Gradusova O.B., Kuz'min S.A. The use of probabilistic and statistical methods to test the significance of scientific evidence: comparative analysis of current forensic practices in russia and abroad. *Theory and Practice of Forensic Science*, 2018, vol. 13, no. 4, pp. 6–15. (in Russian). https://doi.org/10.30764/1819-2785-2018-13-4-6-15
3. Förster E., Rönz B. *Methoden der Korrelations — und Regressionsanalyse*. Berlin, 1979, 324 p.
4. Atlasov I.V., Solodukha R.A. *Steganography of Digital Images: Automatization, Optimization and Reliability*. Voronezh, Voronezhskij institut MVD Rossii. Available at: www.kaggle.com/dataset/c736afc689f328127816c59961677b3106468ce9f8e4399f3d18a12745c9e94c (accessed: 10.11.2021). (in Russian)
5. Atlasov I., Solodukha R. Reduction of steganalytical methods set with determined reliability. *Proc. 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*, 2019, pp. 126–131. https://doi.org/10.1109/SUMMA48161.2019.8947470
6. Wishart J., Bartlett M.S. The distribution of second order moment statistics in a normal system. *Mathematical Proceedings of the*

6. Wishart J., Bartlett M.S. The distribution of second order moment statistics in a normal system // Mathematical Proceedings of the Cambridge Philosophical Society. 1932. V. 28. P. 455–459. https://doi.org/10.1017/S0305004100010690
7. Фишер Р.А. Статистические методы для исследователей. М.: Госстатиздат, 1958. 268 с.
8. Bartlett M.S. On the theory of statistical regression // Proceedings of the Royal Society of Edinburgh. 1934. V. 53. P. 260–283. https://doi.org/10.1017/S0370164600015637
9. Ker A., Böhme R. Revisiting weighted stego-image steganalysis // Proceedings of SPIE. 2008. V. 6819. P. 681905. https://doi.org/10.1117/12.766820
10. Dumitrescu S., Wu X., Memon D. On steganalysis of random LSB embedding in continuous-tone images // IEEE International Conference on Image Processing. V. 3. 2002. P. 641–644.
11. Ker A. A general framework for structural steganalysis of LSB replacement // Lecture Notes in Computer Science. 2005. V. 3727. P. 296–311. https://doi.org/10.1007/11558859_22
12. Fillatre L. Adaptive steganalysis of Least Significant Bit replacement in grayscale natural images // IEEE Transactions on Signal Processing. 2012. V. 60. N 2. P. 556–569. https://doi.org/10.1109/TSP.2011.2174231
13. Westfeld A., Pfitzmann A. Attacks on steganographic systems: Breaking the steganographic utilities EzStego, Jsteg, Steganos and S-Tools-and Some Lessons Learned // Lecture Notes in Computer Science. 2000. V. 1768. P. 61–76. https://doi.org/10.1007/10719724_5
14. Pevný T., Bas P., Fridrich J. Steganalysis by subtractive pixel adjacency matrix // IEEE Transactions on Information Forensics and Security. 2010. V. 5. N 2. P. 215–224. https://doi.org/10.1109/TIFS.2010.2045842
15. Fridrich J., Kodovský J. Rich models for steganalysis of digital images // IEEE Transactions on Information Forensics and Security. 2012. V. 7. N 3. P. 868–882. https://doi.org/10.1109/TIFS.2012.2190402
16. Chen C., Shi Y.Q. JPEG image steganalysis utilizing both intrablock and interblock correlations // Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS). 2008. P. 3029–3032. https://doi.org/10.1109/ISCAS.2008.4542096

Cambridge Philosophical Society, 1932, vol. 28, pp. 455–459. https://doi.org/10.1017/S0305004100010690
7. Fisher R.A. *Statistical methods for research workers*. Oliver and Boyd. 12th ed., revised. Edinburg, London, Oliver and Boyd, 1954.
8. Bartlett M.S. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 1934, vol. 53, pp. 260–283. https://doi.org/10.1017/S0370164600015637
9. Ker A., Böhme R. Revisiting weighted stego-image steganalysis. *Proceedings of SPIE*, 2008, vol. 6819, pp. 681905. https://doi.org/10.1117/12.766820
10. Dumitrescu S., Wu X., Memon D. On steganalysis of random LSB embedding in continuous-tone images. *IEEE International Conference on Image Processing*, vol. 3, 2002, pp. 641–644.
11. Ker A. A general framework for structural steganalysis of LSB replacement. *Lecture Notes in Computer Science*, 2005, vol. 3727, pp. 296–311. https://doi.org/10.1007/11558859_22
12. Fillatre L. Adaptive steganalysis of Least Significant Bit replacement in grayscale natural images. *IEEE Transactions on Signal Processing*, 2012, vol. 60, no. 2, pp. 556–569. https://doi.org/10.1109/TSP.2011.2174231
13. Westfeld A., Pfitzmann A. Attacks on steganographic systems: Breaking the steganographic utilities EzStego, Jsteg, Steganos and S-Tools-and Some Lessons Learned. *Lecture Notes in Computer Science*, 2000, vol. 1768, pp. 61–76. https://doi.org/10.1007/10719724_5
14. Pevný T., Bas P., Fridrich J. Steganalysis by subtractive pixel adjacency matrix. *IEEE Transactions on Information Forensics and Security*, 2010, vol. 5, no. 2, pp. 215–224. https://doi.org/10.1109/TIFS.2010.2045842
15. Fridrich J., Kodovský J. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 2012, vol. 7, no. 3, pp. 868–882. https://doi.org/10.1109/TIFS.2012.2190402
16. Chen C., Shi Y.Q. JPEG image steganalysis utilizing both intrablock and interblock correlations. *Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS)*, 2008, pp. 3029–3032. https://doi.org/10.1109/ISCAS.2008.4542096

Авторы

Солодуха Роман Александрович — кандидат технических наук, доцент, доцент, Воронежский институт МВД России, Воронеж, 394065, Российская Федерация, SC 57201778092, https://orcid.org/0000-0002-3878-4221, standartal@list.ru
Перминов Геннадий Вадимович — кандидат технических наук, доцент, Российский государственный университет правосудия (центральный филиал), Воронеж, 394006, Российская Федерация, https://orcid.org/0000-0003-4170-2861, perminovgv@mail.ru
Атласов Игорь Викторович — доктор физико-математических наук, профессор, профессор, Московский университет МВД России имени В.Я. Кикотя, Москва, 117997, Российская Федерация, SC 57201775678, https://orcid.org/0000-0001-6270-6787, mathematic1@rambler.ru

Authors

Roman A. Solodukha — PhD, Associate Professor, Associate Professor, Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation, SC 57201778092, https://orcid.org/0000-0002-3878-4221, standartal@list.ru
Gennadiy V. Perminov — PhD, Associate Professor, Russian State University of Justice (Central Branch), Voronezh, 394006, Russian Federation, https://orcid.org/0000-0003-4170-2861, perminovgv@mail.ru

Igor V. Atlasov — D.Sc., Full Professor, Kikot Moscow University of the Ministry of Internal Affairs of Russia, Moscow, 117997, Russian Federation, SC 57201775678, https://orcid.org/0000-0001-6270-6787, mathematic1@rambler.ru

Статья поступила в редакцию 08.12.2021
Одобрена после рецензирования 20.12.2021
Принята к печати 28.01.2022

Received 08.12.2021
Approved after reviewing 20.12.2021
Accepted 28.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2022 Том 22 Номер 1

doi: 10.17586/2226-1494-2022-22-1-82-92
УДК 004.932

Классификация объектов на изображениях с учетом искажений на основе двухэтапного топологического анализа

Сергей Владимирович Еремеев[1]✉, Артём Владимирович Абакумов[2]

[1,2] Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, 602252, Российская Федерация

[1] sv-eremeev@yandex.ru✉, https://orcid.org/0000-0001-8482-1479
[2] artem210966@yandex.ru, https://orcid.org/0000-0001-5784-7147

Аннотация

Предмет исследования. Предложен метод автоматической классификации пространственных объектов на изображениях в условиях ограниченного набора данных. Исследована устойчивость метода к искажениям, которые часто возникают на изображениях в силу природных явлений и частичном перекрытии объектов городской инфраструктуры. Для получения высокой точности классификации с помощью существующих подходов требуется большая обучающая выборка, включая наборы данных с искажениями, что значительно увеличивает вычислительную сложность. **Метод.** Предложен метод двухэтапного топологического анализа изображений. Топологические признаки первоначально извлекаются при анализе изображения в диапазоне яркостей от 0 до 255, а затем от 255 до 0. Эти признаки дополняют друг друга и отражают топологическую структуру объекта. При определенных деформациях и искажениях структура объекта в виде извлеченных признаков сохраняется. Преимущество метода — небольшое количество эталонов, что снижает вычислительную нагрузку при обучении по сравнению с нейросетевым подходом. **Основные результаты.** Выполнено исследование и сравнение предложенного метода с современным нейросетевым подходом. Исследование проведено на наборе данных DOTA (Dataset for Object deTection in Aerial images), содержащем снимки пространственных объектов нескольких классов. При отсутствии искажений на изображении нейросетевой подход показал точность классификации свыше 98 %, а при использовании предложенного метода около 82 %. Также в работе применены искажения изображений: поворот на 90°, сужение и обрезание края на 50 %, а также их комбинации. Предложенный метод показал свою устойчивость и превзошел нейросетевой подход. В самой сложной комбинации теста снижение точности классификации нейросетевого подхода составило 46 %, в то время как для предложенного метода только 12 %. **Практическая значимость.** Предложенный метод целесообразно использовать при большой вероятности появления искажений на изображениях. Такие искажения возникают в сфере геоинформатики при анализе объектов разных масштабов, при различных погодных условиях, частичном перекрытии одного объекта другим, при наличии тени и других факторов. Возможно использование предложенного метода в системах технического зрения промышленных предприятий при автоматической классификации типа деталей наложенных объектов.

Ключевые слова

топологический анализ, персистентная гомология, искажение изображений, классификация объектов, нейронные сети

Благодарности

Исследование выполнено в рамках Программы развития ЯрГУ, проект № П2-К-1-Г-3/2021.

Ссылка для цитирования: Еремеев С.В., Абакумов А.В. Классификация объектов на изображениях с учетом искажений на основе двухэтапного топологического анализа // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 82–92. doi: 10.17586/2226-1494-2022-22-1-82-92

© Еремеев С.В., Абакумов А.В., 2022

Classification of objects in images with distortions based on a two-stage topological analysis

Sergey V. Eremeev[1]✉, Artyom V. Abakumov[2]

[1,2] Murom Institute (Branch) of Vladimir State University, Murom, 602252, Russian Federation

[1] sv-eremeev@yandex.ru✉, https://orcid.org/0000-0001-8482-1479

[2] artem210966@yandex.ru, https://orcid.org/0000-0001-5784-7147

Abstract

The authors propose a method for automatic classification of spatial objects in images under conditions of a limited data set. The stability of the method to distortions appearing in images due to natural phenomena and partial overlap of urban infrastructure objects is investigated. High classification accuracy, when using existing approaches, requires a large training sample, including data sets with distortions, which significantly increases computational complexity. The paper proposes a method for a two-step topological analysis of images. Topological features are initially extracted by analyzing the image in the brightness range from 0 to 255, and then from 255 to 0. These features complement each other and reflect the topological structure of the object. Under certain deformations and distortions, the object preserves its structure in the form of extracted features. The advantage of the method is a small number of patterns, which reduces the computational complexity of training compared to neural networks. The proposed method is investigated and compared with the modern neural network approach. The study was performed on a DOTA dataset (Dataset for Object deTection in Aerial images) containing images of spatial objects of several classes. In the absence of distortion in the image, the neural network approach showed a classification accuracy of over 98 %, while the proposed method achieved about 82 %. Further distortions such as 90 degree rotation, 50 % narrowing and 50 % edge truncation and their combinations were applied in the experiment. The proposed method showed its robustness and outperformed the neural network approach. In the most difficult combination of the test, the decrease in classification accuracy of the neural network was 46 %, while the described method showed 12 %. The proposed method can be applied in cases with a high probability of distortion in the images. Such distortions arise in the field of geoinformatics when analyzing objects of various scales, under different weather conditions, partial overlap of one object with another, in the presence of shadows, etc. It is possible to use the proposed method in vision systems of industrial enterprises for automatic classification of the parts that belong to superimposed objects.

Keywords

topological analysis, persistent homology, image distortion, object classification, neural networks

Acknowledgements

The reported study was funded by the YSU Programme (the research project No. П2-К-1-Г-3/2021).

For citation: Eremeev S.V., Abakumov A.V. Classification of objects in images with distortions based on a two-stage topological analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 82–92 (in Russian). doi: 10.17586/2226-1494-2022-22-1-82-92

Введение

Проблема классификации объектов на изображениях в настоящее время является актуальной в самых различных сферах. Под классификацией изображений понимают соотнесение объектов на этих изображениях к одному из классов. Интерес исследователей к этой проблеме с каждым годом возрастает.

В первую очередь это связано с тем, что увеличивается сложность исходной информации, объекты на изображении подвергаются различным искажениям. Это могут быть шумовые факторы, включая освещение, наличие теней, погодные условия [1]. Также могут быть различные аффинные преобразования, получение изображений с разных масштабов и углов съемки в 3D-пространстве, частичное перекрытие объектов на картах городской инфраструктуры и др. Таким образом, увеличение сложности классифицируемых объектов на изображении влечет за собой совершенствование и разработку новых подходов для решения проблемы их классификации.

Цель работы заключается в разработке метода для извлечения топологических особенностей объекта на изображении, а также в исследовании устойчивости этих признаков при классификации с учетом внесения различных искажений в проверочную выборку.

Существующие методы и подходы

В мировом сообществе сформировалась общая концепция для классификации данных на изображениях, которая включает в себя следующие основные этапы: предварительная обработка изображений, извлечение признаков и непосредственно классификация по этим признакам. Каждый этап играет важную роль и вносит свой определенный вклад в итоговый результат.

В настоящее время существует большое количество разработок в области классификации объектов на изображениях, которые используют самые передовые технологии. При этом их можно условно разделить на две большие группы: традиционные и нейросетевые подходы. Последние исследования показывают, что нейросетевые подходы опережают другие методы.

В работе [2] отмечено, что в данном направлении также возникают следующие проблемы:

— переобучения;
— для каждой ситуации требуется своя выборка;
— увеличение вычислительной сложности.

Для решения этих проблем перспективным считается направление GAN (Generative Adversarial Networks), основная суть которого заключается в том, что одна сеть обучает другую.

Рассмотрим традиционные подходы, которые наиболее успешно справляются с наличием искажений при классификации изображений.

Наиболее популярный метод для извлечения признаков — SIFT (Scale Invariant Feature Transform), который позволяет выделить ключевые точки на изображении [3]. Набор признаков из этих точек инвариантен к аффинным преобразованиям. Автоматический поиск на изображении оптимальных ключевых точек является сложной задачей, одно из ее решений показано в работе [4].

Также популярны геометрические подходы, которые позволяют сформировать из ключевых точек графовые модели объекта [5]. При этом образуются топологические связи, которые сохраняются, например, при движении человека на видеопоследовательности. Еще один смежный, широко используемый подход — вычисление геодезических расстояний [6] между объектами интереса на некоторой поверхности, а также использование дифференциальной геометрии [7].

Текстурные признаки также содержат много полезной информации, которая может быть использована для учета различных искажений. К таким методам относятся подходы на основе локальных бинарных шаблонов [8], а также близкие к этой тематике новые направления в виде знаковых представлений изображений [9]. Эти подходы показали свою устойчивость к шуму. Также широко используются методы на основе гистограммы направленных градиентов (Histogram of Oriented Gradients, HOG) [1], которые сохраняют устойчивость признаков при разном освещении. Кроме того, часто для анализа текстурных признаков изображений используют Wavelet [10] и Фурье-преобразование [11], где обработка информации ведется в пространстве низких и высоких частот.

В отдельную группу можно выделить методы, которые используют семантическую информацию о близко расположенных объектах. Так, например, в работе [12] показан подход для контекстного анализа объектов на изображении. При этом заранее создаются правила, которые описывают, как объекты могут быть связаны между собой. В частности, площадка находится напротив дома и около дороги. Такой подход является достаточно перспективным для классификации объектов, особенно в условиях пересечения объектов (например, дерево закрывает часть дома). Однако создание правил является очень трудоемкой операцией, и количество этих правил будет резко увеличиваться при появлении новых типов объектов.

Непосредственно для классификации применяют уже ставшие классическими следующие широко используемые подходы: метод опорных векторов (Support Vector Machine, SVM) [13], дерево решений (Random Forest) [14], сверточные нейронные сети (Convolutional Neural Network, CNN) [15]. Зачастую, для методов машинного обучения требуется вектор признаков заданного размера, что вызывает определенные трудности при неопределенном количестве признаков.

В работе [16] выполнен обширный обзор самых современных методов для извлечения признаков. Вместе с этим продемонстрировано использование различных классификаторов в сочетании с полученными признаками. На разных наборах данных наилучшие результаты показали нейронные сети, однако без учета искажений.

На взгляд авторов настоящей работы, для классификации объектов с учетом их искажений целесообразно использовать топологические методы, так как они сохраняют структуру объектов при различных поворотах, небольших деформациях, изменениях масштаба, а также при появлении шума.

Топология для анализа объектов применяется уже достаточно давно. Сначала это был подсчет количества компонент связности и дыр, которые используются для анализа бинарных изображений. Каждая компонента связности представляет собой отдельный объект из черных пикселов. Дыра характеризуется множеством белых пикселов, которые расположены внутри объекта. Сложный объект может содержать в себе несколько дыр. Относительно недавно появились новые методы алгебраической топологии [17], что вызвало интерес у многих научных лабораторий по всему миру. Основная идея топологического анализа данных заключается в вычислении и анализе глобальных структур данных, которые образуются множеством точек в n-мерном пространстве. В отличии от кластерного анализа учитывается продолжительность существования каждой такой структуры, что отражается в виде топологической характеристики, называемой баркодом [18]. Баркод представляет собой множество параллельных отрезков разной длины, которые содержат информацию о всех сформированных структурах, а также об их объединении. Отдельное направление занимает топологический анализ изображений [19, 20], где также можно анализировать топологические особенности. В этом случае баркод хранит информацию о компонентах связности и дырах, которые формируются на каждом уровне яркости.

Методология топологического анализа данных

В работе использована персистентная гомология, которая является подразделом топологического анализа данных. Построение топологических признаков из множества точек на плоскости с помощью персистентной гомологии заключается в следующем. Сначала соединяются наиболее близкие точки, затем расстояние между точками постепенно увеличивается. К топологическим характеристикам относят компоненты связности и дыры. Изначально каждая отдельная точка представляет собой компоненту связности, далее они объединяются в более крупные компоненты связности.

Наибольший интерес представляет информация о дырах, которые формируются при соединении точек между собой и образуют замкнутый контур. Если все точки контура будут соединены между собой, то дыра исчезает. Персистентная гомология фиксирует начало и окончание существования каждой дыры, а также систематизирует эту топологическую информацию в виде баркода.

Описанный принцип построения баркодов придает топологическим признакам устойчивость к изменению

объекта. В качестве множества точек можно использовать ключевые точки, выделенные на изображении, как, например, показано в работе [21].

Таким образом, сравнив два объекта разной целостности, можно узнать степень их отличия.

В настоящей работе рассмотрен каждый пиксел на изображении. Для формирования признаков постепенно будет увеличиваться не расстояние между объектами, а яркость, которая для изображений в градациях серого изменяется от 0 до 255. Для цветных изображений необходимо анализировать три составляющих: R (Red), G (Green) и B (Blue), каждая из которых имеет диапазон изменения яркостей от 0 до 255.

Сначала анализируется бинарное изображение с порогом 0, затем с порогом 1, 2, …, 255. На каждом бинарном изображении вычисляются топологические особенности.

Под компонентой связности считается любой связанный между собой набор пикселов. Для образования компоненты связности, как и в случае с множеством точек, достаточно всего одного пиксела. Компоненты могут объединяться. В итоге после анализа всех бинарных изображений останется лишь одна компонента связности.

Определение дыр сложнее, так как они начинают существовать, если три рядом стоящие точки соединяются в треугольник. Один несвязанный ни с кем пиксел или два рядом стоящих пиксела считаются потенциальными дырами. Если добавляется пиксел, находящийся рядом с дырой, то он присоединяется к ней. Если же дыра появляется у границы изображения, то она исчезает. Таким образом, не останется ни одной дыры.

Под началом существования дыры понимается яркость пиксела, после добавления которого образуется дыра, а конец существования — яркость пиксела, после добавления которого дыра исчезает.

На рис. 1 показан пример процесса формирования двух дыр при изменении яркости от максимального пиксела к минимальному. Также продемонстрировано поглощение одной дыры другой (рис. 1, *e*).

Первая дыра образована при яркости 5 (рис. 1, *b*), на котором также отдельно выделены две потенциальные дыры. Первая дыра поглощает вторую при яркости 2, тем самым вторая дыра заканчивает свое существование. В итоге топологическая информация о дырах содержится в баркоде, который для приведенного примера состоит из двух отрезков и показывает сведения о появлении и исчезновении каждой дыры (рис. 1, *g*).

Отметим, что на практике компоненты связности являются более слабой характеристикой, чем дыры. Это следует из того, что у всех баркодов, построенных по компонентам связности, всегда будет один большой отрезок, протяженностью во весь баркод.

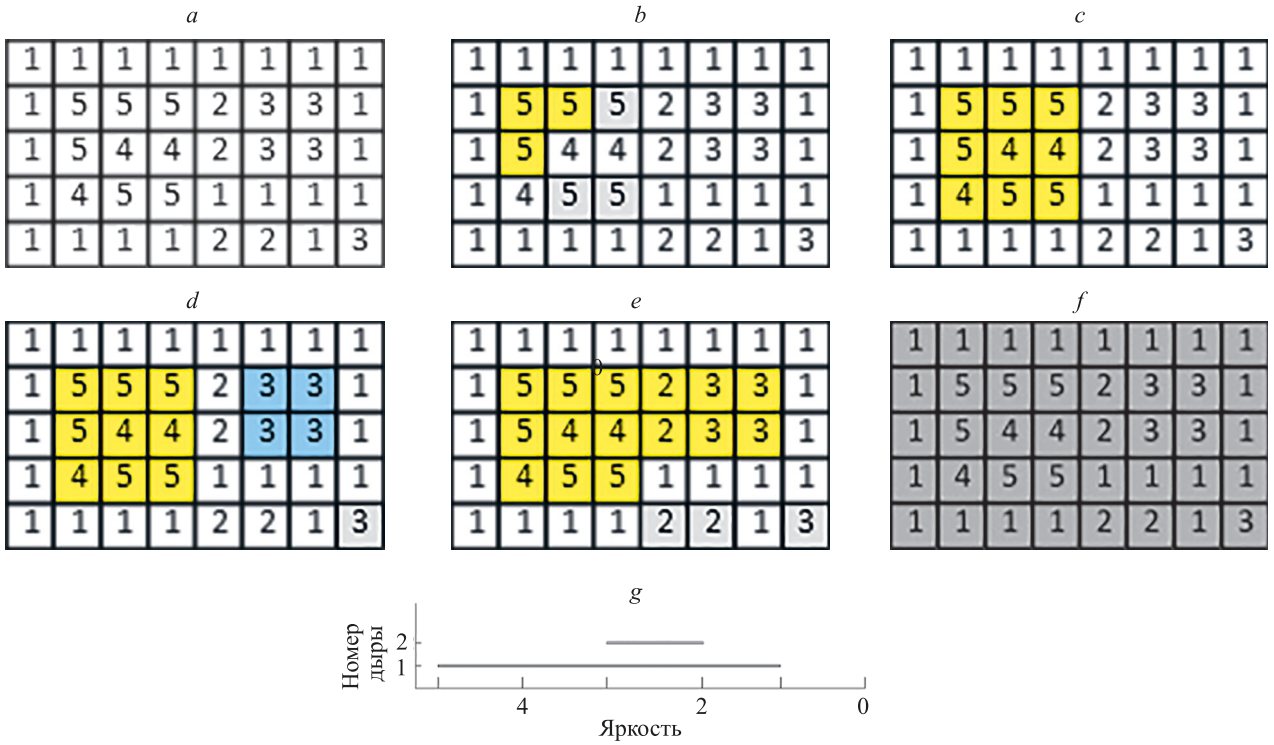


Рис. 1. Процесс формирования дыр: исходное изображение (*a*); выделение всех пикселов, которые больше или равны: 5 (*b*), 4 (*c*), 3 (*d*), 2 (*e*), 1 (*f*); систематизация информации о дырах в виде баркода (*g*).

Серым цветом обозначены потенциальные дыры, желтым и синим — сформированные дыры, темно-серым на рис. 1, *f* — исчезнувшая дыра

Fig. 1. Holes formation process: the source image (*a*); selection of all pixels that are greater than or equal to: 5 (*b*), 4 (*c*), 3 (*d*), 2 (*e*), 1 (*f*); systematization of information about holes is shown in the barcode (*g*).

Potential holes are marked in gray, formed holes are marked in yellow and blue, disappeared holes are marked in dark gray

Предлагаемый подход

В работе рассмотрен двухэтапный анализ изображения для извлечения топологических признаков в виде информации о дырах. Схема предлагаемого подхода показана на рис. 2.

Пусть $I = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}$ — исходное изображение в градациях серого размером $m \times n$.

На первом шаге координаты точек изображения сортируются в порядке возрастания значений их яркостей, т. е. $I(x_{\alpha 1}, y_{\beta 1}) \le I(x_{\alpha 2}, y_{\beta 2}) \le \dots \le I(x_{\alpha m \times n}, y_{\beta m \times n})$, где $\alpha_1, \alpha_2, \dots, \alpha_k, \dots, \alpha_{m \times n}$ и $\beta_1, \beta_2, \dots, \beta_k, \dots, \beta_{m \times n}$ — номера координат точек изображения I по строкам и столбцам соответственно; $I(x_{\alpha k}, y_{\beta k})$ — значение яркости в точке с координатами (α_k, β_k).

На втором шаге формируются топологические признаки по дырам при анализе изображений в диапазоне яркостей от 0 до 255, а затем от 255 до 0. При этом для каждого из двух этапов анализа изображений создается свой баркод, представляющий собой набор отрезков, каждый из которых описывает поведение отдельной дыры. Для первого этапа имеем баркод $B_0^{255} = \{h_1, h_2, \dots, h_s\}$, где s — количество дыр. В свою очередь, $h_i = (b_i, l_i)$, где b_i ($i = 1, 2, \dots, s$) — яркость, на которой появляется дыра; l_i — длина отрезка, который соответствует периоду существования дыры. Для второго этапа при анализе изображения в диапазоне яркостей от 255 до 0 — баркод: $B_{255}^0 = \{h_1', h_2', \dots, h_t'\}$. В результате получим следующий набор топологических признаков: $V = \{h_1, h_2, \dots, h_s, h_1', h_2', \dots, h_t'\}$.

На этапе предобработки баркодов выполним отсев шумов – коротких отрезков на баркоде, экспериментально полученные длины которых составляют 10 значений яркости. Это означает, что те отрезки h_i и h_j', которые имеют длину $l_i < 10$ и $l_j' < 10$ соответственно, отсекаются и не несут существенной информации о структуре объекта на изображении. Это позволяет исключить влияние шума на точность классификации.

Приведем примеры исходных баркодов и после отсеивания шума (рис. 3).

Отфильтрованный набор признаков V поступает на заключительный этап классификации объектов. Сравнение объектов происходит в пространстве топологических признаков. Для соотнесения объекта к определенному классу использован принцип наложения одного баркода на другой [22]. Пусть даны два набора топологических признаков $V_1 = \{h_1, h_2, \dots, h_{s1}, h_1', h_2', \dots, h_{t1}'\}$ и $V_2 = \{g_1, g_2, \dots, g_{s2}, g_1', g_2', \dots, g_{t2}'\}$ таких, что длина каждого отрезка $h \in V_1$ и $g \in V_2$ больше или равна 10. Сначала вычисляется длина общей части при пересечении двух самых больших по длине отрезков из подмножеств $\{h_1, h_2, \dots, h_{s1}\}$ и $\{g_1, g_2, \dots, g_{s2}\}$ с учетом параметра яркости, на которой появилась каждая из дыр. Затем вычисляется пересечение следующих по максимальной длине двух дыр и т. д. Тем самым определяется вклад в общую схожесть каждой пары отрезков. Аналогичная процедура выполняется для подмножеств $\{h_1', h_2', \dots, h_{t1}'\}$ и $\{g_1', g_2', \dots, g_{t2}'\}$. Чем больше суммарная площадь пересечения, тем больше схожи баркоды и, значит, больше схожи объекты между собой.

Для каждого класса используется набор из эталонных объектов. Для них также вычисляется набор топологических признаков. Проверка на схожесть баркодов происходит с каждым эталоном. После процедуры сравнения выдается результат о принадлежности объекта к определенному классу.

Программный код для классификации объектов на основе двухэтапного топологического анализа разработан на языке C++[1].

Результаты экспериментов и их обсуждение

Проведем ряд экспериментов и сравним результаты классификации предлагаемого подхода с одной из нейронных сетей. Для сравнения точности классификации и устойчивости подходов выполним тестирование на наборе данных DOTA v1.5[2].

DOTA представляет собой набор спутниковых снимков и файл с зонами интереса конкретных объектов. В дополнении к координатам зоны интереса, запись об объекте содержит показатель сложности (0 — не сложно, 1 — сложно).

[1] [Электронный ресурс]. Режим доступа: https://github.com/Noremos/barclassifier (дата обращения: 18.12.2021).

[2] Официальная страница набора данных DOTA [Электронный ресурс]. Режим доступа: https://captain-whu.github.io/DOTA/dataset.html (дата обращения: 18.12.2021).

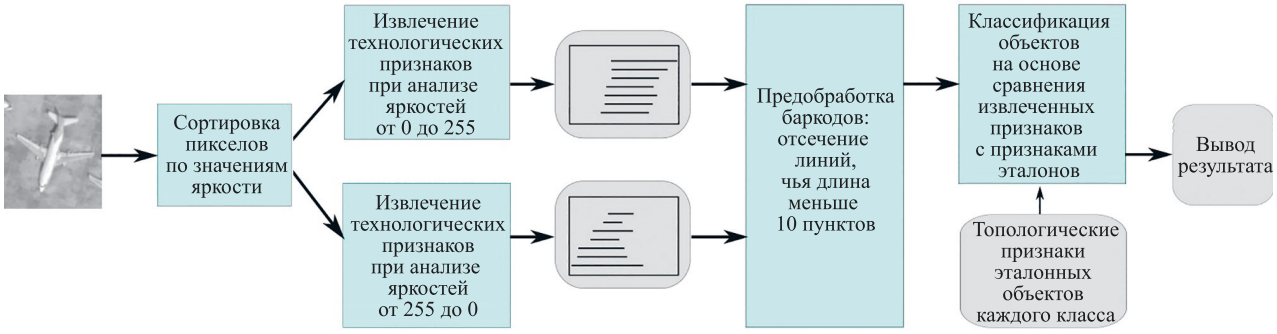


Рис. 2. Схема двухэтапного топологического анализа для классификации объектов на изображении

Fig. 2. Scheme for a two-step topological analysis to classify objects in an image

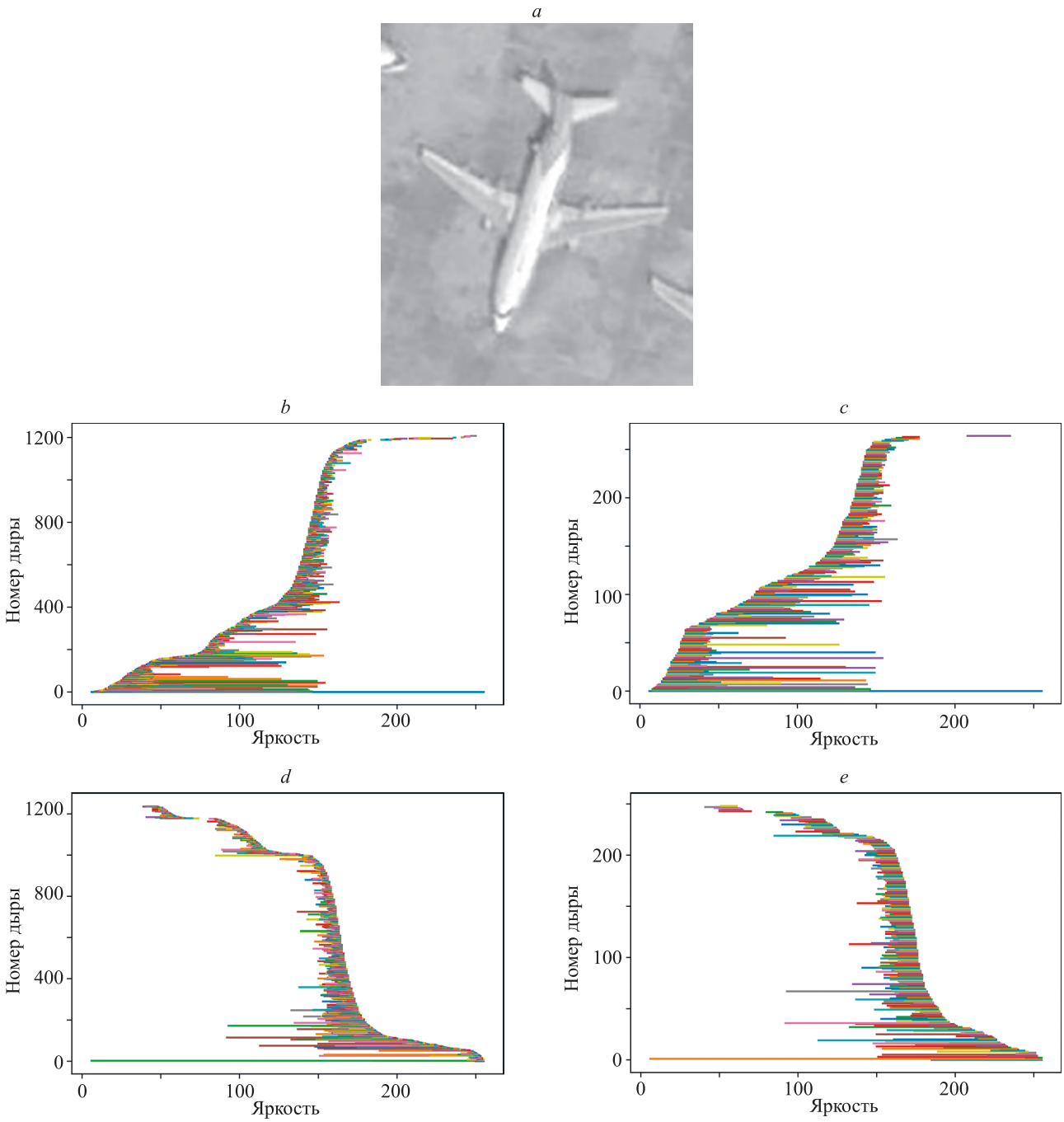


Рис. 3. Топологические признаки изображения в виде баркодов: исходное изображение (*a*); баркод B_0^{255} при анализе изображения в диапазоне яркостей от 0 до 255 с шумом (*b*) и без шума (*c*); баркод B_{255}^0 при анализе изображения в диапазоне яркостей от 255 до 0 с шумом (*d*) и без шума (*e*)

Fig. 3. Topological features of the image in the form of barcodes: the original image (*a*), barcode B_0^{255} after image analysis with brightness from 0 to 255 with noise (*b*) and without noise (*c*), barcode B_{255}^0 after image analysis with brightness from 255 to 0 with noise (*d*) and without noise (*e*)

Главное ограничение при использовании нейронных сетей – необходимость иметь большой набор изображений, нужных для обеспечения максимальной адаптивности. При этом нет гарантии, что нейронная сеть будет корректно работать на новых, отличных от исходного набора данных. Отметим, что чем больше данных было при обучении, тем сложнее будет дообучить сеть.

Сравнение с помощью баркодов работает иначе. Схожесть двух изображений выражается в процентном соотношении. Для использования баркодов в задаче классификации изображений необходимо отобрать ряд эталонов для каждого класса и сравнивать каждый из них с классифицируемым объектом. Причем качество эталонов гораздо важнее количества, поэтому для корректной работы достаточно отобрать несколько наиболее уникальных объектов каждого класса.

По сравнению с новой областью топологического анализа данных, нейронные сети сильно развиты, осо-

бенно в области обработки изображений. Речь идет о CNN. Они способны показывать хорошие результаты при низком количестве и качестве обучаемых данных. Возможно провести сравнение с классической архитектурой нейронной сети, но больший интерес представляет сравнение баркодов с CNN, архитектура которой показана на рис. 4.

Из набора данных для исследования выбраны изображения со следующими классами: самолет, корабль, резервуар для хранения, теннисный корт, баскетбольная площадка, мост, небольшое транспортное средство, вертолет и контейнерный кран. Параметр сложности: 0.

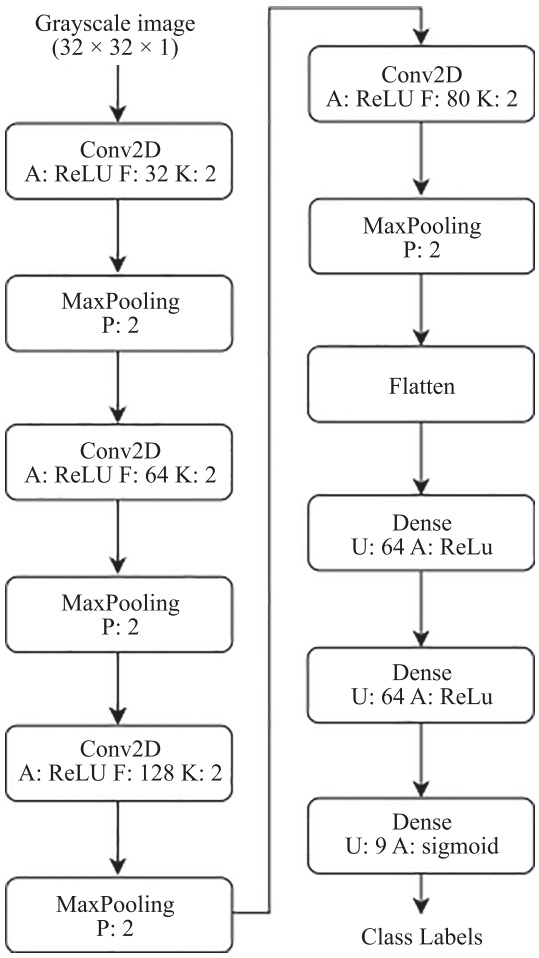


Рис. 4. Архитектура сверточной нейронной сети. Conv2D — слой свертки с параметрами: A (activation function) — функция активации, F (filters) — количество выходных фильтров в свертке, K (kernel size) — ядро окна свертки; Max Pooling — слой подвыборки, где P (pool size) — уменьшает входные данные в P раз; Flatten — преобразует входную матрицу в один массив; Dense — полносвязный слой, где U (units) — размерность выходного пространства; Class Labels — метки

Fig. 4. The architecture of the convolutional neural network. Conv2D is a convolution layer with the following parameters: A is the activation function, F is a number of output filters in the convolution, K is a kernel size; Max Pooling is a subsample layer, where P (pool size) reduces the input data by a factor of P; Flatten is a layer that converts the input matrix into a single array; Dense is a deeply connected layer, where U (units) is the dimensionality of the output space; Class Labels are the output labels

Классификатор для оценки баркодов построен по простому принципу: для каждого класса вычисляется набор баркодов, построенных по каждому из изображений, принадлежащих этому классу. Во время валидации по входному изображению строится баркод, который поочередно сравнивается с каждым баркодом из каждого набора, где выбирается пара с наибольшим процентом схожести.

Помимо обычного набора данных для валидации в работе выполнено исследование устойчивости к искажениям. Главная особенность персистентной гомологии в том, что она в большинстве случаев устойчива именно к искажениям, в то время как нейронная сеть не может распознать искажение объекта, которого не было в исходной выборке.

Во время экспериментов в качестве искажений изображений применены три параметра: поворот — α; сужение — β; обрезание края — γ.

Нейронная сеть обучена на выборке из 87 266 экземпляров. Без внесения искажений тест валидации, состоящий из 26 249 объектов, для нейронной сети дал точность 98,95 %. Базовая точность классификатора по баркодам на наборе для валидации составила около 82 %. Для вычисления точности (Precision) используем формулу:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

где TP и FP — число правильно и неправильно сопоставленных классов.

Заметим, что классификатору по баркодам не нужно много однородных изображений, достаточно поместить уникальные случаи. Однако для чистоты эксперимента этот классификатор принимает ту же обучающую выборку, что и нейронная сеть.

Внесем искажения в изображения, которые используются для теста валидации. Примеры полученных искажений показаны в табл. 1.

Результаты экспериментов сведены в табл. 2. Высокое значение точности нейронной сети может означать наличие переобучения, поэтому была проведена повторная проверка с меньшим числом эпох.

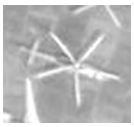
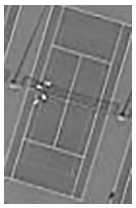
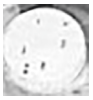
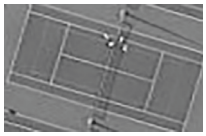
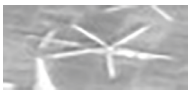
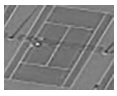
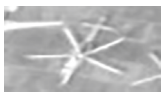
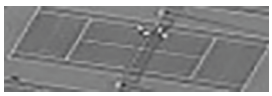
Из полученных результатов видно, что количество эпох практически не влияет на результаты, поэтому подробное сравнение можно опустить. Заметим, что больше всего на нейронные сети влияет обрезание. Из табл. 2 видно, что предложенный метод превосходит нейронную сеть CNN для этого типа искажений.

Приведем результаты, показывающие снижение точности классификации при внесении различных искажений (табл. 3).

Из табл. 3 видно, что при сужении или обрезании края изображения предложенный метод превосходит результаты, чем классификатор на основе нейронной сети, где наихудший результат проявляется при $\gamma = 50$. При добавлении поворота точность у предложенного метода (строки с параметрами 0, 0, 50 и 90, 0, 50 в табл. 2 и 3) несколько возрастает, что может показаться несколько необычным. Однако эта погрешность

Таблица 1. Примеры исходных и искаженных изображений

Table 1. Examples of original and distorted images

Параметры искажения			Визуальное искажение объекта					
α, °	β, %	γ, %	Машина	Вертолет	Теннисный корт	Самолет	Лодка	Резервуар
0	0	0						
90	0	0						
0	50	0						
0	0	50						
90	50	0						
90	0	50						
90	50	50						

возникает из-за изменившегося порядка анализа однотонных пикселов, так как они проверяются с левого верхнего края до правого нижнего построчно. Потому эта погрешность является лишь особенностью реализации.

Подводя итоги результатов экспериментов, можно сказать, что если решается задача, в которой есть вероятность получить на вход данные с непредсказуемыми искажениями, а набор эталонов сильно ограничен, то имеет смысл обратить внимание на классификатор

Таблица 2. Сравнение результатов точности классификации объектов с учетом искажений предложенным методом и с помощью нейронной сети CNN

Table 2. Comparison of the results of classification accuracy object with distortion using the proposed method and the CNN neural network

Параметры искажения			Предложенный метод	Сверточная нейронная сеть (количество эпох обучения)			
α, °	β, %	γ, %		16	10	8	5
0	0	0	0,8193	**0,9895**	0,9885	0,9830	0,9755
90	0	0	0,8069	**0,9169**	0,9139	0,9136	0,9058
0	50	0	0,7790	0,9199	0,9244	0,9111	**0,9252**
0	0	50	**0,7124**	0,5851	0,6282	0,5520	0,5754
90	50	0	0,7818	**0,8654**	0,8633	0,8609	0,8654
90	0	50	**0,7203**	0,5966	0,5870	0,5875	0,5945
90	50	50	**0,7203**	0,5290	0,5364	0,5445	0,5369

Таблица 3. Снижение точности классификации объектов предложенным методом и с помощью нейронной сети CNN при внесении искажений

Table 3. Decreasing the accuracy of object classification using the proposed method and the CNN neural network when introducing distortions

Параметры искажения			Предложенный метод, %	CNN, 16 эпох обучения, %
α, °	β, %	γ, %		
90	0	0	–1,51	–7,34
0	50	0	–4,91	–7,03
0	0	50	–13,05	–40,87
90	50	0	–4,57	–12,54
90	0	50	–12,07	–39,71
90	50	50	–12,07	–46,54

с помощью топологического анализа данных. Если же задача подразумевает анализ и классификацию изображений без искажений, то следует использовать нейросетевые подходы.

Заключение

Разработан метод двухэтапного топологического анализа для классификации объектов на изображениях.

Основу метода составляет персистентная гомология, которая позволяет извлекать топологические признаки и хранить их в виде баркода. Особенность полученных признаков заключается в том, что они устойчивы к искажениям. Проведено исследование работы предложенного метода с нейросетевым подходом. Для этого был выбран набор данных DOTA, который содержит снимки пространственных объектов нескольких классов. Без искажений нейронная сеть показала очень высокие результаты, точность классификации составила свыше 98 %, у предложенного метода около 82 %. Для экспериментов выбраны следующие искажения изображений: поворот на 90°, сужение и обрезание края на 50 %, а также их комбинации. В результате предложенный метод показал свою устойчивость. Особенно резко у нейронной сети теряется точность классификации при обрезании края изображения. В самой сложной комбинации теста снижение точности классификации для нейросетевого подхода составило 46 %, а для предложенного метода 12 %. Тем самым эксперименты подтвердили, что при заданных комбинациях искажений предложенный метод превосходит нейронную сеть CNN. Это достигается за счет того, что признаки, полученные с помощью двухэтапного топологического анализа, сохраняют общую структуру объекта даже при частичной потере исходной информации.

Данный метод целесообразно использовать для классификации пространственных объектов на спутниковых снимках, так как на них часто присутствуют рассмотренные в работе искажения. Также можно применить предложенный метод в системах технического зрения промышленных предприятий для автоматической классификации изделий и в смежных областях.

Исходя из проведенных исследований, можно сказать, что персистентная гомология является перспективной областью для изучения и разработки на ее основе новых методов, в том числе для классификации объектов на изображениях, особенно с учетом искажений. Дальнейшее развитие может открыть новые способы применения топологического анализа и те области, где они наиболее эффективны.

Литература

1. Ярошевич П.В., Богуш Р.П. Алгоритм классификации изображений парковочных мест автостоянки на основе гистограмм ориентированных градиентов и метода опорных векторов // Компьютерная оптика. 2017. Т. 41. № 1. С. 110–117. https://doi.org/10.18287/2412-6179-2017-41-1-110-117
2. Краснов Ф.В., Буторин А.В., Ситников А.Н. Обзор подходов к анализу пространственных изображений высокого разрешения для применения в геофизике // Cloud of Science. 2019. Т. 6. № 1. С. 127–143.
3. Fu J., Rui Y. Advances in deep learning approaches for image tagging // APSIPA Transactions on Signal and Information Processing. 2017. V. 6. P. e11. https://doi.org/10.1017/ATSIP.2017.12
4. Ященко А.В., Беликов А.В., Петерсон М.В., Потапов А.С. Дистилляция нейросетевых моделей для детектирования и описания ключевых точек изображений // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 3. С. 402–409. https://doi.org/10.17586/2226-1494-2020-20-3-402-409
5. Blyumin S., Pogodaev A., Khabibullina E. Graph-structural modeling of some special organizational systems // Proc. of the 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA). 2020. P. 279–283. https://doi.org/10.1109/SUMMA50634.2020.9280724
6. Handrich S. Al-Hamadi A. Localizing body joints from single depth images using geodetic distances and random tree walk // Proc. 24th IEEE International Conference on Image Processing (ICIP). 2017. P. 146–150. https://doi.org/10.1109/ICIP.2017.8296260
7. Karimova L., Terekhov A., Makarenko N., Rybintsev A. Methods of computational topology and discrete Riemannian geometry for the analysis of arid territories // Cogent Engineering. 2020. V. 7. N 1. P. 1808340. https://doi.org/10.1080/23311916.2020.1808340
8. Ojala T., Pietikäinen M., Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. V. 24. N 7. P. 971–987. https://doi.org/10.1109/TPAMI.2002.1017623
9. Мясников В.В. Реконструкция функций и цифровых изображений по их знаковым представлениям // Компьютерная оптика. 2019. Т. 43. № 6. С. 1041–1052. https://doi.org/10.18287/2412-6179-2019-43-6-1041-1052
10. Luo Z., Chen J., Takiguchi T., Ariki Y. Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks // APSIPA Transactions on Signal and Information Processing. 2019. V. 8. P. 1–11. https://doi.org/10.1017/ATSIP.2019.3
11. Ruckdeschel P., Kohl M. General purpose convolution algorithm in S4 classes by means of FFT // Journal of Statistical Software. 2014. V. 59. N 4. P. 1–25. https://doi.org/10.18637/jss.v059.i04
12. Касимов Д.Р. Моделирование рассуждений при поиске и описании объектов на космоснимках // Компьютерная оптика. 2020. Т. 44. № 5. С. 772–781. https://doi.org/10.18287/2412-6179-CO-716
13. Sahu S.K., Pujari A.K., Kagita V.R., Kumar V., Padmanabhan V. GP-SVM: Tree structured multiclass SVM with greedy partitioning // Proc. of the 14th International Conference on Information Technology (ICIT). 2015. P. 142–147. https://doi.org/10.1109/ICIT.2015.24
14. Zhou Q., Lan W., Zhou Y., Mo G. Effectiveness evaluation of anti-bird devices based on random forest algorithm // Proc. of the 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS). 2020. P. 743–748. https://doi.org/10.1109/ICCSS52145.2020.9336891
15. Xu H., Chen Y., Lin R., Kuo J. Understanding convolutional neural networks via discriminant feature analysis // APSIPA Transactions on Signal and Information Processing. 2018. V. 7. P. e20. https://doi.org/10.1017/ATSIP.2018.24
16. Рюмина Е.В., Карпов А.А. Аналитический обзор методов распознавания эмоций по выражениям лица человека // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 2. С. 163–176. https://doi.org/10.17586/2226-1494-2020-20-2-163-176
17. Edelsbrunner H., Letscher H., Zomorodian A. Topological persistence and simplification // Discrete and Computational Geometry. 2002. V. 28. N 4. P. 511–533. https://doi.org/10.1007/s00454-002-2885-2

References

1. Yarashevich P.V., Bohush R.P. Classification algorithm of parking space images based on a histogram of oriented gradients and support vector machines. *Computer Optics*, 2017, vol. 41, no 1, pp. 110–117. https://doi.org/10.18287/2412-6179-2017-41-1-110-117. (in Russian)
2. Krasnov F.V., Butorin A.V., Sitnikov A.N. Review of approaches to the analysis of high-resolution spatial images for Geophysics. *Cloud of Science*, 2019, vol. 6, no. 1, pp. 127–143. (in Russian)
3. Fu J., Rui Y. Advances in deep learning approaches for image tagging. *APSIPA Transactions on Signal and Information Processing*, 2017, vol. 6. pp, e11. https://doi.org/10.1017/ATSIP.2017.12
4. Yashchenko A.V., Belikov A.V., Peterson M.V., Potapov A.S. Distillation of neural network models for detection and description of image key points. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 3, pp. 402–409. (in Russian). https://doi.org/10.17586/2226-1494-2020-20-3-402-409
5. Blyumin S., Pogodaev A., Khabibullina E. Graph-structural modeling of some special organizational systems. *Proc. 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, 2020, pp. 279–283. https://doi.org/10.1109/SUMMA50634.2020.9280724
6. Handrich S. Al-Hamadi A. Localizing body joints from single depth images using geodetic distances and random tree walk. *Proc. 24th IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 146–150. https://doi.org/10.1109/ICIP.2017.8296260
7. Karimova L., Terekhov A., Makarenko N., Rybintsev A. Methods of computational topology and discrete Riemannian geometry for the analysis of arid territories. *Cogent Engineering*, 2020, vol. 7, no. 1, pp. 1808340. https://doi.org/10.1080/23311916.2020.1808340
8. Ojala T., Pietikäinen M., Mäenpää T. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, vol. 24, no. 7, pp. 971–987. https://doi.org/10.1109/TPAMI.2002.1017623
9. Myasnikov VV. Reconstruction of functions and digital images using sign representations. *Computer Optics*, 2019, vol. 43, no. 6, pp. 1041–1052. (in Russian). https://doi.org/10.18287/2412-6179-2019-43-6-1041-1052
10. Luo Z., Chen J., Takiguchi T., Ariki Y. Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks. *APSIPA Transactions on Signal and Information Processing*, 2019, vol. 8, pp. 1–11. https://doi.org/10.1017/ATSIP.2019.3
11. Ruckdeschel P., Kohl M. General Purpose Convolution Algorithm in S4 Classes by Means of FFT. *Journal of Statistical Software*, 2014, vol. 59, no. 4, pp. 1–25. https://doi.org/10.18637/jss.v059.i04
12. Kasimov D.R. Searching and describing objects in satellite images on the basis of modeling reasoning. *Computer Optics*, 2020, vol. 44, no. 5, pp. 772–781. (in Russian). https://doi.org/10.18287/2412-6179-CO-716
13. Sahu S.K., Pujari A.K., Kagita V.R., Kumar V., Padmanabhan V. GP-SVM: Tree structured multiclass SVM with greedy partitioning. *International Conference on Information Technology (ICIT)*, 2015, pp. 142–147. https://doi.org/10.1109/ICIT.2015.24
14. Zhou Q., Lan W., Zhou Y., Mo G. Effectiveness evaluation of anti-bird devices based on random forest algorithm. *Proc. of the 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, 2020, pp. 743–748, https://doi.org/10.1109/ICCSS52145.2020.9336891
15. Xu H., Chen Y., Lin R., Kuo J. Understanding convolutional neural networks via discriminant feature analysis. *APSIPA Transactions on Signal and Information Processing*, 2018, vol. 7, pp. e20. https://doi.org/10.1017/ATSIP.2018.24
16. Ryumina E.V., Karpov A.A. Analytical review of methods for emotion recognition by human face expressions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 2, pp. 163–176. (in Russian). https://doi.org/10.17586/2226-1494-2020-20-2-163-176
17. Edelsbrunner H., Letscher H., Zomorodian A. Topological persistence and simplification. *Discrete and Computational Geometry*, 2002, vol. 28, no. 4, pp. 511–533. https://doi.org/10.1007/s00454-002-2885-2
18. Carlsson E., Carlsson G., de Silva V. An algebraic topological method for feature identification. *International Journal of Computational*

18. Carlsson E., Carlsson G., de Silva V. An algebraic topological method for feature identification // International Journal of Computational Geometry and Applications. 2006. V. 16. N 4. P. 291–314. https://doi.org/10.1142/S021819590600204X
19. Макаренко Н.Г., Уртьев Ф.А., Князева И.С., Малкова Д.Б., Пак И.Т., Каримова Л.М. Распознавание текстур на цифровых изображениях методами вычислительной топологии // Современные проблемы дистанционного зондирования Земли из космоса. 2015. Т. 12. № 1. С. 131–144.
20. Gonzalez-Diaz R., Jimenez M.-J., Medrano B. Spatiotemporal barcodes for image sequence analysis // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. V. 9448. P. 61–70. https://doi.org/10.1007/978-3-319-26145-4_5
21. Еремеев С.В., Андрианов Д.Е., Титов В.С. Алгоритм совмещения пространственных объектов разномасштабных карт на основе топологического анализа данных // Компьютерная оптика. 2019. Т. 43. № 6. С. 1021–1029. https://doi.org/10.18287/2412-6179-2019-43-6-1021-1029
22. Еремеев С.В. Абакумов А.В. Программный комплекс для обнаружения и классификации природных объектов на основе топологического анализа // Программные продукты и системы. 2021. Т. 34. № 1. С. 201–208. https://doi.org/10.15827/0236-235X.133.201-208

Geometry and Applications, 2006, vol. 16, no. 4, pp. 291–314. https://doi.org/10.1142/S021819590600204X
19. Makarenko N.D, Urtiev F.A, Knyazeva I.S, Malkova D., Pak I.T, Karimova L.M. Texture recognition in digital images by computational topology methods. *Sovremennye Problemy Distantsionnogo Zondirovaniya Zemli iz Kosmosa*, 2015, vol. 12, no. 1, pp. 131–144. (in Russian)
20. Gonzalez-Diaz R., Jimenez M.-J., Medrano B. Spatiotemporal barcodes for image sequence analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9448, pp. 61–70. https://doi.org/10.1007/978-3-319-26145-4_5
21. Eremeev S.V., Andrianov D.E., Titov V.S. An algorithm for matching spatial objects of different-scale maps based on topological data analysis. *Computer Optics*, 2019, vol. 43, no. 6, pp. 1021–1029. (in Russian). https://doi.org/10.18287/2412-6179-2019-43-6-1021-1029
22. Eremeev S.V., Abakumov A.V. Software complex for detection and classification of natural objects based on topological analysis. *Software & Systems*, 2021, vol. 34, no. 1, pp. 201–208. (in Russian). https://doi.org/10.15827/0236-235X.133.201-208

Авторы

Еремеев Сергей Владимирович — кандидат технических наук, доцент, доцент, Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, 602252, Российская Федерация, SC 56673892800, https://orcid.org/0000-0001-8482-1479, sv-eremeev@yandex.ru

Абакумов Артём Владимирович — студент, Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, 602252, Российская Федерация, https://orcid.org/0000-0001-5784-7147, artem210966@yandex.ru

Authors

Sergey V. Eremeev — PhD, Associate Professor, Associate Professor, Murom Institute (Branch) of Vladimir State University, Murom, 602252, Russian Federation, SC 56673892800, https://orcid.org/0000-0001-8482-1479, sv-eremeev@yandex.ru

Artyom V. Abakumov — Student, Murom Institute (Branch) of Vladimir State University, Murom, 602252, Russian Federation, https://orcid.org/0000-0001-5784-7147, artem210966@yandex.ru

Статья поступила в редакцию 07.12.2021
Одобрена после рецензирования 20.12.2021
Принята к печати 26.01.2022

Received 07.12.2021
Approved after reviewing 20.12.2021
Accepted 26.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-93-100

Dimensionality reduction of the attributes using fuzzy optimized independent component analysis for a Big Data Intrusion Detection System

Rajan Aswanandini[1]✉, Chandran Deepa[2]

[1] KG College of Arts and Science, Coimbatore, 641035, India
[1,2] Sri Ramakrishna College of Arts and Science, Coimbatore, 641006, India

[1] aswanandini1981@gmail.com✉, https://orcid.org/0000-0002-6968-4653
[2] deepapkd@gmail.com, https://orcid.org/0000-0002-1681-9059

Abstract
Big data cybersecurity has garnered more attraction in recent years with the development of advanced machine learning and deep learning classifiers. These new classifier algorithms have significantly improved Intrusion Detection Systems (IDS). In these classifiers, the performance is positively influenced by high relevant features while less relevant features negatively influence the performance. However, considering all the attributes, especially the high dimensional attributes, increases computational complications. Hence it is essential to diminish the dimensionality of the attributes to improve the classifier performance. To achieve this objective, an efficient dimensionality reduction approach is presented through the development of the Fuzzy Optimized Independent Component Analysis (FOICA) technique. The standard Independent Component Analysis (ICA) is coupled with the fuzzy entropy to transform the high dimension attributes into low dimension attributes and helps in selecting high informative low-dimensional attributes. These selected features are fed to efficient hybrid classifiers namely Hyper-heuristic Support Vector Machines (HH-SVM), Hyper-Heuristic Improved Particle Swarm Optimization based Support Vector Machines (HHIPSO-SVM) and Hyper-Heuristic Firefly Algorithm based Convolutional Neural Networks (HHFA-CNN) to classify the cybersecurity data to identify the intrusions. Experiments are conducted over two cybersecurity datasets and real-time laboratory data whose outcomes specify the supremacy of the suggested IDS model based on FOICA dimensionality reduction.

Keywords
big intrusion data, cybersecurity, intrusion detection system, independent component analysis, dimensionality reduction, hyper-heuristic firefly algorithm, convolutional neural networks, NSL-KDD

For citation: Aswanandini R., Deepa Ch. Dimensionality reduction of the attributes using fuzzy optimized independent component analysis for a Big Data Intrusion Detection System. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 93–100. doi: 10.17586/2226-1494-2022-22-1-93-100

УДК 004.048

Снижение размерности атрибутов с использованием нечетко оптимизированного независимого компонентного анализа для системы обнаружения вторжений в большие данные

Раджан Асуанандини[1]✉, Чандран Дипа[2]

[1] КГ Колледж искусств и науки, Коимбатур, 641035, Индия
[1,2] Колледж искусств и наук Шри Рамакришны, Коимбатур, 641006, Индия

[1] aswanandini1981@gmail.com✉, https://orcid.org/0000-0002-6968-4653
[2] deepapkd@gmail.com, https://orcid.org/0000-0002-1681-9059

Аннотация
Исследования кибербезопасности больших данных в последние годы стали привлекать большое внимание благодаря разработке передовых классификаторов машинного и глубокого обучения. Новые алгоритмы классификаторов значительно улучшили системы обнаружения вторжений. На производительность

© Aswanandini R., Deepa Ch., 2022

классификаторов положительно влияют наиболее релевантные функции, в то время как наличие менее релевантных функций отрицательно влияет на их производительность. Учет всех атрибутов, особенно атрибутов высокой размерности, увеличивает вычислительную сложность. По этой причине, важно уменьшить размерность атрибутов для повышения производительности классификатора. Для достижения этой цели представлен эффективный подход к снижению размерности атрибутов посредством разработки метода нечеткого оптимизированного анализа независимых компонентов (Fuzzy Optimized Independent Component Analysis, FOICA). Стандартный независимый компонентный анализ (Independent Component Analysis, ICA) сочетается с нечеткой энтропией для преобразования атрибутов высокой размерности в атрибуты низкой размерности и помогает в выборе высокоинформативных атрибутов низкой размерности. Выбранные функции передаются в эффективные гибридные классификаторы, а именно в гиперэвристические машины опорных векторов (Hyper-heuristic Support Vector Machines, HH-SVM), гиперэвристические машины опорных векторов с улучшенной оптимизацией роя частиц (Hyper-Heuristic Improved Particle Swarm Optimization based Support Vector Machines, HHIPSO-SVM) и сверточные нейронные сети на основе гиперэвристического алгоритма светлячков (Hyper-Heuristic Firefly Algorithm based Convolutional Neural Networks, HHFA-CNN) для классификации данных кибербезопасности и выявления вторжений. Проведены эксперименты с использованием двух наборов данных о кибербезопасности и лабораторных данных в реальном времени. Полученные результаты подтвердили превосходство предложенной модели Intrusion Detection Systems на основе уменьшения размерности FOICA.

Ключевые слова

большие данные о вторжениях, кибербезопасность, система обнаружения вторжений, независимый компонентный анализ, уменьшение размерности, гиперэвристический алгоритм светлячка, сверточные нейронные сети, NSL-KDD

Ссылка для цитирования: Асуанандини Р., Дипа Ч. Снижение размерности атрибутов с использованием нечетко оптимизированного независимого компонентного анализа для системы обнаружения вторжений в большие данные // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 93–100 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-1-93-100

Introduction

Advancements in designing malicious software programs have led to serious challenges for the improvement of Intrusion Detection Systems (IDS) [1]. Hazardous attacks are progressively refined, and their recognitions are currently one of the challenging areas of research in recent years [2]. In technology-related applications and concerns, data security is a prime requirement, and hence the security systems are incorporated with the possibility of eliminating all the security threats [3].

The common problem about these IDS models is their low detection rate, mainly due to the ineffectiveness of the classifiers to classify the intrusion events. The false alarm rate in the existing methodologies is also considerably beyond the acceptable range, leading to further degradation of detection rates [4]. Machine learning succeeded in improving intrusion detection, which has to be updated [5]. Therefore, machine learning (ML) and deep learning (DL) models were employed to provide high detection accuracy when provided with sufficiently high-quality intrusion data for training [6]. The ML-based IDS are often used for intrusion detection while the DL-based IDS models are preferred for the intrusion data [7]. With multiple applications, the Support Vector Machine (SVM), Artificial Neural Networks (ANN), random forests, Extreme Learning Machines (ELM), [8] and various deep neural networks [9] have been broadly implemented for detection and prevention of intrusion detections over the network.

In this paper, the Fuzzy Optimized Independent Component Analysis (FOICA) technique has been designed by merging the Independent Component Analysis with the fuzzy entropy estimation process for optimal performance. Using this hybrid approach, we reduced the dimensionality of the features and selected the best features for the classifier. Previously, the Hyper-heuristic Support Vector Machines (HH-SVM) was developed to detect intrusion into big data effectively. To overcome the complexity issues in this model, the HH-SVM [10], Hyper-Heuristic Improved Particle Swarm Optimization based Support Vector Machines (HHIPSO-SVM) [11] and Hyper-Heuristic Firefly Algorithm based Convolutional Neural Networks (HHFA-CNN) [12] models were considered to enhance the classification accuracy in the traffic of big data. These three hybrid classifiers are chosen and the FOICA has been applied for improving their performance for intrusion detection. We performed tests over NSL-KDD and ISCX-IDS datasets to evaluate the results of the suggested FOICA technique. The main results of the work are the developed FOICA technique for feature dimensionality reduction by integration of fuzzy logic and ICA method and boosting the training time by dimensionality reduction and diminishing the problem of the hybrid classifiers.

Proposed methodology

The proposed methodology includes the FOICA and the hybrid classifiers. The framework consists of functioning blocks of the pre-processing unit, features extraction and dimensionality reduction using the proposed FOICA technique, and the classification using three classifiers. The intention is to use the FOICA to diminish the dimensionality of the features and improve the performance of three classifiers. The function flow of the proposed framework is shown in Fig. 1.

Pre-processing

Two network intrusion datasets, namely NSL-KDD and ISCX-IDS are employed in this study. The datasets are preprocessed using the normalization technique to scale down or scale up the attribute data or features. The attribute values of the intrusion dataset vary from $-\infty$ to

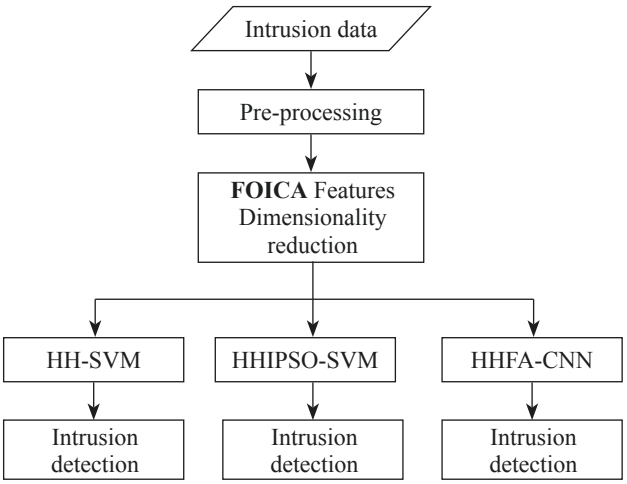


Fig. 1. Functioning of the proposed IDS model using FOICA

+∞, which are not often recognized effectively by the ML classifiers. The ML classifiers require more training time to understand these attribute values in different ranges. Hence it is a must to either scale down or scale up these values to maintain at a specified range such as 0.0 to 1.0, or –1.0 to 1.0, etc. so that they reduce the training time of the classifiers and enable the data classifying process to be performed easily. This process of normalization is also utilized to reduce the overestimation (false positives) and the underestimation (false negatives). In this study, the normalization is implemented to scale down/ up the values of the attributes to regulate their values in the range from 0.0 to 1.0.

FOICA for feature dimensionality reduction

ICA is a multivariate data processing method that is an extended generalization of the popular Principal Component Analysis (PCA) method. The main purpose of the ICA is to decompose the input data into statistically independent components for reducing dimensionality. While PCA projects the data into new solution space, the ICA determines the linear description of the non-Gaussian data as statistically independent components. The ICA can effectively identify the functional relationships and clusters of the data. In this study, fuzzy entropy is used to select the prominent independent components for the classifier and hence reduce the complexity associated with high dimension features.

Initially, FOICA follows the simplest functions of ICA where the attribute data are taken as scalar random variables denoted as $\{x_1, x_2, \ldots, x_n\}$. These variables are assumed as linear combinations of the m unknown independent components denoted as $\{S_1, S_2, \ldots, S_m\}$ which possesses zero mean and are statistically conjoint and self-regulating. When the attribute data is denoted by x_j are rearranged as a Vector $X = (x_1, x_2, \ldots, x_n)^T$, they can be modelled as the linear combination of m random variables denoted as $S = (S_1, S_2, \ldots, S_m)^T$. Appraising these terms into the matrix form leads to $X = AS$, A denoting the mixing matrix formed by the ratio of intensities in X.

To estimate the independent components, the FOICA uses the fixed point iteration scheme that estimates the components faster than the standard PCA and ICA. This fixed point scheme is based on the maximization of negentropy (reverse entropy) which is utilized as the divergence function for approximation. It is denoted as $J(u)$.

$$J(u) = H(u_G) - H(u). \quad (1)$$

Here u_G denotes a Gaussian random vector of the same covariance matrix as u, H denotes the marginal entropy that can be computed as $H(u_i) = -\int p(s_i)\log p(s_i)ds_i$. G denotes any non-quadratic function, u_i and s_i are the independent component and linear combination elements, respectively. $p(.)$ denotes the probability density function.

The mutual information metric can be used as a natural measure to determine independence between two random variables. It is given by

$$MI = J(u) - \sum_i J(u_i). \quad (2)$$

Minimizing this mutual information in FOICA can provide the estimation of the independent components. It is noted that diminishing the mutual information is comparable to maximizing the negentropy. Hence it can be approximated as

$$J_G(u_i) = [E\{G(u_i) - E\{G(y)\}]^2. \quad (3)$$

Here $E\{.\}$ denotes the expectation values and represents a Gaussian variable with zero mean and unit variance. Fuzzy entropy is used as a selection measure for the independent components. The best independent components act as an effective solution for the feature selection and thus improvise the dimensionality reduction. Fuzzy entropy differentiates the pattern distribution better than the ICA alone and appraises the separability of each independent element of the attribute data. Using ICA alone will cause increasing the number of components equal to the number of observed variables that leads to computational complexity in the classification process. This can be overcome by using fuzzy entropy to tune FOICA for optimal dimensionality reduction. The separability property of each component will enhance the classification, and it depends on the fuzzy entropy values. The lower the fuzzy entropy values, the higher the separability of the components. The fuzzy entropy is computed based on the Shannon probabilistic entropy.

$$F(A) = -\sum_{i=1}^{n}(\mu_A(u_i)\log\mu_A(u_i) + (1 - \mu_A(u_i))\log(1 - \mu_A(u_i))). \quad (4)$$

Here $\mu_A(u_i)$ are the fuzzy values of the independent components. This fuzzy entropy is a measure of fuzziness that is used to estimate the global discrimination from the normal fuzzy sets of independent components. The fuzzy set of A with $\mu_A(u_i) = 0.5$ provides the maximum ordering capability, and hence the values are adjusted based on $\mu_A(u_i)$.

In this method, the class ideal vectors representing the class i are created initially as $V_i = (v_1(f_1), v_2(f_2), \ldots, v_i(f_i))$ which can be user-defined. A generalized mean is used to create the class ideal vectors. Then the similarity $S(x, V_i)$ between the x samples and V_i are computed to obtain the j similarities. These similarities are gathered

in a similarity matrix and the fuzzy entropy is computed using Eq. (4) to estimate the similarity between the independent components. The highly similar independent components are selected and used as the vehicle for the selection of the best features to feed the classifiers. In this stage, the forward selection is used to select the relevant features, while backward elimination is used to terminate the irrelevant features until the function termination is obtained. Thus the performance of ICA is improved through fuzzy entropy to optimize the dimensionality reduction and feature selection process.

Algorithm 1: FOICA

Begin
Obtain normalized intrusion data
Initialize attribute data as n scalar random variables
Assign m unknown independent components
Rearrange the attribute data as X with $n >> m$
Estimate the mixing element A using X and S
Set $n = m$
Invert the mixing element A to form U *or* S
Apply the fixed point iteration scheme
Compute the negentropy ($J(u)$) and mutual information (MI) using Eq. (1) & (2)
Maximize MI to approximate independent components using Eq. (3)
Create class ideal vectors using the generalized mean
Estimate similarities between components
Form similarity in $S(x, V_i)$
For each component
    Compute fuzzy entropy using Eq. (4)
    If $F(A)$ is lower, then
        Determine as higher separability
    Else
        Determine as low separability
    End if
End for
Return high similarity independent components
Assign features for components
Apply forward selection & backward elimination
Return best features for classification
End

Classification

After the process of FOICA is complete, the classification is performed using three different classifiers. HH-SVM [13], HHIPSO-SVM [14] and HHFA-CNN are three hybrid classifiers.

— **HH-SVM**: Introduced by Sabar et al. [13], this model is designed by combining the hyper-heuristic optimization with the standard SVM to form hyper-heuristic SVM for the effective intrusion classification. The high-level strategy is to control the selection of low-level heuristics to generate a new optimal SVM configuration.

— **HHIPSO-SVM**: An enhancement of HH-SVM was developed in [14] using a hybrid optimization model of hyper-heuristic particle swarm optimization for tuning the hyper-parameters of SVM to determine its optimal configuration. This hybrid model utilized multiple objectives for tuning the parameters without violating the overall classifier process.

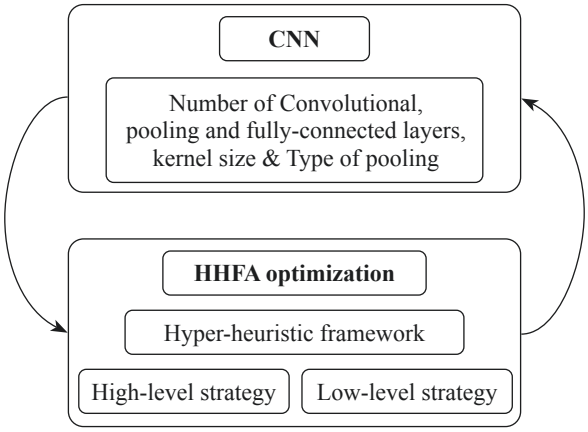


Fig. 2. HHFA-CNN classifier

— **HHFA-CNN**: In this model, a hybrid optimization model of the Hyper-Heuristic Firefly Algorithm is presented to optimize the hyper-parameters of the CNN for enhancing the classifier performance. This model tunes the number of convolution layers, number of pooling layers, pooling type, number of fully connected layers and kernel size to optimize the configuration of CNN. Fig. 2 shows the architecture of HHFA-CNN. The three classifiers are individually utilized to classify the intrusion data to estimate the performance of the proposed FOICA.

Experimental results and discussion

The evaluation of the proposed FOICA model for feature dimensionality reduction is performed using two benchmark instances of cyber security problems, NSL-KDD and ISCX-IDS datasets. NSL-KDD problem instance consists of 311,027 training samples and 77,289 testing samples which are classified as either normal or malicious. ISCX-IDS consists of 208,667 training samples and 78,400 testing samples that are classified as either normal or attack activities are used for this evaluation. Both benchmark instances are available at https://www.unb.ca/cic/datasets/index.html. The experiments are conducted in MATLAB 2016b (version 9.1). The system configurations are given in Table 1.

In addition to the benchmark instances, the real big data traffic instances are also collected using a Local Area Network (LAN) based cloud system through the Wire shark tool with 200 fake source IP addresses in the

Table 1. System configurations

Operating System	Windows 10, 32 bit
Processor	Intel core i5 3470 3.2 GHz
RAM	4 GB DDR3
Storage	500 GB Intel SSD SC2CT060A3 ATA device
Network bandwidth	1 Gbps
Simulation tool	MATLAB v.2016b
Simulation time, s	180

attacker host and 50 trusted sources in the laboratory hosts to simultaneously generate attack and normal traffic. In this analysis, two typical attacks, namely ICMP flooding (Type I) and TCP SYN flooding (Type II), were tested with an observation time of 5 s. The total data samples collected during the testing time included 190,355 samples with 176 legitimate features and one class label. The evaluations are made in the order of classification without dimensionality reduction and with the use of FOICA. Once the best classifier is known, FOICA is compared to existing models with the previously selected best classifier. The performance metrics are accuracy, precision, recall, f-measure and processing time to detect the intrusions.

Performance evaluation over NSL-KDD

The performance of the proposed FOICA is estimated using three classifiers over the NSL-KDD. The performance of the classifiers without FOICA is also evaluated. Table 2 displays the performance comparison of FOICA on three classifiers against the classifiers without dimensionality reduction on the NSL-KDD testing dataset for 25 independent runs.

From Table 2, it can be seen that the performance values of HHFA-CNN, HHIPSO-SVM and HH-SVM are improved by the use of FOICA. When the classifiers are used without FOICA, the resulting accuracy is varies the range of 89 to 97 %. In similar cases, when FOICA is utilized with these classifiers, the accuracy was from 91 to 98 %. Similar results are also obtained for other performance metrics. Among the compared methods, HHFA-CNN with FOICA has achieved high accuracy, precision, recall, and f-measure and minimum processing time. It achieved 1.08 % of high accuracy, 1.96 % of high precision, 8.22 % of high recall, 11.07 % of high f-measure and 0.28 s lesser processing time than the HHFA-CNN model without FOICA.

The performance improvement can be attributed to the fuzzy processing of the features which improved the classifier performance by feeding the vital features and reducing the less relevant ones. The training time is reduced through this deliberated processing in the IDS.

Performance evaluation over ISCX-IDS

Table 3 shows the performance comparison of FOICA on three classifiers against the classifiers without dimensionality reduction on the ISCX-IDS testing dataset for 25 independent runs.

From Table 3, it is seen that the performance of HHFA-CNN, HHIPSO-SVM and HH-SVM are improved by the use of FOICA in ISCX-IDS. All three classifiers with FOICA have achieved better accuracy, precision, precision, recall, and f-measure and minimum processing time than the classifiers without FOICA. The HHFA-CNN classifier with the FOICA model has outperformed the other compared models. It achieved 2.34 % of high accuracy, 0.1 % of high precision, 2.34 % of high recall, 0.78 % of high f-measure and 7.7 s lesser processing time than the HHFA-CNN model without FOICA. This shows the effectiveness of using the FOICA dimensionality reduction approach for intrusion detection.

FOICA has provided this enhanced performance due to their fuzziness in selecting the vital features without any compromise. The transformation of the features through FOICA has also reduced the selection of irrelevant features and the subsequent classifier is trained effectively through this process.

Comparison with existing methods

Since the performance of FOICA is evaluated in Tables 2 and 3, the HHFA-CNN has been the best classifier when using FOICA for dimensionality reduction. Hence in this section, popular algorithms from literature are evaluated and compared with the proposed FOICA with the

Table 2. Performance comparison on NSL-KDD

Algorithm	Metrics, %				Time, s
	Accuracy	Precision	Recall	F-measure	
HH-SVM [13]	89.76	67.10	62.81	62.22	4.65
HHIPSO-SVM [14]	93.33	73.99	64.29	68.37	2.55
HHFA-CNN	96.67	93.94	74	82.79	1.38
HH-SVM + FOICA	91.67	81.52	69.83	89.82	4.02
HHIPSO-SVM + FOICA	94.45	92.03	78.32	92.88	2.16
HHFA-CNN + FOICA	97.75	95.90	82.22	93.86	1.10

Table 3. Performance comparison on ISCX-IDS

Algorithm	Metrics, %				Time, s
	Accuracy	Precision	Recall	F-measure	
HH-SVM [13]	86.60	63.30	60.00	56.19	126.00
HHIPSO-SVM [14]	92.40	69.65	61.10	59.82	49.50
HHFA-CNN	93.33	99.70	93.33	96.55	48.20
HH-SVM + FOICA	88.33	71.87	69.87	79.65	102.00
HHIPSO-SVM + FOICA	94.50	74.78	75.58	82.21	42.88
HHFA-CNN + FOICA	95.67	99.80	95.67	97.33	40.50

Table 4. Performance comparison on NSL-KDD

Algorithm	Metrics, %				Time, s
	Accuracy	Precision	Recall	F-measure	
PCA [15]	86.00	64.50	66.76	78.97	6.30
IG-PCA [16]	97.24	73.56	86.56	81.67	9.40
SCNN [17]	86.64	78.67	82.34	90.23	4.80
CFS-BA [18]	96.11	92.47	90.60	92.30	9.50
PCA-GWO [19]	91.67	90.32	89.18	91.82	7.56
SMO [20]	96.40	93.67	91.33	92.48	3.60
ICA	89.67	71.20	73.50	82.30	3.11
Proposed FOICA	97.75	95.90	82.22	93.86	1.10

Table 5. Performance comparison on laboratory traffic data

Algorithm	Metrics, %				Time, s
	Accuracy	Precision	Recall	F-measure	
HH-SVM [13]	86.67	84.56	86.21	85.36	167.54
HHIPSO-SVM [14]	89.20	87.65	89.98	88.72	155.34
HHFA-CNN	94.56	91.89	93.22	92.50	138.60
HH-SVM + FOICA	91.67	89.67	90.16	89.82	152.50
HHIPSO-SVM + FOICA	93.74	93.66	94.75	94.11	141.67
HHFA-CNN + FOICA	98.89	97.91	98.90	98.25	127.15

HHFA-CNN classifier model acting as the base classifier. Table 4 shows the comparison results of existing models with FOICA over the NSL-KDD data.

The performance results in Table 4 showed that the proposed FOICA has produced better results when used with the HHFA-CNN classifier. It has achieved significantly better performance than the other compared methods in terms of all performance metrics. Although methods like PCA [15], IG-PCA [16], SCNN [17], CFS-BA [18], PCA-GWO [19] and SMO [20] provided comparatively equivalent accuracy, the processing time complexity of these models distances them from the proposed model of FOICA. The utilization of fuzzy entropy to enhance the ICA performance is to be accredited for this performance enhancement. Also, the CNN classifier through deep feature learning has contributed to this effective performance.

Evaluation of laboratory traffic data

Since the laboratory traffic data has been generated over a LAN network setup in real-time, the traffic samples are still raw and extensive preprocessing is required. After the necessary preprocessing, the data is fed to the proposed IDS model. Table 5 displays the performance comparison of FOICA on the three classifiers against the classifiers without dimensionality reduction on the laboratory traffic data.

From Table 5, it can be seen that the use of FOICA has enhanced the performance of the classifiers, as in the case of benchmark data. HHFA-CNN, HHIPSO-SVM and HH-SVM are improved by the use of FOICA with the accuracy increased by about 4–5 %. Similar results are also obtained for other performance metrics. Among the compared methods, HHFA-CNN with FOICA has achieved high accuracy, precision, recall, and f-measure and minimum processing time. It achieved 4.33 % of high accuracy, 6.02 % of high precision, 5.68 % of high recall, 5.75 % of high f-measure and 11.45 s lesser processing time than the HHFA-CNN model without FOICA.

Conclusion

In this paper, we presented an advanced feature dimensionality reduction approach by combining the fuzzy entropy with the standard ICA. This hybrid FOICA model effectively reduced the feature dimension and selected the best features for the classification of big intrusion data. Evaluations were conducted on NSL-KDD and ISCX-IDS datasets and the laboratory data, and FOICA improved the performance of the classifiers. Also, when compared with other existing dimensionality reduction models, the proposed FOICA provided significantly better results. HHFA with the proposed FOICA provided superior performance than the other compared IDS algorithms. In future, other cyber security instances such as UNSW-NB15 will be tested. Moreover, the impact of data sparsity in some data instances will also be studied.

References

1. Liao H.J., Lin C.H.R., Lin Y.C., Tung K.Y. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 2013, vol. 36, no. 1, pp. 16–24. https://doi.org/10.1016/j.jnca.2012.09.004
2. Ashoor A.S., Gore S. Importance of intrusion detection system (IDS). *International Journal of Scientific and Engineering Research*, 2011, vol. 2, no. 1, pp. 1–4.
3. Wang Q., Lu P. Research on application of artificial intelligence in computer network technology. *International Journal of Pattern Recognition and Artificial Intelligence*, 2019, vol. 33, no. 5, pp. 1959015. https://doi.org/10.1142/S0218001419590158
4. Chen T.M., Walsh P.J. Guarding against network intrusions. *Computer and Information Security Handbook*, 2013, pp. 81–95. https://doi.org/10.1016/B978-0-12-394397-2.00005-2
5. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, 2019, vol. 2, no. 1, pp. 20. https://doi.org/10.1186/s42400-019-0038-7
6. Liu H., Lang B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 2019, vol. 9, no. 20, pp. 4396. https://doi.org/10.3390/app9204396
7. Sultana N., Chilamkurti N., Peng W., Alhadad R. Survey on SDN based network intrusion detection system using machine learning approaches. *Peer-to-Peer Networking and Applications*, 2019, vol. 12, no. 2, pp. 493–501. https://doi.org/10.1007/s12083-017-0630-0
8. Sandhu U.A., Haider S., Naseer S., Ateeb O.U. A survey of intrusion detection & prevention techniques. *Proc. of the 2011 International Conference on Information Communication and Management (IPCSIT)*. Vol. 16, 2011, pp. 66–71.
9. Aldweesh A., Derhab A., Emam A.Z. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, 2020, vol. 189, pp. 105124. https://doi.org/10.1016/j.knosys.2019.105124
10. Reddy G.T., Reddy M.P.K., Lakshmanna K., Kaluri R., Rajput D.S., Srivastava G., Baker T. Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 2020, vol. 8, pp. 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942
11. Varma P.R.K., Kumari V.V., Kumar S.S. A survey of feature selection techniques in intrusion detection system: A soft computing perspective. *Advances in Intelligent Systems and Computing*, 2018, vol. 710, pp. 785–793. https://doi.org/10.1007/978-981-10-7871-2_75
12. Almusallam N.Y., Tari Z., Bertok P., Zomaya A.Y. Dimensionality reduction for intrusion detection systems in multi-data streams—A review and proposal of unsupervised feature selection scheme. *Emergence Complexity and Computation*, 2017, vol. 24, pp. 467–487. https://doi.org/10.1007/978-3-319-46376-6_22
13. Sabar N.R., Yi X., Song A. A bi-objective hyper-heuristic support vector machines for big data cyber-security. *IEEE Access*, 2018, vol. 6, pp. 10421–10431. https://doi.org/10.1109/ACCESS.2018.2801792
14. Aswanandini R., Muthumani N. Multi-objective hyper-heuristic improved particle swarm optimization based configuration of support vector machines for big data cyber security. *International Journal of Innovative Technology and Exploring Engineering*, 2019, vol. 8, no. 12, pp. 3892–3897. https://doi.org/10.35940/ijitee.L3401.1081219
15. Vasan K.K., Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science*, 2016, vol. 8, pp. 510–512. https://doi.org/10.1016/j.pisc.2016.05.010
16. Salo F., Nassif A.B., Essex A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks*, 2019, vol. 148, pp. 164–175. https://doi.org/10.1016/j.comnet.2018.11.010
17. Moustakidis S., Karlsson P. A Novel feature extraction methodology using Siamese convolutional neural networks for intrusion detection. *Cybersecurity*, 2020, vol. 3, no. 1, pp. 16. https://doi.org/10.1186/s42400-020-00056-4
18. Zhou Y., Cheng G., Jiang S., Dai M. Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer Networks*, 2020, vol. 174, pp. 107247. https://doi.org/10.1016/j.comnet.2020.107247
19. Swarna Priya R.M., Maddikunta P.K.R., Parimala M., Koppu S., Gadekallu T.R., Chowdhary C.L., Alazab M. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection

Литература

1. Liao H.J., Lin C.H.R., Lin Y.C., Tung K.Y. Intrusion detection system: A comprehensive review // Journal of Network and Computer Applications. 2013. V. 36. N 1. P. 16–24. https://doi.org/10.1016/j.jnca.2012.09.004
2. Ashoor A.S., Gore S. Importance of intrusion detection system (IDS) // International Journal of Scientific and Engineering Research. 2011. V. 2. N 1. P. 1–4.
3. Wang Q., Lu P. Research on application of artificial intelligence in computer network technology // International Journal of Pattern Recognition and Artificial Intelligence. 2019. V. 33. N 5. P. 1959015. https://doi.org/10.1142/S0218001419590158
4. Chen T.M., Walsh P.J. Guarding against network intrusions // Computer and Information Security Handbook. 2013. P. 81–95. https://doi.org/10.1016/B978-0-12-394397-2.00005-2
5. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges // Cybersecurity. 2019. V. 2. N 1. P. 20. https://doi.org/10.1186/s42400-019-0038-7
6. Liu H., Lang B. Machine learning and deep learning methods for intrusion detection systems: A survey // Applied Sciences. 2019. V. 9. N 20. P. 4396. https://doi.org/10.3390/app9204396
7. Sultana N., Chilamkurti N., Peng W., Alhadad R. Survey on SDN based network intrusion detection system using machine learning approaches // Peer-to-Peer Networking and Applications. 2019. V. 12. N 2. P. 493–501. https://doi.org/10.1007/s12083-017-0630-0
8. Sandhu U.A., Haider S., Naseer S., Ateeb O.U. A survey of intrusion detection & prevention techniques // Proc. of the 2011 International Conference on Information Communication and Management (IPCSIT). V. 16. 2011. P. 66–71.
9. Aldweesh A., Derhab A., Emam A.Z. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues // Knowledge-Based Systems. 2020. V. 189. P. 105124. https://doi.org/10.1016/j.knosys.2019.105124
10. Reddy G.T., Reddy M.P.K., Lakshmanna K., Kaluri R., Rajput D.S., Srivastava G., Baker T. Analysis of dimensionality reduction techniques on big data // IEEE Access. 2020. V. 8. P. 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942
11. Varma P.R.K., Kumari V.V., Kumar S.S. A survey of feature selection techniques in intrusion detection system: A soft computing perspective // Advances in Intelligent Systems and Computing. 2018. V. 710. P. 785–793. https://doi.org/10.1007/978-981-10-7871-2_75
12. Almusallam N.Y., Tari Z., Bertok P., Zomaya A.Y. Dimensionality reduction for intrusion detection systems in multi-data streams—A review and proposal of unsupervised feature selection scheme // Emergence Complexity and Computation. 2017. V. 24. P. 467–487. https://doi.org/10.1007/978-3-319-46376-6_22
13. Sabar N.R., Yi X., Song A. A bi-objective hyper-heuristic support vector machines for big data cyber-security // IEEE Access. 2018. V. 6. P. 10421–10431. https://doi.org/10.1109/ACCESS.2018.2801792
14. Aswanandini R., Muthumani N. Multi-objective hyper-heuristic improved particle swarm optimization based configuration of support vector machines for big data cyber security // International Journal of Innovative Technology and Exploring Engineering. 2019. V. 8. N 12. P. 3892–3897. https://doi.org/10.35940/ijitee.L3401.1081219
15. Vasan K.K., Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection // Perspectives in Science. 2016. V. 8. P. 510–512. https://doi.org/10.1016/j.pisc.2016.05.010
16. Salo F., Nassif A.B., Essex A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection // Computer Networks. 2019. V. 148. P. 164–175. https://doi.org/10.1016/j.comnet.2018.11.010
17. Moustakidis S., Karlsson P. A Novel feature extraction methodology using Siamese convolutional neural networks for intrusion detection // Cybersecurity. 2020. V. 3. N 1. P. 16. https://doi.org/10.1186/s42400-020-00056-4
18. Zhou Y., Cheng G., Jiang S., Dai M. Building an efficient intrusion detection system based on feature selection and ensemble classifier // Computer Networks. 2020. V. 174. P. 107247. https://doi.org/10.1016/j.comnet.2020.107247
19. Swarna Priya R.M., Maddikunta P.K.R., Parimala M., Koppu S., Gadekallu T.R., Chowdhary C.L., Alazab M. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection

in IoMT architecture. *Computer Communications*, 2020, vol. 160, pp. 139–149. https://doi.org/10.1016/j.comcom.2020.05.048
20. Khare N., Devan P., Chowdhary C.L., Bhattacharya S., Singh G., Singh S., Yoon B. SMO-DNN: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection. *Electronics*, 2020, vol. 9, no. 4, pp. 692. https://doi.org/10.3390/electronics9040692

in IoMT architecture // Computer Communications. 2020. V. 160. P. 139–149. https://doi.org/10.1016/j.comcom.2020.05.048
20. Khare N., Devan P., Chowdhary C.L., Bhattacharya S., Singh G., Singh S., Yoon B. SMO-DNN: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection // Electronics. 2020. V. 9. N 4. P. 692. https://doi.org/10.3390/electronics9040692

Authors

Rajan Aswanandini — PhD, Assistant Professor, KG College of Arts and Science, Coimbatore, 641035, India; PhD Student, Sri Ramakrishna College of Arts and Science, Coimbatore, 641006, India, SC 57211403371, https://orcid.org/0000-0002-6968-4653, aswanandini1981@gmail.com
Chandran Deepa — PhD, Associate Professor, Sri Ramakrishna College of Arts and Science, Coimbatore, 641006, India, SC 56218174800, https://orcid.org/0000-0002-1681-9059, deepapkd@gmail.com

Авторы

Асуанандини Раджан — PhD, доцент, КГ Колледж искусств и науки, Коимбатур, 641035, Индия; докторант, Колледж искусств и наук Шри Рамакришны, Коимбатур, 641006, Индия, SC 57211403371, https://orcid.org/0000-0002-6968-4653, aswanandini1981@gmail.com
Дипа Чандран — PhD, доцент, Колледж искусств и наук Шри Рамакришны, Коимбатур, 641006, Индия, SC 56218174800, https://orcid.org/0000-0002-1681-9059, deepapkd@gmail.com

Received 30.09.2021
Approved after reviewing 25.12.2021
Accepted 29.01.2022

Статья поступила в редакцию 30.09.2021
Одобрена после рецензирования 25.12.2021
Принята к печати 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-101-113

An optimal swift key generation and distribution for QKD

Mallavalli Raghavendra Suma[1]✉, Perumal Madhumathy[2]

[1] Dayananda Sagar College of Engineering, Bengaluru, 560078, India
[2] RV Institute of Technology and Management, Bengaluru, 560076, India

[1] sumamrvp@gmail.com✉, https://orcid.org/0000-0002-2842-0942
[2] Madhumathyp.rvitm@rvei.edu.in, https://orcid.org/0000-0003-2803-3712

Abstract
Secured transmission between users is essential for communication system models. Recently, cryptographic schemes were introduced for secured transmission and secret transmission between cloud users. In a cloud environment, there are many security issues that occur among the cloud users such as, account hacking, data breaches, broken authentication, compromised credentials, and so on. Quantum mechanics has been implemented in cryptography that made it efficient for strong security concerns over outsourced data in a cloud environment. Therefore, the present research focuses on providing excellent security for cloud users utilizing a swift key generation model for QKD cryptography. The Quantum Key Distribution (QKD) is an entirely secure scheme known as Cloud QKDP. Initially, a random bit sequence is generated to synchronize the channel. An eavesdropper will not permit to synchronize parameters between them. In this key reconciliation technique, the random bit sequence is concatenated with the photon polarisation state. BB84 protocol is improved by optimizing its bit size using FireFly Optimization (FFO) at the compatibility state, and in the next state, both transmitter and receiver generate a raw key. Once the key is generated, it is then used for the transmission of messages between cloud users. Furthermore, a Python environment is utilized to execute the proposed architecture, and the accuracy rate of the proposed model attained 98 %, and the error rate is 2 %. This proves the performance of the proposed firefly optimization algorithm based swift key generation model for QKD performs better than previous algorithms.

Keywords
cryptography, quantum mechanics, quantum key distribution (QKD), eavesdropper: BB84 protocol, reconciliation and firefly optimization

For citation: Suma M.R., Madhumathy P. An optimal swift key generation and distribution for QKD. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 101–113. doi: 10.17586/2226-1494-2022-22-1-101-113

УДК 004.056

Оптимальная быстрая генерация и распределение квантовых ключей

Маллавалли Рагхавендра Сума[1]✉, Перумал Мадхумати[2]

[1] Инженерный колледж Даянанда Сагар, Бангалор, 560078, Индия
[2] RV Институт технологий и менеджмента, Бангалор, 560076, Индия

[1] sumamrvp@gmail.com✉, https://orcid.org/0000-0002-2842-0942
[2] Madhumathyp.rvitm@rvei.edu.in, https://orcid.org/0000-0003-2803-3712

Аннотация
Защищенная передача данных между пользователями важна для систем связи. Большое распространение получили криптографические схемы для защищенной и скрытой передач информации в облачной среде между пользователями. В данной среде возникает множество проблем с безопасностью, такие как взлом учетных записей, утечка данных, нарушение аутентификации, скомпрометированные учетные данные и др. Применение принципов квантовой механики в криптографии повысило ее эффективность для решения проблем безопасности данных, передаваемых на аутсорсинг в облачной среде. В работе предложено решение обеспечения повышенной безопасности для пользователей облачных сервисов за счет применения модели быстрой генерации ключей

© Suma M.R., Madhumathy P., 2022

для криптографии Quantum Key Distribution (QKD). Квантовое распределение ключей представляет собой безопасную схему, известную как Cloud QKDP. Для синхронизации канала генерируется случайная битовая последовательность при этом злоумышленник не может синхронизировать параметры между каналами. В методе согласования ключей случайная битовая последовательность объединена с состоянием поляризации фотона. Первый протокол квантового распределения ключей BB84 улучшен за счет оптимизации его битового размера с помощью FireFly Optimization в состоянии совместимости. В следующем состоянии передатчик и приемник генерируют необработанный ключ. Далее с помощью ключа выполняется передача сообщений между пользователями облака. Для реализации предложенной архитектуры использована среда Python. Точность представленной модели достигает 98 %, а уровень ошибок не превышает 2 %. Выполненные эксперименты показали, что модель генерации ключей Swift на основе алгоритма оптимизации Firefly для QKD работает эффективнее, чем известные алгоритмы.

Ключевые слова

криптография, квантовая механика, квантовое распределение ключей (QKD), перехватчик: протокол BB84, согласование и оптимизация Firefly

Ссылка для цитирования: Сума М.Р., Мадхумати П. Оптимальная быстрая генерация и распределение квантовых ключей // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 101–113 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-1-101-113

Introduction

Communication, the ability to interact in a secure manner, is an essential feature of humanity. Traditional communications fall on the computational complexity of mathematical models. A major problem that is involved here is the factorization of Prime numbers, and researchers work on it to develop a solution for this issue in less exponential time [1]. Quantum computers are used for solving this issue in polynomial time and made it feasible that encryption depends on hardware difficulty of mathematical tasks [2]. This resulted in turning the focus over physical laws to encode the messages, and a foundation type of key distribution algorithm has been developed previously. By establishing a key between a transmitter and receiver, the communication was made efficient and secure. Security of the protocol is realized only when linear and diagonal methods are implemented and analyzed for transmitting information [3]. In order to ensure security, different algorithms such as Elliptic Curve Cryptography (ECC) [4], Hash function cryptography [5], symmetric or asymmetric type cryptography [6] schemes use selective access control and digital signatures. These schemes were found to have some drawbacks such as being difficult to access by the user, and more methods are added to protect vulnerability. These controls are left only to admins for increasing the computational cost by solving mathematical issues, which can also increase computational power [7].

The requirement of a communication model is based on the idea that a user allows parties to establish keys and track the security of the communication channel. Tracking was done by estimating error and noise values caused during communication [8]. Active hackers were capable to catch transmitted messages, and therefore the only way was to secure the message during transmission. Security of cryptography has been measured as strong or weak with respect to time and resources that is required to recover a plain text. Analyzing a ciphertext obtained by a strong cryptography technique, it might be very difficult to decipher it without an appropriate decoding tool [9] that requires computing power and time, with many computers doing analysis in a second. It is impossible to decipher the result of finite cryptography that will not allow intruders to determine messages [10].

Many researchers have been focused on building the strongest encryption methods with respect to computing power. In a cryptographic algorithm, a cypher is a mathematical function used for the encryption and decryption process. It works with a combination of the key, and the key may be in the format of word, number or phrase for encrypting the plaintext [11]. Plaintext encrypt different ciphertext in different techniques using keys. Security of encrypted data is based on the strength or length of the algorithm and the secrecy of the key. Conventional algorithms such as Rivest-Shamir-Adleman (RSA), ECC and Financial Service Authority (FSA) algorithms use the key for encrypting data by an agreement protocol. The key is shared between the transmitter and receiver, and then the message is conventionally encrypted in the communication channel [12]. The sender and the receiver must keep the secret key within themselves. To prevent the secret key from getting disclosed during transmission, different neural network schemes have been implemented previously. An intruder might be able to intercept with the key and read, modify and collapse the data format at the receiver. An authenticated key gets the most attention in the conventional methods. In order to overcome the drawbacks caused by using key encryption stages, a request on security-based transmission protocol have been raised [13]. As the key size and message size length took over all the memory. However, there remains a public key for user access which is mathematically related to deriving message bits when given enough time and computation power.

Cryptography is the method of acquiring confidential data with the help of mathematical equations. Intruders might find different ways to conquer secrets, and instead, researchers developed powerful techniques to compromise the intruders via secured communication. Security threat pushed over Quantum mechanics on Cryptography. Quantum computation is posed on cryptographic schemes that are based on assumptions like factorization problem and discrete logarithm problem. RSA technique is used for purposes in e-commerce [14]. Some basic physics used in cryptographic security made Quantum Cryptography a promising scheme. In quantum key distribution, and ideally

in other quantity encryption applications, any assaults permitted under quantum mechanics are assumed to have a security result. Quantum Key Distribution (QKD) secures information by sharing secured key to the sender and receiver. It becomes crucial to prevent guessing of the key from an eavesdropper. Proper encryption and decryption are essential to prevent Eve from guessing the key since it requires a lot of effort to overcome the drawbacks of QKD to make it practically applicable. A change in photon might affect the message passing through certain materials, and digital signature authenticates the communication channel at the receiver side.

The digital signature process is composed of three key generations, key verification and signature algorithms. These algorithms are essential for real-time tools to be implemented. QKD requires many features for certifying a channel for communication before transmission. Recent research has shown that QKD is efficient at a speed of 16 bits per second at a distance of 256 km, which demands many repeaters and increases the communication costs. Thus, an accurate communication channel for long-distance that minimized repeaters is required. Exchanging data using a single photon requires a dedicated, high-quality channel, and so the transmission of keys from the Quantum channel to two or more sites is impossible because they contradict Quantum principles. The number of tests required to transmit a photon without absorption and depolarisation increases and consumes much time exponentially with regard to the channel length. Error probabilities require a measurement scheme unless this execution takes a long time and power. The physical application of the protocol is a major step towards verifying a communication system. This will include the analysis of state preparation uncertainties, which will result in additional errors. Thus, these drawbacks in the QKD scheme motivated to design an optimal solution to measure the error between a sender and receiver, thus enhancing the channel efficiency for secured communication.

The main contributes of the proposed model are given below.

— To attain secure data transmission in cloud computing, an effective optimal swift key generation technique is introduced.
— To achieve improved key generation and distribution techniques for cloud users, a QKD protocol with its quantum mechanism is developed.
— To attain a fast key generation, the model utilizes Firefly Optimization Algorithm (FOA) optimization.
— For selecting the appropriate block size, an optimization scheme of Frequency Fitness Assignment (FFA) is implemented in this model to make it useful for obtaining the best key pair.
— To verify and evaluate the proposed model, the simulation is accompanied to prove the proposed model.

Literature review

Different Cryptographical schemes are presented in various works in order to provide secure communication. Below, we discuss some of them and their drawbacks.

Chan et al. [15] have presented Identity Based Cryptography that have public key certification. It was used in a Mobile Adhoc Network (MANET). This method has been designed particularly for bi directional channels as they rarely exist in MANET. However, this algorithm requires a centralized server for private key issues for various identities and gives a protocol for sharing this task between users. This remains a drawback for using it in MANETs. Jin et al. [16] have investigated pairwise key generation in dynamic wireless networks with a central node and random user arrival. The establishment of a key generation model for this type of network was done depending on the discrete Markov chain to compute the average time per user spent on completing and waiting for key generation. It was able to tackle both parallel and serial key generation scheduling under different conditions. However, it exploited various wireless broadcast characteristics to reduce probing energy was a drawback.

Wang et al. [17] have described multivariate public key cryptosystems that depend on Clipped Hopfield Neural Network and use them in video framework executed at search space. In the matrix field, Diffie-Hellman key exchange algorithm is lengthened into a matrix field that illustrated feasibility unwanted for specific applications for post-quantum cryptography and classic cryptography. However, it required multivariate public key cryptosystems that permit hardware realization in real-time applications. Xu et al. [18] have investigated secret key generation issues for various types of wireless networks by exploring them in physical layer features of wireless channels. Key generation parameters with low complexity have been proposed that combined point-to-point pairwise key generation method, multi-segment method and onetime pad. Specifically, group key generation methods were studied for three types of networks such as three-node, multimode and multimode mesh. However, it has been reduced by the maximum time allocation issue and reformulated into series programming.

Howe et al. [19] have presented discrete Gaussian Samplers for lattice-based cryptography that targeted physical devices. Discrete Gaussian sampler hardware has been used as the main sampling method, and it aims to offer security against side-channel timing attacks using discrete Ziggurat sampler hardware designs. However, due to non-constant time characteristics of rejection, sampling buffers are implemented between each element that permits parallel execution that acquires maximum computation cost of every component.

Dey et al. [20] have presented a lightweight and secure session key establishment scheme for smart home networks and used Diffie Hellman key exchange as an alternative technique. It established a public key between the home gateway and smart device. However, it completely integrated society into its security concerns, and its measurements were noted rather found that it exhibits very contrasting security protocol that was hard to be implemented in real-time applications.

Zhang et al. [21] have symmetrically investigated the secret key capacity of key generation from wireless channels by considering sampling delay impacts from intruder's source model and found that secret key

capacity is predicted by cross-correlation coefficients of channel measurements as it does not require a legitimate channel. By analyzing Doppler spread, pilot length and key generation parameters, it attempts to achieve secret key security. However, if the message data get leaked to eavesdroppers, it might decrease key capacity and also fails to provide a legitimate communication channel to users.

He et al. [22] have presented an in-depth security protocol of ECC based Radio-Frequency Identification (RFID) schemes that predicted some of the security requirements that satisfied authentication schemes. It analyzed the communication cost and security requirements of this protocol and found that it required authentication schemes that satisfy security requirements. However, this method of ECC-based RFID scheme has unacceptable communication cost and computation complexity that was left unused in IoT applications. Furqan et al. [23] have presented a secret key generation algorithm for wireless channels in multicarrier systems for ensuring confidentiality in wireless communication systems. Secret bits are generated randomly from both magnitudes of orthogonal frequency division multiplexing sub channels also from its position with respect to maximum gains. Mismatch rate and key generation rate have been evaluated in this algorithm. However, unfortunately, it extended dimensions for a secret key generation that, in turn, increased the cost of the algorithm.

Almajed et al. [24] have introduced Se-Enc Scheme that was based on ECC that encodes messages effectively to map them in an elliptic curve. It skipped the encoding phase that resisted encryption attacks, and its security analysis illuminated the degree to which it remained secure. It analyzed padding sizes, number of encoding operations and attacks with respect to decoding operations. However, the performance evaluation of this scheme has failed to provide efficient security over transmission.

Problem statement

Quantum computers are extremely complex to design, manufacture and operate. Consequently, they are hobbled by flaws such as noise, malfunctions, and the loss of quantum coherence, which is critical to their operation but breaks down before any non-trivial process is complete. Due to the various random channels linked with these terminals, the key generation among a group of terminals is more complicated. Hence, numerous tree-based techniques have been devised to achieve the multi-terminal pairwise independent network's group secret key capacity. Quantum cryptography can be used to solve this problem. The Quantum Key Distribution Protocol for Secure Cloud Computing, which uses quantum principles to secure cloud storage and data dynamics, was previously introduced with certain disadvantages, including formal key creation that takes years and an agreement foundation. Cloud computing security does not have to be confined by quantum cryptography. During data transfer, there is a high risk of data being intercepted by an eavesdropper. Therefore, research focuses on the development of an improved cryptography technique for secure and secret transmission in a cloud environment.

Furthermore, this research demonstrates the application of quantum computing features, which is a domain that uses photonics to communicate via protocols. This computer technology provides a security feature that is free of all risks and far more powerful than the protection approaches utilized in ubiquitous computing previously. Because data encryption in any device wraps confidential and personal data with a higher level of protection and security, data should be safe from attacks. Cryptographic methods such as ECC, Chaotic secured Cryptography and RSA algorithm used a key to make transmission secure. However, these algorithms have the necessity of sharing a key between a sender and receiver to make the transmission. This remains a way for intruders to attack. The Quantum Key Distribution Method is a cryptography system that leverages unusual features to share sensitive and confidential data. This key is used to encrypt data that will be communicated between concerned parties in an insecure manner.

Proposed FOA-QKD model

FOA based QKD (FOA-QKD) is a secured communication model that uses proper channel synchronization is a major objective of this work. By implementing QKD protocol with its quantum mechanics, it has been acquired to gain its benefits for efficient communication. Message bits are encrypted using encryption operators and converted to *qbits* as per BB84 protocol. This protocol works as per cryptography rules and also performs analysis over the transmission in order to check whether it is appropriate to transfer data with the obtained key. The key is obtained by estimating errors and optimizing them below the threshold. The length of key bits is minimized twice in order to use it as a key for encryption during further transmission. All other values are eliminated from the sender and receiver sides for security purposes.

Quantum cryptographic constructions for FOA-QKD

QKD is constructed using a random photon generator that emits photons in free space where they acquire different polarisation states. The polarisation states are measured using emitter state analysis. In this proposed model, a one-photon source model with radiative cascade has been used, and this is generated using a single-photon source emitter. The ability to produce single photons are described at relevant degrees. Using probabilistic photon sources, photon pairs with spontaneous non-linear optical processes are evaluated. QKD first develops a set of binary values such as 1's and 0's randomly, and every value from these binary digits is chosen from the transmitter for sending data from two polarisation bases. The polarisation bases can be given as $\{|0>, |+>\} = 1$; $\{|1>, |->\} = 0$. The basic structure of the proposed quantum cryptography is show in Fig. 1.

Photons are randomly measured, and their polarisation state is monitored from one of the Deux bases. By defining a key, the receiver uses a traditional communication system to recognize transmitters. Then it analyzes each photon of the transmitter's, which has been measured in polarisation basis, as well as analyzes the base agreement determined

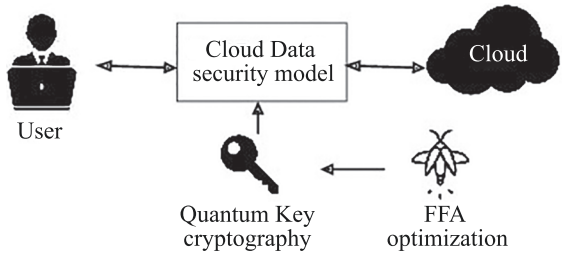


Fig. 1. Quantum Cryptography basic structure

by the protocol in its initial state. The quantitative portion of the QKD protocol is called Alice and Bob encoding and decoding. By assuming that Alice uses *m* different encodings with index from BB84 protocol and each state is denoted as encoded bit values of 0 and 1. In the first step of the protocol, Alice randomly selected *n* bits and *n qbits* prepared in the states to Bob for randomly chosen encoding technique. Each photon contains a quantized amount of energy determined by the photon wavelength. The total energy of 2.5 J is obtained by the addition of many photons.

$$E = hf = h\frac{c}{\eta}. \quad (1)$$

In equation (1), *E* represents the energy of a single-photon, and its wavelength is denoted as η. *c* is the number of photons and *h* gives Planck's constant, *f* is frequency range. By receiving these states that were perturbed received or attacked, Bob applies its measurements to gather classical bits. At last, Alice and Bob employ measurement operations that are compatible. Photons were then measured with the correct benchmark, and the raw key for encryption was generated. This benchmark allows the measurement and communication to establish key and track the security of communication channels.

BB84 protocol

Encode each secret key bit into the single-photon polarisation state. Due to the polarisation states of a single photon, this data was left "fragile" and will be unavailable for eavesdroppers unless this photon was destroyed. Every eavesdropper detects photons and then sends them back to the receiver. But, resendding of this photon, the eve will send them with a wrong polarisation state. This helps to reveal that the eve is present at this communication. Thus, while Alice sends a sequence of pulses that contain single-photon polarised differently, Alice encodes them as Zeroes into *H*-polarised photons and unit as *V*, other half bits are selected randomly and encoded in diagonal polarisation basis, as *D* represents zero and *A* as unity.

State of photon polarisation enhanced in FOA-QKD.

Alice produces an unknown quantic state sequence $\{|x, + | > (k = 1, 2, \ldots, m)\}$. Next, Alice performs a unique transformation in this sequence in all quantum conditions. It is defined as shown in the below equation:

$$\begin{bmatrix} D & H \\ V & A \end{bmatrix}.$$

Apparently, Alice gets a qubit state sequence for unitary information of 1 qubit, then Alice might share the unknown states sequence by following steps.

1. Alice selects and inserts certain particles into the sequence to represent the states of polarisation. Initial state records have been made regarding each decoy particle's position. Alice sends the sequence to Bob, then reports the position of the respective bases and publishes its results.
2. In comparison with the initial states, Alice calculates the probability of an error, and the value of an error is lower than the threshold value. Alice declares that the process is efficient, and this process continues or that the sequencing process is discarded and a new one begins. This process is called an analysis of security.
3. After sequence extraction from the received sequence was performed, the deletion of decoy particles took place, and the sequence took place in a quantum state.
4. Bob randomly picks and inserts some decoy particles in the sequence from the state.

Every decoy particle transmits sequence with respect to its state by recording its initial state and its respective places. Similar to Alice, Bob performs security checking. Comparison between these security analyses results is shown below in Table 1.

This polarisation state, as shown in Table 1, is measured at the receiver side and constructed with its respective volume. Horizontal (*H*), vertical (*V*), diagonal (+45°) (*D*) and anti-diagonal (–45°) (*A*) are the four equal random polarisation states that Alice can select for each bit. On the receiver side, Bob measures polarisation using a standard set up and this way, Bob distinguishes between *H* and *V* polarisation that is used on an *HV* basis. In half of the cases, Bob randomly changes his basis to other random formations. After transmitting a certain number of bits, Bob announces the basis he used for each bit. Then Alice says the cases that are on the same basis. Different bits are then thrown out for optimal key generation.

Key generation

Distributing the photon pairs ensures the highest possible compatibility to the polarisation state. In this case, synchronization between Alice and Bob, knowledge about photons arrival time at nodes and distance between Alice and Bob is essentially known. The strong polarisation correlation between emitted photons and the influence of noise in photons is clearly understandable for key generation. In order to make key generation efficient, Polarisation Dependent Loss (PDL) and Polarisation Mode

Table 1. Polarisation state and qubit assignation

Alice's random bit	0	1	1	0	1	0	0	1	0	1
Alice's random sending basis	×	+	×	+	+	×	+	×	+	+
Photon Polarisation Alice sends	*D*	*H*	*A*	*H*	*V*	*D*	*H*	*A*	*H*	*V*
Qubit generation	↑	→	←	→	↓	↑	→	←	→	↓

Dispersion are assumed to be null. PDL, due to its coherent properties, does not affect key generation as QKD chooses a time bin equal to 1ns and is used for determining the decoy period. This process appears to filter out the majority of background photons and ambient light, but it still considers 97 per cent of photons and does not filter the signal.

Fig. 2 displays the raw key rate, which averaged 135 bits per second during a 13-hour period. The extraction and collecting efficiency of the photons, as well as other concessions made within the experimental setup, limit the bit rate in our experiment. Blinking is another factor, as it reduces the per cent of the time that the quantum dot (QD) is optically active to roughly 0.3 for the QD employed here. Initially, in the key generation phase, the sequence of the polarisation state is shared over the quantum channel, and it results in an array of polarisation states on the receiver side. This state is then used for raw key generation. Raw key generation occurs at both Alice side and Bob's side, error values that occur during this synchronization time are then predicted at the error presumption phase and eliminated assuming that other bits are subjected to the risk state of an eavesdropper.

Error presumption

Quantum mechanics provides actual randomness derived from physical rules. Sending a photon through a 50:50 beam splitter and placing two single-photon detectors on the two outgoing arms is a straightforward technique to make a quantum random number generator. The generated (0 or 1) bit value is determined by which the detector detects a photon. Post-production procedures are used to repair faults in the quantum transmission and erase any leftover information that the eve could have on the raw key,

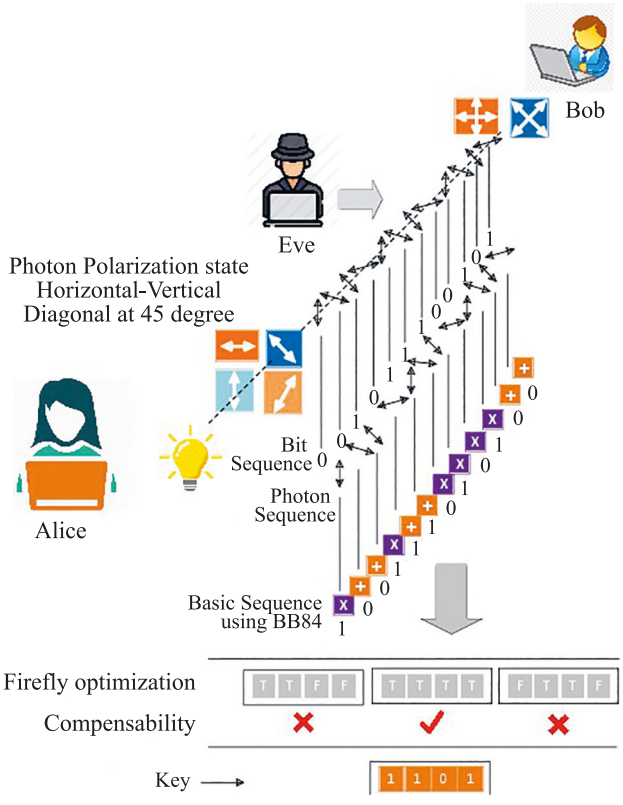


Fig. 2. Key generation architecture

processes including post-selection of data, error correction, and privacy amplification. The end result is a key that Alice and Bob share, but which Eve very probably has no knowledge of. The computational complexity of traditional post-processing techniques, as well as the necessity to process large amounts of raw data in a short amount of time, are currently hurdles in high-speed QKD. As seen in pseudocode 1, a comparison of Alice's raw key and Bob's raw key was performed.

Pseudocode 1. Error Presumption

```
Alice's key = hex1.(into binary)
Bob's Key = hex2.(into binary)
long int i = 0; j= key length;
get (hex1,hex2)
get("%s",hex1);
Out ("\n Equivalent binary value: ");
while (hex1[i])
if
hex1[i]= hex2[i]
Out > True
else
Out> False
```

Error elimination using optimization

Alice and Bob have a secret channel through which Alice sends Bob an *n-bit* string. Following that, Bob receives an *n-bit* string *B* that an eavesdropper with unlimited computing power may obtain by listening to the public channel. Alice and Bob's algorithms define the reconciliation protocol. It runs between strings to establish a public channel for sharing information. The amount of leaked information is the amount of data that an Eavesdropper could gather on *S* with *Q*. To accomplish key reconciliation, Alice and Bob pick a random function from a list of *n* functions with an unknown *m* parameter. This function is shared with the public. On the public channel, Alice transmits this function to Bob. When Bob decodes this string, he gets a string with the shortest distance between all strings. Alice and Bob decide on $k1$ and partition their string into $k1$ bit blocks. In pass 1, the bits in each position are in the form of block *v*. Alice sends Bob parties made out of all of her blocks. Each block whose parity differs from Alice's block is corrected by using Binary Bob. All of Bob's blocks have the even number of mistakes at this point, particularly zero. In this publication, leaked information about the secret string is removed during execution by eliminating one bit from each subset for which parity is known. This information can be saved for each transfer and used to fix errors. Specific block size is chosen so that the probability of a block having one or more errors decreases exponentially as the number of pathways increases. On blocks generated by concatenating a large number of blocks, large enough blocks are employed to completely eliminate errors.

1. Objective function

The objective function is to find the best block size for portioning the message bits for comparison in order to obtain the best key pair. To find a solution that includes both the error value and the optimal key combination, the Firefly algorithm (FA) is employed to

optimize this multi-objective function. The objective function is given as below:

$$F = F_{\min}[F_1]. \quad (2)$$

From equation (2), F_1 represents sub-objective functions and the main objective function is represented by F.

2. Reduction of F_1

The first sub-objective function is represented in the following equation:

$$P(e) = C_k \mathcal{E}(1 - \mathcal{E})^{k-e}. \quad (3)$$

In equation (3), $\mathcal{E}$ denotes error rate, C_k is the current error rate and e is the exact errors. Possibility for an odd number of errors can be given as:

$$P_{odd} = \sum_{i=1}^{\alpha} P(2i - 1) = \frac{1 - (1 - 2\mathcal{E})^k}{2}. \quad (4)$$

There are proportions for the total number of errors in equation (4) and the expectation of the number of errors can be given as

$$\frac{N}{K} P_{odd} = \mathrm{E}\ .$$

The number of errors expected here is as follows:

$$F_1 = \frac{\mathrm{E}}{N\varepsilon} = \frac{1 - (1 - 2\mathcal{E})^k}{2\varepsilon k}. \quad (5)$$

In equation (5), ε is the error rate, K is the block size, N is the key length and errors are assumed to be below the threshold value. The following is explained by the detailed application of the proposed Firefly algorithm.

3. Firefly algorithm

The FA is a search algorithm that uses flashing lights released by a biological process to identify fireflies insects. The intensity of the shining light can be thought of as one of the key enticing communication signals used to attract other fireflies. The process flow of the firefly optimization algorithm is illustrated in Fig. 3. The characteristics of fireflies are summarized here.

— Regardless of their nature, all extant fireflies aim to attract other fireflies.
— The attractiveness of a firefly is proportional to the amount of light it emits and the intensity of that light.
— When two glowing fireflies are present, the one with the less lighting firefly travels toward the one with the better glowing firefly.
— If their intensities are the same, they will move around in a different search space at random.
— The objective function determines the intensity of the firefly light.

The FA technique has two main issues: light intensity discrepancy (I) and attractiveness quantity (β). The light intensity $I(r)$ fluctuates non-linearly (exponentially) with the distance r as follows:

$$I(r) = I_0 e^{-\gamma r}, \quad (6)$$

In equation (6), $I_0 \rightarrow$ initial light intensity; $\gamma \rightarrow$ light absorption coefficient; $\beta \rightarrow$ attractiveness.

The β is given as:

$$\beta = \beta_0 e^{-\gamma r},$$

where attractiveness is given as β_0 at $r = 0$. The displacement of any two fireflies S_i and S_j is given by an empirical formula known as Euclidean distance as follows:

$$r_{ij} = \|S_i - S_j\| = \sqrt{\sum_{k=1}^{k=n} (S_{ik} - S_{jk})^2}.$$

The dimension of the problem is defined by the number 'n'. As a result, an expression is used to characterize the total movement:

$$S_i = S_i + \beta_0 e^{-\gamma r_{ij}^2}(S_i - S_j) + \alpha\varepsilon_i. \quad (7)$$

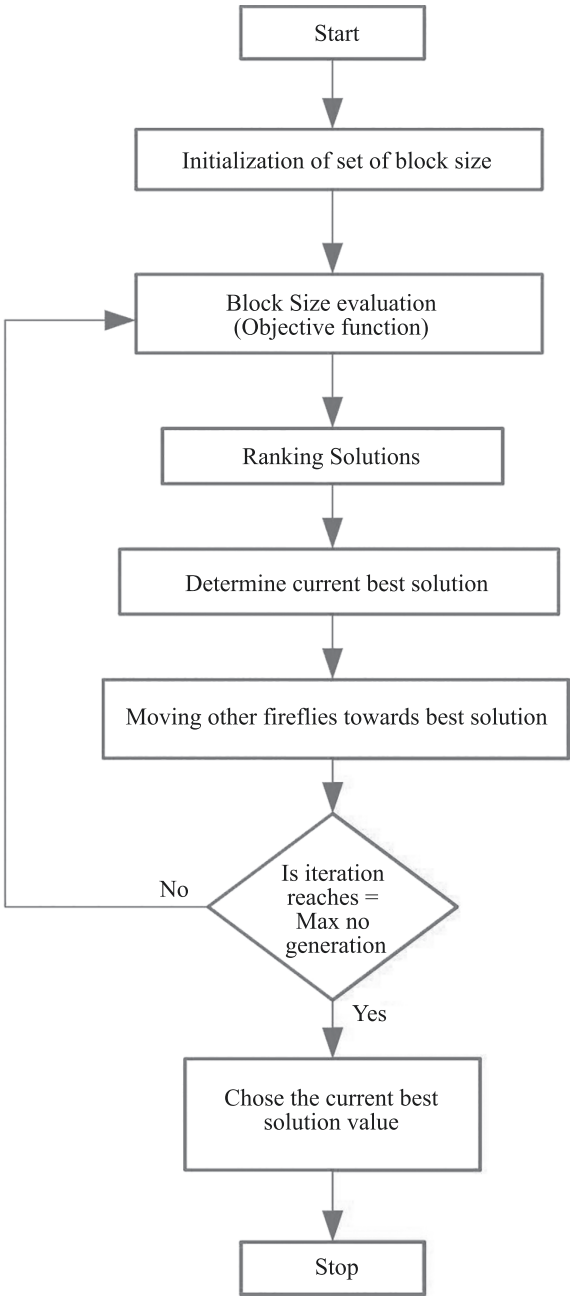


Fig. 3. Flow-chart of the firefly optimization algorithm

From equation (7), γ represents the most important parameter in the firefly algorithm, ε_i is a vector of random numbers being drawn from a Gaussian distribution or uniform distribution and a random walk partial towards the brighter fireflies like if $\beta_0 = 0$. Another more beautiful firefly *j* directs the movement of the i^{th} firefly. Firefly movements can have one of three forms:

The current position in the firefly i^{th}.

Migrating to a more appealing firefly.

Random motion depending on the number and α and εi randomly created by the parameter [0; 1].

Privacy amplification

At this point, Alice and Bob both have the same strings, but they are not private in the entire. There is a chance for the eve to gain some message about them by beam splitting or in the way of intercepting the data. Later one causes an error in Bob's string, and the eve might be able to use a small number of bits and induced errors get eliminated in error caused by physical devices and noise detectors present. Hence, during reconciliation the eve was not be able to get any information. Alice and Bob can use the beam intensity *m* and the bit error rate to calculate the expected fraction of F_1 that Eve has learned. If they are conservative in their assumptions and add several standard deviations to their results, they will have a safe upper bound on the number of bits leaked to Eve. Assuming that Eve knows only deterministic bits, so another issue is whether it might be more useful to her to obtain probabilistic information about *M* instead. Instead of measuring photons keeping the same base as Alice and Bob, a halfway can be picturized out by Eve. It might help her to find matches. Alice, with a maximum probability of about 85 %, is pointless in the case of Bob announcing that his measurement choices were left probabilistic rather than deterministic. The privacy amplification stage is measured using errors caused at Bob's side so as to predict the deterministic bits and thus to informally get the final key.

Proposed FOA-QKD protocol

The proposed model sends a message across a quantum channel, which is then utilized to run a system model to replicate and decode the message. In QKD cryptography, the transmitter is Alice, the receiver is Bob, and the eavesdropper is Eve, which are illustrated in Fig. 4.

— For synchronization, a set of possible parameters is chosen from the codebook, and Alice and Bob produce a set of subset parameters using a pre-shared secret.

— Following the development of a key agreement, Alice and Bob begin synchronizing the transmitter and reception stations in order to begin communication. During the execution of this process, essential system parameters such as channel propagation length are predicted. Synchronization between devices is left independent and used to compare photon states of them.

— With the use of the BB84 protocol, Alice begins key establishment and prepares a bitstream that is indicated as *Set A* with regard to the polarisation states of every photon.

— By establishing a time-bin system, Alice generated a random binary message, $M = \{m_0, m_1, \ldots, m_l\}$, *where* $m_{1\ldots l} = \{0 \text{ or } 1\}$, encodes them using polarisation basis $\{A\}$ and its procedure is given in algorithm 1. The output of this encoding is the set $\{B\}$.

— Then Bob produced a set of measuring bases at random and labelled set *C* for each photon. During computation, the state of each photon is measured and recorded as $\{D\}$.

— This recorded state was then transmitted back to Alice. With the help of $\{A\}$ and $\{C\}$, Alice used certain protocols to estimate values between them, and Alice used these bits to establish the raw key $\{E_{Alice}\}$.

— Bob subsequently selects components of *D* indicated in Alice's communication and creates a raw key, then selects a tiny subgroup from *E* and communicates their states and locations back to Alice.

— Alice forecasts the key bits and relates them to Bob with the aid of this subgroup. Error estimation is performed at this stage by setting up of threshold value. If an error is below the threshold, the communication channel is then established; otherwise, an optimization is performed for the estimating error.

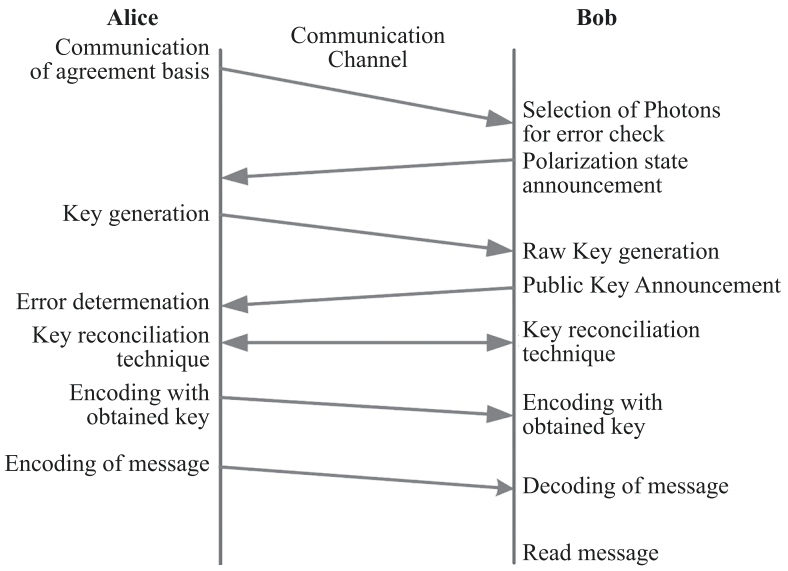


Fig. 4. Communication protocol

— Additionally, a reconciliation technique to correct these errors was performed. A protocol, namely firefly optimization, is established to reduce this error between keys.
— Mayfly optimization presupposes that Alice's key, *EAlice*, and Bob's estimate of Alice's key, *EBob*, each have a source coding. Alice and Bob share *F*, where the comparison was made to check for error.
— Furthermore, the exchange of *F* between Alice and Bob causes inaccuracies that should be corrected. *FAlice* and *FBob* are the only ones left after the comparison. By continuously comparing subdivided blocks of communication, optimization reconciles mistakes between them. With a memory-less channel, a joint probability of *FAlice*, *Bob*=*y* occurs.

$$y = p(FAlice|FBob)p(FBob)$$

— Because *FAlice* and *FBob* are the same lengths, no additional coding bits are required in *F*, and the conditional data entropy between Alice and Bob on message *F* is *H*(*FAlice*|*FBob*), the reconciliation operation is efficient.
— Every iteration takes *F* and is subdivided into the blocks *K*, in which iteration is referred to by the subscript. Alice and Bob compute the copy of their i^{th} message on *n* blocks in every step. Blocks that vary via at least one error are divided continuously in the block between two users and calculated until all blocks are resolved. In the end, key *F* is used for cryptography by both users.
— Alice uses this encrypted solution after setting up a key. This key gain is applied by Bob to a binary string, converted back to a floating-point number to get the encrypted message.
— Following receipt of the encrypted message, Bob uses this key to obtain an appropriate binary string, converts it to a float number, and then obtains the answer.
— Privacy amplification can be achieved by appropriate random data compression or random universal hashing, which reduces the string to a sufficiently shorter one.

Results

Secured key generation and its performance are presented in this paper. Results were obtained by implementing the proposed scheme for selecting the optimal size of the block incompatibility state in order to obtain key values fast and thus improve the secrecy of the channel as far as possible. The proposed scheme is implemented in the Python environment of processor Intel ® Core TM i5-330S CPU @ 2.70GHz and 8GB RAM. Results are compared with previous methods such as C-QKD (Chaotic Quantum Key Distribution), CH-QKD (Chip-based QKD), Full Quantum one way QKD and Post Quantum Cryptography. The analysis involves measuring Accuracy, Error, Key generation rate, and eavesdropping rate. Table 2 illustrates the comparison between the proposed and existing algorithms.

Accuracy

Accuracy is measured with respect to the number of bits encoded and decoded appropriately. It can be described as a number of bits/messages transmitted with the fraction of a number of bits/messages received appropriately. Results are compared with previous methods and shown in Fig. 5.

From Fig. 5, one can see that the proposed model gets a higher accuracy rate of 98 % in the encoding and decoding stage analysis. By sending the number of bits at the encoding process and decoding those using the same technique, the accuracy level was reached at its maximum range. Compared with previous techniques, P-QKD reached 94 %, and C-QKD reached 89 % that was due to its key reconciliation properties. These methods use the traditional channel for transferring messages and thus fail to obtain a higher accuracy range. F-QKD reached 83 % of accuracy value, and CH-QKD reached 72 % of the accuracy range. Thus, the proposed technique has been efficient in terms of accuracy values.

Error

Error-values are measured as false identified bits at the decoder side with respect to encoder values. It can be given as a fraction between original values to the false identified values in the message. The values are measured and plotted in the graph, as shown in Fig. 6.

The graph suggests that the proposed FOA-QKD model exhibits less error value of 2 % due to its improvement in the key reconciliation scheme. P-QKD exhibits 6 %, and

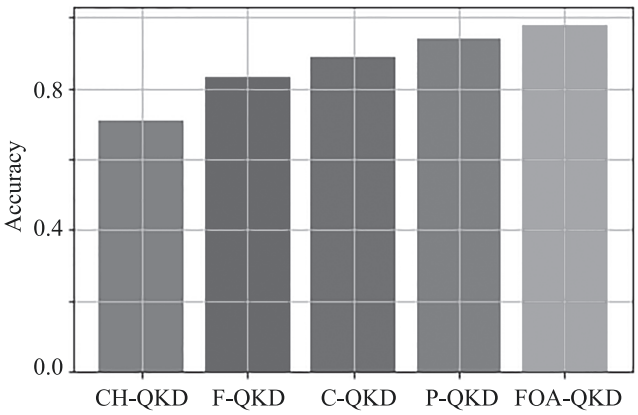


Fig. 5. Accuracy analysis of the proposed and previous methods

Table 2. Comparison between the proposed and existing algorithms

Performance metrics	CH-QKD	F-QKD	C-QKD	P-QKD	FOA-QKD
Accuracy	0.70	0.85	0.89	0.93	0.98
Error	0.28	0.17	0.11	0.06	0.02
Eve's dropping rate	0.81	0.78	0.75	0.68	0.65
Key Generation Time	170	168	167	163	162

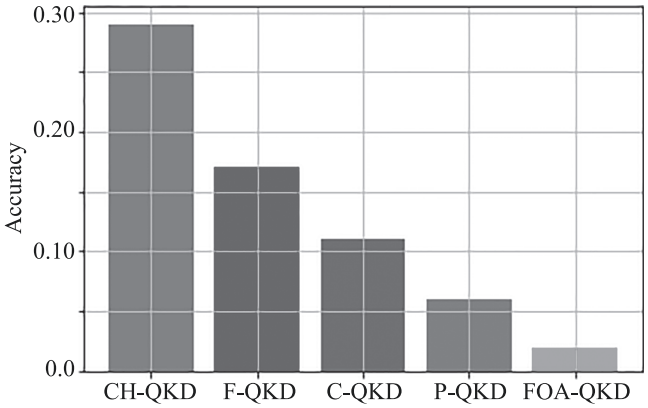


Fig. 6. Error analysis between the proposed and previous models

C-QKD gives 11 % of error. This was due to its traditional approaches over channel synchronization. F-QKD exhibits an error rate of 16 %, and CH-QKD gives an error value of 28 %. The proposed model was found to be error-resistant compared to other QKD techniques. Error-values are minimized due to properly secured transmission between encoding and decoding strategies of the proposed model. Similarly, a higher value of error in CH-QKD is due to its chaotic channel synchronization techniques that failed to minimize the error between information transferred.

Eve's dropping rate

Eve dropping rate was an important concern in the QKD protocol, as they are measured when errors are generated during raw key generation between Alice and Bob with respect to their shared information at the initial stage in order to synchronize channel and transmission rate. It is the ratio of the number of mismatching bits to the total number of extracted bits. To limit eavesdropper's knowledge of the key, both parties discard the last bit of their respective blocks. However, if the parity does not match, they discard the block entirely.

With respect to noise detectors and intruders attack like times attack and Brute-Force attack, eavesdropping rate was measured. From Fig. 7, the proposed model was found to be resistant to the attack and gives a minimum eavesdropping rate due to its enhanced method of announcing the polarisation state between Alice and Bob. FOA-QKD model generates an eavesdropping rate of 0.6 for 2 bits and similarly extends below 0.4 for 8-bits. It reaches 0.5 at a range of 125. Similarly, for previous methods such as P-QKD, C-QKD and CH-QKD, the values range between 0.8 to 0.65 rates and thus evaluated for 2-bit, 4-bit and continuously up to 125-bit.

Key generation time

The raw key was generated as per the proper synchronization of Alice and Bob during key reconciliation time. In the proposed FOA-QKD protocol, key generation time is measured from the time of sending random bit sequences to bob and correspondingly generating key values by eliminating errors between them.

Key generation rate is the measurement of the generation of keys and the fraction of time between them. It was measured for the proposed method as well as for existing methods. Fig. 8, *a* shows that the FOA-QKD model other exhibits very low key generation time for input message bytes of 2 to 125 bytes. Existing methods such as P-QKD exhibits a difference of 0.02 rate values with respect to FOA-QKD. C-QKD and F-QKD need a long key generation time due to their complex protocols for key generation. The rather proposed model used an enhanced format of BB84 protocol for exhibiting key values. CH-QKD protocol then founds its key generation time far greater than all other methods for key generation. In Fig. 8, *b*, the proposed model key generation time was shown. From the figure, it was examined that key generation was found to be linear throughout all key bit size values.

Threshold measurement

The value of the threshold is set as 500 for the generation of photons in order to measure the state of photon polarisation at Bob's side. It remains constant throughout the measurement and, hence, provides a '1' value so that the key generation bit is available with respect to the polarisation state received without error at the receiver side.

From Fig. 9, it is predictable that by setting the threshold value at Bob's side and during synchronization. Synchronization resulted in photons emitted to the detector with appropriate polarisation state at the receiver remaining 1 and others as 0. Values that are categorized as 0 get eliminated as it was the most significant bit and got eliminated simultaneously. Hence, an effective method of setting the threshold was analyzed and presented.

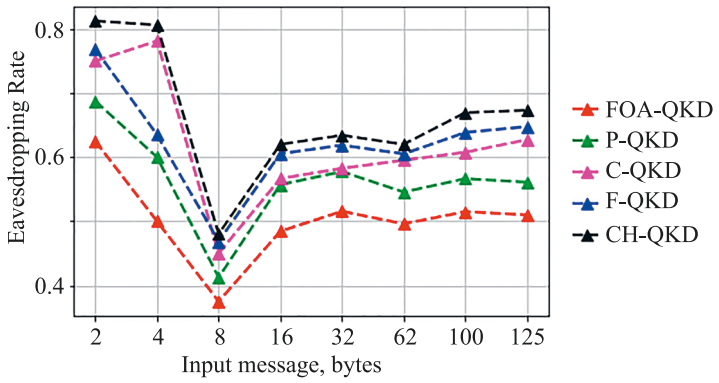


Fig. 7. Eve's dropping rate

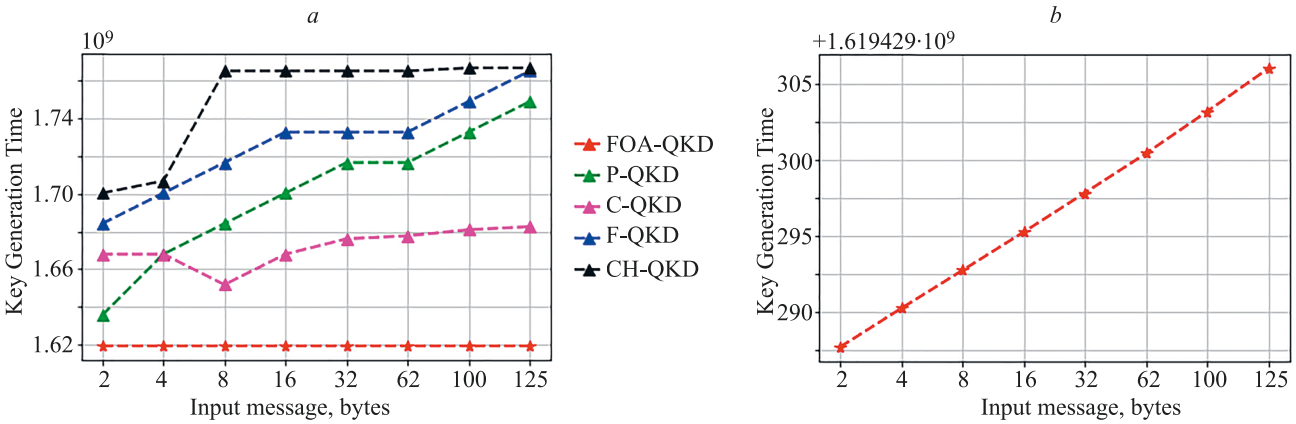


Fig. 8. Key generation rate comparative analysis (*a*); key generation rate for FOA-QKD protocol (*b*)

Table 3. Comparison among the proposed and existing techniques with performance metrics

Algorithm	Encryption time, ms	Decryption time, ms	Packet loss, %	Connection error, %
FOA-QKD	0.200	0.289	0.10	0.20
P-QKD	0.423	0.479	0.27	0.30
C-QKD	0.387	0.400	0.40	0.35
F-QKD	0.520	0.590	0.30	0.53

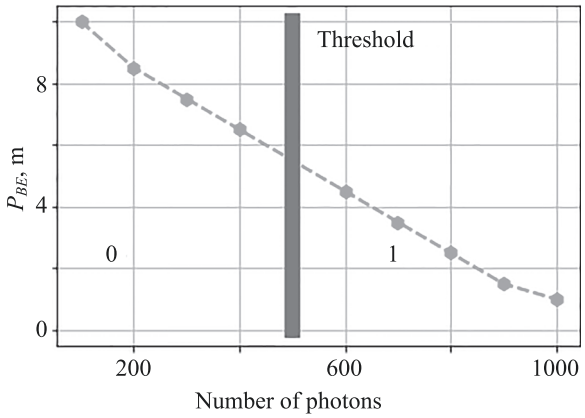


Fig. 9. Threshold rate analysis

Encryption and decryption time

The encryption and decryption times are shown in Table 3. It demonstrates that the proposed model takes less encryption and decryption times than other existing methods. The proposed FOA-QKD have an Encryption time of 0.2 ms and a Decryption time of 2.89 ms. The existing methods such as P-QKD, C-QKD and F-QKD have encryption time of 0.423 ms, 0.387 ms and 0.52 ms, as well as decryption time of 0.479 ms, 0.4 ms and 0.59 ms. The packet loss of the proposed model has 0.1 %, and the connection error is 0.2 %. This proves the proposed model performs better than existing techniques.

Conclusion

QKD protocol with improved key generation and distribution techniques for cloud users have been presented in this paper. Quantum mechanics of photon generation and its polarisation state measurement have been found useful at the end of secret communication model development. This method was widely used in satellite communication for secrecy maintenance. There were certain drawbacks of measuring receiver side bit sequence for generation of the raw key. This key generation was left unwillingly, taking much time and cost than by implementing a number of techniques that improved the complexity and cost of the model. Thus, an optimization scheme of FFA is implemented in this model to make it useful for selecting an appropriate block size that helps a conciliation strategy to achieve a fast key generation and helps a wide range of security with low cost as compared to others. Thus, the proposed model compares and predicts errors, eliminates them with respect to the threshold, generates key values and uses them for further transmission. Results were analyzed and compared with previous techniques, suggesting that the proposed model overcomes all those previous methods and performs efficiently.

References

1. Kiktenko E.O., Malyshev A.O., Gavreev M.A., Bozhedarov A.A., Pozhar N.O., Anufriev M.N., Fedorov A.K. Lightweight authentication for quantum key distribution. *IEEE Transactions on Information Theory*, 2020, vol. 66, no. 10, pp. 6354–6368. https://doi.org/10.1109/TIT.2020.2989459
2. Chitambar E., Fortescue B., Hsieh M.H. The conditional common information in classical and quantum secret key distillation. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 11, pp. 7381–7394. https://doi.org/10.1109/TIT.2018.2851564
3. Ji Z., Yeoh P.L., Zhang D., Chen G., Zhang Y., He Z., Yin H. Secret key generation for intelligent reflecting surface assisted wireless communication networks. *IEEE Transactions on Vehicular Technology*, 2021, vol. 70, no. 1, pp. 1030–1034. https://doi.org/10.1109/TVT.2020.3045728
4. Zhang W.R. From equilibrium-based business intelligence to information conservational quantum-fuzzy cryptography – a cellular transformation of bipolar fuzzy sets to quantum intelligence machinery. *IEEE Transactions on Fuzzy Systems*, 2018, vol. 26, no. 2, pp. 656–669. https://doi.org/10.1109/TFUZZ.2017.2687408
5. Koziel B., Azarderakhsh R., Kermani M.M., Jao D. Post-quantum cryptography on FPGA based on isogenies on elliptic curves. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2017, vol. 64, no. 1, pp. 86–99. https://doi.org/10.1109/TCSI.2016.2611561
6. Yang Y.G., Xu P., Yang R., Zhou Y.H., Shi W.M. Quantum Hash function and its application to privacy amplification in quantum key distribution, pseudo-random number generation and image encryption. *Scientific Reports*, 2016, vol. 6, no. 1, pp. 19788. https://doi.org/10.1038/srep19788
7. Chen Z., Zhou K., Liao Q. Quantum identity authentication scheme of vehicular ad-hoc networks. *International Journal of Theoretical Physics*, 2019, vol. 58, no. 1, pp. 40–57. https://doi.org/10.1007/s10773-018-3908-y
8. Xu F., Curty M., Qi B., Lo H.K. Measurement-device-independent quantum cryptography. *IEEE Journal of Selected Topics in Quantum Electronics*, 2015, vol. 21, no. 3, pp. 148–158. https://doi.org/10.1109/JSTQE.2014.2381460
9. Dong T., Huang T. Neural cryptography based on complex-valued neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, vol. 31, no. 11, pp. 4999-5004. https://doi.org/10.1109/TNNLS.2019.2955165
10. Bai Z., Yang S., Li Y. High-efficiency reconciliation for continuous variable quantum key distribution. *Japanese Journal of Applied Physics*, 2017, vol. 56, no. 4, pp. 044401. https://doi.org/10.7567/JJAP.56.044401
11. Shang T., Chen R., Lei Q. Quantum random oracle model for quantum public-key encryption. *IEEE Access*, 2019, vol. 7, pp. 130024–130031. https://doi.org/10.1109/ACCESS.2019.2940406
12. Zoni D., Galimberti A., Fornaciari W. Efficient and scalable FPGA-oriented design of QC-LDPC bit-flipping decoders for post-quantum cryptography. *IEEE Access*, 2020, vol. 8, pp. 163419–163433. https://doi.org/10.1109/ACCESS.2020.3020262
13. Broadbent A., Schaffner C. Quantum cryptography beyond quantum key distribution. *Designs, Codes, and Cryptography*, 2016, vol. 78, no. 1, pp. 351–382. https://doi.org/10.1007/s10623-015-0157-4
14. Shenoy-Hejamadi A., Pathak A., Radhakrishna S. Quantum cryptography: key distribution and beyond. *Quanta*, 2017, vol. 6, no. 1, pp. 1–47. https://doi.org/10.12743/quanta.v6i1.57
15. Chan A.C. Distributed private key generation for identity based cryptosystems in ad hoc networks. *IEEE Wireless Communications Letters*, 2012, vol. 1, no. 1, pp. 46–48. https://doi.org/10.1109/WCL.2012.120211.110130
16. Jin R., Du X., Zeng K., Huang L., Xiao L., Xu J. Delay analysis of physical-layer key generation in dynamic roadside-to-vehicle networks. *IEEE Transactions on Vehicular Technology*, 2017, vol. 66, no. 3, pp. 2526–2535. https://doi.org/10.1109/TVT.2016.2582853
17. Wang J., Cheng L.M., Su T. Multivariate cryptography based on clipped hopfield neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, vol. 29, no. 2, pp. 353–363. https://doi.org/10.1109/TNNLS.2016.2626466
18. Xu P., Cumanan K., Ding Z., Dai X., Leung K.K. Group secret key generation in wireless networks: algorithms and rate optimization. *IEEE Transactions on Information Forensics and Security*, 2016, vol. 11, no. 8, pp. 1831–1846. https://doi.org/10.1109/TIFS.2016.2553643

Литература

1. Kiktenko E.O., Malyshev A.O., Gavreev M.A., Bozhedarov A.A., Pozhar N.O., Anufriev M.N., Fedorov A.K. Lightweight authentication for quantum key distribution // IEEE Transactions on Information Theory. 2020. V. 66. N 10. P. 6354–6368. https://doi.org/10.1109/TIT.2020.2989459
2. Chitambar E., Fortescue B., Hsieh M.H. The conditional common information in classical and quantum secret key distillation // IEEE Transactions on Information Theory. 2018. V. 64. N 11. P. 7381–7394. https://doi.org/10.1109/TIT.2018.2851564
3. Ji Z., Yeoh P.L., Zhang D., Chen G., Zhang Y., He Z., Yin H. Secret key generation for intelligent reflecting surface assisted wireless communication networks // IEEE Transactions on Vehicular Technology. 2021. V. 70. N 1. P. 1030–1034. https://doi.org/10.1109/TVT.2020.3045728
4. Zhang W.R. From equilibrium-based business intelligence to information conservational quantum-fuzzy cryptography – a cellular transformation of bipolar fuzzy sets to quantum intelligence machinery // IEEE Transactions on Fuzzy Systems. 2018. V. 26. N 2. P. 656–669. https://doi.org/10.1109/TFUZZ.2017.2687408
5. Koziel B., Azarderakhsh R., Kermani M.M., Jao D. Post-quantum cryptography on FPGA based on isogenies on elliptic curves // IEEE Transactions on Circuits and Systems I: Regular Papers. 2017. V. 64. N 1. P. 86–99. https://doi.org/10.1109/TCSI.2016.2611561
6. Yang Y.G., Xu P., Yang R., Zhou Y.H., Shi W.M. Quantum Hash function and its application to privacy amplification in quantum key distribution, pseudo-random number generation and image encryption // Scientific Reports. 2016. V. 6. N 1. P. 19788. https://doi.org/10.1038/srep19788
7. Chen Z., Zhou K., Liao Q. Quantum identity authentication scheme of vehicular ad-hoc networks // International Journal of Theoretical Physics. 2019. V. 58. N 1. P. 40–57. https://doi.org/10.1007/s10773-018-3908-y
8. Xu F., Curty M., Qi B., Lo H.K. Measurement-device-independent quantum cryptography // IEEE Journal of Selected Topics in Quantum Electronics. 2015. V. 21. N 3. P. 148–158. https://doi.org/10.1109/JSTQE.2014.2381460
9. Dong T., Huang T. Neural cryptography based on complex-valued neural network // IEEE Transactions on Neural Networks and Learning Systems. 2020. V. 31. N 11. P. 4999–5004. https://doi.org/10.1109/TNNLS.2019.2955165
10. Bai Z., Yang S., Li Y. High-efficiency reconciliation for continuous variable quantum key distribution // Japanese Journal of Applied Physics. 2017. V. 56. N 4. P. 044401. https://doi.org/10.7567/JJAP.56.044401
11. Shang T., Chen R., Lei Q. Quantum random oracle model for quantum public-key encryption // IEEE Access. 2019. V. 7. P. 130024–130031. https://doi.org/10.1109/ACCESS.2019.2940406
12. Zoni D., Galimberti A., Fornaciari W. Efficient and scalable FPGA-oriented design of QC-LDPC bit-flipping decoders for post-quantum cryptography // IEEE Access. 2020. V. 8. P. 163419–163433. https://doi.org/10.1109/ACCESS.2020.3020262
13. Broadbent A., Schaffner C. Quantum cryptography beyond quantum key distribution // Designs, Codes, and Cryptography. 2016. V. 78. N 1. P. 351–382. https://doi.org/10.1007/s10623-015-0157-4
14. Shenoy-Hejamadi A., Pathak A., Radhakrishna S. Quantum cryptography: key distribution and beyond // Quanta. 2017. V. 6. N 1. P. 1–47. https://doi.org/10.12743/quanta.v6i1.57
15. Chan A.C. Distributed private key generation for identity based cryptosystems in ad hoc networks // IEEE Wireless Communications Letters. 2012. V. 1. N 1. P. 46–48. https://doi.org/10.1109/WCL.2012.120211.110130
16. Jin R., Du X., Zeng K., Huang L., Xiao L., Xu J. Delay analysis of physical-layer key generation in dynamic roadside-to-vehicle networks // IEEE Transactions on Vehicular Technology. 2017. V. 66. N 3. P. 2526–2535. https://doi.org/10.1109/TVT.2016.2582853
17. Wang J., Cheng L.M., Su T. Multivariate cryptography based on clipped hopfield neural network // IEEE Transactions on Neural Networks and Learning Systems. 2018. V. 29. N 2. P. 353–363. https://doi.org/10.1109/TNNLS.2016.2626466
18. Xu P., Cumanan K., Ding Z., Dai X., Leung K.K. Group secret key generation in wireless networks: algorithms and rate optimization // IEEE Transactions on Information Forensics and Security. 2016. V. 11. N 8. P. 1831–1846. https://doi.org/10.1109/TIFS.2016.2553643

19. Howe J., Khalid A., Rafferty C., Regazzoni F., O'Neill M. On practical discrete Gaussian samplers for lattice-based cryptography. *IEEE Transactions on Computers*, 2018, vol. 67, no. 3, pp. 322–334. https://doi.org/10.1109/TC.2016.2642962
20. Dey S., Hossain A. Session-key establishment and authentication in a smart home network using public key cryptography. *IEEE Sensors Letters*, 2019, vol. 3, no. 4, pp. 8667393. https://doi.org/10.1109/LSENS.2019.2905020
21. Zhang J., He B., Duong T.Q. Woods R. On the key generation from correlated wireless channels. IEEE Communications Letters, 2017, vol. 21, no. 4, pp. 961–964. https://doi.org/10.1109/LCOMM.2017.2649496
22. He D., Zeadally S. An analysis of RFID authentication schemes for internet of things in healthcare environment using elliptic curve cryptography. *IEEE Internet of Things Journal*, 2015, vol. 2, no. 1, pp. 72–83. https://doi.org/10.1109/JIOT.2014.2360121
23. Furqan H.M., Hamamreh J.M., Arslan H. New Physical layer key generation dimensions: Subcarrier indices/positions-based key generation. *IEEE Communications Letters*, 2021, vol. 25, no. 1, pp. 59–63. https://doi.org/10.1109/LCOMM.2020.3025262
24. Almajed H.N., Almogren A.S. SE-ENC: A secure and efficient encoding scheme using elliptic curve cryptography. *IEEE Access*, 2019, vol. 7, pp. 175865–175878. https://doi.org/10.1109/ACCESS.2019.2957943

19. Howe J., Khalid A., Rafferty C., Regazzoni F., O'Neill M. On practical discrete Gaussian samplers for lattice-based cryptography // IEEE Transactions on Computers. 2018. V. 67. N 3. P. 322–334. https://doi.org/10.1109/TC.2016.2642962
20. Dey S., Hossain A. Session-key establishment and authentication in a smart home network using public key cryptography // IEEE Sensors Letters. 2019. V. 3. N 4. P. 8667393. https://doi.org/10.1109/LSENS.2019.2905020
21. Zhang J., He B., Duong T.Q. Woods R. On the key generation from correlated wireless channels // IEEE Communications Letters. 2017. V. 21. N 4. P. 961–964. https://doi.org/10.1109/LCOMM.2017.2649496
22. He D., Zeadally S. An analysis of RFID authentication schemes for internet of things in healthcare environment using elliptic curve cryptography // IEEE Internet of Things Journal. 2015. V. 2. N 1. P. 72–83. https://doi.org/10.1109/JIOT.2014.2360121
23. Furqan H.M., Hamamreh J.M., Arslan H. New Physical layer key generation dimensions: Subcarrier indices/positions-based key generation // IEEE Communications Letters. 2021. V. 25. N 1. P. 59–63. https://doi.org/10.1109/LCOMM.2020.3025262
24. Almajed H.N., Almogren A.S. SE-ENC: A secure and efficient encoding scheme using elliptic curve cryptography // IEEE Access. 2019. V. 7. P. 175865–175878. https://doi.org/10.1109/ACCESS.2019.2957943

Authors

Mallavalli Raghavendra Suma — Assistant Professor, Dayananda Sagar College of Engineering, Bengaluru, 560078, India, https://orcid.org/0000-0002-2842-0942, sumamrvp@gmail.com
Perumal Madhumathy — PhD, Associate Professor, RV Institute of Technology and Management, Bengaluru, 560076, India, https://orcid.org/0000-0003-2803-3712, Madhumathyp.rvitm@rvei.edu.in

Авторы

Сума Маллавалли Рагхавендра — доцент, Инженерный колледж Даянанда Сагар, Бангалор, 560078, Индия, https://orcid.org/0000-0002-2842-0942, sumamrvp@gmail.com
Мадхумати Перумал — доктор наук, доцент, RV Институт технологий и менеджмента, Бангалор, 560076, Индия, https://orcid.org/0000-0003-2803-3712, Madhumathyp.rvitm@rvei.edu.in

Received 09.08.2021
Approved after reviewing 11.01.2022
Accepted 31.01.2022

Статья поступила в редакцию 09.08.2021
Одобрена после рецензирования 11.01.2022
Принята к печати 31.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2022 Том 22 Номер 1

doi: 10.17586/2226-1494-2022-22-1-114-119
УДК 004.912

Исследование способов векторизации неструктурируемых текстовых документов на естественном языке по степени их влияния на качество работы различных классификаторов

Виктор Викторович Шадский[1]✉, Александр Борисович Сизоненко[2], Максим Алексеевич Чекмарев[3], Андрей Васильевич Шишков[4], Даниил Андреевич Исакин[5]

[1,2,3,4] Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация
[5] Новосибирский государственный технический университет, Новосибирск, 630087, Российская Федерация

[1] vdvryazan57@yandex.ru✉, https://orcid.org/0000-0002-9221-2283
[2] siz_al@mail.ru, https://orcid.org/0000-0001-8201-9159
[3] max.chek13@gmail.com, https://orcid.org/0000-0002-6832-9991
[4] shishkov-andrey00@mail.ru, https://orcid.org/0000-0002-1841-8750
[5] pm11.isakin@gmail.com, https://orcid.org/0000-0001-7307-6258

Аннотация

Предмет исследования. Повсеместное увеличение объемов обрабатываемой информации на объектах критической информационной инфраструктуры, представленной в текстовой форме на естественном языке, создает проблему ее классификации по степени конфиденциальности. Успех решения данной задачи зависит как от самой модели-классификатора, так и от выбранного способа извлечения признаков (векторизации). Требуется максимально полно передать модели-классификатору свойства исходного текста, содержащие всю совокупность демаркационных признаков. В работе представлена эмпирическая оценка эффективности алгоритмов линейной классификации, основанная на выбранном способе векторизации, а также значении количества настраиваемых параметров в случае применения векторизатора хеширования (Hash Vectorizer). **Метод.** В качестве датасета для обучения и тестирования алгоритмов классификации использованы государственные текстовые документы, условно выступающие в роли конфиденциальных. Выбор подобного текстового массива обусловлен наличием специфической терминологии, повсеместно встречающейся в рассекреченных документах. Терминированность, являясь примитивной демаркационной границей и выступая в роли классификационного признака, облегчает работу алгоритмов классификации, что в свою очередь позволяет сконцентрировать внимание на той доли вклада, которую вносит выбранный способ векторизации. Метрикой оценки качества работы алгоритмов выступает величина ошибки классификации. За величину ошибки принята величина, обратная доле правильных ответов алгоритма (accuracy). Проведена оценка алгоритмов по времени обучения. **Основные результаты.** Полученные гистограммы отражают величину ошибки алгоритмов и время обучения. Выделены наиболее и наименее эффективные алгоритмы для конкретно заданного способа векторизации. **Практическая значимость.** Результаты работы позволяют повысить эффективность решения реальных практических классификационных задач текстовых документов небольшого объема со свойственной специфической терминологией.

Ключевые слова

способ векторизации, TF-IDF, Hash Vectorizer, алгоритм классификации, accuracy

Благодарности

Работа выполнена в Краснодарском высшем военном училище им. С.М. Штеменко в рамках диссертационного исследования в области обработки естественного языка.

Ссылка для цитирования: Шадский В.В., Сизоненко А.Б., Чекмарев М.А., Шишков А.В., Исакин Д.А. Исследование способов векторизации неструктурируемых текстовых документов на естественном языке по степени их влияния на качество работы различных классификаторов // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 114–119. doi: 10.17586/2226-1494-2022-22-1-114-119

© Шадский В.В., Сизоненко А.Б., Чекммарев М.А., Шишков А.В., Исакин Д.А., 2022

A study of vectorization methods for unstructured text documents in natural language according to their influence on the quality of work of various classifiers

Viktor V. Shadsky[1]✉, Alexander B. Sizonenko[2], Maxim A. Chekmarev[3], Andrey V. Shishkov[4], Daniil A. Isakin[5]

[1,2,3,4] Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation
[5] Novosibirsk State Technical University, Novosibirsk, 630087, Russian Federation

[1] vdvryazan57@yandex.ru✉, https://orcid.org/0000-0002-9221-2283
[2] siz_al@mail.ru, https://orcid.org/0000-0001-8201-9159
[3] max.chek13@gmail.com, https://orcid.org/0000-0002-6832-9991
[4] shishkov-andrey00@mail.ru, https://orcid.org/0000-0002-1841-8750
[5] pm11.isakin@gmail.com, https://orcid.org/0000-0001-7307-6258

Abstract
The widespread increase in the volume of processed information at the objects of critical information infrastructure, presented in text form in natural language, causes a problem of its classification by the degree of confidentiality. The success of solving this problem depends both on the classifier model itself and on the chosen method of feature extraction (vectorization). It is required to transfer to the classifier model the properties of the source text containing the entire set of demarcation features as fully as possible. The paper presents an empirical assessment of the effectiveness of linear classification algorithms based on the chosen method of vectorization, as well as the number of configurable parameters in the case of the Hash Vectorizer. State text documents are used as a dataset for training and testing classification algorithms, conditionally acting as confidential. The choice of such a text array is due to the presence of specific terminology found everywhere in declassified documents. Termination, being a primitive demarcation boundary and acting as a classification feature, facilitates the work of classification algorithms, which in turn allows one to focus on the share of the contribution that the chosen method of vectorization makes. The metric for evaluating the quality of algorithms is the magnitude of the classification error. The magnitude of the error is the inverse of the proportion of correct answers of the algorithm (accuracy). The algorithms were evaluated according to the training time. The resulting histograms reflect the magnitude of the error of the algorithms and the training time. The most and least effective algorithms for a given vectorization method are identified. The results of the work make it possible to increase the efficiency of solving real practical classification problems of small-volume text documents characterized by their specific terminology.

Keywords
vectorization method, TF-IDF, Hash Vectorizer, classification algorithm, accuracy

Acknowledgements
The work was carried out at the Krasnodar Higher Military School as part of a dissertation in the field of natural language processing.

For citation: Shadsky V.V., Sizonenko A.B., Chekmarev M.A., Shishkov A.V., Isakin D.A. A study of vectorization methods for unstructured text documents in natural language according to their influence on the quality of work of various classifiers. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 114–119 (in Russian). doi: 10.17586/2226-1494-2022-22-1-114-119

Введение

Увеличение объемов обрабатываемой информации на объектах критической информационной инфраструктуры, в том числе представленной в текстовой форме на естественном языке, непременно влечет за собой целый ряд трудностей, связанных, в первую очередь, с ее обработкой. Одна из таких задач обработки – задача классификации документов [1].

Электронные текстовые документы на естественном языке, независимо от степени конфиденциальности, имеют достаточно широкий спектр классификационных признаков, начиная от тематической составляющей, степени формализации и шаблонизации, и заканчивая тональностью и стилем изложения [2, 3]. Из данного факта следует вывод: для каждой отдельно взятой задачи классификации и соответствующего ей датасета существует детерминированный и не всегда очевидный метод ее решения. Таким образом, перед решением классификационной задачи необходимо провести ее детальный анализ с целью выделения наиболее важных демаркационных признаков и максимального сужения «диапазона» возможных методов ее решения [4].

Например, выберем для составления датасета материалы рассекреченных документов о положении дел в стране[1], выступающие в роли конфиденциальных, и открытых нормативно-правовых актов[2]. Документы представлены в текстовой форме на естественном языке. Отметим, что для более качественной их демаркации необходимо обратить внимание на такой классификационный признак как уровень терминированности [5], который у рассекреченных документов носит ярко выраженный и специфический характер [6, 7].

Для успешного решения задачи классификации указанных документов достаточно использовать классификатор, способный разделить пространство признаков обычной разделяющей гиперплоскостью [8]. В связи с этим в настоящей работе приведены результаты работы простейших алгоритмов классификации [9].

[1] Рассекреченные источники // Исторические материалы [Электронный ресурс]. URL: https://istmat.info/node (дата обращения: 20.12.2021).

[2] Указы и распоряжения Президента Российской Федерации // Кодификация РФ [Электронный ресурс]. URL: https://rulaws.ru/president (дата обращения: 20.12.2021).

Один из ключевых факторов, определяющих успешность решения классификационной задачи — выбор наилучшего способа представления знаний [10], содержащихся в текстовых документах, при котором максимально сохраняются свойства исходного текста и содержащиеся в нем закономерности [9], что в свою очередь отражается на результатах работы моделей-классификаторов.

Выбор способа векторизации

Самые применяемые в настоящее время способы векторизации текста: TF-IDF (Term Frequency-Inverse Document Frequency) и Hash Vectorizer [11]. Выполним сравнение данных способов по величине ошибки, а также времени обучения алгоритмов классификации.

Векторизатор TF-IDF часто используется для создания векторов слов из набора документов. Он масштабирует частоту термов, придавая меньшую значимость наиболее часто встречаемым из них [12]. Однако в таком случае заранее требуется наличие полного словаря, что для большинства практических задач невозможно ввиду периодического его пополнения. Вследствие чего постоянный пересчет векторов TF-IDF крайне неэффективен в вычислительном плане, учитывая при этом широкую применимость векторизатора на больших объемах корпусов[1].

Для словаря с динамически изменяющимся количеством термов более эффективен векторизатор Hash

[1] [Электронный ресурс]. https://habr.com/ru/post/515036/, https://habr.com/ru/post/515084 (дата обращения: 20.12.2021).

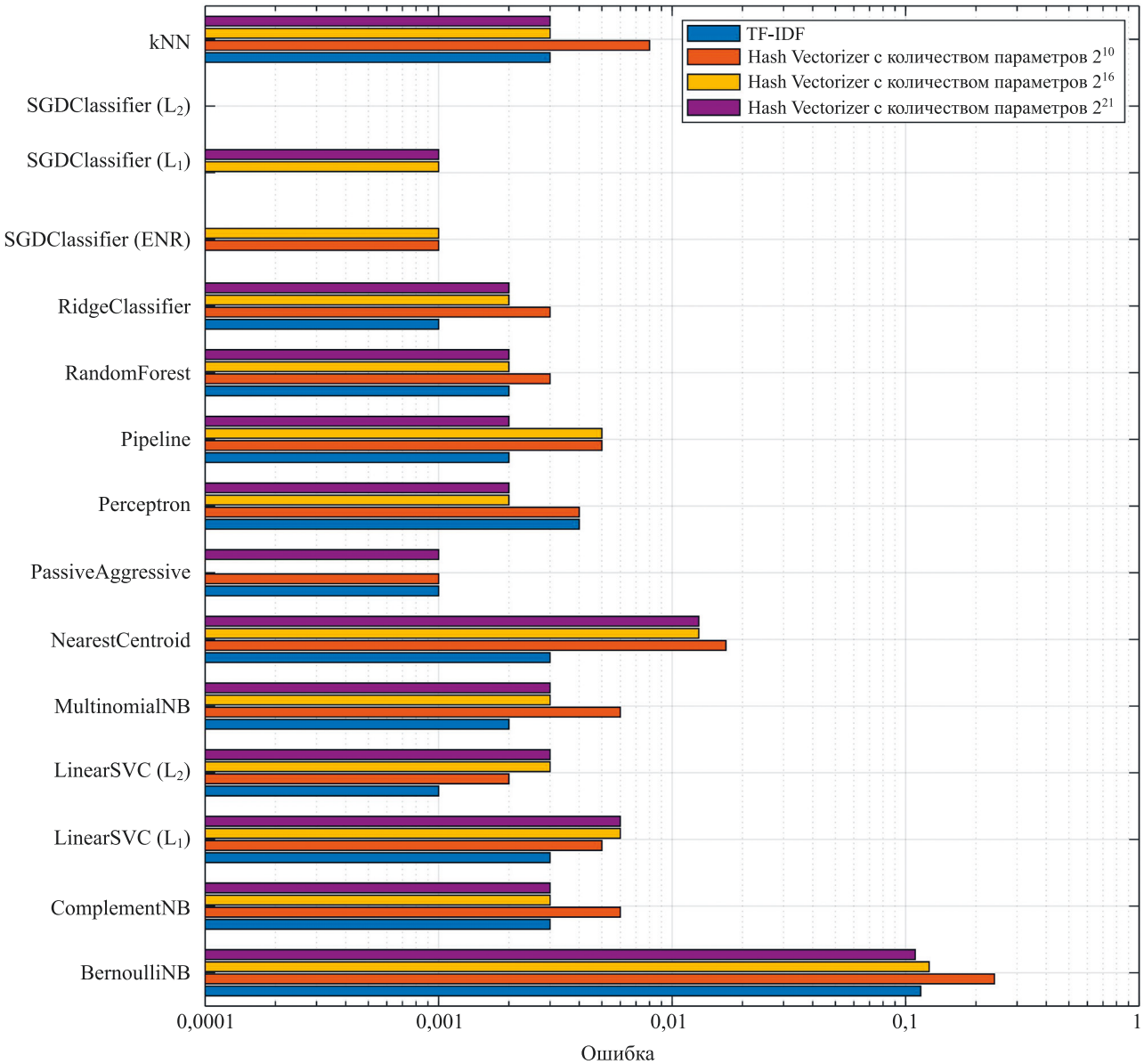


Рис. 1. Величина ошибки алгоритмов классификации при использовании векторизаторов TF-IDF и Hash Vectorizer с различным числом параметров

Fig. 1. The magnitude of the error of classification algorithms when using TF-IDF and Hash Vectorizer with different number of parameters

Vectorizer, популярность которого для большинства практических задач все чаще возрастает. Используя 32-битную функцию MurmurHash3[1], разработанную Остином Эпплби, этот векторизатор позволяет проводить непрерывное обучение по мере доступности данных, не уточняя заранее объем словаря [13].

Результаты работы классификаторов

Ввиду того, что задача классификации в данной работе носит достаточно примитивный характер, большинство алгоритмов классификации показывают практически идеальные результаты по показателю accuracy. По этой причине, для более наглядного представления результатов работы и качественной их оценки, перейдем к рассмотрению величины ошибки, являющейся обратной величиной для доли правильных ответов алгоритма (accuracy). Гистограммы, представленные на рис. 1 в логарифмическом масштабе, показывают, что для векторизатора Hash Vectorizer при увеличении количества параметров с 2^{10} до 2^{21} прослеживается тенденция к уменьшению величины ошибки.

Переходя к рассмотрению такого показателя, как время обучения, подчеркнем, что в зависимости от выбранного способа векторизации данная величина варьируется в широких пределах и увеличивается с ростом количества параметров Hash Vectorizer (рис. 2). Для более наглядного представления шкала, отображающая показатель времени, представлена в логарифмическом масштабе.

[1] MurmurHash [Электронный ресурс]. URL: https://www.sites.google.com/site/murmurhash (дата обращения: 20.12.2021).

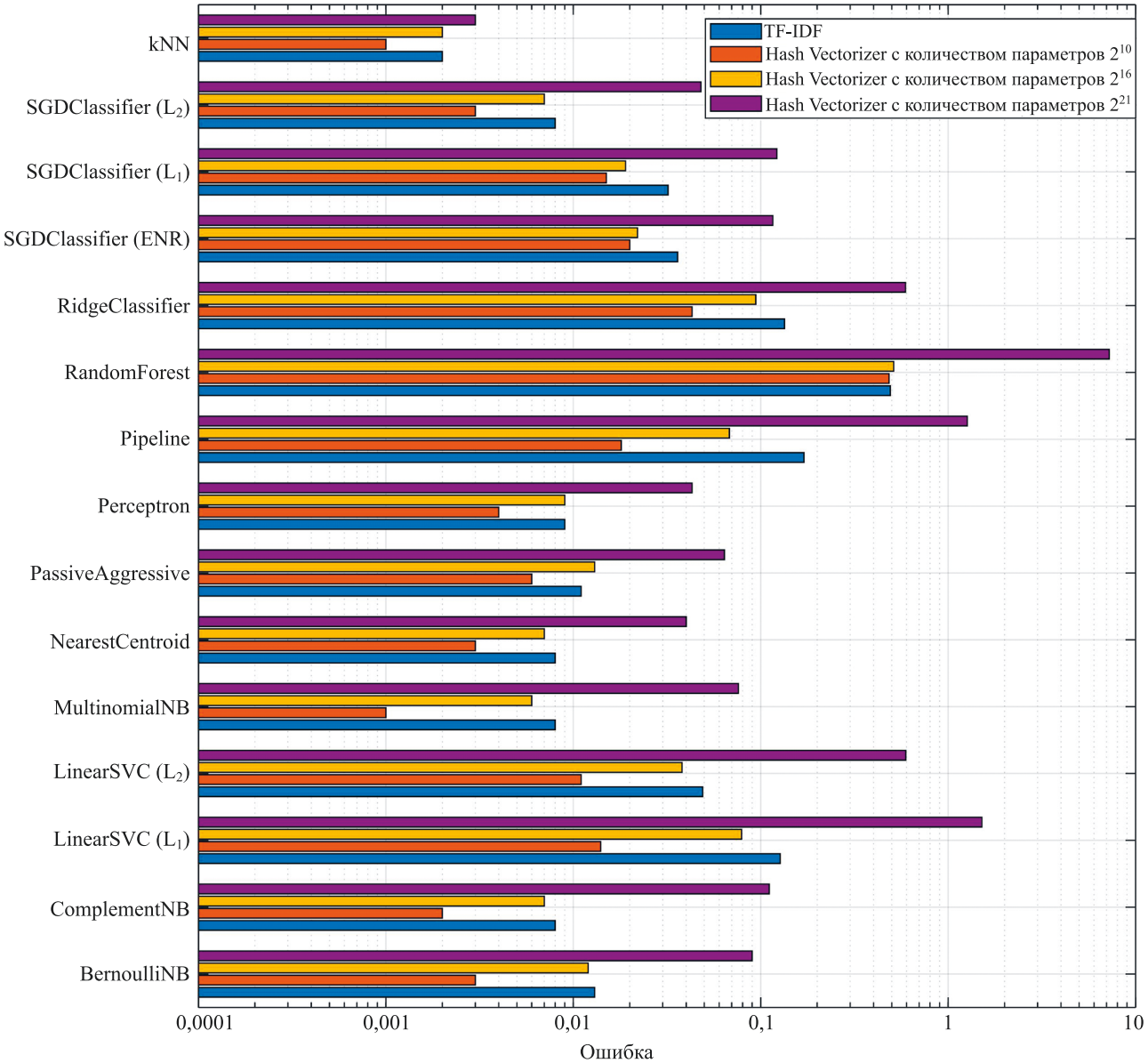


Рис. 2. Время обучения алгоритмов классификации при использовании векторизаторов TF-IDF и Hash Vectorizer с различным числом параметров

Fig. 2. Training time of classification algorithms when using TF-IDF and Hash Vectorizer with different number of parameters

Более детальное сравнение рассмотренных алгоритмов классификации с результатами их работы, отражающими время обучения, тестирования и долю правильных ответов алгоритма размещено в документе[1].

Отметим, что при увеличении числа настраиваемых параметров векторизатора Hash Vectorizer с 2^{10} до 2^{21} алгоритмы классификации показывают лучшие результаты, чем при использовании TF-IDF векторизатора. При этом значительно увеличиваются временные затраты на обучение модели, что связано с увеличением размерности целочисленных индексов, присваиваемых соответствующим токенам словаря. Это приводит к уменьшению числа коллизий [14, 15] и значительному увеличению времени обучения.

Оценивая эффективность алгоритмов путем отношения доли правильных ответов к времени обучения (accuracy/train-time) видно, что для всех исследуемых способов векторизации самый лучший алгоритм классификации — kNN, показавший лучшие результаты по доли правильных ответов и времени обучения. Наиболее высокие результаты (0,992/0,001) данный алгоритм показал при использовании векторизатора Hash Vectorizer с количеством параметров 2^{10}.

Наихудший результат показал алгоритм классификации Random Forest Classifier с результатами (0,998/7,239) для случая применения векторизатора Hash Vectorizer с количеством параметров 2^{21}.

Заключение

В работе выполнена сравнительная оценка качества работы алгоритмов линейной классификации в зависимости от выбранного способа векторизации (представления) исходного текста.

Несмотря на очевидное преимущество векторизатора Hash Vectorizer по сравнению с TF-IDF применительно к документам небольшого объема, проявляющееся в увеличении доли правильных ответов алгоритмов и, соответственно, уменьшении величины ошибки, показатель времени обучения в общей тенденции также возрастает.

Полученный эмпирический опыт по выявлению соответствия величин показателей ошибки и времени обучения алгоритмов классификации количеству настраиваемых параметров способа векторизации Hash Vectorizer позволит существенно повысить качество решения подобных практических классификационных задач путем сужения диапазона выбора настраиваемых параметров моделей-классификаторов.

[1] [Электронный ресурс]. https://docs.google.com/document/d/1pSdWQ8E7H-CBeuTZPxoQPf-yFsLWiXxy/edit (дата обращения: 20.12.2021).

Литература

1. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. № 1. С. 85–99. https://doi.org/10.15827/0236-235X.030.1.085-099
2. Бортников В.И., Михайлова Ю.Н. Документная лингвистика: учебно-методическое пособие / Министерство образования и науки Российской Федерации, Уральский государственный юридический университет. Екатеринбург: Изд-во Уральского университета, 2017. 132 с.
3. Роготнева Е.Н. Документная лингвистика: сборник учебно-методических материалов. Томск: Изд-во Томского политехнического университета, 2011. 784 с.
4. Орлов А.И. Математические методы теории классификации // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 95. С. 23–45.
5. Косова М.В., Шарипова Р.Р. Терминированность как основа классификации документных текстов // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2016. Т. 15. № 4. С. 245–252. https://doi.org/10.15688/jvolsu2.2016.4.26
6. Терских Н.В. Термин как единица специального знания // Система ценностей современного общества. 2008. № 3. С. 97–104.
7. Розенталь Д.Э., Голуб И.Б., Теленкова М.А. Современный русский язык. 13-е изд. М.: АЙРИС-пресс, 2014. 448 с.
8. Крашенинников А.М., Гданский Н.И., Рысин М.Л. Линейная классификация объектов с использованием нормальных гиперплоскостей // Инженерный вестник Дона. 2012. № 4-1 (22). С. 94–99.
9. Dan Nelson. Overview of Classification Methods in Python with Scikit-Learn // Stack Abuse [Электронный ресурс]. URL: https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/ (дата обращения: 20.12.2021).
10. Woods W. Important issues in knowledge representation // Proceedings of the IEEE. 1986. V. 74. N 10. P. 1322–1334. https://doi.org/10.1109/PROC.1986.13634
11. Рашка С., Мирджалили В. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и

References

1. Batura T.V. Automatic text classification methods. *Software & Systems*, 2017, no. 1, pp. 85–99. (in Russian). https://doi.org/10.15827/0236-235X.030.1.085-099
2. Bortnikov V.I., Mikhailova Iu.N. *Documentary Linguistics*. Ekaterinburg, Izdatel'stvo Ural'skogo universiteta Publ., 2017, 132 c. (in Russian)
3. Rogotneva E.N. *Documentary Linguistics*. Teaching materials. Tomsk, Tomsk Polytechnic University Publ., 2011, 784 c. (in Russian)
4. Orlov A.I. Mathematical methods of classification theory. *Polythematic online scientific journal of Kuban State Agrarian University*, 2014, no. 95, pp. 23–45. (in Russian)
5. Kosova M.V., Sharipova R.R. Termination as the basis for classification of document texts. *Science Journal of Volgograd State University. Linguistics*, 2016, vol. 15, no. 4, pp. 245–252. (in Russian). https://doi.org/10.15688/jvolsu2.2016.4.26
6. Terskikh N.V. Term as a unit of specialized knowledge . *Sistema cennostej sovremennogo obshhestva*, 2008, no. 3, pp. 97–104. (in Russian)
7. Rozental D.E., Golub I.B., Telenkova M.A. *Contemporary Russian Language*. Moscow, AJRIS-press Publ., 2014, 448 p. (in Russian)
8. Krasheninnikov A.M., Gdanskiy N.I., Rysin M.L. Linear classification of objects using normal hyperplanes. *Engineering journal of Don*, 2012, no. 4-1 (22), pp. 94–99. (in Russian)
9. Dan Nelson. Overview of Classification Methods in Python with Scikit-Learn. *Stack Abuse*. Available at: https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/ (accessed: 20.12.2021).
10. Woods W. Important issues in knowledge representation. *Proceedings of the IEEE*, 1986, vol. 74, no. 10, pp. 1322–1334. https://doi.org/10.1109/PROC.1986.13634
11. Raschka S., Mirjalili V. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019, 770 p.
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 2018, vol. 181, no. 1, pp. 25–29. https://doi.org/10.5120/ijca2018917395

TensorFlow 2 / пер. с англ. – 3-е изд. СПб.: ООО «Диалектика», 2020. 848 с.
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. V. 181. N 1. P. 25–29. https://doi.org/10.5120/ijca2018917395
13. Kavita Ganesan. HashingVectorizer vs. CountVectorizer [Электронный ресурс]. URL: https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/#.YcGOyavP2Ul (дата обращения: 20.12.2021).
14. Jason Brownlee. How to Encode Text Data for Machine Learning with scikit-learn // Machine learning mastery [Электронный ресурс]. URL: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/ (дата обращения: 20.12.2021).
15. Max Pagels. Introducing One of the Best Hacks in Machine Learning: the Hashing Trick // Medium [Электронный ресурс]. URL: https://medium.com/value-stream-design/introducing-one-of-the-best-hacks-in-machine-learning-the-hashing-trick-bf6a9c8af18f (дата обращения: 20.12.2021).

13. Kavita Ganesan. *HashingVectorizer vs. CountVectorizer*. Available at: https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/#.YcGOyavP2Ul (accessed: 20.12.2021).
14. Jason Brownlee. How to Encode Text Data for Machine Learning with scikit-learn. *Machine learning mastery*. Available at: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/ (accessed: 20.12.2021).
15. Max Pagels. Introducing One of the Best Hacks in Machine Learning: the Hashing Trick. *Medium*. Available at: https://medium.com/value-stream-design/introducing-one-of-the-best-hacks-in-machine-learning-the-hashing-trick-bf6a9c8af18f (accessed: 20.12.2021).

Авторы

Шадский Виктор Викторович — адъюнкт, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-9221-2283, vdvryazan57@yandex.ru

Сизоненко Александр Борисович — доктор технических наук, доцент, начальник кафедры, Краснодарское высшее военное училище им. С.М. Штеменко, https://orcid.org/0000-0001-8201-9159, siz_al@mail.ru

Чекмарев Максим Алексеевич — адъюнкт, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-6832-9991, max.chek13@gmail.com

Шишков Андрей Васильевич — студент, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-1841-8750, shishkov-andrey00@mail.ru

Исакин Даниил Андреевич — студент, Новосибирский государственный технический университет, Новосибирск, 630087, Российская Федерация, https://orcid.org/0000-0001-7307-6258, pm11.isakin@gmail.com

Authors

Viktor V. Shadsky — PhD Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-9221-2283, vdvryazan57@yandex.ru

Alexander B. Sizonenko — D.Sc., Associate Professor, Head of Department, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0001-8201-9159, siz_al@mail.ru

Maxim A. Chekmarev — PhD Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-6832-9991, max.chek13@gmail.com

Andrey V. Shishkov — Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-1841-8750, shishkov-andrey00@mail.ru

Daniil A. Isakin — Student, Novosibirsk State Technical University, Novosibirsk, 630087, Russian Federation, https://orcid.org/0000-0001-7307-6258, pm11.isakin@gmail.com

Статья поступила в редакцию 06.12.2021
Одобрена после рецензирования 29.12.2021
Принята к печати 29.01.2022

Received 06.12.2021
Approved after reviewing 29.12.2021
Accepted 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-120-126
УДК 004.021

Распознавание эмоционального состояния человека на основе сверточной нейронной сети

Гедеш Мануэл Сома[1]✉, Айжана Михайловна Каднова[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2] Воронежский институт МВД России, Воронеж, 394065, Российская Федерация

[1] guedes.soma@mail.ru✉, https://orcid.org/0000-0001-9004-3554
[2] aizhana_kadnova@mail.ru, https://orcid.org/0000-0002-7758-6578

Аннотация

Предмет исследования. Предложено новое решение распознавания эмоционального состояния человека (радость, удивление, грусть, гнев, отвращение, страх и нейтральное состояние) по выражению лица. Наряду с традиционной вербальной коммуникацией эмоции играют значительную роль при определении истинных намерений для реализации коммуникативного акта в различных практических сферах деятельности. Существует большое количество моделей и алгоритмов распознавания эмоций человека по классам и их применения для сопровождения коммуникативного акта. Известные модели дают невысокую точность распознавания эмоциональных состояний. **Методы.** Для классификации выражений лиц построены два классификатора, реализованные в библиотеке Keras: ResNet50, MobileNet, и предложена новая архитектура классификатора сверточной нейронной сети. Обучение классификаторов осуществлено на наборе данных FER 2013. **Основные результаты.** Сравнение результатов работы выбранных классификаторов показало, что предложенная модель имеет наилучший результат по точности валидации (60,13 %) и размеру (15,49 МБ), при этом функция потерь составляет 0,079 для точности и 2,80 — для валидации. **Практическая значимость.** Результаты исследования могут быть использованы для распознавания признаков стресса и агрессивного поведения человека в системах обслуживания населения и в сферах, при наличии общения с большим количеством людей.

Ключевые слова

ResNet50, MobileNet, выражение лица, глубокая нейронная сеть, эмоция, классификация

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 22-21-00604).

Ссылка для цитирования: Сома Г.М., Каднова А.М. Распознавание эмоционального состояния человека на основе сверточной нейронной сети // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 120–126. doi: 10.17586/2226-1494-2022-22-1-120-126

Recognition of the emotional state based on a convolutional neural network

Guedes M. Soma[1]✉, Aizhana M. Kadnova[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation
[2] Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation

[1] guedes.soma@mail.ru✉, https://orcid.org/0000-0001-9004-3554
[2] aizhana_kadnova@mail.ru, https://orcid.org/0000-0002-7758-6578

Abstract

The paper proposes a new solution for recognizing the emotional state of a person (joy, surprise, sadness, anger, disgust, fear, and neutral state) by facial expression. Along with traditional verbal communication, emotions play a significant role in determining true intentions during a communicative act in various areas. There is a large number of models and algorithms for recognizing human emotions by class and applying them to accompany a communicative act. The known models show a low accuracy in recognizing emotional states. To classify facial expressions, two classifiers were

© Сома Г.М., Каднова А.М., 2022

built and implemented in the Keras library (ResNet50, MobileNet) and a new architecture of a convolutional neural network classifier was proposed. The classifiers were trained on the FER 2013 dataset. Comparison of the results for the chosen classifiers showed that the proposed model has the best result in terms of validation accuracy (60.13 %) and size (15.49 MB), while the loss function is 0.079 for accuracy and 2.80 for validation. The research results can be used to recognize signs of stress and aggressive human behavior in public service systems and in areas characterized by the need to communicate with a large number of people.

Keywords
ResNet50, MobileNet, facial expressions, deep neural network, emotion, classification

Acknowledgements
This research is financially supported by the Russian Science Foundation, Agreement No. 22-21-00604.

For citation: Soma G.M., Kadnova A.M. Recognition of the emotional state based on a convolutional neural network. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 120–126 (in Russian). doi: 10.17586/2226-1494-2022-22-1-120-126

Введение

Важный фактор человеческого общения — эмоциональное состояние собеседников, помогающее понять их намерения. В настоящее время существует большое количество алгоритмов, способных автоматически распознавать эмоции человека по выражению лица [1]. Однако качество таких систем распознавания эмоций снижается из-за таких проблем [2] как: некорректный выбор архитектуры нейронной сети; небольшой объем данных для обучения; наличие притворных эмоций, внутриклассовых различий и межклассового сходства; разные углы поворота головы; уровень освещенности; различия в пропорциях лица.

Системы распознавания эмоций человека по лицу включают в себя основные этапы:
— предварительная обработка изображения;
— извлечение визуальных признаков;
— классификация эмоций [2].

Предварительная обработка изображения заключается в поиске области лица, обрезке и масштабировании данной области, изменении яркости изображения, выравнивания лица и регулировки контраста. Извлечение визуальных признаков основано на использовании геометрии и внешнего вида лица, на котором происходит поиск ключевых точек выделенных элементов. Под геометрией понимаются такие элементы лица (их форма и расположение) как: глаза, нос, губы, рот и др. Как правило, определение формы элементов лица выполняется по ключевым точкам. Например, на рис. 1 такие точки представлены для рта и бровей [3]. К внешнему виду лица относится форма и текстура кожи. Классификация извлеченных признаков на выражении лица осуществляется с использованием традиционных и нейросетевых методов машинной классификации.

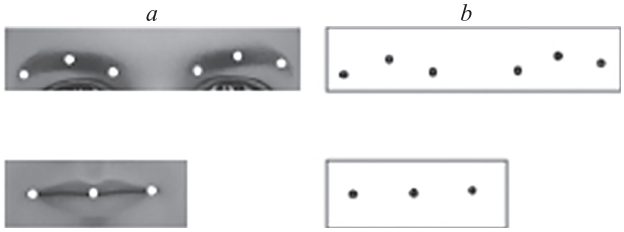


Рис. 1. Основные ключевые точки на лице [3]: исходные изображения (*a*); соединение ключевых точек, соответствующее их изображениям (*b*)

Fig. 1. Main key points on the face [3]: source images (*a*); connection of key points corresponding to their images (*b*)

Материалы и методы

Анализ выражения лица в процессе распознавания эмоций направлен на решение трех основных задач: нахождение лица, извлечение признаков и классификация выражений лица из неподвижных или последовательных изображений.

Точное местоположение лица на изображении или видеокадре позволяет определить метод Виолы–Джонса. Его основной принцип заключается в сканировании подокна, способного обнаружить лицо на входном изображении или видеокадрах. В данном методе изображение представлено в интегральной форме и позволяет оперативно вычислить необходимые объекты [4]. Интегральным представлением изображения является матрица с такими же размерами, что и исходное изображение. Каждый элемент матрицы хранит сумму яркостей пикселов, расположенных левее и выше этого элемента [5–11]. Элементы матрицы рассчитываются следующим образом:

$$L(x, y) = \sum_{i=0, j=0}^{i \le x, j \le y} I(i, j),$$

где $L(x, y)$ — интегральное изображение; $I(i, j)$ — исходное изображение.

Пример расчета интегрального и исходного изображений представлен в табл. 1.

В зависимости от выполняемых функций методы извлечения признаков реализуются на основе: геометри-

Таблица 1. Пример вида матриц исходного и интегрального изображений

Table 1. Example of the source and integral image matrices

Исходное изображение $I(i, j)$				Интегральное изображение $L(x, y)$			
0	1	1	1	0	1	2	3
1	2	2	3	1	4	7	11
1	2	1	1	2	7	11	16
1	3	1	0	3	11	16	21

ческих объектов, позволяющих извлекать информацию о таких объектах как рот, нос, брови и других, и определять их местоположение; моделей внешнего вида, которые извлекают информацию о текстурных особенностях лица; глобальных и локальных объектов, наиболее эффективным из которых [5] является метод главных компонентов (Principal Component Analysis, PCA).

Несмотря на высокую эффективность, PCA требует для своего применения идеализированных условий, таких как единые параметры освещенности, нейтральное выражение лица, отсутствие помех в виде очков и бород. При их несоблюдении главные компоненты не будут отражать межклассовые вариации [7].

Классификация извлеченных признаков на выражении лица (счастье, удивление, гнев, страх, отвращение, грусть и нейтральное) осуществляется с использованием методов машинной классификации (рис. 2). Были подробно рассмотрены несколько методов. Метод опорных векторов (Support Vector Machine, SVM) — обучается на ранее известных данных, для получения максимизации границы принятия решения между группами [8, 9]. Скрытые марковские модели (Hidden Markov models, HMM) сканирующего окна [12] извлекают пикселы, трансформируя их в векторы наблюдений, при этом полученные векторы классифицируются по мимике (выражениям лица) [1]. Многослойный перцептрон (Multilayer perceptron, MPL) состоит из системы простых взаимосвязанных нейронов, или узлов, реализующих нелинейное отображение входного вектора в выходной [13]. Глубокие нейронные сети (Deep neural network, DNN), позволяют изучать сложные нелинейные функции заданного входа для минимизации ошибок [14]. Сверточная нейронная сеть (Convolutional neural network, CNN) — специальная архитектура искусственных нейронных сетей, которая наиболее точно распознает объекты на изображениях, так как, в отличие от многослойного персептрона, учитывается двухмерная топология изображения. При этом сверточные сети устойчивы к небольшим смещениям, изменениям масштаба и поворотам объектов на входных изображениях [10]. В результате анализа в настоящей работе выбрана CNN.

Для распознавания эмоционального состояния человека использована CNN с параметрами: входным слоем (96 × 96 × 3), на который подается цветное изображение (BGR image) размером (96 × 96); 30 скрытых слоев, один их которых — выходной слой, содержащий 7 выходов, определяющих вероятность принадлежности распознанной эмоции к каждому из заданных классов (счастье, удивление, гнев, страх, отвращение, грусть и нейтральное состояние.

На рис. 3 показаны взаимосвязи скрытых слоев. Conv2D – слой свертки с параметрами: F (filters) — количество выходных фильтров в свертке; K (kernel_size) — длина окна свертки. ReLu — функция активации, Batch Normalization — преобразование, которое поддерживает среднее значение выхода около 0 и стандартное отклонение выхода, стремящееся к 1. MaxPooling2D — слой подвыборки, где P (pool_size) позволяет уменьшать входные данные в *n* раз; Dropout обеспечивает случайное выпадение нейронов для избегания перенасыщения нейросети. Flatten трансформирует входную матрицу в один массив. Dense — полносвязный слой, где U (units) — размерность выходного пространства. A:'relu' — функция активации ReLu, Class Labels — метки классов. ρ — слой, обнуляющий в модели нейронной сети часть весов.

Для обучения и валидации разработанной CNN использован набор данных FER 2013, состоящий из 20 888 изображений (рис. 4) с реальными выражениями лица при различных условиях освещения [15]. Набор данных содержит две выборки с изображениями: для обучения «Train» (гнев — 782, страх — 3263, отвращение — 327, счастье — 515, грусть — 956, удивление —



Рис. 2. Обобщение методов машинной классификации эмоций [2]

Fig. 2. Generalization of the methods for machine classification of emotions [2]

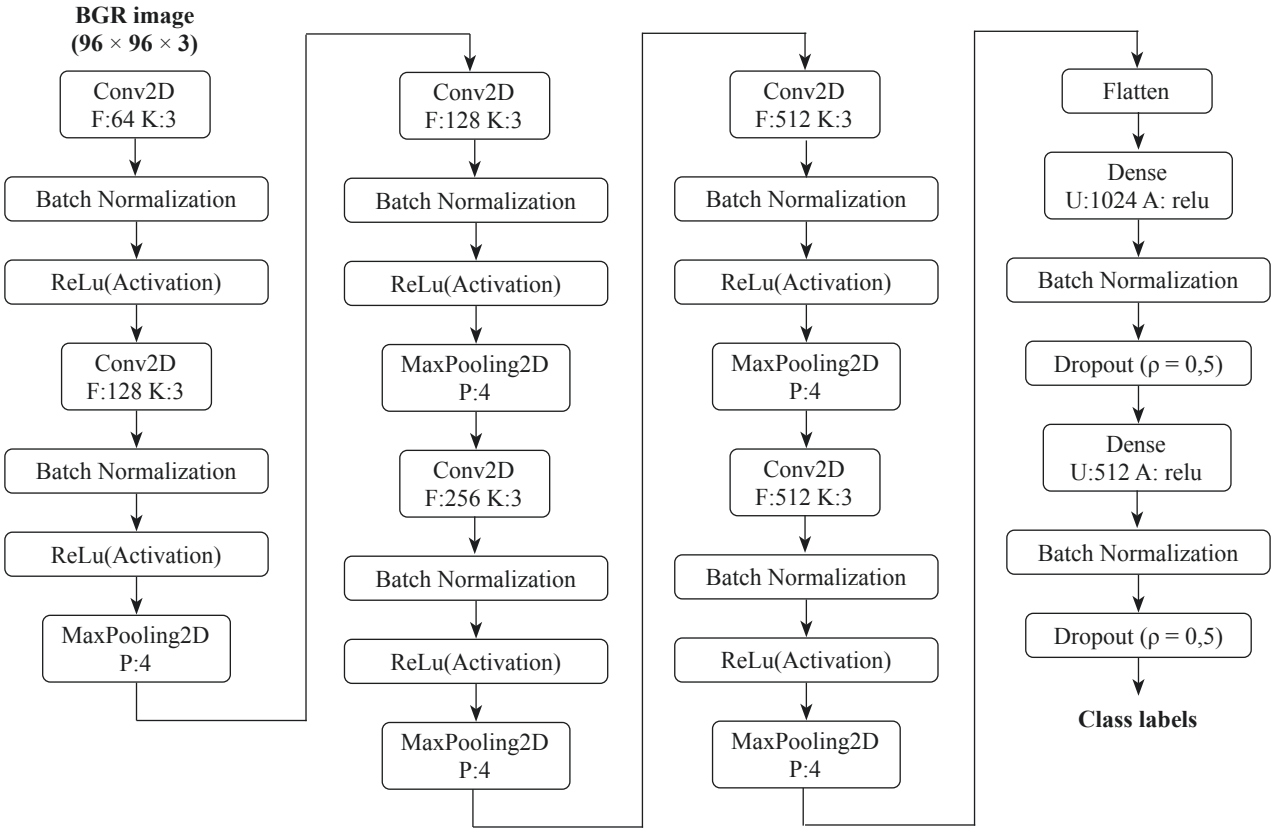


Рис. 3. Архитектура нейронной сети для распознавания эмоционального состояния человека

Fig. 3. Neural network architecture in the system for recognition of a human emotional state



Рис. 4. Примеры изображений из набора данных FER 2013

Fig. 4. Sample images from the FER 2013 dataset

2141, нейтральный — 3955) и валидации «Validation» (гнев — 494, страх — 136, отвращение — 440, счастье — 1121, грусть — 462, удивление — 351, нейтральное — 737.

Эксперименты и результаты

Для оценки качества работы классификатора, разработанного на основе CNN, выполнено сравнение результатов его работы с параметрами известных моделей классификации, также реализованных в библиотеке Keras (часть библиотеки «TensorFlow»): ResNet50 и MobileNet и обученных на наборе данных (FER 2013), по основным параметрам (табл. 2): размер модели; время выполнения — измеряемого на GPU «NVIDIA Tesla P100-PCIE» (GPU Google kaggle); точность — отношение количества правильных классификаций к общему количеству изображений; потери — значение функции потерь, которая используется для оценивая качества принимаемых моделью решений, при обучении модели данный показатель сводится к минимизации на каждой итерации обучения; тренировка — процесс обучения нейронной сети, в котором изменяются веса сети путем градиентного спуска, на каждой итерации минимизируя функцию потерь.

Полученные результаты свидетельствуют о том, что лучшим по точности валидации является разработанный классификатор (60,13 %). Классификатор имеет меньшее время выполнения (0,001 с), чем другие модели, а также небольшой размер (15,49 МБ).

Из представленных графиков на рис. 5 можно увидеть хороший результат — при обучении разработанного классификатора на основе CNN функция потери уменьшается на 0,079, а точность увеличивается до 97,3 %.

Заметим, что по полученным матрицам спутывания (рис. 6) разработанный классификатор превосходит MobileNet по значению полноты (Recall) по всем классам (кроме отвращения и страха), что сказывается на приросте значения невзвешенной средней полноты (Unweighted Average Recall, UAR). Эффективность разработанного классификатора подтверждается получен-

Таблица 2. Результаты оценки моделей классификации

Table 2. Results of evaluating classification models

Модель	Размер модели, МБ	Время выполнения, с	Точность тренировки, %	Потери тренировки	Точность валидации, %	Потери валидации
ResNet50	98	0,016	72,30	0,150	53,20	2,58
MobileNet	16	0,0025	88,92	0,360	56,37	2,70
Разработанный классификатор	15,49	0,001	97,30	0,079	60,13	2,80

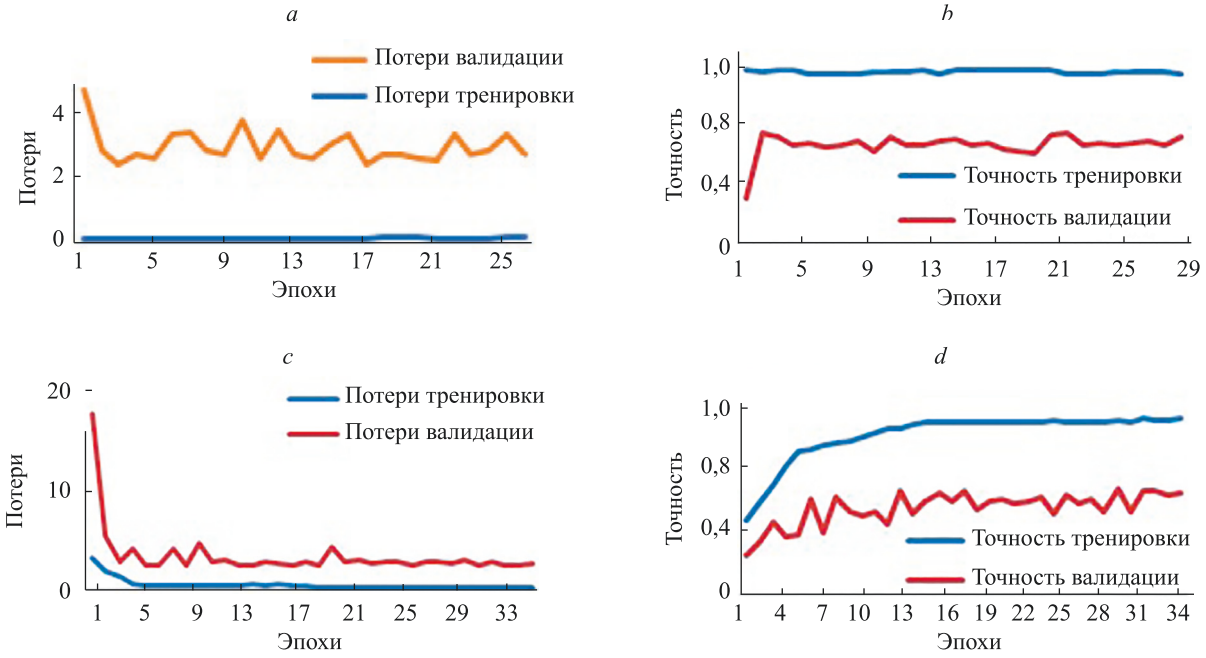


Рис. 5. Обучение и тестирование потерь и точности: разработанного классификатора (*a*, *b*); MobileNet (*c*, *d*)

Fig. 5. Training and testing of loss and accuracy: the developed classifier (*a*, *b*); MobileNet (*c*, *d*)

a

Истинное \ Предсказанное	гнев	страх	отвращение	счастье	грусть	удивление	нейтральное
гнев	**318**	33	19	12	63	15	34
страх	29	**71**	21	4	6	2	3
отвращение	62	40	**199**	14	45	53	27
счастье	28	3	17	**958**	24	43	48
грусть	80	15	26	21	**217**	19	84
удивление	8	3	21	22	3	**289**	5
нейтральное	28	13	35	59	119	39	**444**

b

Истинное \ Предсказанное	гнев	страх	отвращение	счастье	грусть	удивление	нейтральное
гнев	**267**	31	31	16	86	17	46
страх	22	**86**	5	1	11	9	2
отвращение	55	40	**227**	12	40	44	22
счастье	30	12	23	**823**	42	101	90
грусть	89	16	37	23	**192**	20	85
удивление	8	8	28	31	3	**268**	5
нейтральное	33	9	45	62	120	40	**428**

Рис. 6. Матрица спутывания разработанной модели (*a*) и MobileNet (*b*)

Fig. 6. Confusion matrix of the developed model (*a*) and MobileNet (*b*)

ным значением UAR, которое превосходит более чем на 2 % значение, полученное с помощью MobileNet.

Заключение

На основе анализа существующих методов распознавания эмоций по лицу разработана модель классификации соответствующих состояний человека с помощью сверточной нейронной сети. Сравнение результатов работы разработанного классификатора с моделями ResNet50 и MobileNet выявило, что предложенная модель имеет наилучший результат по точности валидации (60,13 %) и размеру (15,49 МБ), при этом функция потерь составляет 0,079 для точности и 2,80 — для валидации.

Литература

1. Varma S., Shinde M., Chavan S.S. Analysis of PCA and LDA features for facial expression recognition using SVM and HMM classifiers // Techno-Societal 2018: Proc. of the 2nd International Conference on Advanced Technologies for Societal Applications. V. 1. 2020. P. 109–119. https://doi.org/10.1007/978-3-030-16848-3_11
2. Рюмина Е.В., Карпов А.А. Аналитический обзор методов распознавания эмоций по выражениям лица человека // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 2. С. 163–176. https://doi.org/10.17586/2226–1494-2020-20-2-163-176
3. Ахметшин Р.И., Кирпичников А.П., Шлеймович М.П. Распознавание эмоций человека на изображениях // Вестник технологического университета. 2015. Т. 18. № 11. С. 160–163.
4. Эрман Е.А., Мамдух Мохаммед Гомаа Мохаммед. Метод обнаружения лиц на изображении с использованием комбинации метода Виолы-Джонса и алгоритмов определения цвета кожи // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2015. № 1. С. 49–55.
5. Мухамадиева К.Б. Сравнительный анализ алгоритмов распознавания лиц // Современные материалы, техника и технологии. 2017. № 7(15). С. 58–62.
6. Амелькин С.А., Захаров А.В., Хачумов В.М. Обобщенное расстояние Евклида Махаланобиса и его свойства // Информационные технологии и вычислительные системы. 2006. № 4. С. 40–44.
7. Karamizadeh S., Abdullah S.M., Manaf A.A., Zamani M., Hooman A. An overview of principal component analysis // Journal of Signal and Information Processing. 2013. V. 4. P. 173–175. https://doi.org/10.4236/jsip.2013.43B031
8. Dino H.I., Abdulrazzaq M.B. Facial expression classification based on SVM, KNN and MLP classifiers // Proc. of the International Conference on Advanced Science and Engineering (ICOASE 2019). 2019. P. 70–75. https://doi.org/10.1109/ICOASE.2019.8723728
9. Greche L., Es-Sbai N., Lavendelis E. Histogram of oriented gradient and multi layer feed forward neural network for facial expression identification // Proc. of the International Conference on Control, Automation and Diagnosis (ICCAD 2017). 2017. P. 333–337. https://doi.org/10.1109/CADIAG.2017.8075680
10. Greche L., Akil M., Kachouri R., Es-Sbai N. A new pipeline for the recognition of universal expressions of multiple faces in a video sequence // Journal of Real-Time Image Processing. 2020. V. 17. N 5. P. 1389–1402. https://doi.org/10.1007/s11554-019-00896-5
11. Kulke L., Brümmer L., Pooresmaeili A., Schacht A. Overt and covert attention shifts to emotional faces: Combining EEG, eye tracking, and a go/no-go paradigm // Psychophysiology. 2021. V. 58. N 8. P. e13838. https://doi.org/10.1111/psyp.13838
12. Aleksic P.S., Katsaggelos A.K. Automatic facial expression recognition using facial animation parameters and multistream HMMs // IEEE Transactions on Information Forensics and Security. 2006. V. 1. N 1. P. 3–11. https://doi.org/10.1109/TIFS.2005.863510
13. Gardner M.W., Dorling S.R. Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences // Atmospheric Environment. 1998. V. 32. N 14-15. P. 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0
14. Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R. CREMA-D: Crowd-sourced emotional multimodal actors dataset // IEEE Transactions on Affective Computing. 2014. V. 5. N 4. P. 377–390. https://doi.org/10.1109/TAFFC.2014.2336244
15. Challenges in Representation Learning Facial Expression Recognition Challenge [Электронный ресурс]. URL: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge (дата обращения: 13.11.2021).

References

1. Varma S., Shinde M., Chavan S.S. Analysis of PCA and LDA features for facial expression recognition using SVM and HMM classifiers. Techno-Societal 2018. *Proc. of the 2nd International Conference on Advanced Technologies for Societal Applications*. Vol. 1, 2020, pp. 109–119. https://doi.org/10.1007/978-3-030-16848-3_11
2. Ryumina E.V. Karpov A.A Analytical review of methods for emotion recognition by human face expressions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 2, pp. 163–176. (in Russian). https://doi.org/10.17586/2226–1494-2020-20-2-163-176
3. Akhmetshin R.I., Kirpichnikov A.P., Shleimovich M.P. Human emotion recognition in images. *Vestnik Technologicheskogo Universiteta*, 2015, vol. 18, no. 11, pp. 160–163. (in Russian)
4. Erman E.A., Mamdouh Mokhammed Gomaa Mokhammed. Face detection in image by using a combination of Viola — Jones method and skin detection algorithms. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Sciences and Informatics*, 2015, no. 1, pp. 49–55. (in Russian)
5. Mukhamadieva K.B. Comparative analysis of facial recognition algorithms. *Modern Materials, Equipment, and Technology*, 2017, no. 7(15), pp. 58–62. (in Russian)
6. Amelkin S. A., Zakharov A.V., Khachumov V.M. Generalized Mahalanobis Euclidean distance and its properties. *Information Technologies and Computer Systems*, 2006, no. 4, pp. 40–44. (in Russian)
7. Karamizadeh S., Abdullah S.M., Manaf A.A., Zamani M., Hooman A. An overview of principal component analysis. *Journal of Signal and Information Processing*, 2013, vol. 4, pp. 173–175. https://doi.org/10.4236/jsip.2013.43B031
8. Dino H.I., Abdulrazzaq M.B. Facial expression classification based on SVM, KNN and MLP classifiers. *Proc. of the International Conference on Advanced Science and Engineering (ICOASE 2019)*, 2019, pp. 70–75. https://doi.org/10.1109/ICOASE.2019.8723728
9. Greche L., Es-Sbai N., Lavendelis E. Histogram of oriented gradient and multi layer feed forward neural network for facial expression identification. *Proc. of the International Conference on Control, Automation and Diagnosis (ICCAD 2017)*, 2017, pp. 333–337. https://doi.org/10.1109/CADIAG.2017.8075680
10. Greche L., Akil M., Kachouri R., Es-Sbai N. A new pipeline for the recognition of universal expressions of multiple faces in a video sequence. *Journal of Real-Time Image Processing*, 2020, vol. 17, no. 5, pp. 1389–1402. https://doi.org/10.1007/s11554-019-00896-5
11. Kulke L., Brümmer L., Pooresmaeili A., Schacht A. Overt and covert attention shifts to emotional faces: Combining EEG, eye tracking, and a go/no-go paradigm. *Psychophysiology*, 2021, vol. 58, no. 8, pp. e13838. https://doi.org/10.1111/psyp.13838
12. Aleksic P.S., Katsaggelos A.K. Automatic facial expression recognition using facial animation parameters and multistream HMMs. *IEEE Transactions on Information Forensics and Security*, 2006, vol. 1, no. 1, pp. 3–11. https://doi.org/10.1109/TIFS.2005.863510
13. Gardner M.W., Dorling S.R. Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences. *Atmospheric Environment*, 1998, vol. 32, no. 14-15, pp. 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0
14. Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 2014, vol. 5, no. 4, pp. 377–390. https://doi.org/10.1109/TAFFC.2014.2336244
15. *Challenges in Representation Learning Facial Expression Recognition Challenge*. Available at: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge (accessed: 13.11.2021).

Авторы

Сома Гедеш Мануэл — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-9004-3554, guedes.soma@mail.ru

Каднова Айжана Михайловна — кандидат технических наук, преподаватель, Воронежский институт МВД России, Воронеж, 394065, Российская Федерация; научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57217115500, https://orcid.org/0000-0002-7758-6578, aizhana_kadnova@mail.ru

Статья поступила в редакцию 21.11.2021
Одобрена после рецензирования 14.01.2022
Принята к печати 29.01.2022

Authors

Guedes M. Soma — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-9004-3554, guedes.soma@mail.ru

Aizhana M. Kadnova — PhD, Lecturer, Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation; Scientific Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57217115500, https://orcid.org/0000-0002-7758-6578, aizhana_kadnova@mail.ru

Received 21.11.2021
Approved after reviewing 14.01.2022
Accepted 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-127-137
УДК 004.85, 378.048.2

Интеллектуализация управления развитием персонала высокотехнологичных сервис-ориентированных компаний

Артем Дмитриевич Береснев

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
artem.beresnev@itmo.ru✉, https://orcid.org/0000-0002-4646-6856

Аннотация

Предмет исследования. Предложен подход к управлению развитием персонала сервис-ориентированных ИТ-компаний. Подход основан на параметрической модели обучения высококвалифицированного персонала и реализован с использованием интеллектуальных алгоритмов. **Метод.** Параметризация модели обучения выполнена на основе теории грубых множеств. Для реализации интеллектуальных алгоритмов применены технологии тематического моделирования с аддитивной регуляризацией (Additive Regularization Thematic Model), использования специальной среды для текущей разработки конфигурации информационных систем и анализа формальных концептов (Formal Concept Analysis). **Основные результаты.** Для разработанного подхода создан набор методик, который реализован в виде библиотеки программных средств. Оценка эффективности предложенного решения реализована на наборе данных, содержащих результаты обработки 2948 заявок на обслуживание, обработанных сотрудниками сервис-ориентированной ИТ-компанией в течение четырех месяцев. Результаты экспериментальной оценки показали, что использование предложенного набора методик и библиотеки программных средств увеличило от 31 до 54 % эффективность работы сервис-инженеров по ключевым показателям. **Практическая значимость.** Применение разработанного подхода позволит оперативно адаптировать квалификацию персонала сервис-ориентированных ИТ-компаний в условиях быстрой смены производственных задач и рабочего окружения без отрыва от рабочего процесса.

Ключевые слова

развитие персонала, сервисная модель предоставления услуг, обучение на рабочем месте, грубые множества, обработка текстов, тематическое моделирование с аддитивной регуляризацией, анализ формальных концептов

Ссылка для цитирования: Береснев А.Д. Интеллектуализация управления развитием персонала высокотехнологичных сервис-ориентированных компаний // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 127–137. doi: 10.17586/2226-1494-2022-22-1-127-137

Intellectualization of personnel development management in high-tech service-oriented companies

Artem D. Beresnev

ITMO University, Saint Petersburg, 197101, Russian Federation
artem.beresnev@itmo.ru✉, https://orcid.org/0000-0002-4646-6856

Abstract

The paper proposes an approach to managing the personnel development in service-oriented IT companies, which is based on a parametric model of training highly qualified personnel and implemented using intelligent algorithms. The parameterization of the training model is carried out on the rough sets theory. The implementation of intelligent algorithms required the following technologies: thematic modeling with additive regularization (Additive Regularization Thematic Model), a special environment for the current development of the configuration of information systems and the analysis of formal concepts (Formal Concept Analysis). The approach was developed as a set of techniques and implemented as a software library. The efficiency assessment was carried out on a dataset that contains the results of processing for 2,948 service requests processed by employees of a service-oriented IT company in 4 months. The results

© Береснев А.Д., 2022

of the experimental evaluation showed that the use of the developed set of methods and the library of software tools increased the efficiency of the work of service engineers in terms of key indicators from 31 to 54 %. Application of the developed approach will make it possible to quickly adapt personnel qualifications in service-oriented IT companies in the context of a rapid change in production tasks and work environment without interrupting the work process.

Keywords
personnel development, service model of service delivery, on-the-job training, rough sets, text processing, thematic modeling with additive regularization, formal concept analysis

For citation: Beresnev A.D. Intellectualization of personnel development management in high-tech service-oriented companies. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 127–137 (in Russian). doi: 10.17586/2226-1494-2022-22-1-127-137

Введение

В настоящее время в области высоких технологий широкое распространение получил сервис-ориентированный подход организации бизнеса, реализуемый высокотехнологичными сервис-ориентированными компаниями (ВСОК) и состоящий в предоставлении заказчику не продукта, а непрерывного сервиса. Такой подход, иначе называемый сервисной моделью предоставления услуг, обеспечивает заказчику быстрый результат, а для ВСОК — увеличение прибыли и постоянный рост [1, 2]. В условиях постоянно меняющихся задач и технологий ключевое значение для успеха сервисной модели приобретает эффективная работа персонала, в связи с чем ВСОК испытывают нарастающую потребность в управлении развитием персонала [3].

Согласно [4–6], персонал ВСОК различается по квалификационным уровням на специалистов начальных («новичок», «младший специалист») и высших уровней («компетентный специалист», «профессионал», «эксперт»), которые соотносятся со степенью формализации решаемых специалистом производственных задач. Наиболее значимая в сервисной модели бизнеса — способность компании решать слабо формализованные проблемы заказчика. В связи с этим конкурентные преимущества ВСОК обеспечиваются в первую очередь, за счет сотрудников высших квалификационных уровней, обладающих техническими знаниями и навыками, инновационной активностью, креативностью, вовлеченностью и ответственностью за результат, достаточными для работы в условиях неопределенности.

На основании выполненного анализа в работах [7–10], задачи управления развитием персонала решены в контексте его обучения профессиональной деятельности. При этом большинство существующих методов управления развитием персонала основано на видоизмененном применении подходов, характерных для профессионального образования (учебные курсы, программы сертификации, тренинги), что предполагает обучение с отрывом от производства по заранее заданной программе. Результаты опроса более чем 6000 специалистов [4] и анализ содержания программ развития персонала [7] показали, что такие подходы плохо адаптируются или практически неприменимы для специалистов высших квалификационных уровней. Требуется интеллектуализация данных подходов, которая обеспечит эффективное, т. е. управляемое и параметризуемое введение в процесс обучения элементов неопределенности.

Таким образом, в настоящей работе решаются следующие задачи:
— построение проблемно-ориентированной модели управления развитием персонала с возможностью ее параметризации;
— разработка набора методик управления развитием персонала высших квалификационных уровней с использованием интеллектуальных алгоритмов;
— программная реализация разработанных методик;
— экспериментальная оценка эффективности разработанного набора методик на примере развития сервис-инженеров конкретной ВСОК.

Проблемно-ориентированная модель управления развитием персонала

Рассмотрим разработанную модель управления развитием персонала ВСОК на основе базовой модели процесса обучения [11], концептуальной модели развития персонала [12] и современных теорий принятия решения в условиях неопределенности [13, 14]. Предложенный кортеж модели имеет вид

$$S = <T, P, C, A, F, I, R>, \quad (1)$$

где T — цель обучения; P — участники процесса; C — содержание обучения; A — средства обучения; F — форма обучения; I — параметризация модели; R — управление моделью. Проблемная ориентация модели связана в первую очередь, с последними двумя компонентами. Компонент I отражает степень интегрированности модели обучения в реальный рабочий процесс и позволяет гибко варьировать параметры неопределенности, вводимой в процесс обучения, такие как: степень соответствия структур модельной и реальной сред; допустимое отклонение текущего поискового поведения от эталонного; выделение значимого для конкретной задачи подграфа из общего когнитивного графа предметной области. Компонент R отражает степень использования результатов текущей профессиональной деятельности в учебном процессе и обеспечивает настраиваемое управление обратной связью.

На основе теории грубых множеств [15] разработан метод проблемно-ориентированного ранжирования средств развития персонала ВСОК, который позволяет выделить подмножества параметров модели (1) и их атрибутов, наиболее значимые для обучения специалистов конкретных квалификационных уровней.

С использованием разработанного метода определены наиболее значимые компетенции для сотрудников

высших квалификационных уровней, такие способности как: работать с большой степенью автономности, основанной на собственных профессиональных знаниях, навыках и опыте; работать в условиях неопределенности и постоянных изменений; быстро осваивать новые технологии и системы. Показано, что для развития этих компетенций наиболее подходящим является непрямое управление обучением персонала на рабочем месте, без отрыва от производства [3, 16].

С помощью многомерного шкалирования определены наиболее востребованные подходы высококвалифицированным персоналом ВСОК как формы обучения на рабочем месте [17]. Для данных подходов имеет большое значение управление по параметрам *I* и *R*:

— самостоятельный поиск в Интернет;
— обучение на основе повседневного опыта профессиональной деятельности;
— обмен знаниями внутри команды.

Набор методик управления развитием персонала высших квалификационных уровней

Для реализации выделенных подходов обучения высококвалифицированного персонала ВСОК разработан набор методик их интеллектуальной поддержки. Содержание и особенности применения методик проиллюстрированы на примере процесса предоставления ИТ-сервиса (рис. 1). Представленный процесс реализации проекта сервиса содержательно соответствует требованиям отраслевых стандартов [18] и состоит из трех этапов: согласование, разработка и эксплуатация сервиса. Циклическое повторение этапов обеспечивает планомерное развитие сервиса на основании опыта эксплуатации и новых потребностей заказчика.

К этапу согласования сервиса относятся подпроцессы: анализ проекта (АП); определение финансовых требований (ФТ); эскизное проектирование (ЭП); структурирование задачи ($Э_1$); выбор принципиального решения и (или) архитектуры, с отбором элементов сервиса и проектированием их взаимодействия ($Э_2$); формализация и описание решения для сервиса в конкретных условиях ($Э_3$; согласование проекта (СП). Разработка (второй этап) включает: планирование проекта (ПП); разработку конфигурации по проекту сервиса (РрП); стабилизацию проекта сервиса (СтП); развертывание сервиса (РП). На третьем этапе эксплуатации выполняются: мониторинг (МК); обслуживание заказчика (ОЗ), включающее выполнение функций трех линий поддержки сервиса ($Л_1$, $Л_2$ и $Л_3$); управление проблемами (УП). На рис. 1, *b* показаны элементы разработанной библиотеки программных средств, реализующие набор методик управления развитием персонала высших квалификационных уровней: поисковая система (ПС), модельная система (МС) и диалоговая справочная система (СС). На рис. 1, *b* зеленым цветом выделены группы подпроцессов, претерпевших изменения в силу применения разработанных программных средств. На первом этапе уменьшается количество итераций *N* проектирования ($N_1 > N_2$). На втором этапе процессы РрП и СтП объединяются в единый процесс модельной разработки проекта сервиса (МРрП), а на третьем — объединяется функционал процессов $Л_1$ и $Л_2$ в единый процесс $Л_{1–2}$.

Первая методика обеспечивает мониторинг текущей активности сотрудника в сети Интернет и предоставляет ему рекомендации по улучшению стратегий информационного поиска. Методика реализована на основе методов машинного обучения и включает: наборы сценариев поведения сотрудника в информационном поиске; обоснование и программную реализацию анализа тем информационного поиска в каждом сценарии на основе модели тематического моделирования с аддитивной регуляризацией (Additive regularization of topic models, ARTM); метрики оценки сдвига темы информационного поиска, выполняемого сотрудником, от базового уровня, такие как когерентность и структура ядра. Последовательность обработки текстовых данных, просматриваемых сотрудником в ходе информационного поиска, в соответствии с разработанной методикой, представлена на рис. 2.

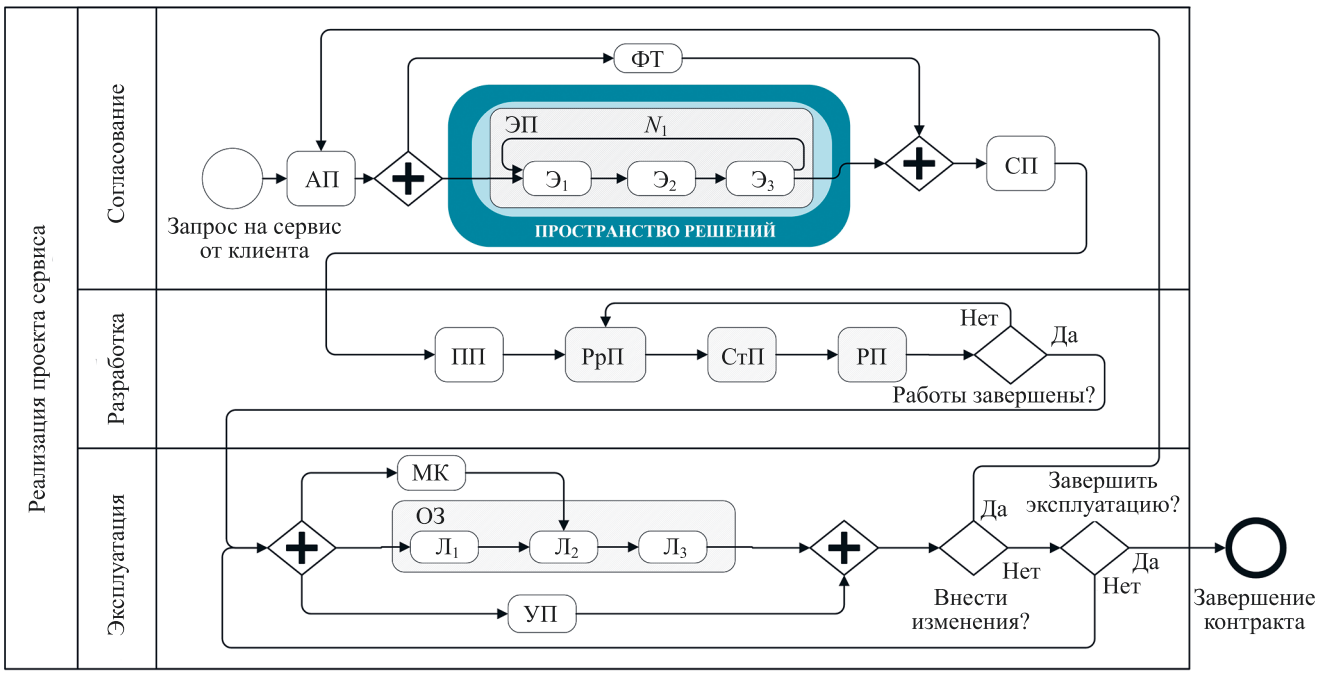
Методика применяется на этапе согласования сервиса (рис. 1, *a*) и реализуется введением специального компонента ПС (рис. 1, *b*). Данный компонент предоставляет сервис-инженеру, осуществляющему эскизное проектирование будущего сервиса, рекомендации по интернет-поиску, сформированные на опыте наиболее эффективных сотрудников, а также сервис-инженеру и его руководителю доступ к данным об общих метриках отклонения поисковой активности сотрудников от требуемой предметной области. Такое непрямое управление, с одной стороны, сохраняет самостоятельность сотрудника в выборе стратегии поиска, а с другой стороны, демонстрирует сотруднику уровень эффективности его поисковой стратегии, что позволяет, в итоге, сократить количество итераций при эскизном проектировании. Использованный подход заключается в сокращении объема пространства возможных решений при сохранении для сотрудника свободы технического творчества при их поиске.

Вторая методика базируется на организации обучения персонала настройке сетевой инфраструктуры. При этом обучение производится при решении реальных производственных задач с использованием специализированной модельной среды, максимально соответствующей реальной рабочей среде (рис. 3).

Методика реализована с помощью использования программного комплекса МС на этапе разработки сервиса (рис. 1, *b*), что позволяет объединить этапы РрП и СтП в единый этап МРрП без потери качества и с сокращением времени разработки. В результате сокращается время на разработку сетевой конфигурации и обеспечивается повышение квалификации персонала за счет неформального обучения на основе повседневного опыта профессиональной деятельности.

Третья методика предполагает применение автоматизированной диалоговой СС для ответа на профессиональные вопросы сотрудников. Система агрегирует в себе информацию из технической документации на обслуживаемые системы, а также экспертную информацию, полученную от наиболее компетентных технических специалистов. Для построения системы в

a



b

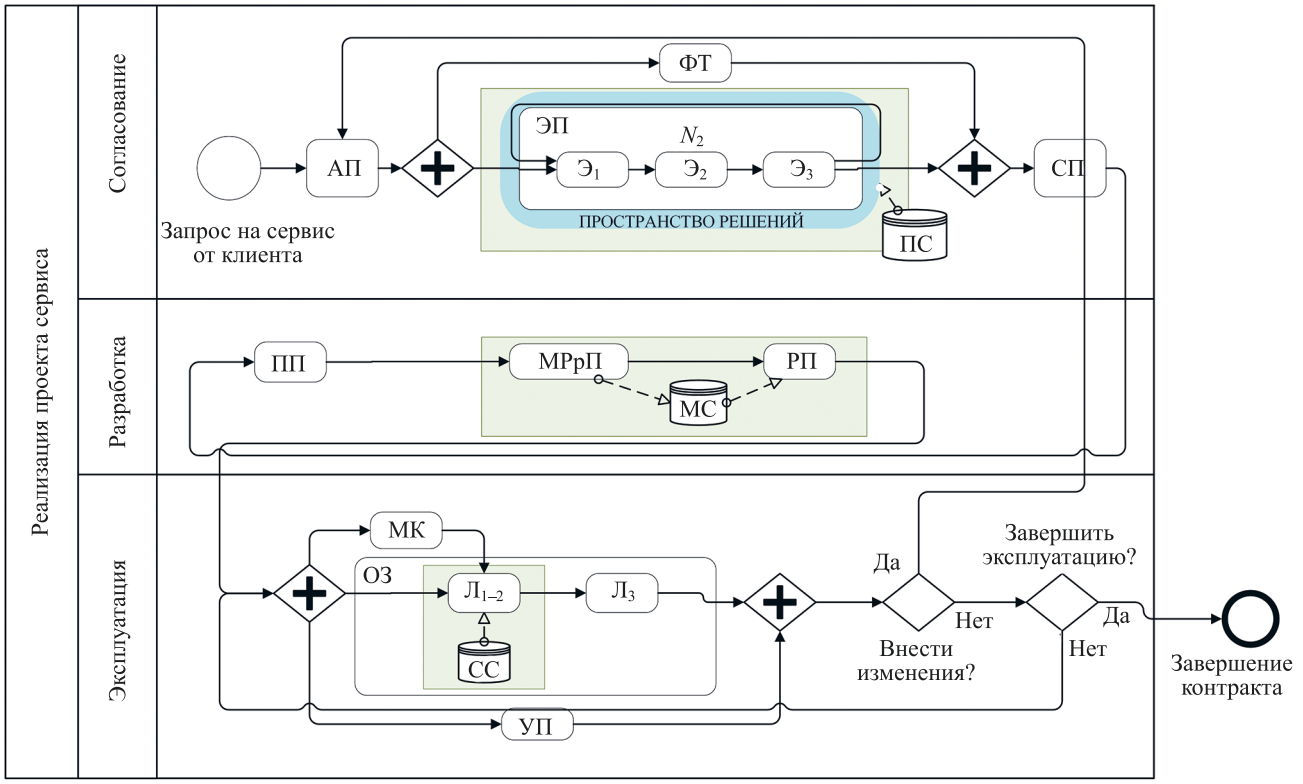


Рис. 1. Типовая схема реализации процесса предоставления IT-сервиса (*a*) и ее модификация с использованием разработанных методик (*b*)

Fig. 1. Typical scheme of implementation of the process of providing IT service (*a*) and its modification using the developed techniques (*b*)

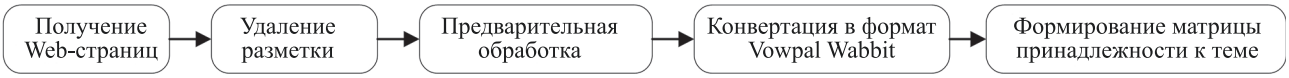


Рис. 2. Последовательность обработки текстовых данных при мониторинге текущей активности сотрудника в сети Интернет

Fig. 2. The sequence of processing text data during monitoring the current activity of an employee on the Internet

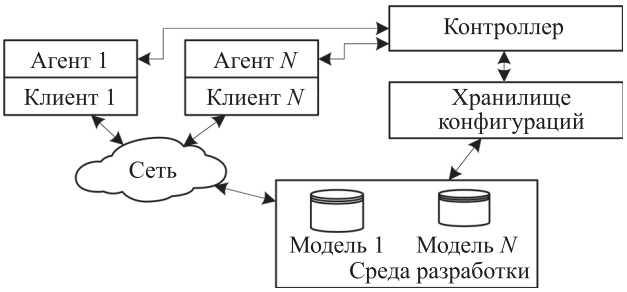


Рис. 3. Архитектура программного комплекса для неформального обучения персонала настройке сетевой инфраструктуры

Fig. 3. Architecture of the software package for informal training of personnel for configuring the network infrastructure

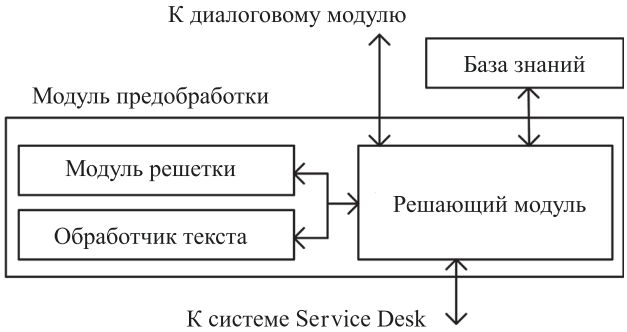


Рис. 4. Принципиальная архитектура компонента диалоговой справочной системы

Fig. 4. Basic architecture of the dialog system component

работе применены методы формального [19] и реляционного [20] концептуальных анализов, на основе которых построена концептуальная решетка, описывающая всю предметную область, в которой работает ВСОК, с учетом весов отношений между понятиями. На базе такой решетки осуществлено управление работой диалоговой СС, которая интегрирована с информационной системой технической поддержки (Service Desk). Принципиальная архитектура компонента диалоговой системы представлена на рис. 4.

Методика используется на этапе эксплуатации сервиса посредством введения диалоговой СС в виде двух модулей: диалогового и предобработки. Модуль предобработки, на основании анализа входящего обращения в службу технической поддержки, предоставляет сервис-инженеру рекомендованные статьи технической документации и инструкции, полезные для обработки заявки. Это позволяет сервис-инженеру выделить из общего когнитивного графа предметной области подграф, соответствующий информации, наиболее значимой для решения конкретной проблемы, и тем самым сократить время обработки заявки.

Программная реализация разработанного набора методик

Первая методика реализована в виде средства интеллектуального мониторинга эффективности персонала на основе учета его активности в сети Интернет, предназначенного для управления развитием сервис-инженеров. В процессе работы сервис-инженер использует доступ к ресурсам сети Интернет через корпоративную сеть для поиска необходимой информации по текущим рабочим проектам. Отметим, что доступ не ограничен тематическими фильтрами. Среди сервис-инженеров выбирается референтный сервис-инженер — сотрудник, на основании поведения которого при поиске информации в сети Интернет строится ядро темы. Выбор сотрудника осуществляется руководителем по признакам, связанным с эффективностью его работы, соблюдением сроков, квалификацией и т. п.

Руководитель отдела сервиса может управлять программным средством и получать информацию о характере поискового поведения сотрудников. На основании этих данных руководитель автоматизировано формирует рекомендации по поисковому поведению для развивающегося сервис-инженера отдела серви-

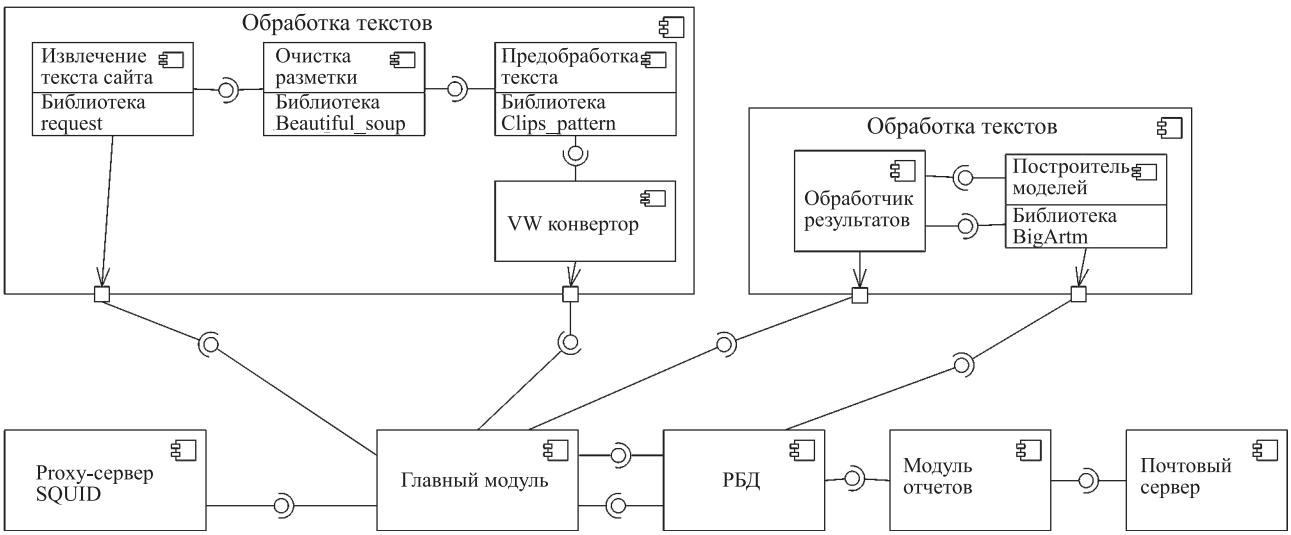


Рис. 5. Программная архитектура средства мониторинга текущей активности сотрудника в сети Интернет

Fig. 5. Software architecture for monitoring the employee's current activity on the Internet

са (сервис-инженера, для которого осуществляется управление развитием). Рекомендации представляют собой перечень тем просмотренных ресурсов с отметкой степени отличия от референтных и перечень рекомендованных Uniform Resource Locator (URL) по актуальным темам.

Программная архитектура средства представлена на диаграмме компонентов (рис. 5). Протоколирование работы осуществлено с помощью proxy-сервера «SQUID». Модуль обработки текстов выполняет извлечение текстов ресурсов с использованием библиотеки requests для Python, очистку разметки от тегов, скриптов и служебной информации с помощью библиотеки Beautiful_Soup, предварительную лингвистическую обработку с помощью библиотек Clips_pattern. Сырые данные для последующей обработки хранятся в формате Vowpal Wabbit (VW), конвертацию в который осуществляет специальный модуль, выполняющий построение моделей на основе библиотеки BigARTM. Для хранения данных используется реляционная база данных (РБД). Модуль обработчика результатов проводит подготовку данных для формирования отчетов. Модуль отчетов позволяет руководителю просмотреть, отредактировать, утвердить и разослать отчеты для каждого инженера отдела сервиса.

Программное исполнение второй методики реализовано в виде средства разработки сетевых конфигураций в специализированной модельной программной среде. Сотрудник может создать, модифицировать, копировать, удалить виртуальную инфраструктуру, провести тестирование инфраструктуры по достижимости узлов и сервисов с помощью набора тестов, получить отчет о тестировании и исправить ошибки. Это позволяет сервис-инженеру работать в безопасной, с точки зрения рисков ошибок, среде и итеративно достигать требуемый результат. Готовую и отлаженную конфигурацию можно экспортировать и передать на оборудование заказчика в ручном или полуавтоматическом режиме.

Программная архитектура средства разработки сетевых конфигураций в модельной среде представлена на диаграмме компонентов (рис. 6). Модуль проектирования позволяет добавить из библиотеки модели коммуникационных и целевых устройств, настроить связь между ними, распределить адресное пространство, указать необходимые сетевые сервисы. Конфигурация сетевых соединений построена с помощью Linux network namespaces (netns) и инфраструктуры виртуальных машин (ВМ). Модуль тестирования позволяет осуществить проверку достижимости сервисов и узлов и получать отчеты о проверке конфигурации. Экспорт конфигурации происходит специальным модулем в формате eXtensible Markup Language (XML). При хранении конфигураций предусмотрено версионирование. Если это возможно и целесообразно, конфигурация может быть перенесена на инфраструктуру заказчика автоматизировано. Для этого модуль переноса соединяется с модулем-агентом в сети заказчика, который конфигурирует целевые устройства. Отметим, что полностью автоматический перенос конфигурации не всегда возможен из-за ограничений целевых устройств и

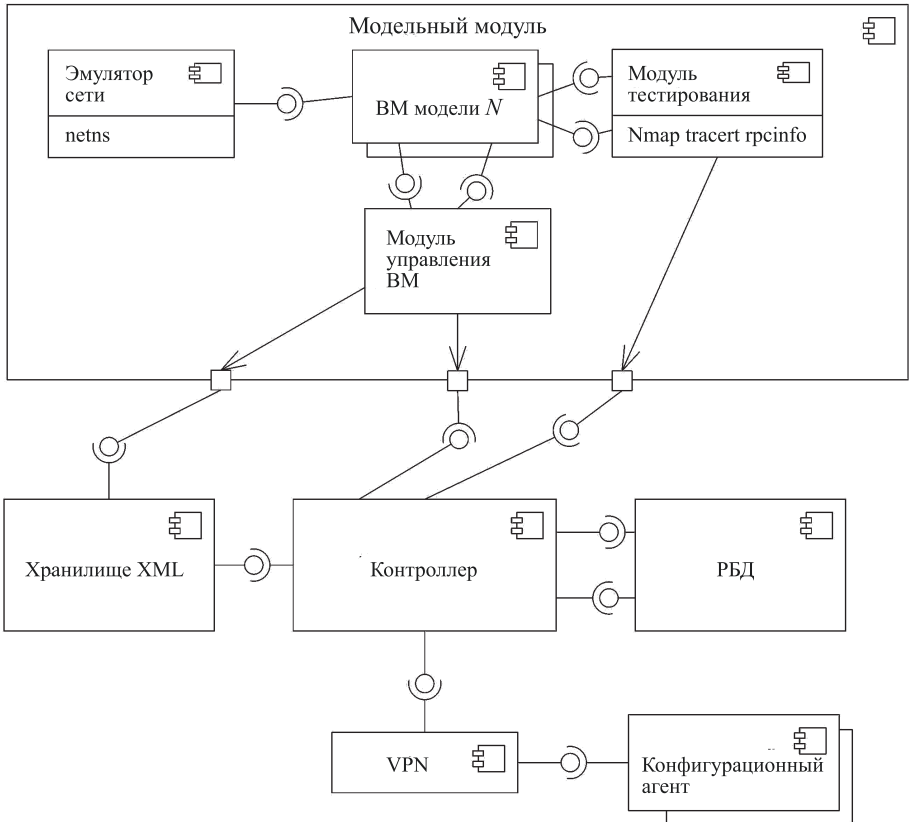


Рис. 6. Архитектура средства разработки сетевых конфигураций в модельной среде

Fig. 6. Architecture of the network configuration development tool in the model environment

политик безопасности заказчика. Коммуникация между контроллером и конфигурационным агентом осуществляется с использованием защищенного канала Virtual Private Network (VPN).

Программная реализация третьей методики представляет собой автоматизированную диалоговую систему для ответа на профессиональные вопросы сотрудников. Эта система предназначена для развития сотрудника технической поддержки и агрегирует в себе информацию из технической документации на обслуживаемые системы, а также экспертную информацию, полученную от наиболее компетентных технических специалистов. Входящие обращения клиентов обрабатываются системой, и сотрудник получает в дополнение к сообщению и его стандартной атрибуции классификацию и рекомендованные материалы из базы знаний, позволяющие быстрее разрешить обращение и повысить квалификацию сотрудника за счет обучения. Подготовку данных для работы диалоговой СС осуществляет эксперт, задача которого сводится к составлению и актуализации кросс-таблицы.

Программная архитектура системы показана на диаграмме компонентов (рис. 7). Работу с кросс-таблицами, включая формирование концептуальной решетки предметной области, поддерживает модуль решетки. Хранение данных о выделенных формальных концептах и их связь со статьями базы знаний реализовано в виде NoSQL базы MongoDB. Входящие обращения клиентов регистрируются в системе Service Desk. Запрос клиента обрабатывается модулем бизнес-логики, который вызывает языковой модуль для выделения набора слов для поиска наиболее подходящего формального концепта и, как следствие, списка актуальных для проблемы статей базы знаний. Взаимодействие с сотрудником технической поддержки реализовано в виде чат-бота, реализованного как часть интерфейса Service Desk системы. Вместе с оповещением о входящем обращении чат-бот сразу рекомендует тему и набор ссылок в базе знаний.

В качестве иллюстрации физического размещения компонентов систем приведена диаграмма размещения для диалоговой СС (рис. 8). Все компоненты развернуты в виртуальном окружении и взаимодействуют через стандартный HTTP. Это позволяет при необходимости масштабировать систему и реализовать независимость от конкретной аппаратной платформы. Данный подход применяется в программных реализациях методик ПС и МС.

Экспериментальная оценка эффективности разработанного набора методик на примере развития сервис-инженеров конкретной ВСОК

Оценка эффективности разработанного набора методик выполнена на примере развития сервис-инженеров реальной ВСОК, основной областью деятельности которой является предоставление сервисов корпоративной печати и управления IT-инфраструктурой заказчиков.

Экспериментальная оценка проведена на основании данных за четыре месяца. Выполнен анализ работы коллектива из 9 сервис-инженеров. Общее количество заявок за выбранный период составило 2948 шт. Заявки имели следующие атрибуты: тип заявки, сроки обработки, количество привлеченных сотрудников, должности сотрудников и количество итераций по заявке до ее закрытия.

В качестве показателей эффективности использованы ключевые показатели (KPI), традиционно рекомендуемые в методологии ITIL для оценки работы службы

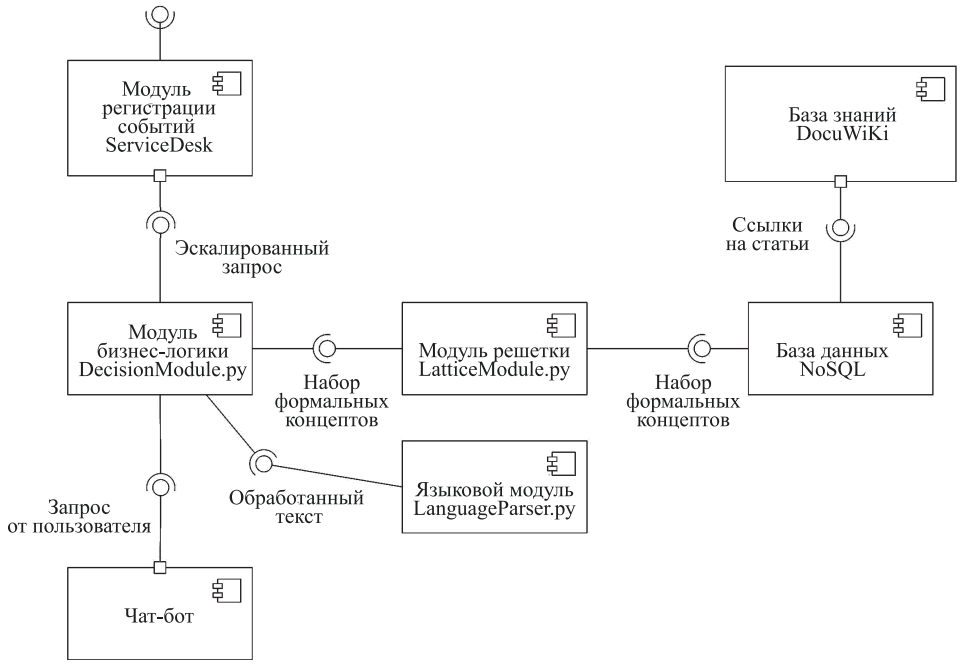


Рис. 7. Программная архитектура диалоговой справочной системы

Fig. 7. Software architecture of the dialog system

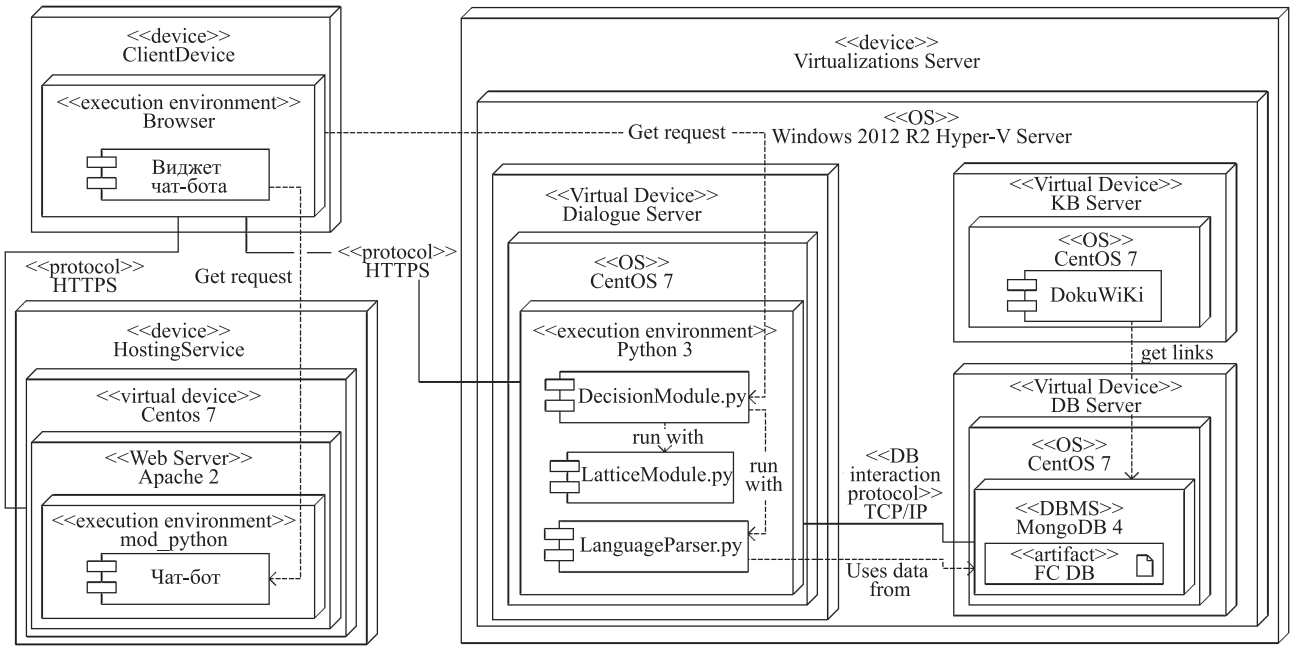


Рис. 8. Диаграмма размещения для диалоговой системы

Fig. 8. Installation diagram for the dialog system

клиентского сервиса и связанные с квалификацией и организацией обработки заявок (табл. 1).

Для оценки проверен анализ результатов работы отдела сервиса после внедрения систем, реализующих разработанные методики. Результаты приведены в табл. 2. В качестве референтных значений использованы усредненные данные за один предыдущий год, в течение которого значения KPI практически не менялись.

Из табл. 2 видно, что по всем использованным KPI наблюдается значительное улучшение показателей. Применение предложенных методик интеллектуализации управления развитием персонала высокотехнологичных сервис-ориентированных компаний и реализующих их программных комплексов позволило повысить эффективность работы сервис-инженеров по ключевым показателям от 31 до 54 %. Эти результаты правомерно связать с развитием личностных и профессиональных качеств сотрудников отдела сервиса.

Для дополнительного подтверждения эффективности предложенных методик выполнен анализ развития персонала отдела сервиса по отдельным сотрудникам. Для каждого сервис-инженера определено количество самостоятельно закрытых сотрудником заявок на обслуживание по отношению к общему количеству направленных ему заявок. В табл. 3 приведены усредненные данные за четыре месяца.

Из представленных экспериментальных данных видно увеличение диапазона технических задач, решаемых отдельным специалистом. Значительный рост личных показателей выявлен для заявок, относящихся к настройкам компьютерных сетей и сетевых серви-

Таблица 1. Ключевые показатели (KPI), использованные для оценки эффективности методик

Table 1. Key performance indicators (KPIs) used to evaluate the effectiveness of methods

Показатель	Содержание и влияющие показатели
Среднее время обработки заявки	Среднее время, затрачиваемое на обработку заявки. На показатель оказывает влияние уровень квалификации инженера, знание сервиса, доступ к информационно-справочной информации, эффективность коммуникации внутри проектной команды
Доля заявок, разрешенных на первом уровне поддержки	Доля заявок, обработанных без привлечения более компетентных специалистов или проектных команд, по отношению к общему числу заявок. На показатель влияет уровень квалификации персонала, его способность принимать квалифицированные решения и локализовать причину проблемы
Доля неудачных изменений от общего числа вносимых изменений	Доля изменений в конфигурацию сервиса закончилась неудачей с первого раза (не были выполнены, привели к потере функциональности). На параметр оказывает влияние качество подготовки изменений, тестирование, квалификация персонала
Соотношение запланированных и реальных сроков проекта	Отступление от запланированных сроков разработки и внедрения сервисов
Процент проваленных тестирований на сдачу услуг	Какой процент сервисов был возвращен на доработку заказчиком или интегратором в результате проведения тестирования на инфраструктуре заказчика

Таблица 2. Изменение средних значений ключевых показателей (KPI) при применении разработанного набора методик
Table 2. Changing the average values of key performance indicators (KPIs) using the developed set of methods

Показатель	Референтные значения	Период наблюдения, месяцы				Среднее значение за четыре месяца	Изменение KPI, %
		1	2	3	4		
Среднее время обработки заявки	2 ч 24 мин	2 ч 12 мин	1 ч 48 мин	1 ч 30 мин	1 ч 6 мин	1 ч 39 мин	31
Доля заявок, разрешенных на первом уровне поддержки	0,31	0,35	0,39	0,52	0,67	0,48	54
Доля неудачных изменений от общего числа вносимых изменений	0,39	0,31	0,27	0,11	0,09	0,19	51
Соотношение запланированных и реальных сроков проекта	0,7	0,7	0,79	0,93	1,4	0,95	35
Число не пройденных тестирований на сдачу услуг, %	30	29	18	16	9	18	40

Таблица 3. Доля самостоятельно закрытых сервис-инженерами заявок, %
Table 3. Share of applications closed by service engineers independently, %

Тип заявки	Период наблюдения, месяцев				Увеличение
	1	2	3	4	
Неисправность печатающей техники	80	81	84	84	4
Неисправность компьютера	35	50	64	72	37
Неисправность сети	20	29	55	65	45
Неисправность системного сервиса (СУБД и т. п.)	10	17	34	40	30
Настройка сетевой печати	60	67	70	71	11
Настройка компьютера	35	38	43	47	12
Настройка сети	20	30	38	48	28
Настройка системного сервиса (СУБД и т. п.)	5	18	39	44	39
Удаленная консультация по печатающей технике	80	81	84	85	5
Удаленная консультация по компьютерной технике	35	35	38	40	5
Удаленная консультация по сетевым настройкам	20	34	43	49	29
Удаленная консультация по системным сервисам	10	21	29	44	34

сов, поскольку это направление на момент внедрения предлагаемых методик было для сервис-инженеров новым, т. е. таким, в котором они изначально не имели достаточного опыта. Отметим, что перечисленные задачи отличаются предметной областью, спецификой и необходимостью использования не только профессиональных, но и личных качеств, что еще раз подтверждает эффективность разработанной методики.

Заключение

В работе рассмотрена проблема управления развитием персонала высокотехнологичных сервис-ориентированных компаний. Показано, что существующие подходы к управлению развитием персонала сложно адаптируются или практически неприменимы для специалистов высших квалификационных уровней. Для обеспечения эффективного, управляемого и параметризуемого введения в процесс обучения элементов неопределенности требуется интеллектуализация подходов обучения персонала.

На основе базовой модели процесса обучения, концептуальной модели развития персонала и современных теорий принятия решения в условиях неопределенности построена проблемно-ориентированная модель управления развитием персонала с возможностью ее параметризации. На основе теории грубых множеств разработан метод проблемно-ориентированного ранжирования средств развития персонала, выявлены наиболее значимые компетенции для сотрудников высших квалификационных уровней, определены подходы обучения на рабочем месте, наиболее востребованные высококвалифицированным персоналом и доступные для параметрического управления.

Разработан набор методик управления развитием персонала высших квалификационных уровней с ис-

пользованием интеллектуальных алгоритмов, в том числе технологии тематического моделирования с аддитивной регуляризацией (Additive Regularization Tematic Model), использования специальной среды для текущей разработки конфигурации информационных систем и анализа формальных концептов (Formal Concept Analysis). Методики реализуют управление поведением сотрудника при веб-поиске, использование специальной среды для текущей разработки конфигурации информационных систем и предоставление корпоративных знаний через автоматизированную диалоговую систему на основе анализа формальных концептов. Разработанные методики интеллектуализации управления развитием персонала высокотехнологичными сервис-ориентированными компаниями обеспечивают необходимые подходы к обучению на рабочем месте и покрывают все типовые этапы жизненного цикла сервиса.

Применение разработанных методик проиллюстрировано на примере процесса предоставления ИТ-сервиса. Подробно рассмотрена программная реализация разработанных методик в виде библиотеки программных средств.

Эффективность разработанного подхода проверена на основании отраслевых показателей оценки качества работы отдела сервиса, а также на базе анализа развития персонала отдела сервиса по отдельным сотрудникам. В условиях реальной производственной деятельности достигнуто повышение эффективности работы сервис-инженеров по ключевым показателям от 31 до 54 %, а также увеличение диапазона технических задач, решаемых отдельным специалистом. Показано, что применение разработанного подхода позволяет оперативно адаптировать квалификацию персонала в условиях быстрой смены производственных задач и рабочего окружения без отрыва от рабочего процесса.

Литература

1. Zakrzewska-Bielawska A. High technology company — concept, nature, characteristics // Proc. of the 8th International Conference on Management, Marketing and Finances. 2010. P. 93–98.
2. Wolf M., Terrell D. The high-tech industry, what is it and why it matters to our economic future // Beyond the Numbers: Employment and Unemployment. 2016. V. 5. N 8. P. 1–7.
3. Cacciattolo K. Defining workplace learning // European Scientific Journal. 2015. Special ed. V. 1. P. 243–250.
4. Hart J. Modern Workplace Learning 2019: A Framework for continuous improvement, learning and development [Электронный ресурс]. URL: https://www.modernworkplacelearning.com/cild/ (дата обращения: 25.02.2020).
5. Sweeney F. Workplace learning: The evolution of IT learning. March 2016 [Электронный ресурс]. URL: https://www.theceomagazine.com/business/management-leadership/workplace-learning-the-evolution-of-it-learning/ , свободный. Яз. англ. (дата обращения: 20.11.2021).
6. Dreyfus S.E., Dreyfus H.L. A five-stage model of the mental activities involved in directed skill acquisition. Washington, DC: Storming Media [Электронный ресурс]. URL: https://www.researchgate.net/publication/235125013_A_Five-Stage_Model_of_the_Mental_Activities_Involved_in_Directed_Skill_Acquisition, свободный. Яз. англ. (дата обращения: 20.11.2021).
7. Галинская Е.В., Иващенко А.А., Новиков Д.А. Модели и механизмы управления развитием персонала. М.: ИПУ РАН, 2005. 68 с.
8. Шестакова Е.В. Инновационные технологии обучения персонала. 2015 [Электронный ресурс]. URL: https://elib.osu.ru/bitstream/123456789/1230/1/2278-2283.pdf, свободный. Яз. рус. (дата обращения: 21.11.2021).
9. Кязимов К.Г. Управление человеческими ресурсами: профессиональное обучение и развитие: учебник для академического бакалавриата. 2-е изд., перераб. и доп. М.: Юрайт, 2019. 202 с.
10. Вялова Е.П., Квашнина Г.А., Федянин В.И. Динамическое моделирование и оптимизация управления производительностью инновационной организации с учетом профессиональной компетенции сотрудников // Системы управления и информационные технологии. 2007. № 4.1(30). С. 134–136.
11. Симонов В.П. Педагогический менеджмент: Ноу-хау в образовании. М.: Высшее образование, 2006. 368 с.
12. Jacobs R.L., Park Y. A proposed conceptual framework of workplace learning: Implications for Theory development and research in human resource development // Human Resource Development Review. 2009. V. 8. N 2. P. 133–150. https://doi.org/10.1177/1534484309334269
13. Принятие решений в неопределенности. Правила и предубеждения / под ред. Д. Канемана, П. Словика, А. Тверски. Харьков: Гуманитарный центр, 2021. 540 с.
14. Трегуб И.В., Горошникова Т.А. Имитационные модели принятия решений: учебное пособие. М.: Инфра-М, 2020. 193 с.

References

1. Zakrzewska-Bielawska A. High technology company – concept, nature, characteristics. *Proc. of the 8th International Conference on Management, Marketing and Finances*, 2010, pp. 93–98.
2. Wolf M., Terrell D. The high-tech industry, what is it and why it matters to our economic future. *Beyond the Numbers: Employment and Unemployment*, 2016, vol. 5, no. 8, pp. 1–7.
3. Cacciattolo K. Defining workplace learning. *European Scientific Journal*, 2015, Special ed., vol. 1, pp. 243–250.
4. Hart J. *Modern Workplace Learning 2019: A Framework for continuous improvement, learning and development*. Available at: https://www.modernworkplacelearning.com/cild/ (accessed: 25.02.2020).
5. Sweeney F. *Workplace learning: The evolution of IT learning. March 2016*. Available at: https://www.theceomagazine.com/business/management-leadership/workplace-learning-the-evolution-of-it-learning/ (accessed: 20.11.2021).
6. Dreyfus S.E., Dreyfus H.L. *A five-stage model of the mental activities involved in directed skill acquisition. Washington, DC: Storming Media*. Available at: https://www.researchgate.net/publication/235125013_A_Five-Stage_Model_of_the_Mental_Activities_Involved_in_Directed_Skill_Acquisition (accessed: 20.11.2021).
7. Galinskaia E.V., Ivashchenko A.A., Novikov D.A. *Models and Mechanisms for Personnel Development Management.* Moscow, ICS RAS, 2005, 68 p. (in Russian)
8. Shestakova E.V. *Innovative technologies for personnel training*. Available at: https://elib.osu.ru/bitstream/123456789/1230/1/2278-2283.pdf (accessed: 21.11.2021). (in Russian)
9. Kiazimov K.G. Human *Resources Management: Professional Training and Development*. Moscow, 2019, 202 p. (in Russian)
10. Vialova E.P., Kvashnina G.A., Fedianin V.I. Dynamic modeling and optimization of performance management in an innovative organization, taking into account professional competences of employees. *Sistemy upravlenija i informacionnye tehnologii*, 2007, no. 4.1(30), pp. 134–136. (in Russian)
11. Simonov V.P. *Pedagogical Management*. Moscow, Vysshee obrazovanie Publ., 2006, 368 p. (in Russian)
12. Jacobs R.L., Park Y. A proposed conceptual framework of workplace learning: Implications for Theory development and research in human resource development. *Human Resource Development Review*, 2009, vol. 8, no. 2, pp. 133–150. https://doi.org/10.1177/1534484309334269
13. *Judgment under uncertainty: heuristics and biases*. Ed. by D. Kahneman, P. Slovic, A. Tversky. Cambridge University Press, 1982, 555 p.
14. Tregub I.V., Goroshnikova T.A. *Decision Simulation Models*. Moscow, Infra-M, 2020, 193 p. (in Russian)
15. Užga-Rebrovs O. Knowledge representing features in rough sets theory. *Proc. 7th International Scientific and Practical Conference on*

15. Užga-Rebrovs O. Knowledge representing features in rough sets theory // Proc. 7th International Scientific and Practical Conference on Environment, Technology and Resources. 2009. P. 169–176. https://doi.org/10.17770/etr2009vol2.1045
16. Федорова Е.С., Лашманова Н.В. Применение комбинированных методов оценки в системе управления компетенциями инновационного предприятия // Дискурс. 2016. № 6. С. 56–59.
17. Hart J. Modern Workplace Learning 2020: How Modern Professionals Prefer to Learn [Электронный ресурс]. URL: www.modernworkplacelearning.com/cild/introduction/1-7-how-modern-professionals-prefer-to-learn/ (дата обращения: 20.06.2021).
18. The Official Introduction to the ITIL Service Lifecycle. London: TSO [Электронный ресурс]. URL: http://wikiitil.ru/books/00%20ITIL3%20Official%20Introduction.pdf, свободный. Яз. англ. (дата обращения: 18.10.2021).
19. Poelmans J., Ignatov D.I., Kuznetsov S., Dedene G. Formal concept analysis in knowledge processing: A survey on applications // Expert Systems with Applications. 2013. V. 40. N 16. P. 6538–6560. https://doi.org/10.1016/j.eswa.2013.05.009
20. Priss U. Associative and formal concepts // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2002. V. 2393. P. 354–368. https://doi.org/10.1007/3-540-45483-7_27

Environment, Technology and Resources, 2009, pp. 169–176. https://doi.org/10.17770/etr2009vol2.1045
16. Fedorova E.S., Lashmanova N.V. Application of combined methods of assessment in competence management system of innovative companies. *Discourse*, 2016, no. 6, pp. 56–59. (in Russian)
17. Hart J. *Modern Workplace Learning 2020: How Modern Professionals Prefer to Learn*. Available at: www.modernworkplacelearning.com/cild/introduction/1-7-how-modern-professionals-prefer-to-learn/ (accessed: 20.06.2021).
18. *The Official Introduction to the ITIL Service Lifecycle. London: TSO*. Available at: http://wikiitil.ru/books/00%20ITIL3%20Official%20Introduction.pdf (accessed: 18.10.2021).
19. Poelmans J., Ignatov D.I., Kuznetsov S., Dedene G. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 2013, vol. 40, no. 16, pp. 6538–6560. https://doi.org/10.1016/j.eswa.2013.05.009
20. Priss U. Associative and formal concepts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2002, vol. 2393, pp. 354–368. https://doi.org/10.1007/3-540-45483-7_27

Автор

Береснев Артем Дмитриевич — старший преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57202210221, https://orcid.org/0000-0002-4646-6856, artem.beresnev@itmo.ru

Author

Artem D. Beresnev — Senior Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57202210221, https://orcid.org/0000-0002-4646-6856, artem.beresnev@itmo.ru

Статья поступила в редакцию 17.12.2021
Одобрена после рецензирования 13.01.2022
Принята к печати 30.01.2022

Received 17.12.2021
Approved after reviewing 13.01.2022
Accepted 30.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

MODELING AND SIMULATION

doi: 10.17586/2226-1494-2022-22-1-138-146
УДК 629.12. 053

Исследование эффективности работы системы коррекции магнитного компаса

Дмитрий Геннадиевич Грязин[1,2], Татьяна Владимировна Падерина[1,2]✉

[1,2] АО «Концерн «ЦНИИ «Электроприбор», Санкт-Петербург, 197046, Российская Федерация
[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] volnagdg@yandex.ru, https://orcid.org/0000-0003-3562-6815
[2] paderinata@gmail.com✉, https://orcid.org/0000-0001-5293-6351

Аннотация

Предмет исследования. В условиях плавания судна при воздействии качки появляются динамические погрешности компаса, обусловленные влиянием перераспределенных магнитных масс судна, а также воздействием центростремительного и тангенциального ускорений. Влияние этих погрешностей может быть скомпенсировано введением в измерительную схему компаса системы коррекции, в которой используется один гироскопический датчик угловой скорости. **Метод.** Метод коррекции основан на применении в схеме гироскопического датчика угловой скорости с вертикально расположенной осью чувствительности. Формируемый сигнал представляет собой разность выходного показания магнитного компаса и проинтегрированного сигнала датчика угловой скорости, который не чувствителен к действию переносного ускорения и перераспределенных магнитных масс при крене судна. Полученная разность будет содержать погрешность компаса от воздействия качки, которая в дальнейшем компенсируется в выходном сигнале компаса. Выбор параметров динамических звеньев схемы для реализации системы коррекции и исследования ее работы выполнен методом имитационного моделирования с использованием предложенного аналитического выражения для погрешности магнитного компаса от воздействия качки. Выполнено сравнение результатов моделирования с результатами экспериментальных исследований компаса, проведенных с использованием специализированного стенда, который имитирует рыскание судна и позволяет изменять значение вектора напряженности магнитного поля Земли при воздействии на магнитную систему компаса. **Основные результаты.** Экспериментальные исследования компаса показали, что коэффициент коррекции погрешности от качки при изменении периодов угловых колебаний объекта от 6 до 28 с, который характеризует степень подавления качки, находится в диапазоне 0,16–0,48 (среднее значение 0,35), а при проведении моделирования среднее значение коэффициента — 0,21. Заниженная оценка коэффициента коррекции при моделировании обусловлена отсутствием учета динамических свойств картушки компаса и зависит от соотношения периодов собственных колебаний картушки и периодов возмущающих воздействий. **Практическая значимость.** Полученные результаты подтвердили высокую эффективность рассматриваемой системы коррекции компаса, а также требуемое качество разработанного специализированного стенда для оценки его работы. Результаты исследования могут найти применение при разработке современных магнитных компасов, для обеспечения высокой точности указания курса за счет использования предлагаемой системы коррекции.

Ключевые слова

магнитный компас, качка судна, система коррекции, гироскопический датчик угловой скорости

Благодарности

Работа выполнена при поддержке гранта РФФИ № 20-08-00265. Авторы выражают благодарность сотрудникам АО «Концерн «ЦНИИ «Электроприбор» К.С. Гороховскому и И.В. Сергачёву, участвовавшим в разработке стенда рыскания и проведении исследования характеристик системы коррекции магнитного компаса на этом стенде.

Ссылка для цитирования: Грязин Д.Г., Падерина Т.В. Исследование эффективности работы системы коррекции магнитного компаса // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 138–146. doi: 10.17586/2226-1494-2022-22-1-138-146

© Грязин Д.Г., Падерина Т.В., 2022

A study of the efficiency of the magnetic compass correction system

Dmitriy G. Gryazin[1,2], Tatiana V. Paderina[1,2]✉

[1,2] Concern CSRI Elektropribor, JSC, Saint Petersburg, 197046, Russian Federation
[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] volnagdg@yandex.ru, https://orcid.org/0000-0003-3562-6815
[2] paderinata@gmail.com✉, https://orcid.org/0000-0001-5293-6351

Abstract

In sailing conditions under the rolling, dynamic compass errors may appear due to the influence of the redistributed magnetic masses of the ship, as well as of centripetal and tangential accelerations. The influence of these errors can be compensated by introducing a correction system into the measuring circuit of the compass, which uses one gyroscopic angular rate sensor. The correction method is based on the use of a gyroscopic angular rate sensor with a vertical axis of sensitivity in the circuit. The generated signal is the difference between the output reading of the magnetic compass and the integrated signal of the angular rate sensor, which is insensitive to the action of translational acceleration and redistributed magnetic masses during the ship's heel. The resulting difference will contain the compass error from the rolling effect, further compensated for in the compass output signal. The choice of the parameters of the dynamic links of the circuit for the implementation of the correction system and the study of its operation was carried out by simulation method using the proposed analytical expression for the error of the magnetic compass from the rolling effect. Comparison of the simulation results with the results of experimental studies of the compass was carried out using a specialized stand that simulates the yaw of a ship and allows one to change the value of the vector of the Earth's magnetic field, when exposed to the magnetic system of the compass. Experimental studies of the compass showed that the pitching error correction coefficient is in the range 0.16-0.48 (average value is 0.35), when the object's angular oscillation periods change from 6 to 28 s, which characterizes the degree of pitching suppression, and when modeling, the average value of the coefficient is 0.21. The underestimation of the correction factor in modeling is due to the lack of taking into account the dynamic properties of the compass rose and depends on the ratio of the periods of natural oscillations of the compass and the periods of disturbing influences. The results confirmed the high efficiency of the considered compass correction system and the required quality of the developed specialized stand for evaluating its work. The results of the study can be used in the development of modern magnetic compasses to ensure high accuracy of directional guidance through the use of the proposed correction system.

Keywords

magnetic compass, ship pitching, compass correction system, gyroscopic angular rate sensor

Acknowledgements

This work was supported by the RFBR grant No. 20-08-00265. The authors express their gratitude to the employees of State Research Center of the Russian Federation Concern CSRI Elektropribor, JSC, K.S. Gorokhovsky and I.V. Sergachev, who participated in the development of the yaw stand and the study of the characteristics of the magnetic compass correction system on this stand.

For citation: Gryazin D.G., Paderina T.V. A study of the efficiency of the magnetic compass correction system. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 138–146 (in Russian). doi: 10.17586/2226-1494-2022-22-1-138-146

Введение

Развитие навигационного приборостроения привело к появлению на кораблях и судах значительного числа навигационных приборов, использующих новейшие технологии. К их числу следует отнести электронные картографические навигационные системы, спутниковые GPS-компасы и др. По требованиям международной конвенции по безопасности человеческой жизни на море — SOLAS, которые отражены в правилах по оборудованию морских судов Российского морского регистра судоходства[1], магнитным компасом (МК) должно быть оборудовано каждое судно. Аналогичные требования выдвигаются и к речным судам. Традиционно эксплуатация МК на судне сопровождается периодическими девиационными работами, направленными на компенсацию влияния намагниченности судовой стали на показания МК. Эти работы проводятся при условии влияния на чувствительный элемент МК конкретного значения горизонтальной и вертикальной составляющих магнитного поля Земли, а их регулярность зависит от квалификации экипажа. Методика девиационных работ хорошо известна[2], а их результаты контролируются инспекторами регистров. Отметим, что работам по устранению девиации посвящено множество исследований, в частности [1–4], однако все они направлены на компенсацию статической погрешности МК. В условиях плавания судна происходит изменение значения составляющих магнитного поля Земли, приводящее к перемагничиванию магнитомягкого судового железа, кроме того, на качке, сопровождающей плавание судна, влияние судовых магнитных сил постоянно изменяется [5–7]. На величину динамической погрешности МК влияют центростремительное и тангенциальное ускорения при воздействии качки, ускорения от поперечно горизонтальных колебаний судна, исследованием влияния которых в свое время занимался Н.Ю. Рыбалтовский [8]. Эту совокупность ускорений далее определим

[1] Правила по оборудованию морских судов. Часть V. Навигационное оборудование. Российский морской регистр судоходства. Санкт-Петербург, 2020, 246 с.

[2] Правила штурманской службы №38. Девиационные работы на кораблях и судах военно-морского флота. Главное управление навигации и океанографии МО РФ. 1994.

как переносные, в основном обусловленные тем, что компас установлен на некотором плече относительно центра качания судна. Указанные погрешности при плавании судна в высоких широтах в условиях, когда горизонтальная составляющая магнитного поля много меньше вертикальной, могут приводить к значительной ошибке в измерении и индикации курса МК [9, 10]. В умеренных широтах динамические погрешности приводят к избыточному рысканию картушки на качке, которое не отражает действительного значения рыскания судна [11].

Если на крупнотоннажных судах главный судовой МК и остается резервным средством, используемым только в аварийных ситуациях, то на судах водоизмещением до 150 т он является основным курсоуказателем и встраивается в систему авторулевого [12–14]. Отметим, что активизация судоходства в арктической зоне[1] делает задачу снижения динамических погрешностей МК еще более актуальной.

Влияние динамических погрешностей может быть скомпенсировано с помощью введения в измерительную схему МК системы коррекции, в которой используется один гироскопический датчик угловой скорости. Цель работы — разработка системы коррекции магнитного компаса и оценка эффективности ее работы с использованием результатов имитационного моделирования и стендовых испытаний. Испытания проведены на специализированном стенде, имитирующем рыскание судна с заданной частотой, а также позволяющем изменить значение естественного вектора напряженности магнитного поля Земли, воздействующего на магнитную систему компаса.

Динамические погрешности магнитного компаса

Магнитная система компаса в общем случае представляет собой систему со сосредоточенными параметрами. Рассмотрим работу судового МК, в котором картушка помещена в котелок и покоится в вязкой жидкости [15, 16] (рис. 1).

Динамические характеристики МК при синусоидальном рыскании судна с частотой ω опишем следующим уравнением второго порядка:

$$(J_{\text{В}} + \lambda_{\text{В}})\ddot{\varphi} + W_{\text{В}}\dot{\varphi} + (M\mu H_{\Gamma} + N)\sin\varphi = A\sin(\omega_{\text{К}} + \varphi_{\Phi}),$$

где φ — угол разворота картушки при рыскании; $J_{\text{В}}$ — момент инерции картушки относительно вертикальной оси; $\lambda_{\text{В}}$ — присоединенный момент инерции жидкости при развороте картушки относительно вертикальной оси; $W_{\text{В}}$ — коэффициент гидродинамического сопротивления развороту картушки при ее развороте относительно вертикальной оси; M — магнитный момент картушки; μ — магнитная (постоянная) проницаемость в вакууме; H_{Γ} — горизонтальная составляющая магнитного поля Земли; N — момент сил сухого трения; A, $\omega = 2\pi/T$, T и φ_{Φ} — амплитуда, круговая частота, период и фаза вынужденных колебаний картушки под воздействием рыскания соответственно. Коэффициент $W_{\text{В}}$ — функция от угловой скорости разворота картушки, однако в гидродинамике его принято обозначать одной переменной. Он определяет демпфирующую силу и всегда имеет знак, противоположный угловой скорости разворота картушки. Горизонтальная составляющая магнитного поля Земли, определяющая направляющую силу МК, изменяется в различных местах Земного шара от нуля на магнитных полюсах до значения 32 А/м у южной оконечности Азии.

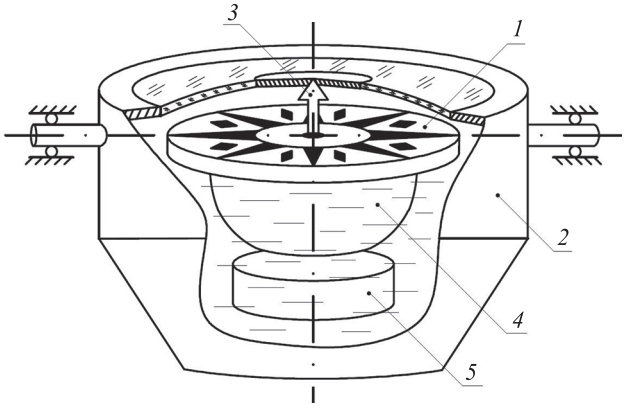


Рис. 1. Котелок судового магнитного компаса:
1 — картушка; *2* — котелок с жидкостью; *3* — игла;
4 — кольцевая магнитная система;
5 — магниторезистивный преобразователь

Fig. 1. Bowler of a ship's magnetic compass: *1* — car; *2* — pot with liquid; *3* — needle; *4* — ring magnetic system; *5* — magnetoresistive transducer

При отклонении судна от горизонтального положения возникает креновая девиация, величина которой зависит от проекции вертикальной составляющей индукции (напряженности) магнитного поля судна $H_{\text{ВС}}$ и определяется величиной крена и дифферента на плоскость картушки МК. Таким образом, при крене судна на угол θ возникает добавочная сила [17]:

$$F_{\text{К}} = ((e - k)H_{\text{В}} - H_{\text{ВС}})\theta,$$

где e и k — постоянные коэффициенты (параметры Пуассона), характеризующие тензор влияния индуктивной намагниченности, которая зависит от широты места и перемагничивания магнитомягкого железа; $H_{\text{В}}$ и $H_{\text{ВС}}$ — проекции вертикальной составляющей вектора магнитного поля Земли и постоянной намагниченности корабельных конструкций в корабельной системе координат.

Добавочная сила $F_{\text{К}}$ вызывает дополнительную погрешность $\Delta H_{\text{К}}$ в измерениях горизонтальной составляющей поля Земли [17]. В случае значительных углов бортовой качки и одновременного воздействия килевой, значения $F_{\text{К}}$ и $\Delta H_{\text{К}}$ могут существенно увеличиться. Значение дополнительной погрешности зависит от знака угла крена и на качке имеет знакопеременный характер, в связи с этим погрешность проявляется в виде дополнительного угла рыскания показаний МК,

[1] Указ Президента РФ от 26 октября 2020 г. №645 «О Стратегии развития Арктической зоны Российской Федерации и обеспечения национальной безопасности на период до 2035 года» [Электронный ресурс]. Режим доступа: https://www.garant.ru/products/ipo/prime/doc/74710556, своб., яз. рус. (дата обращения: 08.09.2021).

пропорционального углу качки. На курсах 0 и 180° она максимальна, и может достигать 3° на каждый градус крена [18], а на курсах 90 и 270° равна нулю. При изменении магнитной широты места величина $F_К$ будет изменяться. В работе [19] академика А.Н. Крылова креновая девиация пропорциональна тангенсу угла магнитного наклонения в точке местоположения судна и компенсируется специальным магнитом, установленным в нактоузе. Для компенсации креновой девиации по мере изменения широты плавания должно изменяться и положение магнита, однако этой работой экипаж, как правило, не занимается. В случае выполнения текущих работ по компенсации креновой девиации с применением судового инклинатора, ее остаточное значение обязательно будет присутствовать в связи с изменением индуктивной намагниченности судового металла на текущей широте и намагниченности при перемещении груза на палубе.

Конструкция современных МК позволяет ослабить влияние ускорений качки на чувствительный элемент компаса, но не исключает его полностью [16]. Для исключения влияния ускорения котелок МК устанавливается в двуосном карданове подвесе и представляет собой короткопериодный маятник. В свою очередь картушка МК с магнитной системой, помещенная в котелок с жидкостью (рис. 1), также является маятником, период колебаний которого составляет несколько десятков секунд. Под воздействием переносного ускорения a картушка МК отклоняется от плоскости горизонта, и ее наклон вызывает появление проекции ΔB вертикальной составляющей индукции магнитного поля B, действующей на магнитную систему МК [18]:

$$\Delta B = B\sin(a/g)К_М,$$

где g — ускорение силы тяжести; $К_М$ — магнитный курс.

Суммарная динамическая погрешность МК от качки $\delta_р$, обусловленная действием креновой девиации и переносного ускорения, зависит от погрешностей ΔB и $\Delta H_К$ в измерении горизонтальной составляющей магнитного поля Земли.

Учитывая приведенную в [8] формулу для погрешности МК от влияния переносного ускорения, а также от действия $\Delta H_К$, суммарную погрешность $\delta_р$ можно представить в виде:

$$\mathrm{tg}\delta_р = \left(\frac{h}{l_0} + \frac{\Delta H_К}{H_Г}\right)\mathrm{tg}I\cos К_М\cdot\theta, \qquad (1)$$

где h — высота картушки компаса над осью качаний судна; l_0 — длина маятника с периодом колебания $\tau = 2\pi\sqrt{l_0/g}$, соответствующим периоду колебаний судна; I — угол магнитного наклонения.

Выражение (1) использовано при проведении имитационного моделирования работы системы коррекции.

Погрешность $\delta_р$ пропорциональна тангенсу магнитного наклонения, значение которого резко возрастает с увеличением широты места. Рост погрешности $\delta_р$ в высоких широтах на фоне естественного снижения горизонтальной составляющей магнитного поля Земли может привести к недопустимой погрешности компаса.

Влияние погрешности $\delta_р$ может быть скомпенсировано в случае измерения углов рыскания гироскопическим датчиком, не подверженным влиянию переносного ускорения и перераспределенных магнитных масс при крене судна.

Установку гироскопического датчика угловой скорости (ДУС) с вертикальной осью чувствительности по конструктивным соображениям целесообразно производить на котелке компаса. Погрешность измерения угловой скорости рыскания ДУС из-за наклона котелка составит $\Delta\dot{\gamma}_р = \dot{\gamma}_р(-\delta\theta_к\sin\theta_к\cos\psi_к - \delta\psi_к\sin\psi_к\cos\theta_к)$, где $\dot{\gamma}_р$ — угловая скорость рыскания; $\delta\theta_к$ и $\delta\psi_к$ — текущие значения углов наклона котелка по крену и дифференту. Поскольку конструкция МК обеспечивает наклоны котелка на углы $\theta_к = \psi_к = \pm 10°$ [16], максимальное значение погрешности $\Delta\dot{\gamma}_р$ при условии $\delta\theta_к = \theta_к$ и $\delta\psi_к = \psi_к$ будет равно: $\Delta\dot{\gamma}_р = -\dot{\gamma}_р(\theta_к^2 + \psi_к^2)$, откуда следует, что эта погрешность не превышает 3 % от величины $\dot{\gamma}_р$. Погрешность $\Delta\dot{\gamma}_р$ является незначительной по сравнению с погрешностью $\delta_р$, зависящей от тангенса угла магнитного наклонения.

Система коррекции магнитного компаса

На рис. 2 приведена блок-схема системы коррекции МК [20], входными сигналами которой являются сигналы датчика магнитного курса (ДМК) и ДУС.

Идея комплексирования информации инерциальных датчиков и магнитометров встречается во многих современных публикациях, например [21, 22]. В этих публикациях магнитометры используются, как правило, для коррекции ряда выходных параметров объекта, в частности параметров ориентации, выработанных инерциальными датчиками и имеющих, в силу этого, накапливающиеся со временем погрешности.

В рассматриваемой системе коррекции компаса решается другая задача — с помощью ДУС корректируется сигнал МК, при этом для формирования разностного измерения по угловой информации выходной сигнал ДУС интегрируется. Полученное таким образом измерение, содержащее погрешность компаса $\delta_р$, пропускается через входящий в систему коррекции фильтр высоких частот (ФВЧ) с целью устранения, прежде всего, постоянной погрешности ДУС. Результатом работы системы коррекции является определение погрешности и исключение ее из показаний компаса. Благодаря системе коррекции МК приобретает свойства гиромагнитного компаса, представляющего собой магнитный курсоуказатель с гироскопическим устройством, обеспечивающим так же, как и система коррекции в МК, сглаживание показаний магнитного чувствительного элемента [23].

Работа системы коррекции включает следующие действия (рис. 2).

1. Выработку с помощью ДМК сигнала о мгновенных значениях курса в горизонтальной системе координат, состоящего из магнитного курса $К_М$, угла рыскания $\gamma_р$ и погрешности $\delta_р$: $К_М + \gamma_р + \delta_р$, при этом

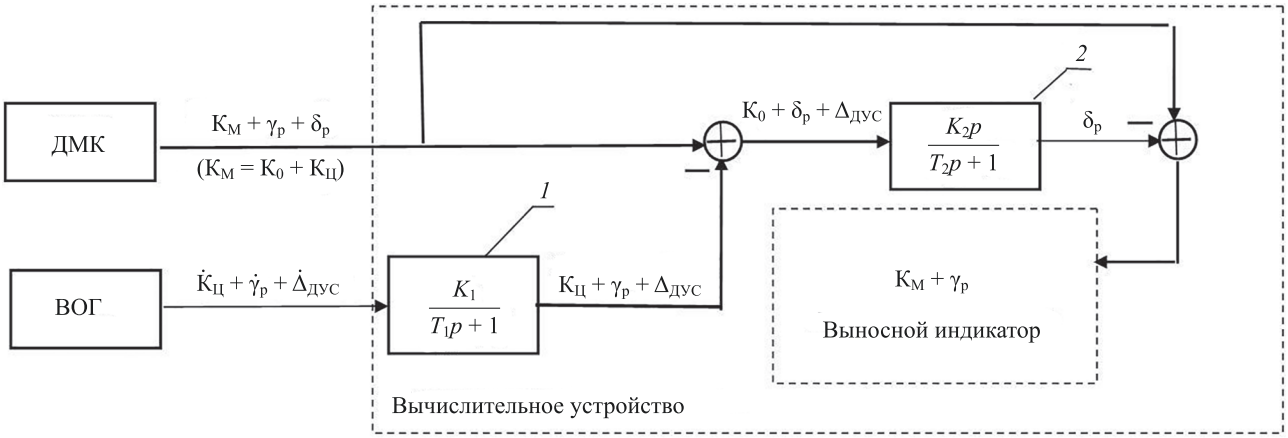


Рис. 2. Блок-схема системы коррекции магнитного компаса: *1* — интегратор $\frac{K_1}{T_1 p + 1}$; *2* — фильтр высоких частот $\frac{K_2 p}{T_2 p + 1}$ (где K_1, K_2 — коэффициенты усиления, T_1, T_2 — постоянные времени)

Fig. 2. Block diagram of the magnetic compass correction system: *1* — integrator $\frac{K_1}{T_1 p + 1}$; *2* — high-pass filter $\frac{K_2 p}{T_2 p + 1}$ (where K_1, K_2 — the gain, T_1, T_2 — the time constant)

$К_М = К_0 + К_Ц$, где $К_0$ — постоянное значение угла магнитного курса, $К_Ц$ — угол циркуляции объекта.

2. Выработку с помощью ДУС сигнала угловой скорости, состоящего из угловых скоростей изменения курса $\dot{К}_Ц$, угла рыскания $\dot{\gamma}_р$, погрешности ДУС $\dot{\Delta}_{ДУС}$, характеризуемой его дрейфом: $\dot{К}_Ц + \dot{\gamma}_р + \dot{\Delta}_{ДУС}$.
3. Интегрирование сигнала ДУС с помощью апериодического звена $\frac{K_1}{T_1 p + 1}$ при нулевом начальном условии и формирование на выходе интегратора сигнала $К_Ц + \gamma_р + \Delta_{ДУС}$.
4. Вычисление разности между выходными сигналами ДМК и интегратора с целью формирования сигнала $К_0 + \delta_р + \Delta_{ДУС}$.
5. Фильтрацию низкочастотной погрешности $\Delta_{ДУС}$ вместе с сигналом $К_0$ с помощью ФВЧ $\frac{K_2}{T_2 p + 1}$.
6. Исключение из значений магнитного курса погрешности $\delta_р$ путем вычисления разности между выходным сигналом ДМК и сигналом с выхода ФВЧ, пропорциональным погрешности $\delta_р$.

Особенность системы коррекции — использование в качестве ДУС волоконно-оптического гироскопа (ВОГ), дрейф смещения нуля которого $\dot{\Delta}_{ДУС}$ от пуска к пуску характеризуется случайной постоянной, в силу чего его влияние может быть устранено с помощью применения ФВЧ.

Параметры динамических звеньев схемы (рис. 2) имеют следующие численные значения:

— T_1 = 7200 с, что определяется необходимостью сокращения времени переходного процесса и величины перерегулирования при интегрировании сигналов ДУС, особенно при маневрировании судна; коэффициент усиления K_1 при таком значении T_1 равен 7200;
— T_2 = 40 с, что определяется необходимостью передачи без искажений сигнала $\delta_р$, зависящего от периодов бортовой и килевой качек, значения которых для большинства типов судов[1] находятся, соответственно, в диапазонах 6–28 с и 3–14 с; коэффициент усиления K_2 также равен 40.

Моделирование работы системы коррекции магнитного компаса

С целью оценки характеристик МК с системой коррекции выполнено моделирование ее работы в среде MATLAB (Simulink). Общая схема моделирования приведена на рис. 3.

При моделировании углы рысканья и крена судна задавались в виде гармонических колебаний [24]:

$$К = К_0 + \gamma_р,\ \gamma_р = A_р \sin[(2\pi/T_р)t + \psi_р],$$
$$\theta = A_\theta \sin[(2\pi/T_\theta)t + \psi_\theta],$$

где $A_р$ и A_θ — амплитуды гармонических колебаний углов рысканья и крена судна; $T_р$, T_θ, $\psi_р$, ψ_θ — периоды и фазы соответствующих колебаний (фазы $\psi_р$, ψ_θ = 0).

Заданные при формировании пропорциональной углам крена погрешности $\delta_р$, значения коэффициентов $\frac{\Delta H_К}{H_Г}$ и $\frac{h}{l_0}$ соответственно составляют 0,05 и 1, $\mathrm{tg}I = 3$ (для широты 60°).

Проведенное моделирование системы коррекции для различных условий углового движения судна подтвердило ее эффективность при использовании ВОГ тактического класса точности с характеристикой смещения нуля от пуска к пуску, заданного случайной постоянной со среднеквадратическим отклонением $\sigma = 3°/ч$.

Характеристикой эффективности системы коррекции является коэффициент подавления погрешности

[1] Благовещенский С.Н., Холодилин А.Н. Справочник по статике и динамике корабля. Т. 2. Динамика (качка) корабля. Л.: Судостроение, 1976. 178 с.



Рис. 3. Общая схема моделирования работы системы коррекции магнитного компаса

Fig. 3. General scheme for modeling the operation of the magnetic compass correction system

от качки $К_П$, который рассчитывается следующим образом:

$$К_П = \max|DK_M|/\max|\delta_p|,$$

где $\max|DK_M|$ — максимальное значение абсолютной величины погрешности системы коррекции DK_M, рассчитанное на интервале времени, составляющем 3–5 периодов колебаний углов крена; DK_M — разность выходного сигнала МК с системой коррекции и истинного значения магнитного курса; $\max|\delta_p|$ — максимальное значение абсолютной величины погрешности δ_p, рассчитанное на интервале времени, составляющем 3–5 периодов колебаний углов крена.

При моделировании задавалось постоянное значение курса $К_0 = 0°$, рыскание судна с периодом 5–10 с и амплитудой — 5°, бортовая качка с периодами 6–28 с и амплитудой — 20°.

Как показывают результаты моделирования, коэффициент подавления погрешности от качки $К_П$ зависит от численных значений периодов крена и имеет детерминированное значение для каждого заданного периода крена. В таблице приведены значения коэффициента $К_П$ для различных численных значений периодов крена.

Таблица. Зависимость коэффициента подавления погрешности от качки $К_П$ от значений периодов крена

Table. Dependence of the coefficient of suppression of error from pitching on the values K_s for the roll periods

Период крена, с	Коэффициент $К_П$
28,0	0,25
20,0	0,24
15,0	0,22
12,0	0,20
10,0	0,18
9,0	0,17
8,0	0,15
7,0	0,13
6,0	0,12

Видно, что значения коэффициентов подавления погрешности, приведенные в таблице, показывают хорошее качество предложенной системы коррекции.

Моделирование работы МК с системой коррекции проведено и в условиях циркуляции судна. Выполнен анализ результатов моделирования работы МК при следующих условиях: в отсутствии качки при повороте судна с начального значения курса 0° до значения курса 90° с постоянной угловой скоростью от 3 до 6°/с. В результате следует, что погрешность работы МК с системой коррекции составляет ±0°22′12″, при этом время переходного процесса не превышает 140 с.

Стенд рысканья для оценки эффективности работы системы коррекции магнитного компаса

Эффективность работы системы коррекции МК, позволяющей снизить влияние динамических погрешностей в условиях качки судна, может быть экспериментально оценена с помощью специализированного стенда. Этот стенд должен имитировать рыскание судна с заданной частотой. Кроме того, стенд должен изменять значение горизонтальной составляющей естественного вектора напряженности магнитного поля Земли, воздействующего на магнитную систему МК.

Кинематическая схема и внешний вид стенда представлены на рис. 4.

Стенд состоит из привода переменного тока, платформы для монтажа котелка МК и двух постоянных магнитов, расположенных на некотором расстоянии от оси разворота привода. Котелок МК установлен на плече l, относительно этой же оси разворота. При развороте котелка МК на ограниченный угол, на картушку МК наводится дополнительное магнитное поле, имитирующее изменяющуюся погрешность δ_p. Таким образом, при воспроизведении рыскания картушка компаса отклонится от значения физического рыскания на дополнительный угол. Система коррекции при этом компенсирует воздействие дополнительной креновой погрешности.

Стенд может быть также использован для оценки точности работы МК при циркуляции судна. Для этого следует демонтировать постоянные магниты и воспроизвести разворот котелка с необходимой угловой скоростью на заданный угол.

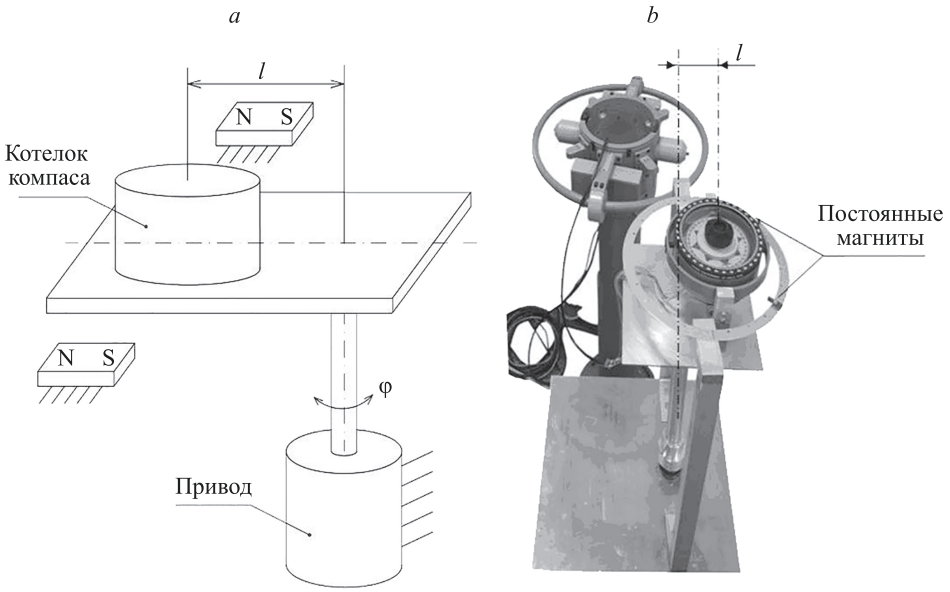


Рис. 4. Кинематическая схема (*a*) и внешний вид (*b*) стенда рыскания
Fig. 4. Kinematic diagram (*a*) and configuration (*b*) of the yaw stand

Выполненное на разработанном стенде исследование макета МК с системой коррекции показало эффективность его применения для оценки работы МК.

Лабораторные исследования режимов функционирования системы коррекции магнитного компаса

Экспериментальные исследования работы системы коррекции проведены при использовании опытного образца котелка МК «Азимут КМ05Д»[1], установленного на стенде рыскания. При этом на котелке был размещен ВОГ ВГ071TS[2].

Исследования выполнены на широте 60° с.ш. с горизонтальной составляющей напряженности магнитного поля 15 ± 1 мкТл в условиях действия нескомпенсированного значения погрешности от качки, составляющего не менее 10°. Задаваемые значения постоянных времени динамических звеньев системы коррекции варьировались: T_1 от 9200 до 7200 с, T_2 от 40 до 60 с. Измерения выполнялись при воспроизведении рыскания в диапазоне углов ±10° с периодами от 5 до 40 с.

На рис. 5 показаны графики результатов моделирования и стендовых испытаний. При моделировании значения коэффициента подавления погрешности от качки $К_П$ получены в зависимости от периодов возмущающих воздействий (углов крена), а при стендовых испытаниях — от заданных при проведении экспериментальных исследований периодов возмущающих воздействий (углов рыскания).

Результаты стендовых испытаний показали, что значения $К_П$ находятся в диапазоне от 0,16 до 0,48 со средним значением 0,35 при изменении периодов рыскания от 6 до 28 с. Для аналогичного диапазона изменения периодов крена среднее значение $К_П$, полученное по результатам моделирования, составляет 0,21. Заниженная оценка значений $К_П$ при моделировании обусловлена тем, что не учтены динамические свойства картушки МК, зависящие от соотношения периода собственных колебаний картушки и периода возмущающих воздействий. Резкое увеличение значения $К_П$ при приближении периода возмущающих воздействий к значению 40 с связано с близостью этого значения к постоянной времени ФВЧ (рис. 2).

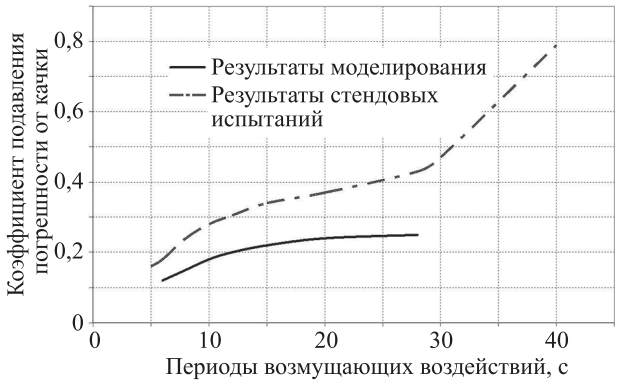


Рис. 5. Зависимость коэффициента подавления погрешности от качки от значений периодов возмущающих воздействий
по результатам моделирования и стендовых испытаний
Fig. 5. Dependence of the coefficient of suppression of the error from the pitching on the values for the periods of disturbing influences according to the results of modeling and bench tests

[1] Азимут КМ05Д. Магнитный компас [Электронный ресурс]. Режим доступа: http://www.elektropribor.spb.ru/katalog/navigatsionnye-pribory/azimut-km05d-vseshirotnyy-magnitnyy-kompas, своб. яз. рус. (дата обращения: 08.09.2021).

[2] Группа ВГ071. Волоконно-оптические гироскопы [Электронный ресурс]. Режим доступа: https://www.fizoptika.ru/catalog/gruppa-vg-071, своб., яз. рус. (дата обращения: 08.09.2021).

Экспериментальные исследования работы МК с системой коррекции на стенде проведены также в режиме циркуляции, т. е. при развороте стенда с МК с постоянной угловой скоростью. Результаты экспериментальных исследований показали, что при развороте судна с угловой скоростью циркуляции 3 °/с и больше, МК с системой коррекции обеспечивает снижение погрешности курсоуказания более чем в два раза по сравнению с МК без системы коррекции. При этом время переходного процесса составляет не более 100–140 с, что хорошо согласуется с результатами проведенного моделирования.

Заключение

Конструкция и тактико-технические характеристики магнитных компасов все время совершенствуются. Как показывает практика судовождения, магнитный компас на многих судах остается резервным средством курсоуказания, к использованию которого прибегают лишь в чрезвычайных ситуациях, связанных, например, с выходом из строя судовых генераторов энергии, с отказом гирокомпаса и в других случаях. Несмотря на все попытки избежать подобных ситуаций, они, к сожалению, еще часто встречаются. Опрос судоводителей на эту тему, проходящих курсы повышения квалификации в Центре дополнительного образования при ГУМРФ им. адмирала С.О. Макарова, показал, что 87,8 % штурманов применяли в своей практике магнитные компасы. Повсеместная экономия судовладельцев на судовом оборудовании также способствует активному использованию магнитных компасов, которыми должно быть оборудовано каждое судно.

Предложенная система коррекции позволяет придать магнитному компасу свойства гиромагнитного компаса, а при установке компаса в систему авторулевого улучшить характеристики этой системы и расширить ее функциональность за счет более точного удержания судна на заданном курсе. Применение данной системы коррекции наиболее эффективно при плавании в высоких широтах, однако и в умеренных широтах она позволяет снизить погрешность от качки не менее, чем в четыре раза. Ее применение обеспечивает также повышение стабильности текущего показания курса на цифровом индикаторе рулевого при воздействии качки, что делает управление судном более эффективным.

Закончим настоящую работу словами академика А.Н. Крылова: «Компас инструмент малый, но, если бы его не было, Америка не была бы открыта».

Литература

1. Кардашинский-Брауде Л.А. Девиация магнитных компасов. Л.: ЦНИИ «Румб», 1986. 81 с.
2. Воронов В.В., Кардашинский-Брауде Л.А. Универсальный широтный компенсатор девиации магнитного компаса // Навигация и гироскопия. 1990. № 2.
3. Arribalzaga J., Martinez A.T., Vila J.A. Analysis and calculation of the magnetic moment of a magnet compensation for type A magnetic needle // Journal of Maritime Research. 2013. V. 10. N 2. P. 80–81.
4. Androjna A., Belev B., Pavic I., Perkovič M. Determining residual deviation and analysis of the current use of the magnetic compass // Journal of Marine Science and Engineering. 2021. V. 9. N 2. P. 1–15. https://doi.org/10.3390/jmse9020204
5. Van Suong Nguyen. Calculation of the deviation coefficients for marine magnetic compass // Journal of International Maritime Safety, Environmental Affairs and Shipping. 2019. V. 2. N 2. P. 112–115. https://doi.org/10.1080/25725084.2019.1569336
6. Meleshko V.V., Lakoza S.L., Sharov S.A. Method of identifying and eliminating magnetic compass deviation // Proc. of the IEEE 1st Ukraine Conference on Electrical and Computer Engineering (UKRCON 2017). 2017. P. 288–291. https://doi.org/10.1109/UKRCON.2017.8100494
7. Iribar I.B., Munoz J.A.V., Labajos C.A.P. Latitude error in compass deviation: Mathematical method to determine the latitude error in magnetic compass deviation // Polish Maritime Research. 2014. V. 21. N 3. P. 25–31. https://doi.org/10.2478/pomr-2014-0026
8. Рыбалтовский Н.Ю. Магнитно-компасное дело. Л.: Водтрансиздат, 1954. 492 с.
9. Грязин Д.Г., Сергачёв И.В. Влияние внешних факторов на переходные процессы магнитных компасов в высоких широтах // Известия высших учебных заведений. Приборостроение. 2020. Т. 63. № 10. С. 921–929. https://doi.org/10.17586/0021-3454-2020-63-10-921-929
10. Грязин Д.Г., Падерина Т.В., Сергачев И.В. О возможности использования судовых магнитных компасов в высоких широтах // Материалы XXXII конференции памяти выдающегося конструктора гироскопических приборов Н.Н. Острякова. СПб.: АО «Концерн» ЦНИИ «Электроприбор», 2020. С. 122–125.
11. Грязин Д.Г., Падерина Т.В. Мехатронные устройства управления современным судовым магнитным компасом // Мехатроника,

References

1. Kardashinskii-Braude L.A. *Deviations of Magnetic Compasses*. Leningrad, CNII Rumb, 1986, 81 p. (in Russian)
2. Voronov V.V., Kardashinskii-Braude L.A. *Universal latitude compensator for magnetic compass deviation*. Navigation and Gyroscopy, 1990, no. 2. (in Russian)
3. Arribalzaga J., Martinez A.T., Vila J.A. Analysis and calculation of the magnetic moment of a magnet compensation for type A magnetic needle. *Journal of Maritime Research*, 2013, vol. 10, no. 2, pp. 80–81.
4. Androjna A., Belev B., Pavic I., Perkovič M. Determining residual deviation and analysis of the current use of the magnetic compass. *Journal of Marine Science and Engineering*, 2021, vol. 9, no. 2, pp. 1–15. https://doi.org/10.3390/jmse9020204
5. Van Suong Nguyen. Calculation of the deviation coefficients for marine magnetic compass. *Journal of International Maritime Safety, Environmental Affairs and Shipping*, 2019, vol. 2, no. 2, pp. 112–115. https://doi.org/10.1080/25725084.2019.1569336
6. Meleshko V.V., Lakoza S.L., Sharov S.A. Method of identifying and eliminating magnetic compass deviation. *Proc. of the IEEE 1st Ukraine Conference on Electrical and Computer Engineering (UKRCON 2017)*, 2017, pp. 288–291. https://doi.org/10.1109/UKRCON.2017.8100494
7. Iribar I.B., Munoz J.A.V., Labajos C.A.P. Latitude error in compass deviation: Mathematical method to determine the latitude error in magnetic compass deviation. *Polish Maritime Research*, 2014, vol. 21, no. 3, pp. 25–31. https://doi.org/10.2478/pomr-2014-0026
8. Rybaltovskii N.Iu. *Magnetic Compass*. Leningrad, Vodtransizdat Publ., 1954, 492 c. (in Russian)
9. Gryazin D.G., Sergachev I.V. Influence of external factors on transient processes of magnetic compass at high latitudes. *Journal of Instrument Engineering*, 2020, vol. 63, no. 10, pp. 921–929. (in Russian). https://doi.org/10.17586/0021-3454-2020-63-10-921-929
10. Gryazin D.G., Paderina T.V., Sergachev I.V. On the possibility of using ship's magnetic compasses at high latitudes. *Proc. of the 32nd conference in memory of N.N. Ostryakov*, St. Petersburg, Concern CSRI Elektropribor, 2020, pp. 122–125. (in Russian)
11. Gryazin D.G., Paderina T.V. Mechatronic control devices for a modern ship's magnetic compass. *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2021, vol. 22, no. 8, pp. 425–432. (in Russian). https://doi.org/10.17587/mau.22.425-432

автоматизация, управление. 2021. Т. 22. № 8. С. 425–432. https://doi.org/10.17587/mau.22.425-432
12. Lushnikov E. Magnetic compass in modern maritime navigation // TransNav. International Journal on Marine Navigation and Safety of Sea Transportation. 2015. V. 9. N 4. P. 539–543. https://doi.org/10.12716/1001.09.04.10
13. Lushnikov E. The reliability of compass information at navigational safety // Scientific Journals of the Maritime University of Szczecin. 2012. N 29. P. 117–121.
14. Basterretxea-Iribal I., Sotés I., Uriarte J.I. Towards an improvement of magnetic compass accuracy and adjustment // Journal of Navigation. 2016. V. 69. N 6. P. 1325–1340. https://doi.org/10.1017/S0373463316000138
15. Казакова Г.Ф., Кардашинский-Брауде Л.А., Фомкин Я.И. Судовые индукционные и гидромагнитные компасы, стрелочные магнитные компасы с индукционными датчиками. Л.: НПО «Азимут», 1991. 62 с.
16. Кардашинский-Брауде Л.А. Современные судовые магнитные компасы. СПб.: ФГУП ГНЦ РФ — ЦНИИ «Электроприбор», 1999. 137 с.
17. Нечаев П.А., Григорьев В.В. Магнитно-компасное дело. 4-е изд. М.: Транспорт, 1983. 240 с.
18. Кардашинский-Брауде Л.А. Повышение динамической устойчивости показаний судовых магнитных компасов при горизонтальных ускорениях [Электронный ресурс]. URL: http://www.navydevices.ru/informations/section-publications/publications/85-2002-09-21-13-29-52, свободный. Яз. рус. (дата обращения: 23.07.2020).
19. Крылов А.Н. О возмущениях показаний компасов, происходящих на качке корабля на волнении // Избранные труды академика А.Н. Крылова. М.: Изд. АН СССР, 1958. С. 115–170.
20. Зиненко В.М., Грязин Д.Г., Молочников А.А., Сергачев И.В., Матвеев Ю.В., Короленко И.В. Способ измерения магнитного курса судна в высоких широтах и устройство для его реализации. Патент RU2688900C1. Бюл. 2019. № 15.
21. Боронахин А.М., Лукомский Ю.А., Шпекторов А.Г., Ха Мань Тханг, Чан Танг Дык. Курсовертикаль для задач управления движением скоростного судна // Известия СПбГЭТУ «ЛЭТИ». 2012. № 4. С. 75–80.
22. Боронахин А.М., Лукомский Ю.А., Шпекторов А.Г., Ха Мань Тханг, Чан Танг Дык. Обоснование требований к системам навигации и ориентации при решении задач управления движением подвижных объектов // Известия СПбГЭТУ «ЛЭТИ». 2012. № 2. С. 77–80.
23. Кожухов В.П., Воронов В.В., Григорьев В.В. Магнитные компасы. М.: Транспорт, 1981. 212 с.
24. Бородай И.К., Моренштильдт В.А., Виленский Г.В. Прикладные задачи динамики судов на волнении. Л.: Судостроение, 1989. 258 с.

12. Lushnikov E. Magnetic compass in modern maritime navigation. TransNav. *International Journal on Marine Navigation and Safety of Sea Transportation*, 2015, vol. 9, no. 4, pp. 539–543. https://doi.org/10.12716/1001.09.04.10
13. Lushnikov E. The reliability of compass information at navigational safety. *Scientific Journals of the Maritime University of Szczecin*, 2012, no. 29, pp. 117–121.
14. Basterretxea-Iribal I., Sotés I., Uriarte J.I. Towards an improvement of magnetic compass accuracy and adjustment. *Journal of Navigation*, 2016, vol. 69, no. 6, pp. 1325–1340. https://doi.org/10.1017/S0373463316000138
15. Kazakova G.F., Kardashinskii-Braude L.A., Fomkin Ia.I. *Ship Induction and Hydromagnetic Compasses, Needle Magnetic Compasses with Inductive Sensors*. Leningrad, NPO "Azimut", 1991, 62 c. (in Russian)
16. Kardashinskii-Braude L.A. *Modern Marine Magnetic Compasses*. St. Petersburg, Concern CSRI Elektropribor, 1999, 137 p. (in Russian)
17. Nechaev P.A., Grigorev V.V. Magnetic Compass. Moscow, Transport Publ., 1983, 240 p. (in Russian)
18. Kardashinskii-Braude L.A. *Increasing the dynamic stability of ship's magnetic compass readings at horizontal accelerations*. Available at: ttp://www.navydevices.ru/informations/section-publications/publications/85-2002-09-21-13-29-52 (accessed: 23.07.2020). (in Russian)
19. Krylov A.N. On perturbations of compass readings that occur on the rolling of a ship in waves. *Izbrannye trudy akademika A.N. Krylova*. Moscow, Academy of Sciences of the Soviet Union, 1958, pp. 115–170. (in Russian)
20. Zinenko V.M., Gryazin D.G., Molochnikov A.A., Sergachev I.V., Matveev Y.V., Korolenko I.V. Method for measuring ship magnetic course in high latitudes and device for its implementation. *Patent RU2688900C1*, 2019. (in Russian)
21. Boronachin A.M., Lukomski Y.A., Shpektorov A.G., Ha Manh Thang, Tran Tang Duc. Yaw-and-vertical gyro unit for high-speed marine vehicle motion control. *Proceedings of Saint Petersburg Electrotechnical University,* 2012, no. 4, pp. 75–80. (in Russian)
22. Boronachin A.M., Lukomski Y.A., Shpektorov A.G., Ha Manh Thang, Tran Tang Duc. The basis of navigation system requirements for solving a problem of marine vehicles motion control. *Proceedings of Saint Petersburg Electrotechnical University*, 2012, no. 2, pp. 77–80. (in Russian)
23. Kozhukhov V.P., Voronov V.V., Grigorev V.V. *Magnetic Compasses*. Moscow, Transport Publ., 1981, 212 p. (in Russian)
24. Borodai I.K., Morenshildt V.A., Vilenskii G.V. *Applied Problems of Ship Dynamics in Waves*. Leningrad, Sudostroenie Publ., 1989, 258 p. (in Russian)

Авторы

Грязин Дмитрий Геннадиевич — доктор технических наук, профессор, главный метролог – начальник отдела, АО «Концерн «ЦНИИ «Электроприбор», Санкт-Петербург, 197046, Российская Федерация; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 25638150600, https://orcid.org/0000-0003-3562-6815, volnagdg@yandex.ru

Падерина Татьяна Владимировна — кандидат технических наук, старший научный сотрудник, АО «Концерн «ЦНИИ «Электроприбор», Санкт-Петербург, 197046, Российская Федерация; доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 8602442700, https://orcid.org/0000-0001-5293-6351, paderinata@gmail.com

Authors

Dmitriy G. Gryazin — D.Sc., Professor, Chief Metrologist, Head of the Department, Concern CSRI Elektropribor, JSC, Saint Petersburg, 197046, Russian Federation; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 25638150600, https://orcid.org/0000-0003-3562-6815, volnagdg@yandex.ru

Tatiana V. Paderina — PhD, Senior Researcher, Concern CSRI Elektropribor, JSC, Saint Petersburg, 197046, Russian Federation; Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 8602442700, https://orcid.org/0000-0001-5293-6351, paderinata@gmail.com

Статья поступила в редакцию 20.09.2021
Одобрена после рецензирования 24.11.2021
Принята к печати 27.01.2022

Received 20.09.2021
Approved after reviewing 24.11.2021
Accepted 27.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-147-154

A new analytical model of drain current and small signal parameters for AlGaN-GaN high-electron-mobility transistors

Azzeddine Farti[1]✉, Abdelkader Touhami[2]

[1,2] Hassan II University of Casablanca, Casablanca, 20190, Morocco

[1] azzeddine.farti-etu@etu.univh2c.ma✉, https://orcid.org/0000-0002-8694-1471
[2] abdelkadertouhami2016@gmail.com, https://orcid.org/0000-0001-8582-1884

Abstract
The paper proposes a new analytical model of the drain current in AlGaN-GaN high-electron-mobility transistors (HEMT) on the basis of a polynomial expression for the Fermi level as a function of the concentration of charge carriers. The study investigated the influence of parasitic resistances (source and drain sides), high-speed saturation, the amount of aluminum in the AlGaN barrier, and low field mobility. To isolate the output characteristics, cut-off frequency and steepness, the parameters of the hyper frequency signal were developed. Comparison of analytical calculations with experimental measurements confirmed the validity of the proposed model.

Keywords
AlGaN-GaN, HEMTs, 2-DEG two-dimensional electron gas, current–voltage characteristics, cut-off frequencies, Transconductance

Acknowledgements
This work carried out at the Faculty of Sciences Ain Chok Km 8, Hassan II University, Casablanca, Morocco.

For citation: Farti A., Touhami A. A new analytical model of drain current and small signal parameters for AlGaN-GaN high-electron-mobility transistors. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 147–154. doi: 10.17586/2226-1494-2022-22-1-147-154

УДК 621.382.3

Новая аналитическая модель тока стока и параметров малых сигналов AlGaN-GaN транзисторов с высокой подвижностью электронов

Азеддин Фарти[1]✉, Абделькадер Тухами[2]

[1,2] Университет Хасана II, Факультет наук Касабланки, Касабланка, 20190, Марокко

[1] azzeddine.farti-etu@etu.univh2c.ma✉, https://orcid.org/0000-0002-8694-1471
[2] abdelkadertouhami2016@gmail.com, https://orcid.org/0000-0001-8582-1884

Аннотация
Предложена новая аналитическая модель тока стока в устройствах AlGaN-GaN транзисторов с высокой подвижностью электронов (HEMT) на основе полиномиального выражения для уровня Ферми как функции концентрации носителей заряда. В ходе исследования изучено влияние паразитных сопротивлений (стороны истока и стока), высокоскоростного насыщения, количества алюминия в барьере AlGaN и низкой подвижности поля. Для выделения выходных характеристик, частоты среза и крутизны разработаны параметры гиперчастотного сигнала. Сравнение аналитических расчетов с экспериментальными измерениями подтвердило справедливость предложенной модели.

Ключевые слова
AlGaN-GaN, транзистор с высокой подвижностью электронов, HEMT, двумерный электронный газ, вольтамперная характеристика, частота среза, крутизна характеристики

Благодарности
Работа выполнена на факультете наук Айн Чок Km 8, Университет Хасана II, Касабланка, Марокко.

© Farti A., Touhami A., 2022

Ссылка для цитирования: Фарти А., Тухами А. Новая аналитическая модель тока стока и параметров малых сигналов AlGaN-GaN транзисторов с высокой подвижностью электронов // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 147–154 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-1-147-154

Introduction

High-electron-mobility transistors (HEMT) based on the AlGaN-GaN hetero-structure have rapidly emerged in the fabrication of very high-speed integrated circuits [1]. These transistors have taken an important place in the realization of high-power, high-voltage, and high-temperature devices because of their very tall switching speed, poor power consumption and relatively easy manufacturing technology.

Due to their considerable gap energy and significant discontinuity of conduction bands between the AlGaN and GaN layers, HEMT transistors can create a two-dimensional electron gas (2-DEG), with a concentration in the order of 10^{13} cm^{-2} under the spontaneous polarization effect and the piezoelectric effect [2].

To develop a reliable model for HEMT transistors, an accurate estimate of the 2-DEG density at the hetero-interface is significant. A number of charge control models for HEMTs and several theoretical models of drain current in HEMT transistors have been developed to characterize the concentration of the two-dimensional electron's gas (2-DEG) at the AlGaN-GaN interface [2–6].

The previous work on theoretical analysis of HEMT AlGaN-GaN [5] has been done by solving simultaneously the Schrödinger and Poisson equations. According to the assumption's authors, the polarization in nitride alloys interpolates linearly [7]; meanwhile, Fiorentini et al [8] have argued that there is a significant variety in the theoretical results when nonlinear polarization is applied instead of linear polarization. This can be explained by the fact that nonlinearity modifies the interface charge densities induced by the polarization. Analyzing the models involving polarization nonlinearity [9–12], we found that theoretical models are not suitable with the experimental result (the current-voltage characteristic and the transconductance).

In the paper, we present a new model of charge control using the 2-DEG channel current equation developed by Rashmi et al [5], and considering the nonlinearity of the polarization which will produce a good optimization with the experimental result. In this model we also introduce the Fermi level $E_F(T, m)$ in its simple polynomial form as a function of the electron gas density $n_s(T, m)$.

Based on the proposed analytical model, we will extract the current-voltage characteristics and describe the variations of the channel transconductance of the device. To highlight the performance and operating limits of HEMT transistors at microwave frequencies, first, we study the variation of cut-off frequencies f_T and the transcanductance g_m as a function of the drain current at fixed drain voltage, and secondly, we analyze the variation of the cut-off frequency as a function of the bias voltages V_{ds} and V_{gs}.

Model formulation

Charge-control model

The structure of HEMT mainly consists of three different materials: substrate, material with a wide bandgap, and material with a low bandgap. The junction of these two materials leads to the formation of two-dimensional electron gas (2-DEG) at the interface. It can be modulated by the gate-source voltage, while the formation of the electron gas is performed between the spacer and the small gap layer (Fig. 1). The sheet carrier concentration of the two-dimensional electron gas (2-DEG) constituted at the AlGaN-GaN heterojunction is given by [9]:

$$n_s(T, m) = \frac{\varepsilon(m)}{qd}(V_{gs} - V_{th}(T, m) - E_F(T, m)), \qquad (1)$$

where T is the temperature; m is the Al mole fraction in AlGaN-GaN; q is the electron charge; $\varepsilon(m)$ is the dielectric constant of $Al_mGa_{1-m}N$; d is the summation of the doped AlGaN layer thickness (d_d) and the undoped AlGaN spacer layer thickness (d_i). $V_{th}(m)$ is the threshold voltage, $E_F(T, m)$ is the position of the Fermi level which is expressed as a function of $n_s(T, m)$ and it is given by the form of a simple polynomial [13].

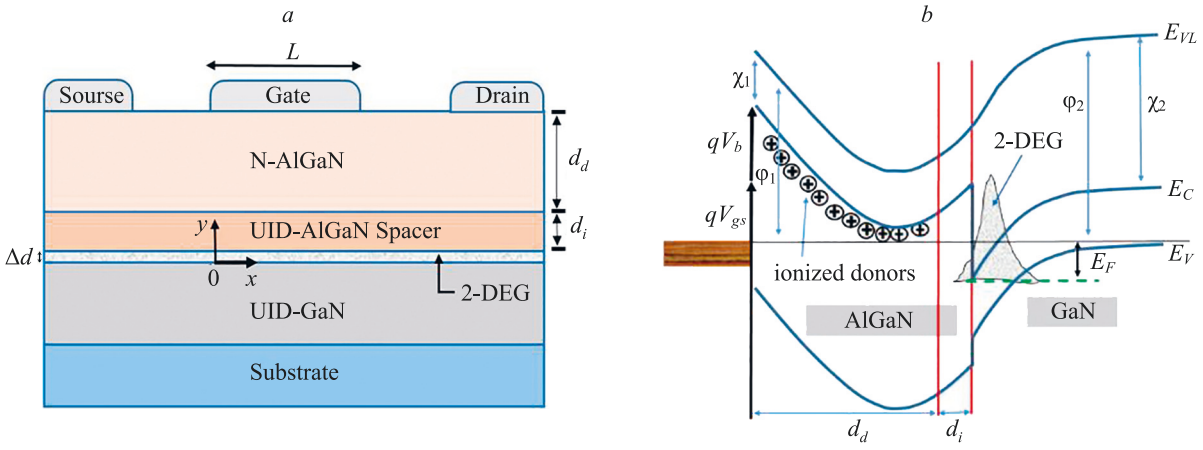


Fig. 1. HEMT transistor AlGaN-GaN in cross-section (*a*) and the band diagram of the AlGaN-GaN structure (*b*)

$$E_F(T, m) = K_1(T) + K_2(T)\sqrt{n_s(T, m)} + K_3(T)n_s(T, m), \quad (2)$$

where K_1, K_2 and K_3 are three different parameters depending on the temperature T.

Where L is the Gate length, x present the position along the transistor channel. V_b presents the Schottky-barrier height, φ_1 and φ_2 are work functions of AlGaN and GaN, respectively. χ_1 (χ_2) is electron affinity of AlGaN (GaN). E_{VL}, E_C and E_V show the vacuum level, the energy of the conduction band, the energy of the valence band, respectively.

By introducing the equation (2) into (1), we obtained a quadratic equation, to demonstrate the concentration of 2-DEG charge carriers at the heterojunction [14].

$$n_s(T, m) = \quad (3)$$
$$= \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2,$$

where

$$K_4(T, m) = K_3(T) + \frac{qd}{\varepsilon(m)}.$$

The threshold voltage $V_{th}(T, m)$ depends on the aluminum mole fraction of MOSFET, and its expression given by [15].

$$V_{th}(T, m) = \varphi(m) - \Delta E_C(T, m) - \frac{qN_d d_d^2}{2\varepsilon(m)} - \frac{\sigma_{PZ}(m)}{\varepsilon(m)}d,$$

where $\varphi(m)$, $\Delta E_C(m)$ are the Schottky barrier height and conduction band discontinuity between GaN and AlGaN, respectively. N_d is the doping density of the Al-Ga-N layer and $\sigma_{PZ}(m)$ is the charge density induced by the total polarization, can be written as follows [8]:

$$\sigma_{PZ}(m) = |P_{PP}(m) + P^{sp}_{Al_mGa_{1-m}N}(m) - P^{sp}_{Al_mGa_{1-m}N}(0)|,$$

where $P_{PP}(m)$ is the piezo-electric polarization with its expression and it varies as a function of the value of m. $P^{sp}_{Al_mGa_{1-m}N}(m)$ and $P^{sp}_{Al_mGa_{1-m}N}(0)$ are the spontaneous polarizations of $Al_mGa_{1-m}N$ and GaN, respectively [8].

$$P_{PP}(m) = \begin{cases} P^{pz}_{Al_mGa_{1-m}N}(m) & \text{For } 0 \le m < 0.38 \\ P^{pz}_{Al_mGa_{1-m}N}(2.33–3.5m) & \text{For } 0.38 \le m \le 0.67, \\ 0 & \text{For } 0.67 \le m \le 1 \end{cases}$$

The expressions of spontaneous and piezo-electric polarizations are given in Table 1 [8].

Current–voltage characteristics ($I_{ds} - V_{ds}$)

By fixing the value of voltage V_{ds} superior of the threshold voltage, the electron movement at the AlGaN/GaN interface is ensured by the application of the source-drain voltage V_{ds} superior to zero. The generated drain current I_{ds} is given by the following equation [5].

$$I_{ds}(T, m, x) = Zq\mu(x)\left(n_s(T, m)\frac{dV_C(x)}{dx} + \frac{K_BT}{q}\frac{dn_s(T, m)}{dx}\right), \quad (4)$$

where Z, $V_C(x)$ and K_B are the gate width, the channel potential at position x, and Boltzmann's constant, respectively. $\mu(x)$ is the field dependent of the mobility given by:

$$\mu(x) = \frac{\mu_0}{\left(1 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}\frac{dV_C(x)}{dx}\right)}, \quad (5)$$

where E_C is the saturation electric field, μ_0 is the low-field mobility, and v_{sat} is the saturation drift velocity.

Depending on the variation of the bias voltage V_{ds}, the HEMT transistor can be function under two different regimes (linear regime and saturation regime).

The linear region ($V_{ds} < V_{dsat}$)

To describe the variation of potential at each position x in the 2-DEG channel, the voltage V_{gs} is replaced by $V_{gs} - V_C(x)$ in the expression (3) of sheet carrier concentration of two-dimensional electron gas formed at the AlGaN/GaN heterojunction.

$$n_s(T, m, x) = \left(\frac{-K_2(T)}{2K_4(T, m)} + \right. \quad (6)$$
$$\left. + \frac{\sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - V_C(x) - K_1(T))}}{2K_4(T, m)}\right)^2.$$

Using the following approximation [16] in equation (6) we get:

$$4K_4(T, m)V_C(x) <<$$
$$<< K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T)).$$

After using the above given approximation in equation (6), a simple expression of the concentration $n_s(T, m)$ have the following form:

$$n_s(T, m) = \quad (7)$$
$$= \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2.$$

Table 1. Parameters of the $Al_mGa_{1-m}N$/GaN heterostructure

Description	Parameter	Expression
Spontaneous polarization of AlGaN	$P^{sp}_{Al_mGa_{1-m}N}(m)$, C/m^2	$–0.09m – 0.034(1 – m) + 0.019m(1 – m)$
Piezo-electric polarization of AlGaN	$P^{pz}_{Al_mGa_{1-m}N}(m)$, C/m^2	$mP^{pz}_{AlN}[\varepsilon_b(m)] + (1 – m)P^{pz}_{GaN}[\varepsilon_b(m)]$
Piezo-electric polarization of AlN	$P^{pz}_{AlN}(m)$, C/m^2	$–1.808\varepsilon_b(m) + 5.624\varepsilon_b^2(m)$
Piezo-electric polarization of GaN	$P^{pz}_{GaN}(m)$, C/m^2	$–0.918\varepsilon_b(m) + 9.541\varepsilon_b^2(m)$
The basal strain	$\varepsilon_b(m)$	$(a(0) – a(m))/a(m)$
The lattice constant	$a(m)$, nm	$0.031986 – 0.000891m$

Consequently, the drain current from equation (4) will be expressed by:

$$I_{ds}(T, m, x) = Zq\mu(x)\left(n_s(T, m)\frac{dV_C(x)}{dx}\right). \quad (8)$$

We substitute the equations (5) and (7) in the drain current equation (8), and a new equation will appear which describes the drain current at each position x of the channel:

$$I_{ds}(T, m, x) = Zq\mu_0 \times$$
$$\times \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(m)(V_{gs} - V_{th}(m) - K_1(T)}}{2K_4(T, m)}\right)^2 \times$$
$$\times \frac{\frac{dV_C(x)}{dx}}{\left(1 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}\frac{dV_C(x)}{dx}\right)}. \quad (9)$$

Fig. 2 shows the equivalent electrical circuit diagram of the high-electron-mobility transistor HEMT (AlGaN-GaN), where R_s and R_d are the parasitic resistances of the source and drain, respectively. R_c represents the resistance of the 2-DEG channel.

From the schema in Fig. 2, the channel potential for $x = 0$ and $x = L$ will be given by:

$$V_C(x)|_{x=0} = I_{ds}R_s,$$
$$V_C(x)|_{x=L} = V_{ds} - (I_{ds}R_d).$$

With the above given boundary conditions, the various steps of integration of the equation (9) along the transistor channel will be written as follows:

$$I_{ds}\int_0^L \left(1 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}\frac{dV_C(x)}{dx}\right)dx =$$
$$= Zqn_s(T, m)\mu_0 \int_0^L \left(\frac{dV_C(x)}{dx}\right)dx,$$
$$I_{ds}\left([x]_0^L + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}[V_C(x)]_0^L\right) = Zqn_s(T, m)\mu_0[V_C(x)]_0^L,$$
$$-I_{ds}^2\left(\frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}(R_s + R_d)\right) + \quad (10)$$
$$+ I_{ds}\left(L + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}V_{ds} + Zqn_s(T, m)\mu_0(R_s + R_d)\right) -$$
$$- Zqn_s(T, m)\mu_0 V_{ds} = 0.$$

From equation (10), by replacing $ns(T, m)$ with its value, the first term is multiplied by I_{ds}^2 into α_1, the second term is multiplied by I_{ds} into α_2 and the last term that does not contain I_{ds} by α_3, we can obtain:

$$\alpha_1 = -\left(\frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}(R_s + R_d)\right),$$
$$\alpha_2 = L + Zq\mu_0(R_s + R_d)\times$$
$$\times \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T)}\right)^2 +$$
$$+ \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}V_{ds}, \quad (11)$$
$$\alpha_3 = -Zq\mu_0 \times \quad (12)$$
$$\times \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2 \times$$
$$\times V_{ds}.$$

The expression of the drain current I_{ds} in linear regime will be given by the root of equation (10) and can be written in the following form:

$$I_{ds} = \frac{-\alpha_2 + \sqrt{\alpha_2^2 - 4\alpha_1\alpha_3}}{2\alpha_1}.$$

The saturation region ($V_{ds} \geq V_{dsat}$)

The application of an electric field in a semiconductor leads to free charge carriers (electrons and holes) to acquire a velocity of displacement υ proportional to the applied field. The following relationship describes this velocity:

$$\upsilon = \mu(x)E(x) = \frac{\mu_0}{\left(1 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}\frac{dV_C(x)}{dx}\right)}\frac{dV_C(x)}{dx}.$$

At high electric field values, the velocity of the carriers is saturated at the value υ_{sat}. This fact has allowed us to replace the velocity υ by υ_{sat} in expression (9) of the drain current I_{ds}:

$$I_{dsat} = Zq\upsilon_{sat} \times \quad (13)$$
$$\times \left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2.$$

In the same way, we replace the voltage V_{ds} by the saturation voltage V_{dsat} in the equations (11) and (12).

In this case, the expression of the drain current in the saturation regime can be written in the form:

$$I_{dsat} = \frac{-\beta_2 + \sqrt{\beta_2^2 - 4\alpha_1\beta_3}}{2\alpha_1}, \quad (14)$$

With,

$$\beta_2 = \delta_2 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}}V_{dsat}, \quad (15)$$

and

$$\beta_3 = \delta_3 V_{dsat}, \quad (16)$$

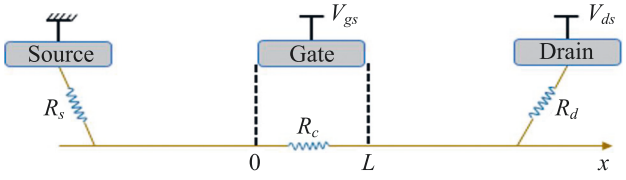


Fig. 2. Equivalent electrical schema of the HEMT

where

$$\delta_2 = L + Zq\mu_0(R_s + R_d) \times$$
$$\times\left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2,$$

$$\delta_3 = -Zq\mu_0 \times$$
$$\times\left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2.$$

To determine the expression of the saturation voltage V_{dsat}, we equalize the equations (13) and (14) of the drain current I_{ds}:

$$\frac{-\beta_2 + \sqrt{\beta_2^2 - 4\alpha_1\beta_3}}{2\alpha_1} = Zq\upsilon_{sat} \times \quad (17)$$
$$\times\left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2.$$

We put:

$$\delta_1 = 2\alpha_1 Zqv_{sat} \times$$
$$\times\left(\frac{-K_2(T) + \sqrt{K_2^2(T) + 4K_4(T, m)(V_{gs} - V_{th}(T, m) - K_1(T))}}{2K_4(T, m)}\right)^2,$$

And equation (17) will be described as:

$$-\beta_2 + \sqrt{\beta_2^2 - 4\alpha_1\beta_3} = \delta_1. \quad (18)$$

According to the expressions of (15) and (16), equation (18) is as follows:

$$\delta_1 = -\left(\delta_2 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}} V_{dsat}\right) +$$
$$+\sqrt{\left(\delta_2 + \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}} V_{dsat}\right)^2 - 4\alpha_1\delta_3 V_{dsat}},$$

As a result, the expression of the saturation voltage V_{dsat} can be deduced and expressed by the formula:

$$V_{dsat} = \frac{-\delta_1(2\delta_2 + \delta_1)}{\left(2\delta_1 \frac{(\mu_0 E_C - v_{sat})}{E_C v_{sat}} + 4\alpha_1\delta_3\right)}.$$

Small signal parameters

At this stage, for better modelling of the HEMT transistor, we need to determine and calculate its small signals parameters like the transconductance g_m and the cut-off frequency f_T, which may reflect the performance of this transistor.

Transconductance g_m

The transconductance is considered as one of the most important parameters in evaluating the performance of transistors for high-frequency applications. The transconductance g_m of the $Al_mGa_{1-m}N/GaN$ HEMT transistor is defined as [17].

$$g_m = \left.\frac{\partial I_{ds}}{\partial V_{gs}}\right|_{V_{ds}=cte}, \quad (19)$$

where:

$$I_{ds} = \frac{-\alpha_2 + \sqrt{\alpha_2^2 - 4\alpha_1\alpha_3}}{2\alpha_1}, \quad (20)$$

From (19) and (20), we obtained the expression of the transconductance written by the equation:

$$g_m = \frac{1}{2\alpha_1}\left(-\frac{\partial\alpha_2}{\partial V_{gs}} + \frac{2\alpha_2\frac{\partial\alpha_2}{\partial V_{gs}} - 4\alpha_1\frac{\partial\alpha_3}{\partial V_{gs}}}{2\sqrt{\alpha_2^2 - 4\alpha_1\alpha_3}}\right).$$

Cut-off frequency f_T

The characterization of the high-frequency performance of the HEMT transistor requires a detailed study of its cut-off frequency f_T. This later determines the factor of the microwave operation of HEMT transistor components. The cut-off frequency can be calculated versus the transconductance g_m as follows [10]:

$$f_T = \frac{g_m\mu_0}{2\pi L\frac{Z\mu_0\varepsilon(m)}{(d_d + d_i + \Delta d)}}, \quad (21)$$

where Δd presents the thickness of two-dimensional electron gas 2-DEG.

Results and discussion

To highlight the models of electrical and technological parameters (I_{ds}, g_m and f_T) of the HEMT transistor, we made a comparison with experimental results [18]. The electrical, geometrical and technological parameters used in the proposed model are presented in Table 2.

The values of the constants K_1, K_2 and K_3 used in the theoretical calculations for different values of the gate voltage V_{gs} with $m = 0.15$ are depicted in Table 3.

Fig. 3 shows the electrical characteristics of the HEMT transistor ($Al_{0.15}Ga_{0.85}N/GaN$) as compared to experimental measurements [18]. The length and the width gate of this transistor are equal to 1 μm and 75 μm, respectively. Fig. 3, *a* shows the variation of the drain current as a function of the drain voltage for different values of the gate voltage. For the calculation, we took a gate voltage sweep between –2 V and 1 V with a step of 1 V.

The results of the simulation indicate that there are two operating regimes. The first one is a linear regime, implying that the drain voltage increases, then the drain current I_{ds}

Table 2. Different parameters used in the proposed model

Parameter	Description	Value
V_{th}, V	Threshold voltage	–2.6
d_d, nm	Thickness of the doped layer	22
d_i, nm	Thickness of the undoped layer	3
Z, μm	Gate width	75
L, μm	Gate length	1
V_{sat}, m/s	Saturation velocity	1.19×10^5
R_s, Ω	Parasitic source resistance	0.6
R_d, Ω	Parasitic drain resistance	0.9

Table 3. Values of the parameters K_1, K_2 and K_3 for different gate voltage values V_{gs}. They are calculated with the same method [19] using the effective mass $m^* = 0.22m_0$ [7]

Gate voltage , V_{gs}, V	K_1, V	$K_2 \times 10^{-8}$, V·m	$K_3 \times 10^{-18}$, V·m^2
1	–0.27750	1.68742	–4.23715
0	–0.27573	1.45776	–3.21042
–1	–0.27423	1.26515	–2.34936
–2	–0.27422	1.26274	–2.33856

also increases. In a saturated regime, the drain current keeps constant and independent to the drain voltage V_{ds}. This can be explained by the pinching of the canal and the saturation of the electron velocity. These results coincide with the experimental measurements and prove the validity of the proposed model.

The curve in Fig. 3, *b* represents the dependence of the drain current vs. the gate voltage for a drain voltage set at the value 5V. We concluded that the intersection of the asymptote of the curve with the axis of the voltages V_{gs} is equal to –2.6 V. This value is equivalent to the threshold voltage V_{th} of the HEMT transistor under study. In other words, the results of our model are in concordance with the experimental data and confirm the validity of our model.

Fig. 3, c displays the variation of the transconductance g_m as a function of the gate voltage V_{gs} with the value of the drain voltage V_{ds} equal to 5 V. This transconductance g_m increases with the voltage V_{gs} till it reaches a maximum g_{mMax} (160 mS/mm) for the value of the gate voltage equal to –0.74 V. Consequently, the HEMT transistor achieves its optimum operating point. After this maximum, the transconductance decreases progressively, which can be attributed to the saturation of the carrier's velocity and the reduction of their mobility in the channel. This is caused by the increase in channel resistance. In addition, the thermal effects and presence of defects can be responsible for this drop transconductance.

Fig. 4, *a* shows the cut-off frequencies f_T and the transcanductance g_m as a function of the drain current I_{ds} with gate width $Z = 2 \times 75$ µm and the fixed drain voltage V_{ds} = 5V. We observe that the maximum of the cut-off frequency is 9.66 GHz which corresponds to a drain current

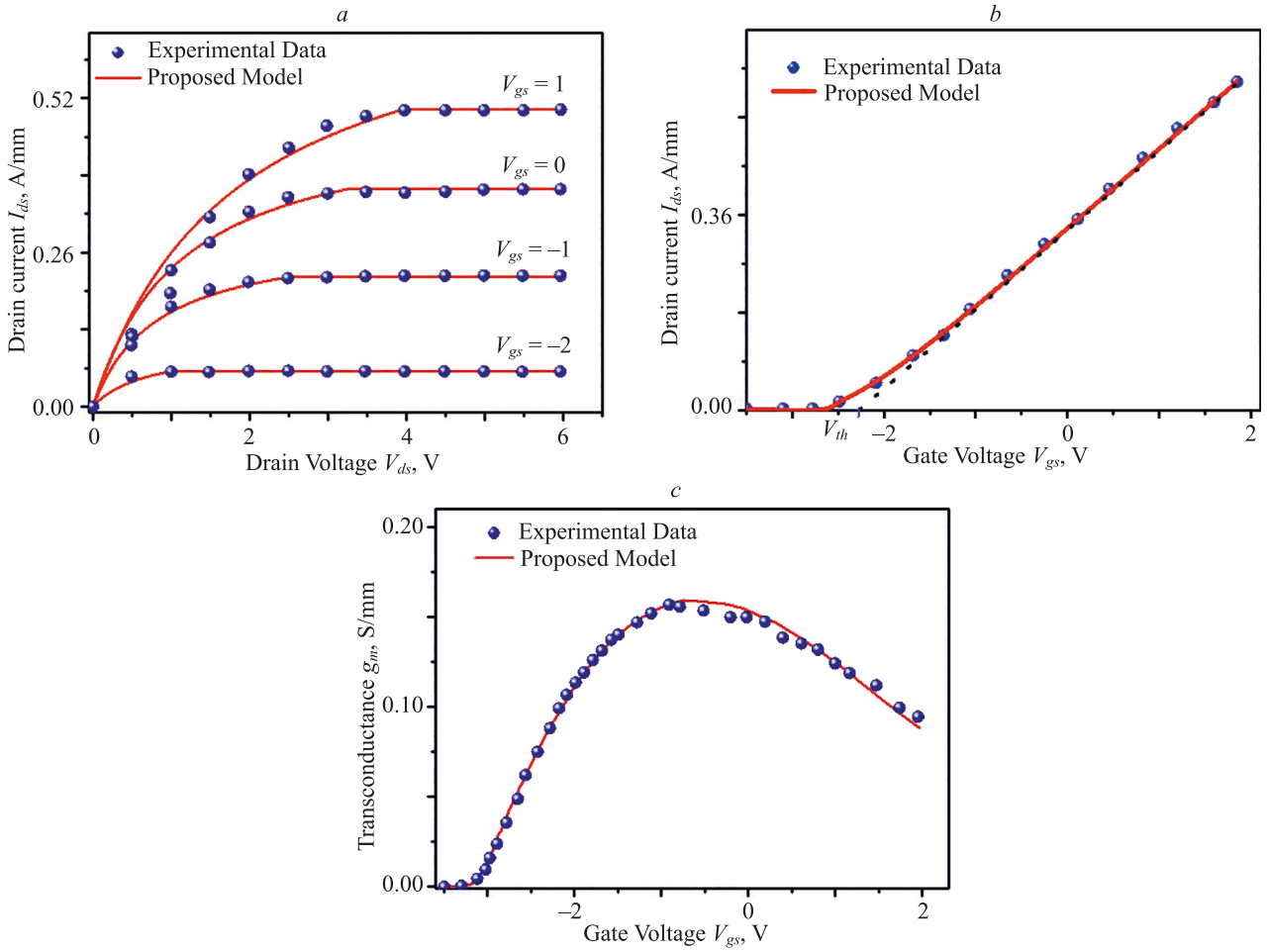


Fig. 3. Variation of the drain current I_{ds} as a function of: the drain voltage V_{ds} for different gate voltage values V_{gs} (*a*) and gate voltage V_{gs} by fixing the drain voltage V_{ds} at 5V (*b*); variation of the transconductance g_m as a function of gate voltage V_{gs} (*c*)

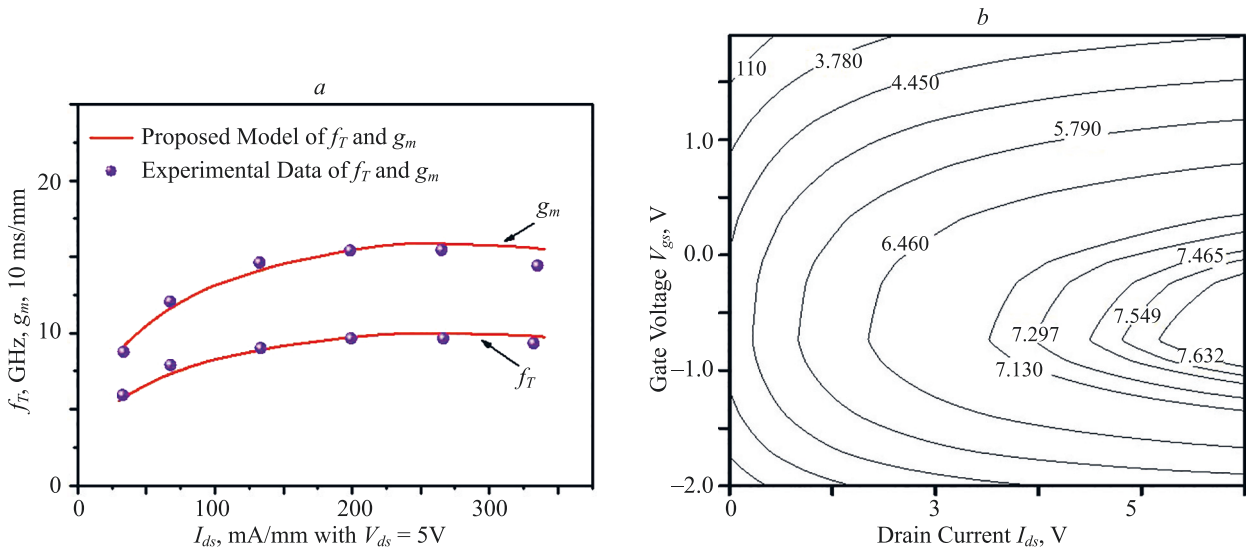


Fig. 4. Cut-off frequencies vs. drain current I_{ds} at drain voltage V_{ds} at 5V ($Z = 2\times75$ μm) (*a*); dependence of the cut-off frequency f_T with polarization voltages V_{gs} and V_{ds} (*b*)

of 250 mA/mm, where the maximum of transcanductance is located.

Fig. 4, *b* presents the dependence of the cut-off frequency f_T on the gate and drain polarization voltages. It is obvious that the cut-off frequency rises with the gate voltage and decreases after reaching a maximum. Also, this cut-off frequency increases with the applied drain voltage. This is due to the impact of the gate voltage on the depth of the potential well and favors important diffusion of free electrons from the semiconductor donors.

Conclusion

In summary, we have studied a new analytical model of the drain current and small signal parameters of AlGaN-GaN HEMTs devices and considered the nonlinear polarization effect using the 2-DEG channel current equation. With the help of the charge control model, a detailed investigation of different functioning regimes of the HEMT transistor has been conducted. In this work, we optimized the difference between the proposed model and the experimental data. This later extracted from the current-voltage characteristics evidence that this model is useful for the prediction of drain current. Basing on this model, we can calculate the values of the transconductance g_m and the cut-off frequency f_T. These results showed that the simulation model is satisfactory and consistent with the experimental measurements, evidencing the validity of our proposed model.

References

1. Liu J., Guo Y., Zhang J., Yao J., Huang X., Huang C., Huang Z., Yang K. Analytical model for the potential and electric field distributions of AlGaN/GaN HEMTs with gate connected FP based on Equivalent Potential Method. *Superlattices and Microstructures*, 2020, vol. 138, pp. 106327. https://doi.org/10.1016/j.spmi.2019.106327
2. Rashmi, Haldar S., Gupta R.S. 2-D analytical model for current-voltage characteristics and output conductance of AlGaN/GaN MODFET. *Microwave and Optical Technology Letters*, 2001, vol. 29, no. 2, pp. 117–123. https://doi.org/10.1002/mop.1102
3. Mohanbabu A., Anbuselvan N., Mohankumar N., Godwinraj D., Sarkar C.K. Modeling of sheet carrier density and microwave frequency characteristics in Spacer based AlGaN/AlN/GaN HEMT devices. *Solid-State Electronics*, 2014, vol. 91, pp. 44–52. https://doi.org/10.1016/j.sse.2013.09.009
4. Jebalin B.K., Rekh A.S., Prajoon P., Godwinraj D., Kumar N.M., Nirmal D. Unique model of polarization engineered AlGaN/GaN based HEMTs for high power applications. *Superlattices and Microstructures*, 2015, vol. 78, pp. 210–223. https://doi.org/10.1016/j.spmi.2014.10.038
5. Rashmi, Kranti A., Haldar S., Gupta R.S. An accurate charge control model for spontaneous and piezoelectric polarization dependent two-dimensional electron gas sheet charge density of lattice-mismatched AlGaN/GaN HEMTs. *Solid-State Electronics*, 2002, vol. 46, no. 5, pp. 621–630. https://doi.org/10.1016/S0038-1101(01)00332-X
6. Kumar S.P., Agrawal A., Kabra S., Gupta M., Gupta R.S. An analysis for AlGaN/GaN modulation doped field effect transistor using

Литература

1. Liu J., Guo Y., Zhang J., Yao J., Huang X., Huang C., Huang Z., Yang K. Analytical model for the potential and electric field distributions of AlGaN/GaN HEMTs with gate connected FP based on Equivalent Potential Method // Superlattices and Microstructures. 2020. V. 138. P. 106327. https://doi.org/10.1016/j.spmi.2019.106327
2. Rashmi, Haldar S., Gupta R.S. 2-D analytical model for current-voltage characteristics and output conductance of AlGaN/GaN MODFET // Microwave and Optical Technology Letters. 2001. V. 29. N 2. P. 117–123. https://doi.org/10.1002/mop.1102
3. Mohanbabu A., Anbuselvan N., Mohankumar N., Godwinraj D., Sarkar C.K. Modeling of sheet carrier density and microwave frequency characteristics in Spacer based AlGaN/AlN/GaN HEMT devices // Solid-State Electronics. 2014. V. 91. P. 44–52. https://doi.org/10.1016/j.sse.2013.09.009
4. Jebalin B.K., Rekh A.S., Prajoon P., Godwinraj D., Kumar N.M., Nirmal D. Unique model of polarization engineered AlGaN/GaN based HEMTs for high power applications. Superlattices and Microstructures. 2015. V. 78. P. 210–223. https://doi.org/10.1016/j.spmi.2014.10.038
5. Rashmi, Kranti A., Haldar S., Gupta R.S. An accurate charge control model for spontaneous and piezoelectric polarization dependent two-dimensional electron gas sheet charge density of lattice-mismatched AlGaN/GaN HEMTs // Solid-State Electronics. 2002. V. 46. N 5. P. 621–630. https://doi.org/10.1016/S0038-1101(01)00332-X
6. Kumar S.P., Agrawal A., Kabra S., Gupta M., Gupta R.S. An analysis for AlGaN/GaN modulation doped field effect transistor using

accurate velocity-field dependence for high power microwave frequency applications. *Microelectronics Journal*, 2006, vol. 37, no. 11, pp. 1339–1346. https://doi.org/10.1016/j.mejo.2006.07.003
7. Ambacher O., Smart J., Shealy J.R., Weimann N.G., Chu K., Murphy M., Schaff W.J., Eastman L.F., Dimitrov R., Wittmer L., Stutzman M., Reiger W., Hilsenbeck J. Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN heterostructures. *Journal of Applied Physics*, 1999, vol. 85, no. 6, pp. 3222–3233. https://doi.org/10.1063/1.369664
8. Fiorentini V., Bernardini F., Ambacher O. Evidence for nonlinear macroscopic polarization in III-V nitride alloy heterostructures. *Applied Physics Letters*, 2002, vol. 80, no. 7, pp. 1204–1206. https://doi.org/10.1063/1.369664 ttps://doi.org/10.1063/1.1448668
9. Chattopadhyay M.K., Tokekar S. Temperature and polarization dependent polynomial based non-linear analytical model for gate capacitance of AlmGa1-mN/GaN MODFET. *Solid-State Electronics*, 2006, vol. 50, no. 2, pp. 220–227. https://doi.org/10.1016/j.sse.2005.10.016
10. Tyagi R.K., Ahlawat A., Pandey M., Pandey S. An analytical two-dimensional model for AlGaN/GaN HEMT with polarization effects for high power applications. *Microelectronics Journal*, 2007, vol. 38, no. 8-9, pp. 877–883. https://doi.org/10.1016/j.mejo.2007.07.003
11. Li M., Wang Y. 2-D Analytical model for current–voltage characteristics and transconductance of AlGaN/GaN MODFETs. *IEEE Transactions on Electron Devices*, 2008, vol. 55, no. 1, pp. 261–267. https://doi.org/10.1109/TED.2007.911076
12. Huque M.A., Eliza S.A., Rahman T., Huq H.F., Islam S.K. Temperature dependent analytical model for current–voltage characteristics of AlGaN/GaN power HEMT. *Solid-State Electronics*, 2009, vol. 53, no. 3, pp. 341–348. https://doi.org/10.1016/j.sse.2009.01.004
13. Rathi S., Jogi J., Gupta M., Gupta R.S. Modeling of hetero-interface potential and threshold voltage for tied and separate nanoscale InAlAs–InGaAs symmetric double-gate HEMT. *Microelectronics Reliability*, 2009, vol. 49, no. 12, pp. 1508–1514. https://doi.org/10.1016/j.microrel.2009.07.044
14. Mukhopadhyay P., Banerjee U., Bag A., Ghosh S., Biswas D. Influence of growth morphology on electrical and thermal modeling of AlGaN/GaN HEMT on sapphire and silicon. *Solid-State Electronics*, 2015, vol. 104, pp. 101–108. https://doi.org/10.1016/j.sse.2014.11.017
15. Gangwani P., Kaur R., Pandey S., Haldar S., Gupta M., Gupta R.S. Modeling and analysis of fully strained and partially relaxed lattice mismatched AlGaN/GaN HEMT for high temperature applications. *Superlattices and Microstructures*, 2008, vol. 44, no. 6, pp. 781–793. https://doi.org/10.1016/j.spmi.2008.07.004
16. Chattopadhyay M.K., Tokekar S. Thermal model for dc characteristics of AlGaN/GaN hemts including self-heating effect and non-linear polarization. *Microelectronics Journal*, 2008, vol. 39, no. 10, pp. 1181–1188. https://doi.org/10.1016/j.mejo.2008.01.043
17. Madhulika, Malik A., Jain N., Mishra M., Kumar S., Rawal D.S., Singh A.K. Nanoscale structural parameters based analytical model for GaN HEMTs. *Superlattices and Microstructures*, 2019, vol. 130, pp. 267–276. https://doi.org/10.1016/j.spmi.2019.04.040
18. Wu Y.F., Keller S., Kozodoy P., Keller B.P., Parikh P., Kapolnek D., Denbaars S.P., Mishra U.K. Bias dependent microwave performance of AlGaN/GaN MODFET's up to 100 V. *IEEE Electron Device Letters*, 1997, vol. 18, no. 6, pp. 290–292. https://doi.org/10.1109/55.585362
19. Dasgupta N., Dasgupta A. An analytical expression for sheet carrier concentration vs gate voltage for HEMT modelling. *Solid-State Electronics*, 1993, vol. 36, no. 2, pp. 201–203. https://doi.org/10.1016/0038-1101(93)90140-L

accurate velocity-field dependence for high power microwave frequency applications // Microelectronics Journal. 2006. V. 37. N 11. P. 1339–1346. https://doi.org/10.1016/j.mejo.2006.07.003
7. Ambacher O., Smart J., Shealy J.R., Weimann N.G., Chu K., Murphy M., Schaff W.J., Eastman L.F., Dimitrov R., Wittmer L., Stutzman M., Reiger W., Hilsenbeck J. Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN heterostructures // Journal of Applied Physics. 1999. V. 85. N 6. P. 3222–3233. https://doi.org/10.1063/1.369664
8. Fiorentini V., Bernardini F., Ambacher O. Evidence for nonlinear macroscopic polarization in III-V nitride alloy heterostructures // Applied Physics Letters. 2002. V. 80. N 7. P. 1204–1206. https://doi.org/10.1063/1.1448668
9. Chattopadhyay M.K., Tokekar S. Temperature and polarization dependent polynomial based non-linear analytical model for gate capacitance of AlmGa1-mN/GaN MODFET // Solid-State Electronics. 2006. V. 50. N 2. P. 220–227. https://doi.org/10.1016/j.sse.2005.10.016
10. Tyagi R.K., Ahlawat A., Pandey M., Pandey S. An analytical two-dimensional model for AlGaN/GaN HEMT with polarization effects for high power applications // Microelectronics Journal. 2007. V. 38. N 8-9. P. 877–883. https://doi.org/10.1016/j.mejo.2007.07.003
11. Li M., Wang Y. 2-D Analytical model for current–voltage characteristics and transconductance of AlGaN/GaN MODFETs // IEEE Transactions on Electron Devices. 2008. V. 55. N 1. P. 261–267. https://doi.org/10.1109/TED.2007.911076
12. Huque M.A., Eliza S.A., Rahman T., Huq H.F., Islam S.K. Temperature dependent analytical model for current–voltage characteristics of AlGaN/GaN power HEMT // Solid-State Electronics. 2009. V. 53. N 3. P. 341–348. https://doi.org/10.1016/j.sse.2009.01.004
13. Rathi S., Jogi J., Gupta M., Gupta R.S. Modeling of hetero-interface potential and threshold voltage for tied and separate nanoscale InAlAs–InGaAs symmetric double-gate HEMT // Microelectronics Reliability. 2009. V. 49. N 12. P. 1508–1514. https://doi.org/10.1016/j.microrel.2009.07.044
14. Mukhopadhyay P., Banerjee U., Bag A., Ghosh S., Biswas D. Influence of growth morphology on electrical and thermal modeling of AlGaN/GaN HEMT on sapphire and silicon // Solid-State Electronics. 2015. V. 104. P. 101–108. https://doi.org/10.1016/j.sse.2014.11.017
15. Gangwani P., Kaur R., Pandey S., Haldar S., Gupta M., Gupta R.S. Modeling and analysis of fully strained and partially relaxed lattice mismatched AlGaN/GaN HEMT for high temperature applications // Superlattices and Microstructures. 2008. V. 44. N 6. P. 781–793. https://doi.org/10.1016/j.spmi.2008.07.004
16. Chattopadhyay M.K., Tokekar S. Thermal model for dc characteristics of AlGaN/GaN hemts including self-heating effect and non-linear polarization // Microelectronics Journal. 2008. V. 39. N 10. P. 1181–1188. https://doi.org/10.1016/j.mejo.2008.01.043
17. Madhulika, Malik A., Jain N., Mishra M., Kumar S., Rawal D.S., Singh A.K. Nanoscale structural parameters based analytical model for GaN HEMTs // Superlattices and Microstructures. 2019. V. 130. P. 267–276. https://doi.org/10.1016/j.spmi.2019.04.040
18. Wu Y.F., Keller S., Kozodoy P., Keller B.P., Parikh P., Kapolnek D., Denbaars S.P., Mishra U.K. Bias dependent microwave performance of AlGaN/GaN MODFET's up to 100 V // IEEE Electron Device Letters. 1997. V. 18. N 6. P. 290–292. https://doi.org/10.1109/55.585362
19. Dasgupta N., Dasgupta A. An analytical expression for sheet carrier concentration vs gate voltage for HEMT modelling // Solid-State Electronics. 1993. V. 36. N 2. P. 201–203. https://doi.org/10.1016/0038-1101(93)90140-L

Authors

Azzeddine Farti — PhD, Associate Professor, Hassan II University of Casablanca, Casablanca, 20190, Morocco, https://orcid.org/0000-0002-8694-1471, azzeddine.farti-etu@etu.univh2c.ma
Abdelkader Touhami — D.Sc, Full Professor, Hassan II University of Casablanca, Casablanca, 20190, Morocco, SC 8158211000, https://orcid.org/0000-0001-8582-1884, abdelkadertouhami2016@gmail.com

Авторы

Фарти Азеддин — PhD, доцент, Университет Хасана II, Факультет наук Касабланки, Касабланка, 20190, Марокко, https://orcid.org/0000-0002-8694-1471, azzeddine.farti-etu@etu.univh2c.ma
Тухами Абделькадер — D.Sc, профессор, профессор, Университет Хасана II, Факультет наук Касабланки, Касабланка, 20190, Марокко, SC 8158211000, https://orcid.org/0000-0001-8582-1884, abdelkadertouhami2016@gmail.com

Received 08.09.2021
Approved after reviewing 03.12.2021
Accepted 28.01.2022

Статья поступила в редакцию 08.09.2021
Одобрена после рецензирования 03.12.2021
Принята к печати 28.01.2022

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-155-166

Imputation and system modeling of acid-base state parameters for different groups of patients

Dmitry I. Kurapeev[1], Mikhail S. Lushnov[2], Tianxing Man[3], Natalia A. Zhukova[4]✉

[1,2] Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation
[3,4] ITMO University, Saint Petersburg, 197101, Russian Federation
[4] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] dkurapeev@gmail.com, https://orcid.org/0000-0002-2190-1495
[2] Lushnov_ms@almazovcentre.ru, https://orcid.org/0000-0002-9683-1858
[3] mantx626@gmail.com, https://orcid.org/0000-0003-2187-1641
[4] nazhukova@mail.ru✉, https://orcid.org/0000-0001-5877-4461

Abstract
The paper investigated the possibility of correct replacement of missing values in sets of acid-base state in the artery and vein in different groups of patients with different outcomes of the disease: “discharged”, “died”, “transferred to another medical institution”, as well as prospects for the application of individual optimization multidimensional estimates of these biomedical parameters in the form of projections on a one-dimensional space. The relevance of the above tasks is determined by the need for the full use of medical data in the analysis of large repositories of information of medical organizations and the provision of verified multidimensional assessments of biomedical systems to doctors from a large range of patient health indicators. A statistical method has been applied to verify the correctness of imputation data sets using discriminant analysis procedures. Further, the imputed data set was processed to obtain a symmetric correlation matrix optimized in a certain way and the accompanying logarithms of criterion functions that are individual system assessments of the condition of each patient in different groups of patients at a certain point in the study. After that, to identify differences in the logarithms of the criterion functions of the acid-base state parameters between groups of patients with different outcomes, the authors used the method of calculating multidimensional Hotelling T^2 statistics. The correctness of the application of discriminant analysis procedures to verify the imputation of data sets is shown. Differences in the logarithms of the criteria functions of the acid-base state indicators between venous and arterial blood by patient outcome groups were revealed. Significant differences in the parameters of acid-base state based on the multidimensional statistics of T^2 Hotelling between groups of patients with different outcomes were revealed. It is found that data imputation significantly increases the volume and representativeness of the sample under study. It is demonstrated that the substituted data make it possible to carry out a systematic statistical assessment of the totality of body parameters based on the calculation of the logarithms of the criterion functions of the acid-base state. Such logarithms make it possible to reliably distinguish patients in different groups of patients by outcomes in three groups: “discharged”, “deceased”, “transferred to another medical institution”. 100 % differences of biochemical parameters according to the multivariate T^2 Hotelling statistics between these three groups of patients with COVID-19 are shown. The results of the study can be applied in the development of information systems of individual medical biochemical and hematological devices and analyzers and corresponding artificial intelligence systems in the future.

Keywords
acid-base state, imputation, medical information system, COVID-19, criterion function

Acknowledgements
The research was supported by the Russian Science Foundation under Grant No. 17-15-01177.
The research was carried out within the framework of the budget theme No. 0060-2019-0011.

For citation: Kurapeev D.I., Lushnov M.S., Man T., Zhukova N.A. Imputation and system modeling of acid-base state parameters for different groups of patients. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 155–166. doi: 10.17586/2226-1494-2022-22-1-155-166

© Kurapeev D.I., Lushnov M.S., Man T., Zhukova N.A., 2022

УДК 616.092+616.98, 303.425.6, 338.45, 519.873, 681.518

Вменение и системное моделирование параметров кислотно-основного состояния различных групп пациентов

Дмитрий Ильич Курапеев[1], Михаил Степанович Лушнов[2], Тяньсин Ман[3], Наталья Александровна Жукова[4]✉

[1,2] Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация
[3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[4] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] dkurapeev@gmail.com, https://orcid.org/0000-0002-2190-1495
[2] Lushnov_ms@almazovcentre.ru, https://orcid.org/0000-0002-9683-1858
[3] mantx626@gmail.com, https://orcid.org/0000-0003-2187-1641
[4] nazhukova@mail.ru✉, https://orcid.org/0000-0001-5877-4461

Аннотация

Предмет исследования. Исследована возможность корректной замены недостающих значений в наборах кислотно-основного состояния в артерии и вене в разных группах пациентов с различными исходами заболевания: «выписан», «умер», «переведен в другое медицинское учреждение». Рассмотрены перспективы применения индивидуальных оптимизационных многомерных оценок данных биомедицинских параметров в виде проекций на одномерное пространство. Решение поставленных задач необходимо для полноценного использования медицинских данных при анализе больших хранилищ информации медицинских организаций и предоставления врачам из большого круга показателей о здоровье пациентов верифицированных многомерных оценок биомедицинских систем. **Методы.** Применен статистический метод для проверки корректности в наборах данных вменения с использованием процедур дискриминантного анализа. Выполнена оценка импутированного набора данных для получения оптимизированной симметричной корреляционной матрицы и сопутствующих логарифмов критериальных функций. Получены индивидуальные системные оценки состояния каждого пациента в разных группах пациентов в определенный момент исследования. Применен метод вычисления многомерной статистики Хотеллинга T^2 для выявления различий в логарифмах критериальных функций параметров кислотно-основного состояния между группами пациентов с различными исходами. **Результаты.** Показана корректность применения процедур дискриминантного анализа для проверки вменения наборов данных. Выявлены статистически значимые отличия логарифмов критериальных функций показателей кислотно-основного состояния между венозной и артериальной кровью и биохимических параметров крови на основе многомерной статистики Хотеллинга T^2 между группами пациентов с различными исходами. **Практическая значимость.** Доказано, что импутация данных значительно увеличивает объем и представительность исследуемой выборки. Продемонстрировано, что замещенные данные позволяют проводить системную статистическую оценку совокупности параметров организма на основе расчета логарифмов критериальных функций кислотно-основного состояния. Такие логарифмы позволяют точно различать пациентов по исходам в трех группах: «выписанные», «умершие», «переведенные в другое медицинское учреждение». Показаны 100 % различия биохимических показателей по многомерной статистике Хотеллинга T^2 между указанными тремя группами пациентов с COVID-19. Результаты исследования могут быть применены при разработке информационных систем отдельных медицинских биохимических и гематологических приборов и анализаторов и в перспективе соответствующих систем искусственного интеллекта.

Ключевые слова

кислотно-основное состояние, вменение, медицинская информационная система, COVID-19, критериальная функция

Благодарности

Исследование выполнено при поддержке Российского научного фонда в рамках гранта № 17-15-01177.
Исследование проводилось в рамках бюджетной темы № 0060-2019-0011.

Ссылка для цитирования: Курапеев Д.И., Лушнов М.С., Ман Т., Жукова Н.А. Вменение и системное моделирование параметров кислотно-основного состояния различных групп пациентов // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 22, № 1. С. 155–166 (на англ. яз.). doi: 10.17586/2226-1494-2021-22-1-155-166

Introduction

Nowadays, there are currently millions of records accumulated in medical information systems (MIS) and databases. Data tend to contain a significant number of omissions. Medical data can be analyzed from physiological system positions [1] and synergetics [2]. In the first case, when considering MIS, the term “system” is understood primarily from the technical point of view as data accumulation and storage, and in the second case, it is used as a medical and physiological concept for the meaningful modeling of the conditions of patients with various diseases. Therefore, for the qualitative modeling of the systemic conditions of patients, it is important to take into account the maximum possible amount of data without their distortion and loss of information since the system model implies the most extensive filling of data sets. The results of the research and scientific ideas of academician

P.K. Anokhin [1] are taken as the basis for the description of physiological systems in this work.

The aim of the work is to conduct systematic studies that are based on correctly replaced missing values and the assessment of the average criterion functions (CF) of a set of parameters of the acid-base state of arterial and venous blood in different groups of patients (DGP) [3–8].

Materials and Methods

The calculation of the criterion functions (CF) is carried out based on the results of measurements for many individual parameters that characterize the state of the system under study. A correlation matrix is constructed for the entire sample of the biosystem, which is subjected to a special transformation using branches and boundaries with the selection of the optimal subset of features and the evaluation of the criterion function for each patient. The method is based on the estimation of some monotone function — CF from some biological set (**A**). The algorithm is based on calculating the maximum CF based on a certain quadratic form and on finding the largest set of n variables that maximizes the CF for the entire subset containing m features.

CF is calculated using the quadratic form: $\mathbf{C}(\mathbf{A}_m) = (\mathbf{X}_m{}^T)\mathbf{S}_m{}^{-1}(\mathbf{X}_m)$, where $\mathbf{A}_m$ is a set of m variables, $\mathbf{X}_m$ is vector of variables (a set of bioparameters — the functional system of a particular patient) and $\mathbf{S}_m$ is symmetric positive definite correlation matrix of size $m \times m$; symbol $\mathbf{X}_m{}^T$ means the operation of transposing a vector, $\mathbf{S}_m{}^{-1}$ is the operation of calculating the inverse matrix [8].

Thus, the above information shows that correctly constructed matrices without missing data can ensure successful modeling.

The relevance of the problem focusing on the influence of omissions in data on the results and conclusions of medical research is noted in a large number of publications [9, 10].

In the field of medical statistics (the generally accepted Russian and English terminology has not been defined yet) the following are the terms used to search for publications: data with omissions, omissions or omitted values (observations), censored values (observations), censorship (the process of forming censored data and data with omissions), random omissions, non-random omissions, non-random censorship (that is, a non-random mechanism for forming omissions and censored observations), offset (in the results and the conclusions of the study), the loss of part of the data, imputation (replacement of the missing value by its assessment) [11]. The omissions in the datasets can be processed using statistics, for example, using multivariate analysis.

The problem of omissions and the resulting shifts in research output is most relevant in the branches of medicine dealing with extreme and terminal conditions, that is, in traumatology [12, 13], resuscitation [14], emergency cardiology [15, 16]. The problem of data loss in experimental and clinical pharmacology studies remains relevant [17]. In the USA, a special regulatory act has been issued [18], which defines 3 gradations of omissions in clinical pharmacology data, depending on the degree of their randomness, and gives instructions for their statistical processing. However, some experts in this field express the opinion that this is not recognized as a serious problem, or consider it a nuisance that should be ignored [16].

Usually, the data of sociological research obtained through interviews are characterized by a large number of omissions [19]. To work effectively with them, it becomes necessary to fill in the gaps in the data since simply discarding observations containing missing values can lead to changes in the statistical characteristics of the sample [20].

Despite many years of development of methods for analyzing data with omissions, satisfactory solutions for many medical problems have not been found yet. Most authors recognized that there is no universal algorithm for statistical and mathematical processing of data with omissions [17, 21]. As a result, for example, in the work [22], 6 methods of data processing with omissions were used in parallel.

One of the most common methodological approaches is the imputation (replacement) of the missing value by its statistical assessment, that is, the value obtained from the preserved (non-missed) values that are closest to the missed one. Among such methods, multiple imputation is very promising, and it is used, in particular, in the works [13, 21, 23, 24].

The implementation of the MICE algorithm (R-package) made it possible to form the database necessary for the study with the restored values [25]. The obtained data sets are used to solve the problem of modeling the dependence of the labor supply on the individual's health characteristics using spatial regression with fixed effects on panel data. It is shown that filling in the gaps in data makes it possible to eliminate some obstacles that arise during econometric modeling. In addition, excluding observations with omissions would probably lead to a bias in parameter estimates [26, 27].

Thus, from the conducted literature review, it can be concluded that in recent years, methods for statistical and mathematical analysis of data with omissions have been intensively developed.

Ethical considerations. In this paper, studies of depersonalized acid-base state (ABS) data of DGP were conducted. The purpose of the study was not to make contacts with patients and to store and process their personal data.

Results

When analyzing the biochemical parameters in DGP (patients with COVID-19), a systematic approach to functional systems was applied according to [2]. At the beginning of March 2021, 1,687 such patients were registered in the MIS qMS of the Almazov National Medical Research Center (Saint Petersburg, Russia). Among them, 3 groups of patients were identified: "discharged", "died", and "transferred to another medical facility (MF)".

The gender and age composition of the patients was as follows. There were about the same number of men and women. Men 21–35 years — 6 %, 35–60 years —

42 %, 60–75 years — 35 %, 75–90 years — 16 %. Women 20–35 years — 6 %, 35–55 years — 28 %, 55–75 years — 49 %, 75–90 years — 16 %.

ABS samples were made to patients as needed according to the indications and prescriptions of the medical staff. From the database, 18,658 studies of ABS of patients with COVID-19 were uploaded to the "Statistica" system of the company "StatSoft-Tibco". These acid-base state (ABS) samples consisted of 21 parameters, e.g. arteries: ABE_(art) — excess of bases, Ca^{2+}_(art) — calcium ion concentration, Cl^{-}_(art) – concentration of the chlorine ion, ctBil_(art) — bilirubin, $ctCO_2(B)$_(art) — total carbon dioxide content (calculated), $ctCO_2(P)$_(art) — total carbon dioxide content in the plasma (calculated), ctHb_(art) — reference hemoglobin level, ctO_2_(art) — the total concentration of oxygen in the blood including the concentration of oxygen dissolved in the plasma, Glu_(art) — glucose concentration, $HCO_3(P)$_(art) — plasma bicarbonate, Hct_(art) — hematocrit, Lac_(art) — lactate content, Na^{+}_(art) — sodium ion concentration, p50_(art) — the affinity of hemoglobin for oxygen, CO_2_(art) — partial pressure of oxygen, pH_(art) — acidity, O_2_(art) — partial pressure of oxygen, SBE_(art) — lack of bases, sO_2_(art) — oxygen saturation, K^{+}_(art) — potassium ion concentration, Osmolarity_(art) — arterial blood osmolarity.

A line-by-line deletion can destroy a noticeable part of the data if missing values are scattered throughout the data table or are located in several variables. And this is true in our case.

In such cases, a deeper analysis is usually made, using additional R-libraries for multiple substitution, such as "Mice". One can look at the relationship between the presence of missing values of one variable and the observed values of other variables and what happens in this case. The imputation for the parameters of the artery and vein was performed separately. 5 imputation sets were generated to select the best set.

The predicted mean matching method (PMM) was set as an imputation model for quantitative variables. PMM is a type of linear regression in which the imputation values calculated from the regression model are compared with the nearest observed values. The method of fully conditional specification of PMM — "predictive mean matching" was used due to the degeneracy of some regression parameters. The minimum and maximum values of the parameters of the 0-data set are used as restrictions on the maximum and minimum values for subsequent imputations (Table 1).

The data group the artery ABS. The table shows the minimum and maximum values of arterial blood parameters. They are the same for all five imputed data sets in order to avoid data distortion and prevent biases of statistical sets. These are the set values for imputation.

In all six data sets, including the initial imputed one, the entire set of 21 parameters for the artery and vein ABS were distributed according to the normal distribution law according to the Kolmogorov-Smirnov and Liliefors criteria with a probability less than 0.01.

The 5th group of the set of ABS for arteria with imputation is presented in Table 2. Similar sets of

Table 1. Average statistical parameters of the artery ABS in the group without imputation (0-data set)

Variable	Number of observations	Mean	Minimum	Maximum	Standard deviation
ABE_art	11871	0.4503	–31.0000	28.0000	5.10135
Ca^{2+}_art	11563	0.9676	0.2500	151.0000	1.40431
Cl^{-}_art	11541	109.9747	79.0000	149.0000	7.34768
ctBil_art	1601	26.8499	0.0000	265.0000	26.36076
$ctCO_2B$_art	11713	48.0726	7.4200	107.9000	12.01896
$ctCO_2P$_art	11513	55.2930	8.6000	121.3000	13.55506
ctHb_art	11885	107.7292	10.0000	195.0000	22.52998
ctO_2_art	11277	14.2235	1.8000	33.3000	2.98734
Glu_art	11647	8.9560	0.1000	37.0000	4.01030
HCO_3P_art	11894	24.7031	2.9000	59.1000	5.10861
Hct_art	2659	1.1126	0.0400	59.0000	5.10570
Lac_art	11684	2.5942	0.0000	30.0000	2.65705
Na_art	11895	142.3490	97.0000	195.0000	7.58373
p50_art	10533	25.7687	11.0000	262.4000	5.15931
pCO_2_art	11901	40.3392	10.0000	192.0000	13.14258
pH_art	11902	7.4116	6.7300	7.7600	0.09963
pO_2_art	11873	96.3650	18.0000	541.0000	41.05349
SBE_art	1013	1.7253	–27.2700	30.0000	7.60385
sO_2_art	11880	95.4498	10.5000	100.9000	6.50825
K^{+}_art	11861	4.1157	1.8000	19.6000	0.75089
Osmolarity_art	11317	293.0776	202.0000	402.0000	16.08119

Table 2. Average statistical parameters of the ABS of the artery in the imputation group 5 (5-data set)

Variable	Number of observations	Mean	Minimum	Maximum	Standard deviation
ABE_art	18658	0.4434	–31.0000	28.0000	5.18306
Ca^{2+}_art	18658	0.9885	0.2500	151.0000	2.20358
Cl^{-}_art	18658	109.9232	79.0000	149.0000	6.76504
ctBil_art	18658	35.1288	0.0000	265.0000	41.66691
$ctCO_2B$_art	18658	47.8398	7.4200	107.9000	9.98395
$ctCO_2P$_art	18658	54.6650	8.6000	121.3000	11.45273
ctHb_art	18658	107.2712	10.0000	195.0000	19.92765
ctO_2_art	18658	14.2043	1.8000	33.3000	2.60291
Glu_art	18658	9.0953	0.1000	37.0000	4.36174
HCO_3P_art	18658	24.6568	2.9000	59.1000	4.36005
Hct_art	18658	1.0042	0.0400	59.0000	4.95924
Lac_art	18658	2.5468	0.0000	30.0000	2.53684
Na_art	18658	142.3128	97.0000	195.0000	6.37804
p50_art	18658	25.0472	11.0000	262.4000	9.41725
pCO_2_art	18658	40.7309	10.0000	192.0000	15.32162
pH_art	18658	7.4134	6.7300	7.7600	0.14171
pO_2_art	18658	100.1013	18.0000	541.0000	46.94539
SBE_art	18658	0.0432	–27.2700	30.0000	5.36797
sO_2_art	18658	95.4232	10.5000	100.9000	6.30708
K^{+}_art	18658	4.1217	1.8000	19.6000	0.82043
Osmolarity_art	18658	293.2238	202.0000	402.0000	13.34610

data imputation were obtained for the models of the ABS vein.

It is also noteworthy that all significantly different parameters have smaller average values compared to the initial indicators and smaller standard deviations (variances), which indicates that at least the imputed values do not fluctuate more significantly in the vicinity of their average values, compared with the initial sample.

To assess the degree of differences between the original array and the imputed data, we performed a discriminant analysis between these samples (sets of imputations). Canonical discriminant functions for imputed artery data sets are obtained (Table 3).

For the analysis, the first 4 of the canonical discriminant functions were used, which 100 % describe the variances and the first of which "explains" the process of "scattering" by 96.3 %, and the second by 2.1 % (in total 98.4 %). The first two canonical discriminant functions for the 1st data set (without imputation) and the fifth set of imputed data are shown in Fig. 1.

According to Table 3, the results of classification and comparison of forecasts for belonging to the implanted artery data were obtained and showed that 21.9 % of the initial grouped observations were classified correctly. This means that only 22 % of the imputed observations differ from each other, which implies that the differences between the imputed sets of arterial blood parameters are not significant. This fact suggests that the imputed sets do not differ from each other, but they allow for the problems of system models to use correctly (in the sense of without omissions) the criteria functions of the ABS, which will be discussed below.

The graphical results of comparing the canonical discriminant functions (for two canonical functions that explain 98.4 % of the variance of the artery parameters in total) are presented in Fig. 1, in which the practical equivalence of the initial and the fifth imputed artery data sets is clearly demonstrated.

Canonical discriminant functions for imputed vein data sets are obtained in the same way (Table 4).

Table 3. Canonical discriminant functions of the ABS artery and their eigenvalues

Function	Eigenvalue	Variances, %	Total, %	Canonical correlation
1	0.123	96.3	96.3	0.331
2	0.003	2.1	98.4	0.052
3	0.002	1.3	99.7	0.041
4	0.000	0.3	100.0	0.021

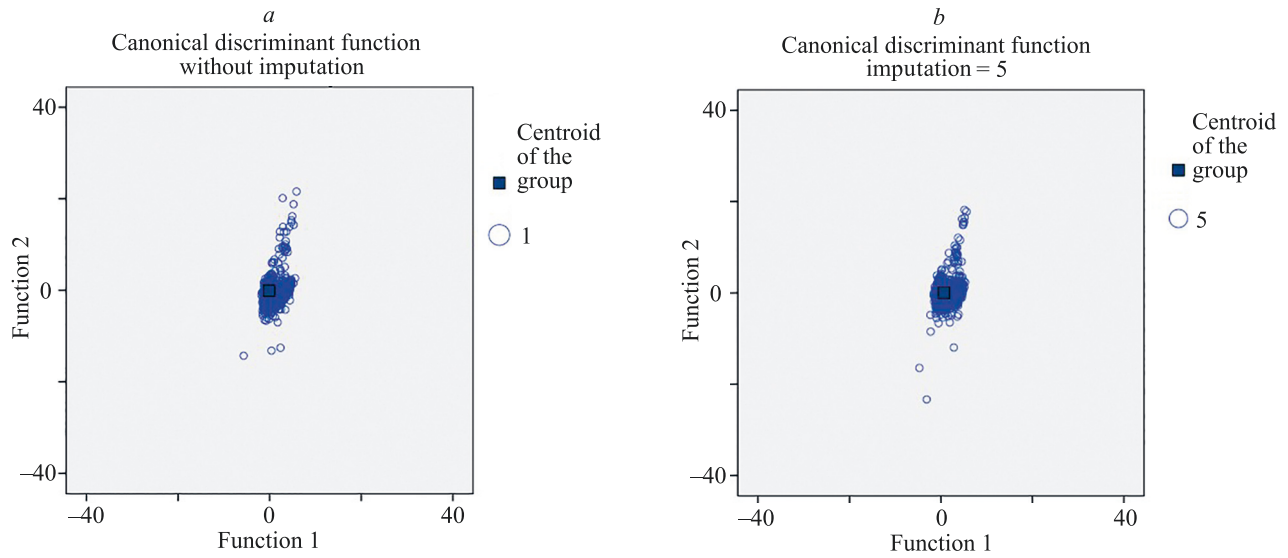


Fig. 1. Comparison of canonical discriminant functions (centroids of groups) of the ABS artery for data sets without imputation (*a*) and in the 5th group of imputation (*b*)

Table 4. Canonical discriminant functions of the ABS vein and their eigenvalues

Function	Eigenvalue	Variances, %	Total, %	Canonical correlation
1	0.101	93.8	93.8	0.303
2	0.004	3.7	97.5	0.063
3	0.002	1.8	99.3	0.043
4	0.001	0.7	100.0	0.028

For the analysis, the first 4 of the canonical discriminant functions were used, which 100 % describe the variances and the first of which "explains" the process of "scattering" by 93.8 %, and the second by 3.7 % (in total 97.5 %). The first two canonical discriminant functions for the 1st data set (without imputation) and the fifth set of imputed vein data are shown in Fig. 2.

The results of classification and comparison of forecasts for belonging to the imputed data of the vein are obtained, from which it follows that 25.9 % of the initial grouped observations are classified correctly. This means that only 26 % of the implanted observations differ from each other, which means that the differences between the implanted sets of venous blood parameters are not significant.

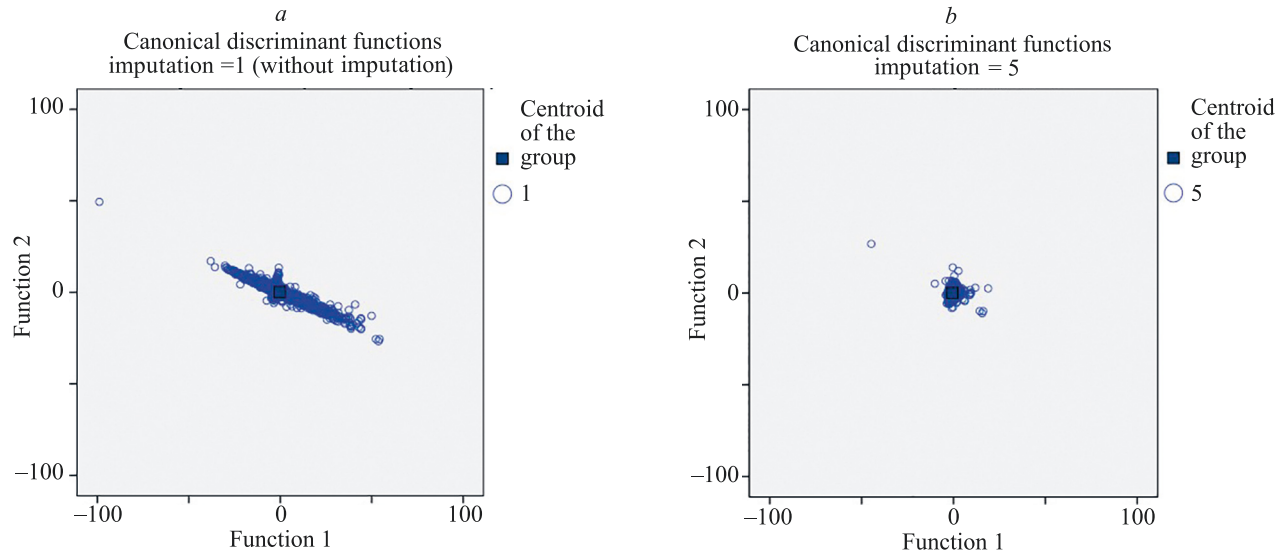


Fig. 2. Comparison of canonical discriminant functions (centroids of groups) of braid veins for data sets without imputation (*a*) and in the 5th group of imputation (*b*)

Table 5. Hotelling T^2 statistics for comparing the parameters of ABS in the artery and vein of sample groups: "Died", "Discharged", "Transferred to another medical facility"

Vessel parameter	Comparison groups	Hotelling T^2
Art	Died – Discharged	1220.570
Art	Died – Transferred to another medical facility	1742.560
Art	Discharged – Transferred to another medical facility	741.517
Vein	Died – Discharged	1696.840
Vein	Died – Transferred to another medical facility	496.160
Vein	Discharged – Transferred to another medical facility	351.272

The graphical results of comparing the canonical discriminant functions (for two canonical functions that explain 97.5 % of the variances of the vein parameters in total) are presented in Fig. 2. The practical equivalence of the original and the fifth imputed data sets is clearly demonstrated.

Let us move on to a multidimensional comparison based on the statistics of Hotelling T^2 of imputed data between 3 groups of patients (Table 5): "discharged", "died", "transferred to another medical facility". A significance level $p = 0.0000$ for F-statistics was obtained for all comparison groups.

From Table 5, presented above, it follows that the 21 ABS parameters in both artery and vein in a multidimensional space differ in pairs for 3 "states" according to the multidimensional statistics of Hotelling T^2. This remarkable fact confirms, firstly, the significant difference between these conditions in the outcomes of treatment of patients with COVID-19, secondly, the adequacy of the applied multidimensional Hotelling T^2 statistics, and thirdly, the correctness of the initial data imputation, along with the results of the discriminant analysis with its canonical discriminant functions and centroids.

Dynamic system statistical analysis of ABS parameters in DGP

Now we present the results of a systematic dynamic analysis of the ABS for each individual patient with COVID-19. System analysis assumes a fully filled data matrix. Hence, the reasons for the above data imputation are clear.

The average statistical indicators of the decimal logarithms of the CF ABS were calculated on an IBM-compatible PC with a clock frequency of 2.5 GHz of the Intel i5-2400S processor for 7 hours. The reliability of the differences in the averages and variances for survivors, deceased, and transferred to other medical institutions of patients is calculated; the results are shown in Fig. 3.

It follows from Fig. 3 that both the mean and variance values of the logarithm of the CF of the ABS in patients with COVID-19 differ for the artery and vein according to the outcomes in the groups "discharged", "died" and "transferred to other medical facilities" significantly. This fact is confirmed by the graphs of the average values of this indicator by groups. The results are noteworthy: despite the very insignificant numerical differences between the averages and variances, they are nevertheless statistically significant. In addition, these are decimal logarithms, and they are the power of the number that needs to be raised to get the true value of the CF of the ABS, which in turn reflects the "level of correlation" within the ABS system of a particular patient. Moreover, since the logarithms have a value greater than 3 in our case, the indicators of CF ABS have an order of magnitude from 1000 to 10 000, and the differences in the average logarithms for the deceased and those discharged from the clinic are even on near 0.12, they can have a statistically significant control effect on the ABS system.

According to the data in Fig. 3, it can be seen that all the compared groups differ significantly in mean values and variances, except for the absence of differences in the average values of the logarithm of the CF ABS vein between the groups "discharged" –"transferred to another medical facility" (Fig. 3, *f*).

Thus, from Fig. 3, a pairwise difference in the logarithms of the CF ABS for all groups of states ("died", "discharged", "transferred to another medical facility") follows, which emphasizes the diagnostic value of the system indicator under consideration.

We will demonstrate how the decimal logarithm of the system criterion function of the ABS behaves for individual groups of patients: "died" (Fig. 4), "discharged" (Fig. 5), "transferred to another medical facility" (Fig. 6).

Discussion

The results of the studies revealed a certain statistically significant determinism in the response of ABS to Covid infection (DGP) from the standpoint of dividing patients into groups: "discharged", "died", and "transferred to another medical facility". There is probably a certain uncertainty in the outcome of the disease, depending on the initial state of the ABS at the time of contact of the coronavirus with the human body. This is indicated by the "levels – groups" of outcomes. This paper presents the results of studies for men and women at all ages. The results for different age categories for men and women may likely differ. Subsequent studies may reveal the peculiarities of the response of gender and age groups of people to contact with the coronavirus.

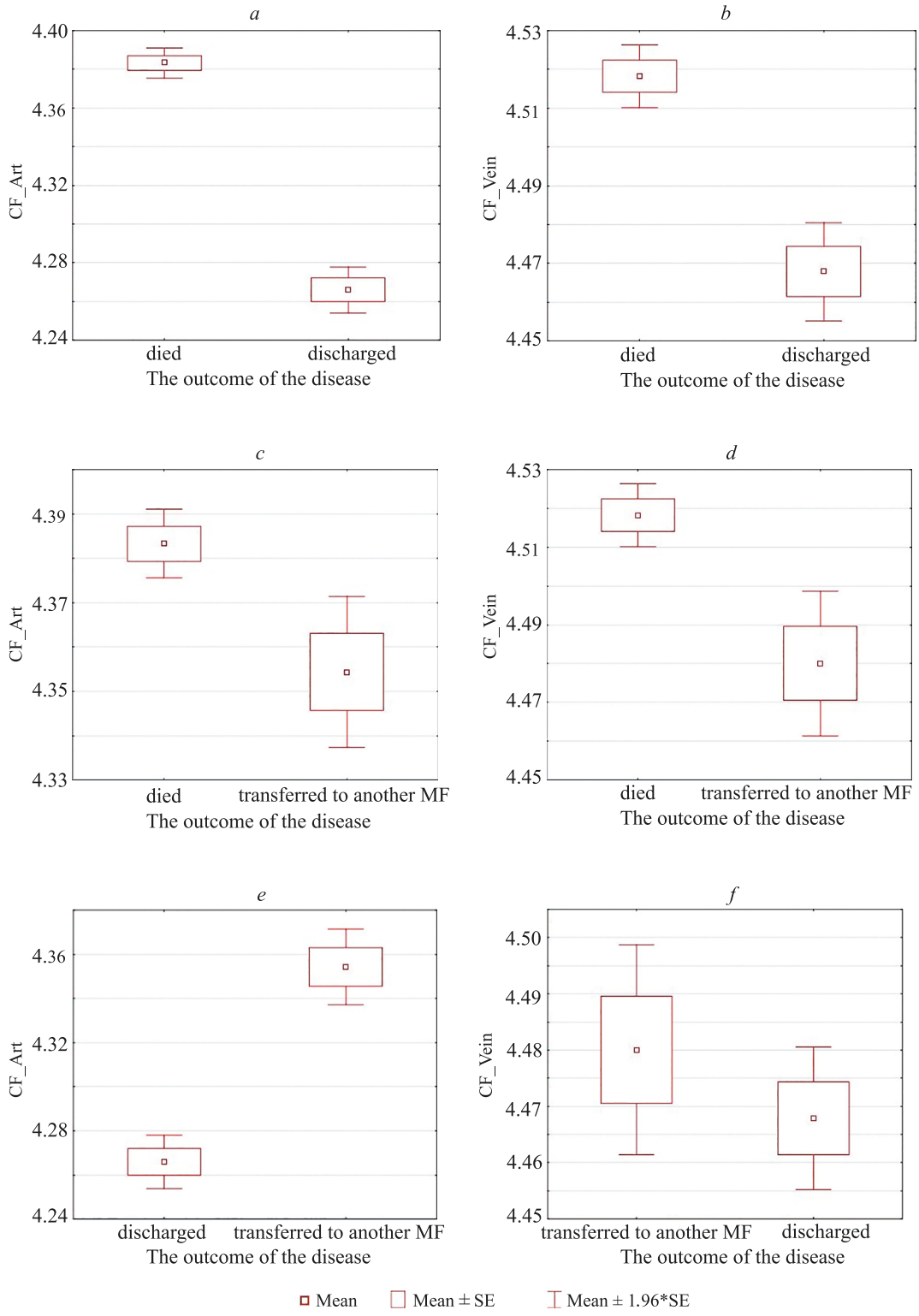


Fig. 3. Graphs of pairwise differences in the average values of the logarithms of the CF ABS of the arteries and veins in groups: “died”, “discharged” (*a*) and (*b*); “died”, “transferred to another medical facility” (*c*) and (*d*); “discharged”, “transferred to another medical facility” (*e*) and (*f*)

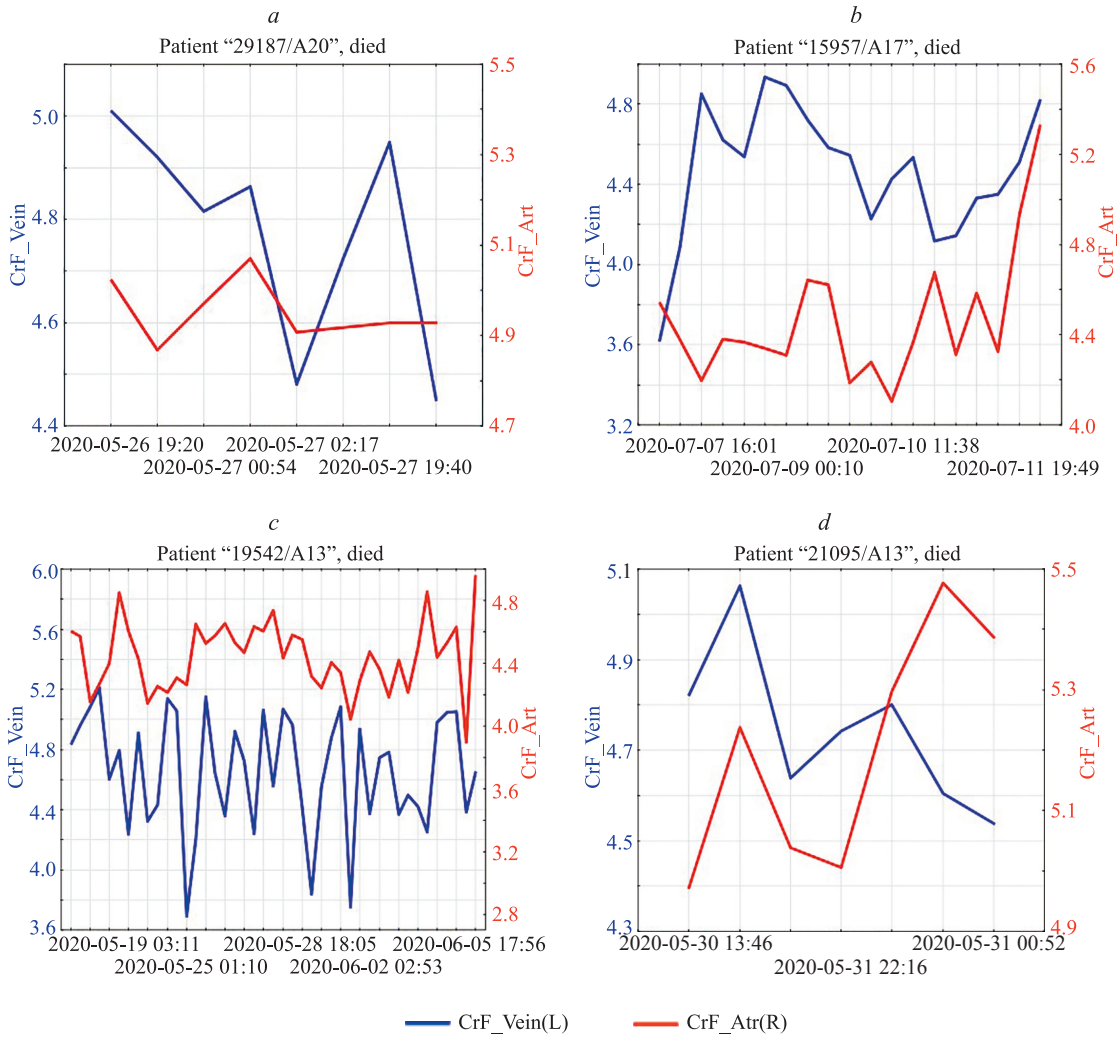


Fig. 4. Examples of the dynamics of logarithms of CF ABS by patient groups: “Died”: severe patient (fatal outcome, arterial and vein indicators are almost synchronous) (*a*) and (*b*); fatal patient (there is a tendency to synchronicity of arterial and vein indicators, at the final stage of the trend (the last third), the opposite phase of the processes is visible) (*c*) and (*d*)

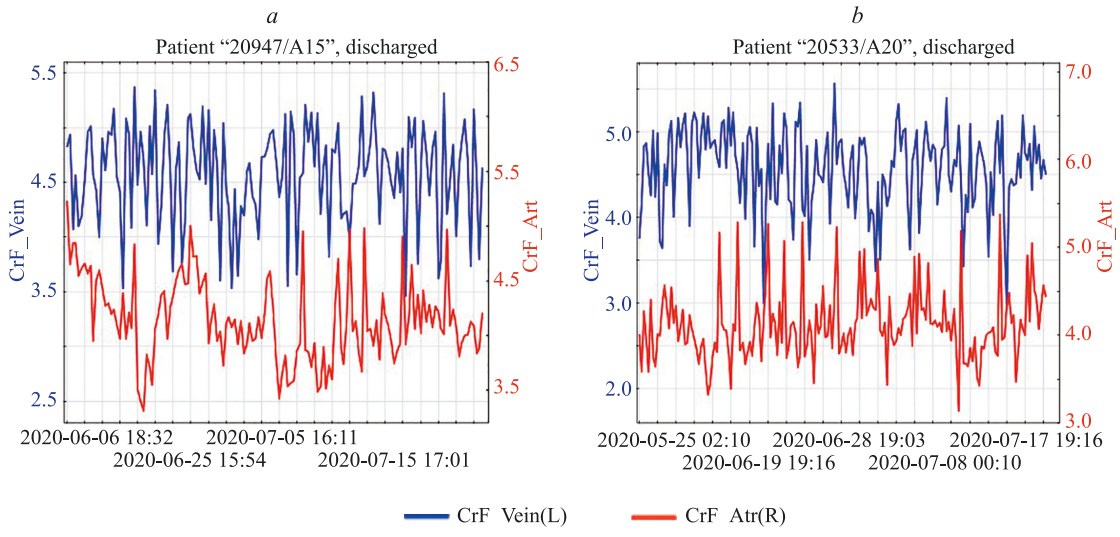


Fig. 5. Dynamics of logarithms of CF ABS by patient groups: “Discharged”. In both cases, the logarithms demonstrate relative stability (*a*) and (*b*)

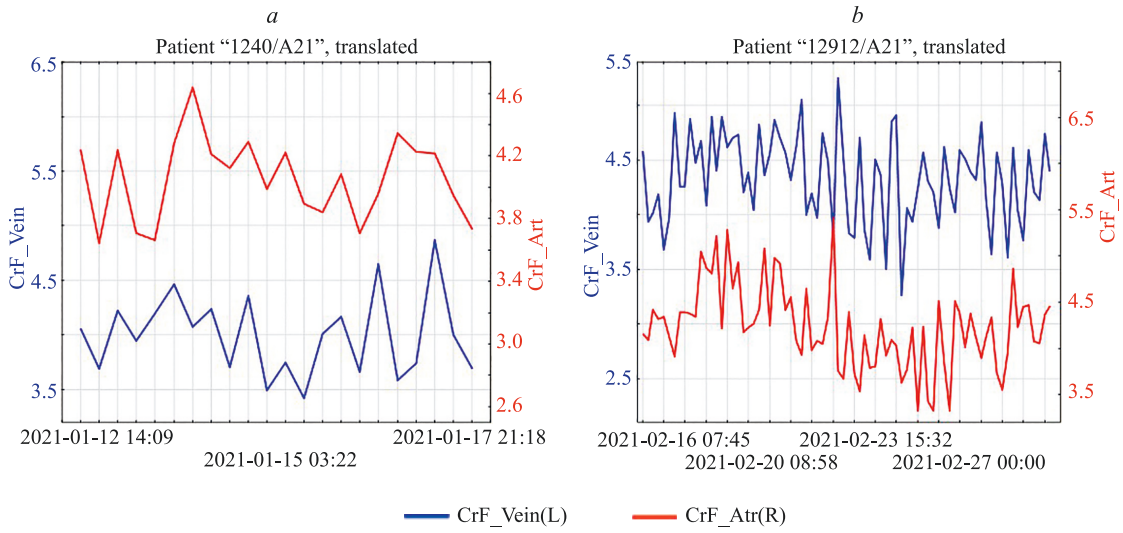


Fig. 6. Examples of the dynamics of logarithms of CF ABS by patient groups: "Transferred to another medical facility". Logarithms demonstrate a long-term trend and synchronicity (*a*) and (*b*)

The clinical implications of this work are that research doctors receive a new analytical tool for evaluating the multidimensional ABS of DGP in the form of a criterion function, by which it is possible to determine the dynamics of a particular patient's system. The social consequences of this scientific work show that when treating and evaluating patients with COVID-19, it is necessary to pay attention to the prognostic value of the dynamics of CF ABS, which can indicate the possibility of a fatal outcome in a certain category of patients.

Conclusion

The work shows:

— Data imputation (substitution) significantly increases the volume of the verified sample.

— The substituted data make it possible to carry out a systematic statistical assessment of the totality of the parameters of the organism based on the calculation of the logarithms of the criteria functions of the ABS.

— The logarithm of CF ABS, being a systematic statistical assessment, allows us to distinguish in DGP by outcomes in three groups reliably: "discharged", "died", and "transferred to another medical facility".

References

1. Anokhin P.K. Theory of a functional system. *Uspekhi Fiziologicheskikh Nauk*, 1970, vol. 1, no. 1, pp. 19–54. (in Russian)
2. Haken G. Information and Self-Organization: *A Macroscopic Approach to Complex Systems*. Springer-Verlag, 1988, 188 p.
3. Lushnov A.M., Lushnov M.S. *Medical Information Systems: Multidimensional Analysis of Medical and Environmental Data*. Saint Petersburg, Helikon Plus, 2013, 458 p. (in Russian)
4. Kurapeev D.I., Lushnov M.S., Osipov V.Yu., Vodyaho A.I., Zhukova N.A. Synthesis of integral models of system dynamics of an Acid-Base State (ABS) of patients at operative measures. *Acta Scientific Medical Sciences*, 2019, vol. 3, no. 3, pp. 16–29.
5. Kurapeev D.I., Lushnov M.S., Zhukova N.A. Comparison of integral models of systemic dynamics of Acid-Base State of venous and arterial circulating blood in patients with surgical interventions. *Acta Scientific Medical Sciences*, 2019, vol. 3, no. 6, pp. 38–48.
6. Lebedev S., Zhukova N., Vodyaho A., Kurapeev D., Lushnov M. An ontology-driven toolset for fast prototyping of medical data processing systems. *International Journal of Biology and Biomedical Engineering*, 2017, vol. 11, pp. 135–142.
7. Narenda P.M., Fukunaga K. A branch and bound algorithm for feature subset selection. *IEEE Transactions Computers*, 1977, vol. C-26, no. 9, pp. 917–922. https://doi.org/10.1109/TC.1977.1674939
8. Ridout M.S. An improved branch and bound algorithm for feature subset-selection. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1988, vol. 37, no. 1, pp. 139–147. https://doi.org/10.2307/2347512
9. Bech C.N., Brabrand M., Mikkelsen S., Lassen A. Risk factors associated with short term mortality changes over time, after arrival to the emergency department. *Scandinavian Journal of Trauma*,

Литература

1. Анохин П.К. Теория функциональной системы // Успехи физиологических наук. 1970. Т. 1. № 1. С. 19–54.
2. Хакен Г. Информация и самоорганизация: Макроскопический подход к сложным системам / пер. с англ. М.: Мир, 1991. 240 с.
3. Лушнов А.М., Лушнов М.С. Медицинские информационные системы: многомерный анализ медицинских и экологических данных. СПб.: Геликон Плюс, 2013. 458 с.
4. Kurapeev D.I., Lushnov M.S., Osipov V.Yu., Vodyaho A.I., Zhukova N.A. Synthesis of integral models of system dynamics of an Acid-Base State (ABS) of patients at operative measures // Acta Scientific Medical Sciences. 2019. V. 3. N. 3. P. 16–29.
5. Kurapeev D.I., Lushnov M.S., Zhukova N.A. Comparison of integral models of systemic dynamics of Acid-Base State of venous and arterial circulating blood in patients with surgical interventions // Acta Scientific Medical Sciences. 2019. V. 3. N 6. P. 38–48.
6. Lebedev S., Zhukova N., Vodyaho A., Kurapeev D., Lushnov M. An ontology-driven toolset for fast prototyping of medical data processing systems // International Journal of Biology and Biomedical Engineering. 2017. V. 11. P. 135–142.
7. Narenda P.M., Fukunaga K. A branch and bound algorithm for feature subset selection // IEEE Transactions Computers. 1977. V. C-26. N 9. P. 917–922. https://doi.org/10.1109/TC.1977.1674939
8. Ridout M.S. An improved branch and bound algorithm for feature subset-selection // Journal of the Royal Statistical Society. Series C (Applied Statistics). 1988. V. 37. N 1. P. 139–147. https://doi.org/10.2307/2347512
9. Bech C.N., Brabrand M., Mikkelsen S., Lassen A. Risk factors associated with short term mortality changes over time, after arrival to the emergency department // Scandinavian Journal of Trauma,

Resuscitation and Emergency Medicine, 2018, vol. 26, no. 1, pp. 29. https://doi.org/10.1186/s13049-018-0493-2
10. Fuchs P.A., Del Junco D.J., Fox E.E., Holcomb J.B., Rahbar M.H., Wade C.A., Alarcon L.H., Brasel K.J., Bulger E.M., Cohen M.J., Myers J.G., Muskat P., Phelan H.A., Schreiber M.A., Cotton B.A. Purposeful variable selection and stratification to impute missing Focused Assessment with Sonography for Trauma data in trauma research. *Journal of Trauma and Acute Care Surgery*, 2013, vol. 75, no. 1 (suppl. 1), pp. S75–S81. https://doi.org/10.1097/TA.0b013e31828fa51c
11. Mantrova A.I. How can omissions in medical data affect the results of research? *Scientific Review. Biological Sciences*, 2019, no. 2, pp. 5–9. (in Russian)
12. Lefering R., Huber-Wagner S., Nienaber U., Maegele M., Bouillon B. Update of the trauma risk adjustment model of the TraumaRegister DGU™: the Revised Injury Severity Classification, version II. *Critical Care*, 2014, vol. 18, no. 5, pp. 476. https://doi.org/10.1186/s13054-014-0476-2
13. Seleno N., Vogel J., Liao M., Hopkins E., Byyny R., Moore E., Gravitz C., Haukoos J. Denver trauma organ failure score outperforms traditional methods of risk stratification in trauma. *Academic Emergency Medicine*, 2012, vol. 19, suppl. 1, pp. S144.
14. Zolin P.P. Statistical processing of censored samples in the study of extreme and terminal states. *Pathogenesis, clinical findings and therapy of extreme and terminal states: Proceedings of scientific and practical conference*. Omsk, 1998, pp. 39–43. (in Russian)
15. Fabian-Jessing B.K., Vallentin M.F., Secher N., Hansen F.B., Dezfulian C., Granfeldt A., Andersen L.W. Animal models of cardiac arrest: A systematic review of bias and reporting. *Resuscitation*, 2018, vol. 125, pp. 16–21. https://doi.org/10.1016/j.resuscitation.2018.01.047
16. Krantz M.J., Kaul S. The ATLAS ACS 2–TIMI 51 trial and the burden of missing data: (Anti-Xa Therapy to Lower Cardiovascular Events in Addition to Standard Therapy in Subjects With Acute Coronary Syndrome ACS 2–Thrombolysis In Myocardial Infarction 51). *Journal of the American College of Cardiology*, 2013, vol. 62, no. 9, pp. 777–781. https://doi.org/10.1016/j.jacc.2013.05.024
17. Little R.J., D'Agostino R., Cohen M.L., Dickersin K., Emerson S.S., Farrar J.T., Frangakis C., Hogan J.W., Molenberghs G., Murphy S.A., Neaton J.D., Rotnitzky A., Scharfstein D., Shih W.J., Siegel J.P., Stern H. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 2012, vol. 367, pp. 1355–1360. https://doi.org/10.1056/NEJMsr1203730
18. *National Research Council. The prevention and treatment of missing data in clinical trials*. Washington, DC, National Academies Press, 2010, 162 p.
19. Fabrykant M.S. Model-oriented approach to missing values: Multiple imputation in multilevel regression using R (on the example of analyzing survey data). *Sociology: methodology, methods, mathematical modeling (Sociology:4M)*, 2015, no. 41, pp. 7–29. (in Russian)
20. Van Buuren S. *Flexible Imputation of Missing Data*. CRC Press, 2012, 342 p. https://doi.org/10.1201/b11826
21. Kulikova K.Yu. Statistical analysis of medical data with missing values. *Management processes and sustainability*, 2014, vol. 1, no. 1, pp. 253–258. (in Russian)
22. Witkiewitz K., Falk D.E., Kranzler H.R., Litten R.Z., Hallgren K.A., O'Malley S.S., Anton R.F. Methods to analyze treatment effects in the presence of missing data for a continuous heavy drinking outcome measure when participants drop out from treatment in alcohol clinical trials. *Alcoholism: Clinical and Experimental Research*, 2014, vol. 38, no. 11, pp. 2826–2834. https://doi.org/10.1111/acer.12543
23. Parvez B., Shah A., Muhammad R., Shoemaker M.B., Graves A.J., Heckbert S.R., Xu H., Ellinor P.T., Benjamin E.J., Alonso A., Shintani A.K., Roden D., Darbar D. Replication of a risk prediction model for ambulatory incident atrial fibrillation using electronic medical record. *Circulation*, 2012, vol. 126, no. 21, meeting abstract 18578.
24. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine*, 2016, vol. 4, no. 2, pp. 30. https://doi.org/10.3978/j.issn.2305-5839.2015.12.63
25. Aladyshkina A.S., Lakshina V.V., Leonova L.A., Maksimov A.G. Working with data on population health: imputation. *Social aspects of population health*, 2020, vol. 66, no. 1, pp. 12. (in Russian). https://doi.org/10.21045/2071-5021-2020-66-1-12

Resuscitation and Emergency Medicine. 2018. V. 26. N 1. P. 29. https://doi.org/10.1186/s13049-018-0493-2
10. Fuchs P.A., Del Junco D.J., Fox E.E., Holcomb J.B., Rahbar M.H., Wade C.A., Alarcon L.H., Brasel K.J., Bulger E.M., Cohen M.J., Myers J.G., Muskat P., Phelan H.A., Schreiber M.A., Cotton B.A. Purposeful variable selection and stratification to impute missing Focused Assessment with Sonography for Trauma data in trauma research // Journal of Trauma and Acute Care Surgery. 2013. V. 75. N 1 (Suppl. 1). P. S75–S81. https://doi.org/10.1097/TA.0b013e31828fa51c
11. Мантрова А.И. Как пропуски в медицинских данных могут влиять на результаты исследований? // Научное обозрение. Биологические науки. 2019. № 2. С. 5–9.
12. Lefering R., Huber-Wagner S., Nienaber U., Maegele M., Bouillon B. Update of the trauma risk adjustment model of the TraumaRegister DGU™: the Revised Injury Severity Classification, version II // Critical Care. 2014. V. 18. N 5. P. 476. https://doi.org/10.1186/s13054-014-0476-2
13. Seleno N., Vogel J., Liao M., Hopkins E., Byyny R., Moore E., Gravitz C., Haukoos J. Denver trauma organ failure score outperforms traditional methods of risk stratification in trauma // Academic Emergency Medicine. 2012. V. 19. Suppl. 1. P. S144.
14. Золин П.П. Статистическая обработка цензурированных выборок при изучении экстремальных и терминальных состояний // Патогенез, клиника и терапия экстремальных и терминальных состояний: Материалы научно-практической конференции, г. Омск, 21 октября 1998 г. Омск: Омская государственная медицинская академия, 1998. С. 39–43.
15. Fabian-Jessing B.K., Vallentin M.F., Secher N., Hansen F.B., Dezfulian C., Granfeldt A., Andersen L.W. Animal models of cardiac arrest: A systematic review of bias and reporting // Resuscitation. 2018. V. 125. P. 16–21. https://doi.org/10.1016/j.resuscitation.2018.01.047
16. Krantz M.J., Kaul S. The ATLAS ACS 2–TIMI 51 trial and the burden of missing data: (Anti-Xa Therapy to Lower Cardiovascular Events in Addition to Standard Therapy in Subjects With Acute Coronary Syndrome ACS 2–Thrombolysis In Myocardial Infarction 51) // Journal of the American College of Cardiology. 2013. V. 62. N 9. P. 777–781. https://doi.org/10.1016/j.jacc.2013.05.024
17. Little R.J., D'Agostino R., Cohen M.L., Dickersin K., Emerson S.S., Farrar J.T., Frangakis C., Hogan J.W., Molenberghs G., Murphy S.A., Neaton J.D., Rotnitzky A., Scharfstein D., Shih W.J., Siegel J.P., Stern H. The prevention and treatment of missing data in clinical trials // New England Journal of Medicine. 2012. V. 367. P. 1355–1360. https://doi.org/10.1056/NEJMsr1203730
18. National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: National Academies Press, 2010. 162 p.
19. Фабрикант М.С. Модель-ориентированный подход к отсутствующим значениям: множественная импутация в многоуровневой регрессии посредством R (на примере анализа опросных данных) // Социология: методология, методы, математическое моделирование (Социология:4М). 2015. № 41. С. 7–29.
20. Van Buuren S. Flexible Imputation of Missing Data. CRC Press, 2012. 342 p. https://doi.org/10.1201/b11826
21. Куликова К.Ю. Статистический анализ медицинских данных при наличии пропусков // Процессы управления и устойчивость. 2014. Т. 1. № 1. С. 253–258.
22. Witkiewitz K., Falk D.E., Kranzler H.R., Litten R.Z., Hallgren K.A., O'Malley S.S., Anton R.F. Methods to analyze treatment effects in the presence of missing data for a continuous heavy drinking outcome measure when participants drop out from treatment in alcohol clinical trials // Alcoholism: Clinical and Experimental Research. 2014. V. 38. N 11. P. 2826–2834. https://doi.org/10.1111/acer.12543
23. Parvez B., Shah A., Muhammad R., Shoemaker M.B., Graves A.J., Heckbert S.R., Xu H., Ellinor P.T., Benjamin E.J., Alonso A., Shintani A.K., Roden D., Darbar D. Replication of a risk prediction model for ambulatory incident atrial fibrillation using electronic medical record // Circulation. 2012. V. 126. N 21. Meeting Abstract 18578.
24. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package // Annals of Translational Medicine. 2016. V. 4. N 2. P. 30. https://doi.org/10.3978/j.issn.2305-5839.2015.12.63
25. Аладышкина А.С., Лакшина В.В., Леонова Л.А., Максимов А.Г. Особенности работы с данными, характеризующими здоровье

26. Little R.J.A., Rubin D.B. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014, 408 p.
27. Seitz C., Lanius V., Lippert S., Gerlinger C., Haberland C., Oehmke F., Tinneberg H.-R. Patterns of missing data in the use of the endometriosis symptom diary. *BMC Women's Health*, 2018, vol. 18, no. 1, pp. 88. https://doi.org/10.1186/s12905-018-0578-0

населения: заполнение пропусков в данных // Социальные аспекты здоровья населения. 2020. Т. 66. № 1. С. 12. https://doi.org/10.21045/2071-5021-2020-66-1-12
26. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. John Wiley & Sons, 2014. 408 p.
27. Seitz C., Lanius V., Lippert S., Gerlinger C., Haberland C., Oehmke F., Tinneberg H.-R. Patterns of missing data in the use of the endometriosis symptom diary // BMC Women's Health. 2018. V. 18. N 1. P. 88. https://doi.org/10.1186/s12905-018-0578-0

Authors

Dmitry I. Kurapeev — PhD, Chief Information Officer, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 57225231263, https://orcid.org/0000-0002-2190-1495, dkurapeev@gmail.com

Mikhail S. Lushnov — D.Sc., Methodologist, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 57190125064, https://orcid.org/0000-0002-9683-1858, Lushnov_ms@almazovcentre.ru

Tianxing Man — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57209975532, https://orcid.org/0000-0003-2187-1641, mantx626@gmail.com

Natalia A. Zhukova — PhD, Associate Professor, Leading Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation; Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56406142300, https://orcid.org/0000-0001-5877-4461, nazhukova@mail.ru

Авторы

Курапеев Дмитрий Ильич — кандидат медицинских наук, заместитель генерального директора, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 57225231263, https://orcid.org/0000-0002-2190-1495, dkurapeev@gmail.com

Лушнов Михаил Степанович — доктор медицинских наук, врач-методист, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 57190125064, https://orcid.org/0000-0002-9683-1858, Lushnov_ms@almazovcentre.ru

Ман Тяньсин — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57209975532, https://orcid.org/0000-0003-2187-1641, mantx626@gmail.com

Жукова Наталия Александровна — кандидат технических наук, доцент, ведущий научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация; доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56406142300, https://orcid.org/0000-0001-5877-4461, nazhukova@mail.ru

Received 17.10.2021
Approved after reviewing 07.12.2021
Accepted 28.01.2022

Статья поступила в редакцию 17.10.2021
Одобрена после рецензирования 07.12.2021
Принята к печати 28.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-167-178
УДК 351.814.334.3

Построение на базе задачи машины Дубинса опорных траекторий движения объектов с учетом постоянных внешних воздействий

Сергей Петрович Хабаров[1], Мария Львовна Шилкина[2]✉

[1,2] Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург, 194021, Российская Федерация

[1] Serg.Habarov@mail.ru, https://orcid.org/0000-0003-1337-0150
[2] mchernobay@inbox.ru✉, https://orcid.org/0000-0001-6501-5789

Аннотация

Предмет исследования. Рассмотрен основанный на решении задачи машины Дубинса подход к формированию опорных траекторий движения подвижных объектов (судов, беспилотников) из начальной в заданную конечную конфигурацию. Движение происходит за минимальное время с учетом ограниченной управляемости объектов по угловым скоростям и действия на них постоянных по величине и направлению внешних возмущений. **Метод.** Предложено вместо аналитического исследования прямой и сопряженной систем, вытекающих из принципа максимума Понтрягина, и оставаясь в парадигме задачи машины Дубинса, воспользоваться простым перебором возможных типов траекторий с целью определения среди них наилучшей по быстродействию. Для каждой траектории расчет точек переключения управления на траектории производится на основе минимизации разности расстояния между координатами конечной заданной точки и точки, в которую траектория приходит в зависимости от выбора параметров двух промежуточных точек переключения управления. **Основные результаты.** Решена задача поиска наилучшей по быстродействию траектории движения объекта из одной точки в другую с использованием подхода Дубинса, причем для обеих точек заданы их координаты и курсовые углы. Все расчеты произведены с учетом искажающих траекторию постоянных по величине и направлению внешних воздействий. Задача поиска кривых Дубинса сводится к поиску параметров двух промежуточных точек, в которых происходит смена управления. Рассмотрены возможные варианты смены управлений с учетом имеющихся ограничений. Вычислены длины траекторий движения и выбрана наилучшая по времени прохождения. Предложенный метод рассматривает нескольких допустимых по ограничениям траекторий, учитывающих влияние внешнего воздействия, из которых перебором выбирается оптимальная. Наличие нескольких допустимых траекторий дает преимущества при выборе траектории в зависимости от окружающей обстановки. **Практическая значимость.** Вместо решения задачи нелинейной оптимизации принципа максимума Понтрягина использован простой перебор возможных типов траекторий с целью определения среди них наилучшей по быстродействию. Поиск каждой из возможных траекторий выполняется из условия минимизации невязки аналитического решения и краевого условия на конце траектории. При поиске возможных траекторий учтены ограничения на управление, влияние постоянных по величине и направлению внешних воздействий и постоянство значения управления на каждом участке траектории. Выполнение данных ограничений позволяет правильно моделировать движение судна. Физически ограничения на управление (радиус поворота) связаны с тем, что не все значения угла поворота руля реализуемы на практике. Ограничения могут быть связаны не только с величиной радиуса поворота, но и с географическими особенностями конкретной местности: для беспилотных летательных аппаратов — здания и рельеф, а для судов — береговая линия, мели, острова и другие. В связи с этим может оказаться, что найденное оптимальное по быстродействию решение не может быть реализуемым на практике. Тогда предложенный в работе метод предоставляет возможность выбора другой траектории среди менее оптимальных по быстродействию.

Ключевые слова

кривая Дубинса, алгоритмы, оптимальное управление, навигация по путевым точкам, беспилотные летательные аппараты, кратчайший путь, оптимизация

Ссылка для цитирования: Хабаров С.П., Шилкина М.Л. Построение на базе задачи машины Дубинса опорных траекторий движения объектов с учетом постоянных внешних воздействий // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 167–178 doi: 10.17586/2226-1494-2022-22-1-167-178

© Хабаров С.П., Шилкина М.Л., 2022

Construction of movement trajectories for objects based on the Dubins car problem, taking into account constant external influences

Sergey P. Khabarov[1], Maria L. Shilkina[2]✉

[1,2] St. Petersburg State Forest Technical University, Saint Petersburg, 194021, Russian Federation

[1] Serg.Habarov@mail.ru, https://orcid.org/0000-0003-1337-0150

[2] mchernobay@inbox.ru✉, https://orcid.org/0000-0001-6501-5789

Abstract

The paper considers an approach in terms of optimal speed problem for Dubins cars to the formation of control trajectories of moving objects (airplanes, ships), that have control restrictions, under external influences that are constant in magnitude and direction and constant control values at each part of the trajectory. Instead of solving the Pontryagin maximum principle, it is proposed to use a simple comparison of possible control strategies in order to determine the best among them in terms of speed. For each strategy, the calculation of control switching points on the trajectory is based on minimizing the difference between the specified coordinates of the endpoint and the coordinates of the point at which the trajectory comes, depending on the choice of the parameters of two intermediate control switching points. The problem of finding the best speed trajectory for an object from one point to another is solved using the Dubins approach, and their coordinates and heading angles are given for both points. All calculations were carried out taking into account wind and water disturbances, which are constant in magnitude and direction and distort the trajectory. The problem of finding the Dubins paths is reduced to finding the parameters of two intermediate points at which the control changes. Different possibilities for changing controls are considered, taking into account the existing restrictions. The lengths of the trajectories are calculated and the best travel time is selected. The proposed method considers several trajectories acceptable in terms of constraints, taking into account the external influences, from which the optimal path is selected by comparison. Having multiple feasible strategies is beneficial when choosing a trajectory depending on the environment. Instead of solving the problem of nonlinear optimization of the Pontryagin maximum principle, a simple comparison of possible control strategies is used in order to determine the best among them in terms of speed, each of the possible strategies is sought from the condition of minimizing the residual of the analytical solution and the boundary condition at the end of the trajectory. When searching for possible trajectories, control constraints, the influence of external influences, that are constant in magnitude and direction, and the constancy of the control value at each part of the trajectory are taken into account. And all these factors make it possible to sufficiently and adequately simulate the movement of the ship. Physically, restrictions on control (turning radius) are associated with a limited steering angle. Restrictions can be associated not only with restrictions on the turning radius, but also with geographical features of a specific area: for unmanned aerial vehicles this may be due to the buildings and terrain, and for ships this may be due to the coastline, shoals, islands, etc. In this regard, it may turn out that the solution found optimal in terms of speed cannot be realizable in practice. Then the method proposed in the work has the ability to choose another trajectory among the less optimal in terms of speed.

Keywords

Dubins path, algorithms, optimal control, waypoint navigation, unmanned aerial vehicles, shortest path, optimization

For citation: Khabarov S.P., Shilkina M.L. Construction of movement trajectories for objects based on the Dubins car problem, taking into account constant external influences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 167–178 (in Russian). doi: 10.17586/2226-1494-2022-22-1-167-178

Введение

Для мониторинга состояния линейных объектов, таких как линии электропередач, газопроводов, границ лесных массивов, а также при исследовании речных и прибрежных акваторий, используются беспилотные летательные аппараты (БПЛА), исследовательские суда и подводные роботы. Они должны передвигаться по заданным программным траекториям и характеризуются ограниченной управляемостью по угловым скоростям. Как правило, объекты, осуществляющие мониторинг, при перемещении в естественной среде испытывают на себе постоянное по силе и направлению воздействие ветра, а водные суда еще и воздействие течений. Программные траектории определяют генеральное направление движения подвижных объектов, и в окрестности точек, задающих эти траектории, должны стабилизироваться все кинематические параметры движения объектов. Программные траектории, как правило, задаются в виде упорядоченного набора путевых точек $\{P_i, i = 1, ..., n\}$, который рассматривается как совокупность прямолинейных участков траектории, соединяющих между собой путевые точки P_i и P_{i+1} для $i = 1, ..., n - 1$. В связи с этим практически значимая задача — гладкое сопряжение всех участков опорной траектории движения между собой так, чтобы она проходила через заданные путевые точки, удовлетворяла ограничениям на угловые скорости и учитывала возмущения, вызванные воздействием ветра и водных течений. Предположение о фиксированности сигналов управления на каждом из участков траектории позволяет для решения поставленной задачи сформировать подход, в основе которого решение известной в теории управления задачи машины Дубинса.

В работе [1] приведено решение проблемы построения множества разрешимости в задаче проводки самолета при ветровом возмущении с помощью нелинейной системы дифференциальных уравнений четвертого порядка, упрощенно описывающей движение самолета в горизонтальной плоскости при наличии заданного уровня возмущений. Отметим, что с точки зрения меньшей сложности расчетов, особенно для случая сложной программной траектории и специфических требований к прохождению каждого из ее участков, полезным является подход к формированию кривых сопряжения как кратчайших траекторий движения объекта из некоторой

исходной точки в заданную конечную точку, которые лежат на сопрягаемых участках программной траектории, с использованием машины Дубинса. Результаты, полученные Л. Дубинсом [2], используются для исследования объектов, двигающихся в горизонтальной плоскости с постоянной по величине линейной скоростью, и с ограничениями на радиус поворота. Такие объекты называют машиной Дубинса, и для них на базе принципа максимума Л.С. Понтрягина Дубинс показал, что наискорейший переход из точки в точку с заданными начальным и конечным направлениями линейной скорости осуществляется при помощи кусочно-постоянного управления не более чем с двумя переключениями. Именно этот подход используется в работах [3–5], но без учета внешних возмущений, воздействующих на движущийся объект, а именно ветра или течений.

Модель Дубинса применяется для решения многих задач, таких как определение множества достижимости при симметричном и несимметричном ограничениях на управление [6–8], поиске оптимального управления для задачи с подвижным концом траектории с использованием генетического алгоритма [9]. Модель используется для решения задач перехвата движущейся по предписанной траектории цели и оптимального управления по критерию быстродействия с произвольным направлением вектора скорости при перехвате [10]. А также для формулировки требований к минимальному объему экспериментов с математическими моделями подвижных объектов, необходимых для создания математических соотношений, определяющих параметры участков искомых траекторий [11]. В общем случае модели машины Дубинса находят применение при построении траекторий движения в горизонтальной плоскости [12–14], также существуют методы распространения подхода Дубинса к построению 3D-траекторий [15].

В данной работе предложен подход с использованием машины Дубинса к задаче о движении объекта в горизонтальной плоскости для формирования режимов маневрирования подвижных объектов при наличии постоянных по величине и направлению внешних воздействий и для поиска гладких кривых сопряжения участков программной траектории. При этом модель динамики движения объекта характеризуется постоянной по величине линейной скоростью и ограничениями на угловую скорость (радиус поворота). Эта модель движения описывается нелинейной системой дифференциальных уравнений третьего порядка, где две фазовые переменные характеризуют геометрическое положение объекта на плоскости, а третья — угол направления вектора скорости.

Аналитическое исследование и решение прямой и сопряженной систем, вытекающих из принципа максимума Понтрягина, предлагается заменить перебором возможных типов траекторий с учетом постоянных по величине внешних возмущений с целью определения среди них оптимальной по быстродействию. Данный подход связан с ограниченным для машин Дубинса количеством возможных траекторий. Все расчеты произведены с учетом постоянных по величине значений управления на каждом участке траектории.

Отметим, что рассматриваемые методы поиска наилучших по быстродействию кривых сопряжения связаны с вычислительной сложностью. По этой причине при реализации методов в информационных системах необходимо учитывать тот факт, что они функционируют в параллельных потоках с алгоритмами других подсистем, и возникает задача обеспечения своевременности реализации вычислительных процессов [16, 17]. Успешной работе выбранных алгоритмов в составе крупных информационных систем во многом может способствовать приоритетное обслуживание с репликацией критически важных запросов.

Постановка задачи

Выполним исследование множества достижимости для нелинейной управляемой системы третьего порядка, называемой машиной Дубинса. Объект перемещается на плоскости с постоянной по величине скоростью и имеет ограничения на угловые скорости поворота вправо и влево. Требуется найти кратчайшую траекторию движения объекта из некоторой исходной точки $A(x_0, y_0, \varphi_0)$ в заданную конечную точку $B(x_k, y_k, \varphi_k)$ при наличии постоянного внешнего воздействия на объект. В этом случае движение управляемого объекта на плоскости описывается системой дифференциальных уравнений вида:

$$\begin{cases} x'(t) = |\mathbf{V}|\sin\varphi(t) + W_x \\ y'(t) = |\mathbf{V}|\cos\varphi(t) + W_y \\ \varphi'(t) = u(t) \\ u(t) \in [-U, U] \\ x(t_0) = x_0, y(t_0) = y_0, \varphi(t_0) = \varphi_0 \\ x(t_k) = x_k, y(t_k) = y_k, \varphi(t_k) = \varphi_k \end{cases}, \qquad (1)$$

где $x(t)$, $y(t)$ — координаты объекта в неподвижной системе координат; $\varphi(t)$ — угол курса (между направлением вектора скорости и направлением на север); $\mathbf{V}$ — постоянная по величине скорость движения объекта; $u(t)$ — скалярный сигнал управления; $W_x = |\mathbf{W}|\sin\Theta$, $W_y = |\mathbf{W}|\cos\Theta$, где $\mathbf{W}$, Θ — постоянные по величине скорость и угол внешнего воздействия.

Известно [2], что при отсутствии внешних воздействий управления $u(t)$, которые приводят систему (1) на границу множества достижимости в момент t_k, удовлетворяют принципу максимума Понтрягина, а соответствующие траектории на плоскости состоят из конечного числа дуг окружностей (при $u(t) = \pm U$) и прямолинейных участков траекторий (при $u(t) = 0$). При этом в работе [18, глава 15] показано, что существует всего три примитива движения, из которых могут быть построены все оптимальные кривые для машины Дубинса, которые имеют вид: R (если $u(t) = U$); S (если $u(t) = 0$) и L (если $u(t) = -U$).

При использовании данных символов каждый возможный вид кратчайшего пути может быть обозначен как последовательность из трех символов, соответствующих порядку, в котором применяются примитивы, тем самым кодируются стратегии управления, используемые на каждом из трех участков кривой Дубинса.

Нет необходимости иметь два последовательных примитива одного и того же типа, потому что их можно объединить в один. При этом возможны двенадцать слов длиной в три символа, но Дубинс показал, что оптимальными являются только следующие шесть слов, кодирующие соответствующие типы траекторий:

$$\{\text{RLR, LRL, LSL, LSR, RSL, RSR}\}. \qquad (2)$$

Кратчайший путь между любыми двумя точками опорной траектории всегда можно охарактеризовать одним из этих слов. Они называются кривыми Дубинса. Таким образом, переход из любой произвольной точки A в точку B может быть осуществлен не более чем тремя переключениями управляющего сигнала в моменты времени t_0, t_1 и t_2. На каждом участке траектории с временем движения $\Delta t_0 = t_1 - t_0$, $\Delta t_1 = t_2 - t_1$ и $\Delta t_2 = t_3 - t_2$ управляющий сигнал $u(t)$ является кусочно-постоянным и принимает значения $u(\Delta t_0) = u_0$, $u(\Delta t_1) = u_1$ и $u(\Delta t_2) = u_2$ в соответствии с одним из возможных типов траекторий (2).

Основываясь на принципе решения классической задачи машины Дубинса, предлагается применить аналогичный подход и в случае наличия внешних постоянных воздействий на объект. Тогда исходная задача может быть сведена к поиску параметров двух промежуточных точек P и Q (рис. 1). Это позволит определить не только вид отдельных участков кривой Дубинса, но и время, затрачиваемое на прохождение каждого из трех участков.

Методика исследования

Для решения поставленной задачи применим аналитическое решение системы (1) для поиска параметров точек P, Q и B на основе данных об исходной точке $A(x_0, y_0, \varphi_0)$ и заданном курсовом угле $\varphi_3 = \varphi_k$ в конечной точке. Это решение будет иметь вид:

$$\begin{aligned}
\varphi_1 &= \varphi(t_1) = \int_{t_0}^{t_1} u(t)dt = \varphi_0 + u_0(t_1 - t_0),\\
x_1 &= x(t_1) = \int_{t_0}^{t_1} (V\sin\varphi(t) + W_x)dt =\\
&= x_0 + \frac{V}{u_0}\cos\varphi_0 - \frac{V}{u_0}\cos\varphi_1 + W_x(t_1 - t_0),
\end{aligned} \qquad (3.1)$$

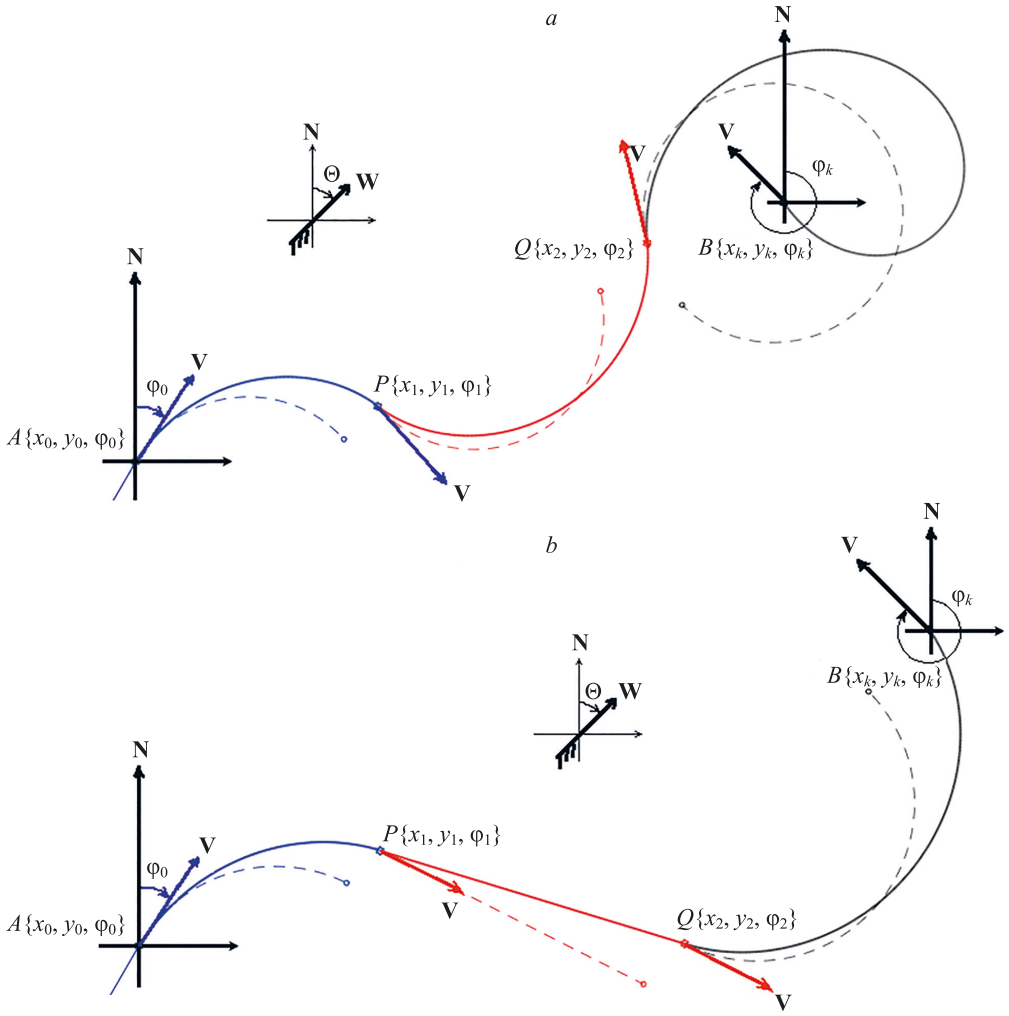


Рис. 1. Пример реализации траекторий RLR (*a*) и RSL (*b*). Пунктиром показан вид участков траектории в условиях отсутствия внешнего воздействия **W**

Fig. 1. An example of the implementation of trajectories RLR (*a*) and RSL (*b*). The dotted line shows the view of the trajectory sections in the absence of external influence **W**

$$y_1 = y(t_1) = \int_{t_0}^{t_1} (V\cos\varphi(t) + W_y)dt = \\ = y_0 - \frac{V}{u_0}\sin\varphi_0 + \frac{V}{u_0}\sin\varphi_1 + W_y(t_1 - t_0). \quad (3.2)$$

При $t_0 = 0$ сигнал управления u_0 может принимать одно из двух возможных значений в зависимости от выбранного значения управления ($u_0 = u(0) = \pm U$). Так как u_0 представляет собой угловую скорость перемещения объекта, то выражение V/u_0 является радиусом циркуляции (R_0) объекта на данном участке кривой Дубинса. Тогда, введя новые обозначения вида

$$a = x_0 + R_0\cos\varphi_0, \\ b = y_0 - R_0\sin\varphi_0 \quad (4)$$

уравнения (3.1) и (3.2), характеризующие точку P, можно переписать в виде:

$$x_1 = a - R_0\cos\varphi_1 + W_x\Delta t_0, \\ y_1 = b + R_0\sin\varphi_1 + W_y\Delta t_0. \quad (5)$$

Траектории RLR и LRL

У траекторий типов RLR и LRL на интервалах времени Δt_0 и Δt_2 сигнал управления имеет один и тот же знак, а на интервале Δt_1 он противоположного знака. Иначе говоря, $R_2 = R_0$, а $R_1 = -R_0$. Различие траекторий данных типов в том, что при выборе RLR $u_0 = U$, а при LRL $u_0 = -U$. Решение системы (1) на интервале Δt_0 имеет вид (5). Используя уравнения (5), для траекторий RLR и LRL можно записать выражения для вычисления параметров точки Q:

$$x_2 = a - R_0(2\cos\varphi_1 - \cos\varphi_2) + W_x(\Delta t_0 + \Delta t_1), \\ y_2 = b + R_0(2\sin\varphi_1 - \sin\varphi_2) + W_y(\Delta t_0 + \Delta t_1) \quad (6)$$

и для точки B:

$$x_3 = (a - R_0\cos\varphi_3) - 2R_0(\cos\varphi_1 - \cos\varphi_2) + \\ + W_x(\Delta t_0 + \Delta t_1 + \Delta t_2), \\ y_3 = (b + R_0\sin\varphi_3) + 2R_0(\sin\varphi_1 - \sin\varphi_2) + \\ + W_y(\Delta t_0 + \Delta t_1 + \Delta t_2). \quad (7)$$

Движение по трем фрагментам траекторий RLR и LRL осуществляется по дугам, длина которых при заданной угловой скорости позволяет определить время прохождения каждого участка траектории:

$$\Delta t_i = \Delta f_i/u_i,$$

где $\Delta f_i = \varphi_{i+1} - \varphi_i$, $i = 0, 1, 2$ — разность курсовых углов в начале и в конце дуги.

Однако изменение курсового угла на каждом из фрагментов траектории зависит не только от начального и конечного значений курсового угла, но и от того, в каком направлении (по или против часовой стрелки) осуществляется это движение.

Считая, что оба значения курсовых углов φ_i и φ_{i+1} приведены к значениям в диапазоне от 0 до 2π, вычисление значения $\Delta f_i = \varphi_{i+1} - \varphi_i$ должно быть скорректировано с учетом направления движения на конкретном i-ом фрагменте траектории:

$$\Delta f_i = \begin{cases} 2\pi - \Delta f_i, \text{ если } \Delta f_i > 0 \text{ и } \Delta f_i < 2\pi \text{ и } R_i < 0, \\ -\Delta f_i, \text{ если } \Delta f_i < 0 \text{ и } R_i < 0, \\ 2\pi + \Delta f_i, \text{ если } \Delta f_i < 0 \text{ и } R_i > 0, \\ \Delta f_i, \text{ во всех иных случаях.} \end{cases}$$

После вычисления Δf_i и соответствующих временных интервалов Δt_i, которые могут быть подставлены в (7), можно заключить, что координаты точки $B(x_3, y_3)$ являются нелинейными функциями от φ_1 и φ_2 имеют вид:

$$x_3 = F_1(\varphi_1, \varphi_2, C_1), \\ y_3 = F_2(\varphi_1, \varphi_2, C_2),$$

где C_1 и C_2 соответствуют постоянным параметрам, определяемым набором начальных данных решаемой задачи.

Для решения исходной двухточечной краевой задачи надо потребовать, чтобы функции F_3 и F_4, описывающие степень приближения (x_3, y_3) к (x_k, y_k), были бы равны нулю, тогда:

$$\begin{cases} F_3(\varphi_1, \varphi_2, C_1) = \sqrt{(x_k - x_3)^2} = 0 \\ F_4(\varphi_1, \varphi_2, C_1) = \sqrt{(y_k - y_3)^2} = 0 \end{cases}. \quad (8)$$

Таким образом, перебирая значения φ_1, φ_2, определяющие все возможные траектории движения типа RLR и LRL, можно определить среди траекторий кратчайшую, решив систему из двух нелинейных уравнений.

В частности, на рис. 2 приведен скриншот, соответствующий построению траекторий перехода динамического объекта, двигающегося с постоянной скоростью 10 узлов и допустимым радиусом циркуляции 500 м из точки A (0 м, 0 м, 30°) в точку B (2500 м, 1000 м, –45°) при скорости течения 1 м/с и углом между направлением скорости течения и направлением на север $\Theta = 45°$.

Видно, что из траекторий типа RLR и LRL более короткой является траектория, полученная при решении системы (8) для типа RLR (при $u = [U, -U, U]$), при значениях $\varphi_1 = 138°$ и $\varphi_2 = 346°42'$ (рис. 2, *а*). Полученное решение нелинейной системы уравнений позволяет на основе формул (4)–(7) определить все параметры оптимальной траектории:

— точку P с координатами ($x_1 = 933{,}8$ м, $y_1 = 214{,}4$ м, $\varphi_1 = 138°$);
— точку Q с координатами ($x_2 = 1973{,}1$ м, $y_2 = 845{,}2$ м, $\varphi_2 = 346°42'$);
— точку B с координатами ($x_3 = 2500$ м, $y_3 = 1000$ м, $\varphi_3 = 315°$);
— общее время движения по траектории RLR $T = 996$ с и время пребывания на каждом из фрагментов траектории: $\Delta t_0 = 183$ с, $\Delta t_1 = 256$ с и $\Delta t_2 = 557$ с.

Траектории RSR, RSL, LSL, LSR

Подход к расчету и построению кривых Дубинса для траекторий типов RSR, RSL, LSL, LSR аналогичен описанному в разделе «Траектории RLR и LRL».

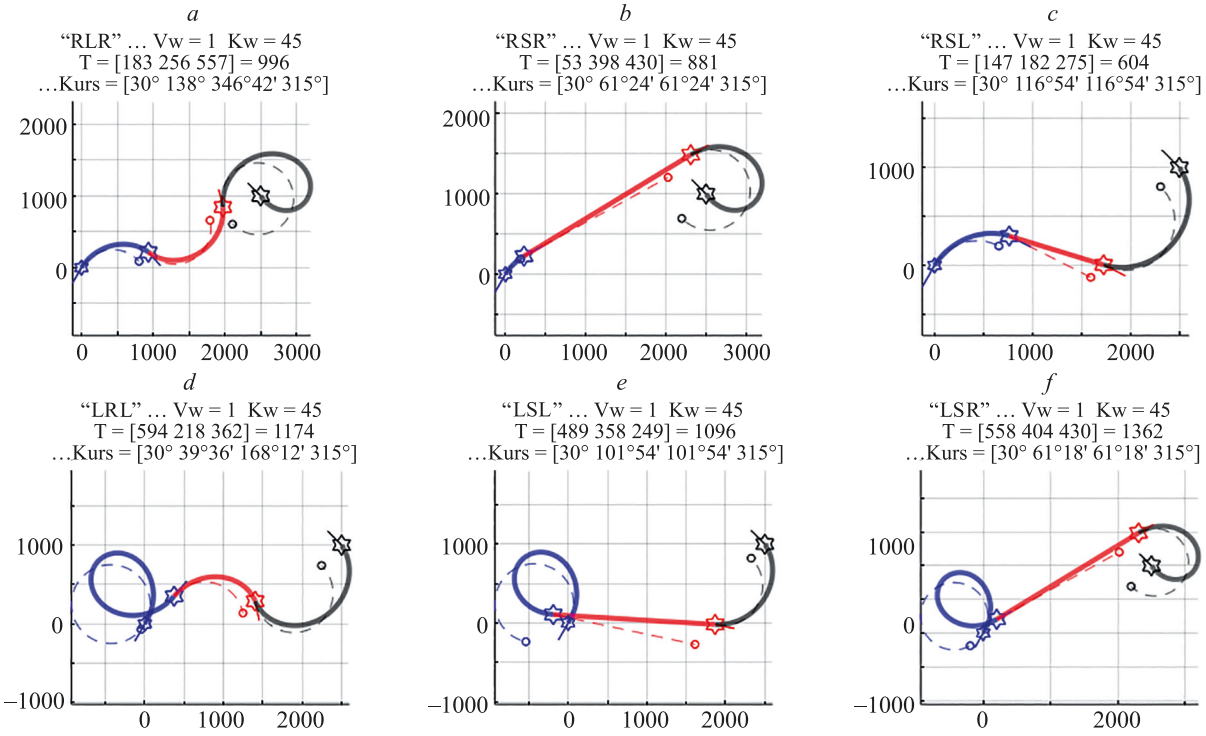


Рис. 2. Кривые Дубинса для 6 видов траекторий: RLR (*a*), RSR (*b*), RSL (*c*), LRL (*d*), LSL (*e*), LSR (*f*). Пунктиром показан вид участков траектории при отсутствии внешних воздействий

Fig. 2. Dubins paths for 6 types of trajectories: RLR (*a*), RSR (*b*), RSL (*c*), LRL (*d*), LSL (*e*), LSR (*f*). The dotted line shows how the sections of the trajectory would look in the absence of external influences

Данным типам траекторий соответствуют четыре варианта (таблица) управляющих воздействий и соответствующих им радиусов циркуляции, равных $R_i = V/u_i$, $i = 0, 1, 2$.

Основные различия данных типов траекторий:

— для траекторий RSR и LSL значение $R_2 = R_0$, в то время как для RSL и LSR $R_2 = -R_0$;

— на втором временном интервале радиус циркуляции $R_1 = \infty$, т. е. выполняется прямолинейное движение с курсовым углом φ_1.

Особенность траекторий типа LSL, LSR, RSL, RSR в том, что начав свое движение по прямолинейному участку из точки $P\{x_1, y_1, \varphi_1\}$, объект должен попасть в точку Q с координатами $\{x_2, y_2, \varphi_2\}$, причем в точку Q движущийся объект придет с тем же курсовым углом, с которым началось его движение из точки P — $\varphi_1 = \varphi_2$.

При реализации траекторий LSL, LSR, RSL, RSR курсовой угол φ_2 в точке Q должен быть всегда равен курсовому углу φ_1 в точке P. Данное утверждение положено в основу предложенного алгоритма, основной целью которого является определение значения угла φ_1 и длины прямолинейного участка траектории L, зная которые можно однозначно вычислить значения координат точек P и Q. Определение и подбор значений φ_1 и L осуществляется из условия выполнения требования, что вся траектория должна прийти в точку с заданными по условию задачи координатами (x_k, y_k) и курсовым углом $\varphi_3 = \varphi_k$.

Таблица. Значения управляющих сигналов на различных участках кривой Дубинса

Table. Control values at different parts of the Dubins path

Тип траектории	Значения сигнала управления		
	$u_0(t_1 - t_0)$	$u_1(t_2 - t_1)$	$u_2(t_k - t_2)$
RSR	U	0	U
RSL	U	0	$-U$
LSL	U	0	$-U$
LSR	U	0	U

Оставаясь в заявленной ранее парадигме поиска траекторий аналогично траекториям Дубинса вида LSL, LSR, RSL, RSR, выражения координат точки B, полученные из решения нелинейных уравнений модели (1), зависящие от двух параметров φ_1 и L, имеют вид:

$$\begin{aligned} x_3 &= c - (R_0 - R_2)\cos\varphi_1 + |\mathbf{V}|\cdot\Delta t_2 \sin\varphi_1 + \\ &+ W_x(\Delta t_0 + \Delta t_1 + \Delta t_2), \\ y_3 &= d - (R_0 - R_2)\sin\varphi_1 + |\mathbf{V}|\cdot\Delta t_2 \cos\varphi_1 + \\ &+ W_y(\Delta t_0 + \Delta t_1 + \Delta t_2), \end{aligned} \tag{9}$$

где

$$c = x_0 + R_0\cos\varphi_0 - R_2\cos\varphi_3,\ d = y_0 - R_0\sin\varphi_0 + R_2\sin\varphi_3,$$

$$V = UR_0,\ \Delta t_0 = \Delta f_0/U,\ \Delta t_2 = \Delta f_2/U,$$

$$\Delta t_1 = \frac{L}{\sqrt{(V\sin\varphi_1 + W_x)^2 + (V\cos\varphi_1 + W_y)^2}},$$

а значения Δf_i вычисляются на основе курсовых углов φ_i и φ_{i+1} с учетом того, в каком направлении осуществляется движение. Значения φ_1, L подберем таким об-

разом, чтобы вычисленные значения координат (x_3, y_3) конечной точки программной траектории совпали со значениями координат (x_k, y_k), требуемыми по краевому условию задачи.

После вычисления изменения курсового угла Δf_i, подставим его в уравнение (9). Тогда можно заключить, что координаты точки $B(x_3, y_3)$ являются нелинейными функциями от φ_1 и L имеют вид:

$$x_3 = F_1(\varphi_1, L, C_1),$$
$$y_3 = F_2(\varphi_1, L, C_2),$$

где C_1 и C_2 — набор постоянных параметров, определяемых начальными данными решаемой задачи. Для решения исходной двухточечной краевой задачи требуется, чтобы функции невязки F_3 и F_4, описывающие степень приближения (x_3, y_3) к (x_k, y_k), были равны нулю, тогда

$$\begin{cases} F_3(\varphi_1, L, C_1) = \sqrt{(x_k - x_3)^2} = 0 \\ F_4(\varphi_1, L, C_1) = \sqrt{(y_k - y_3)^2} = 0 \end{cases}. \qquad (10)$$

После решения системы (10) и определения всех возможных значений φ_1, L для различных типов траекторий появляется возможность на основании (4) и (5) определить координаты точек $P(x_1, y_1, \varphi_1)$ и $Q(x_2, y_2, \varphi_1)$, используя следующие выражения:

$$x_2 = a - R_0\cos\varphi_1 + |\mathbf{V}|\cdot\Delta t_2\sin\varphi_1 + W_x\Delta f_0/U,$$
$$x_2 = b + R_0\sin\varphi_1 + |\mathbf{V}|\cdot\Delta t_2\cos\varphi_1 + W_y\Delta f_0/U.$$

При этом, если время пребывания на начальном и конечном фрагментах траекторий оценивается, как и ранее, на основе длины пройденных дуг ($\Delta t_i = \Delta f_i/u_i$, $i = 0,2$), то время движения на прямолинейном участке траектории можно определить из выражения:

$$\Delta t_1 = L/\sqrt{(V\sin\varphi_1 + W_x)^2 + (V\cos\varphi_1 + W_y)^2}.$$

Таким образом, решая систему (10) из двух нелинейных уравнений, можно определить набор значений φ_1, L, который позволит найти все возможные траектории движения среди траекторий типа RSR, RSL, LSL, LSR, и определить среди них кратчайшую. В общем случае, при определенных заданных краевых условиях, некоторые типы траекторий (2) оказываются недостижимы.

Из рис. 2 видно, что для выбранных краевых условий алгоритм обеспечил построение всех четырех траекторий RSR, RSL, LSL, LSR. Самая короткая траектория, полученная при решении системы (10), соответствует траектории RSL (при $u = [-U, 0, U]$), для которой значения $\varphi_1 = 116°54'$ и $L = 1007{,}3$ м. Полученное решение нелинейной системы уравнений позволило на основе выражений (4), (5), (9) и (10) определить все параметры оптимальной траектории:

— точку P с координатами ($x_1 = 763$ м, $y_1 = 300{,}3$ м, $\varphi_1 = 116°54'$);
— точку Q с координатами ($x_2 = 1726{,}5$ м, $y_2 = 6{,}2$ м, $\varphi_2 = 116°54'$);
— точку B с координатами ($x_2 = 2500$ м, $y_2 = 1000$ м, $\varphi_2 = 315°$);
— общее время движения по траектории RSL $T = 604$ с и время пребывания на каждом из фрагментов траектории: $\Delta t_0 = 147$ с, $\Delta t_1 = 182$ с и $\Delta t_2 = 275$ с.

Обсуждение результатов

Особенностью программной реализации данного алгоритма является то, что, оставаясь в парадигме поиска траекторий в виде траекторий Дубинса, она позволяет не только получить наилучшую по времени траекторию перехода из точки в точку, но и при желании просмотреть все возможные варианты, что может быть очень полезно с практической точки зрения при наличии внешних ограничений на возможные движения объекта.

Для проверки адекватности работы предложенного алгоритма проведено исследование его работы в случае отсутствия внешних воздействий на переходе из точки A в точку B. На рис. 3 приведены графики, соответствующие построению траекторий перехода динамического объекта, двигающегося с постоянной скоростью 10 узлов и допустимым радиусом циркуляции 500 м из точки A (0 м, 0 м, 30°) в точку B (700 м, 700 м, –45°). Выполнено сравнение с одним из известных ранее аналитических алгоритмов [3, 4, 19], для которого получен набор возможных траекторий — кривых Дубинса — для всех возможных траекторий типа RLR и LRL (рис. 3, *a*, *b*, *c*, *d*).

На рис. 3, *e*, *f*, *g*, *h* представлены результаты расчета, выполненного для тех же типов траекторий по описанному в настоящей работе алгоритму. Рис. 3 демонстрируют практически полное совпадение результатов, что доказывает адекватность работы предложенного алгоритма и его совпадение с аналитическим решением, по крайней мере при отсутствии внешних воздействий.

Для вывода систем (8) и (10) в случае наличия внешних воздействий применен строгий математический аппарат. Отметим, что аналитическое решение нелинейных систем (8) и (10) или нахождение областей существования их решения представляет собой трудную задачу. Потому в данной работе применен численный метод поиска решений данных систем, имеющих разрывы первого рода и множество локальных минимумов. В этих условиях уделено внимание не только поиску глобального минимума, но и проверке наличия решения выбранных нелинейных систем. В связи с этим до момента запуска основного алгоритма поиска проведено эвристическое определение областей изоляции корней (рис. 4), а в процессе реализации алгоритма в программном коде предусмотрен перехват ошибочных ситуаций, вызванных отсутствием решений этих систем. Определение факта отсутствия областей изоляции или возникновение исключений позволяет принять решение о реализуемости того или иного типа траектории при соответствующих наборах исходных данных. При этом для некоторых исследуемых наборов данных возможна одна из следующих ситуаций:

— точки A и B находятся достаточно близко между собой, и ни одна из траекторий LSR, RSL, LSL и RSR не может быть реализована;

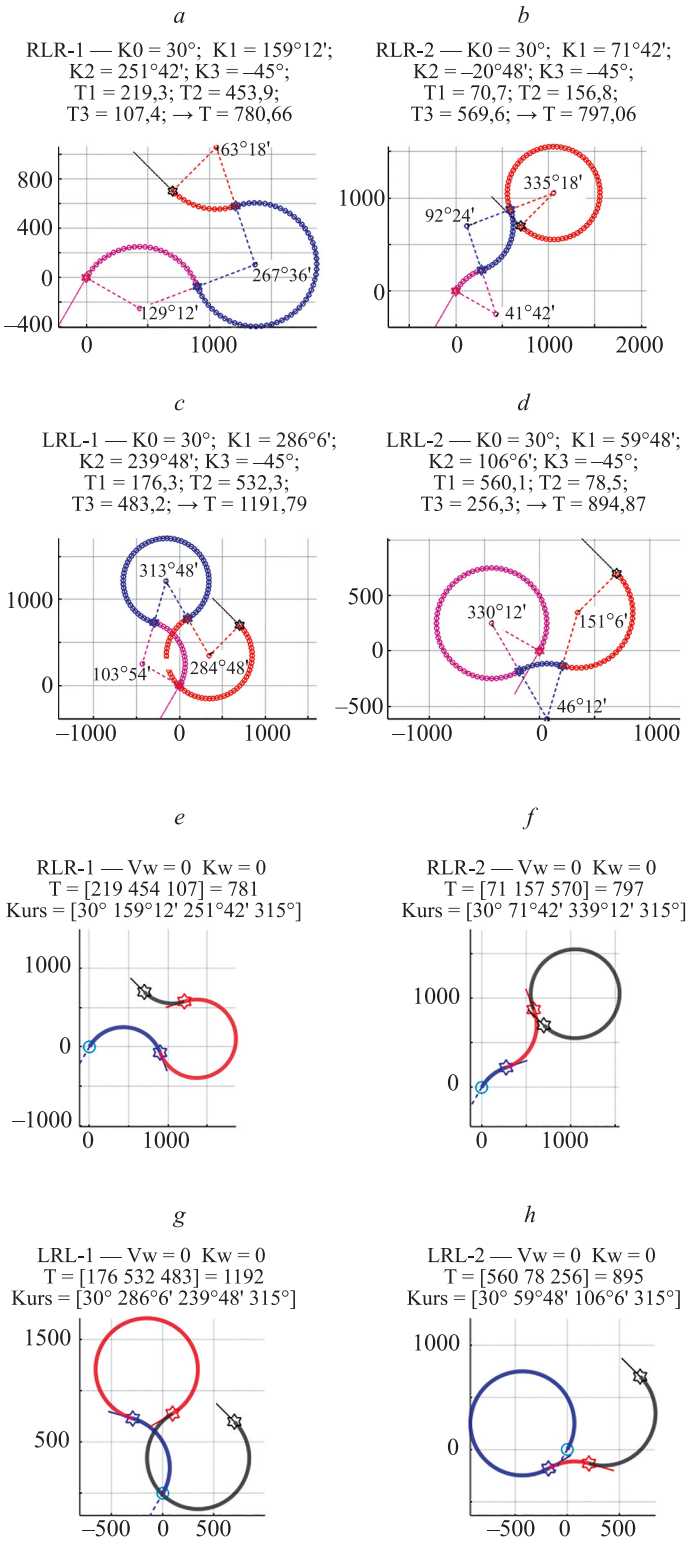


Рис. 3. Кривые Дубинса, полученные с использованием геометрического подхода: для случаев, когда центр средней окружности справа и слева от прямой, проходящей через центры крайних окружностей, — для RLR (*a*) и (*b*), для LRL (*d*) и (*c*) соответственно; а также кривые Дубинса, полученные по предложенному алгоритму при отсутствии внешних возмущений: для случаев, когда центр средней окружности справа и слева от прямой, проходящей через центры крайних окружностей, — для RLR (*e*) и (*f*), для LRL (*h*) и (*g*) соответственно

Fig. 3. Dubins paths obtained using the geometric approach: for the cases when the center of the middle circle is to the right and to the left of the line passing through the centers of the first and last circles — for RLR (*a*) and (*b*), for LRL (*d*) and (*c*), respectively; also Dubins paths obtained by the proposed algorithm in the absence of external disturbances: for the cases when the center of the middle circle is to the right and to the left of the line passing through the centers of the first and last circles — for RLR (*e*) and (*f*), for LRL (*h*) and (*g*), respectively

— точки A и B достаточно удалены, и траектории RLR и LRL нереализуемы, но даже в этих условиях всегда находилась наилучшая траектория среди, как минимум, одной или нескольких реализуемых траекторий.

Для траекторий типа RLR и LRL исследована зависимость расстояния между заданной конечной точкой траектории (x_k, y_k) и точкой (x_3, y_3), в которую приходит программная траектория, в зависимости от выбора программных точек переключения, т. е. от выбора расчетных значений курсовых углов φ_1 и φ_2. На рис. 4 представлен трехмерный график функции невязки вида

$$\varepsilon(\varphi_1, \varphi_2) = \sqrt{(x_k - x_3)^2} + \sqrt{(y_k - y_3)^2},$$

где x_3, y_3 рассчитаны по формулам (7), а параметры φ_1 и φ_2 изменяются в пределах от 0 до 350°.

Красным цветом на рис. 4 выделены точки локальных минимумов функции $\varepsilon(\varphi_1, \varphi_2)$, для которых $\varepsilon(\varphi_1, \varphi_2) < 100$ м. Расчет выполнен для тех же начальных и конечных условий, что и на рис. 2.

Ввиду того, что для решения исходной двухточечной краевой задачи требуется, чтобы функции F_3 и F_4 из системы (8), описывающие степень приближения (x_3, y_3) к (x_k, y_k), были бы равны нулю, то определение областей возможных локальных минимумов способствует более обоснованному выбору начальных приближений для решения систем нелинейных уравнений типа (8) и (10).

На рис. 5 представлен ряд графиков, которые показывают, как траектории двигающегося со скоростью 10 узлов объекта при переходе из точки A (0 м, 0 м, 0°) в точку B (–500 м, 0 м, 180°) с радиусом циркуляции 500 м деформируются при изменении угла $Kw = \Theta$ между направлением постоянного внешнего воздействия и направлением на север. Показано, как меняется внешний вид траекторий в зависимости от угла $Kw = \Theta$ (рис. 1), изменяющегося в диапазоне от 0 до ±180° с шагом ±45°.

На рис. 5 представлены наилучшие по времени программные траектории движения объекта для режимов возврата на обратный курс (с курса 0 на 180°), при ограничениях: внешнее воздействие постоянно по величине и направлению; радиус циркуляции задан; между тремя точками переключения управление остается постоянным. При этом рассмотрены два варианта требуемого отклонения от начальной точки разворота: 500 м на запад (рис. 5, *a*) и 1500 м на восток (рис. 5, *b*).

Видно, что в результате расчетов наилучшими по времени оказались разные типы траекторий: в первом случае это RLR, а во втором — RSR. Самая короткая траектория на рис. 5, *a* получается при боковом течении с запада, которое помогает осуществить крутой разворот на противоположное направление. Самая длинная траектория при боковом течении с востока, которое сильно сносит объект от заданной точки.

На рис. 5, *b* наоборот, траектория разворота является более сглаженной, поэтому наискорейшая траектория для такого маневра получается при течении, направленном в сторону требуемого отклонения от исходной точки, и самая долгая — при направлении течения в сторону, противоположную требуемому отклонению от исходной точки. Таким образом, из проведенного моделирования следует, что траектории типа

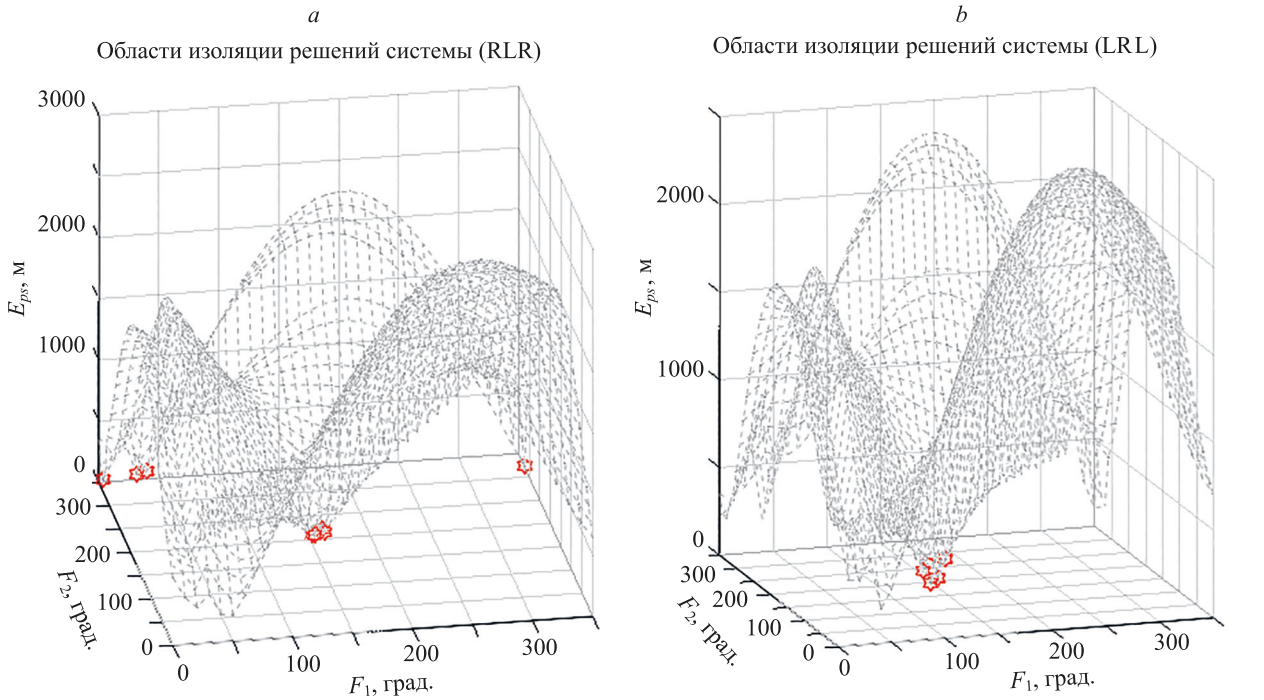


Рис. 4. Зависимость расстояния (E_{ps}) между заданной конечной точкой траектории и точкой, в которую приходит программная траектория в зависимости от значения курсовых углов в точках переключения управления, для траекторий типа RLR (*a*) и LRL (*b*)

Fig. 4. Dependence of the distance between the specified end point of the trajectory and the point at which the programmed trajectory comes for RLR (*a*) and LRL (*b*) strategies, depending on the value of the heading angles at points where the control is switched

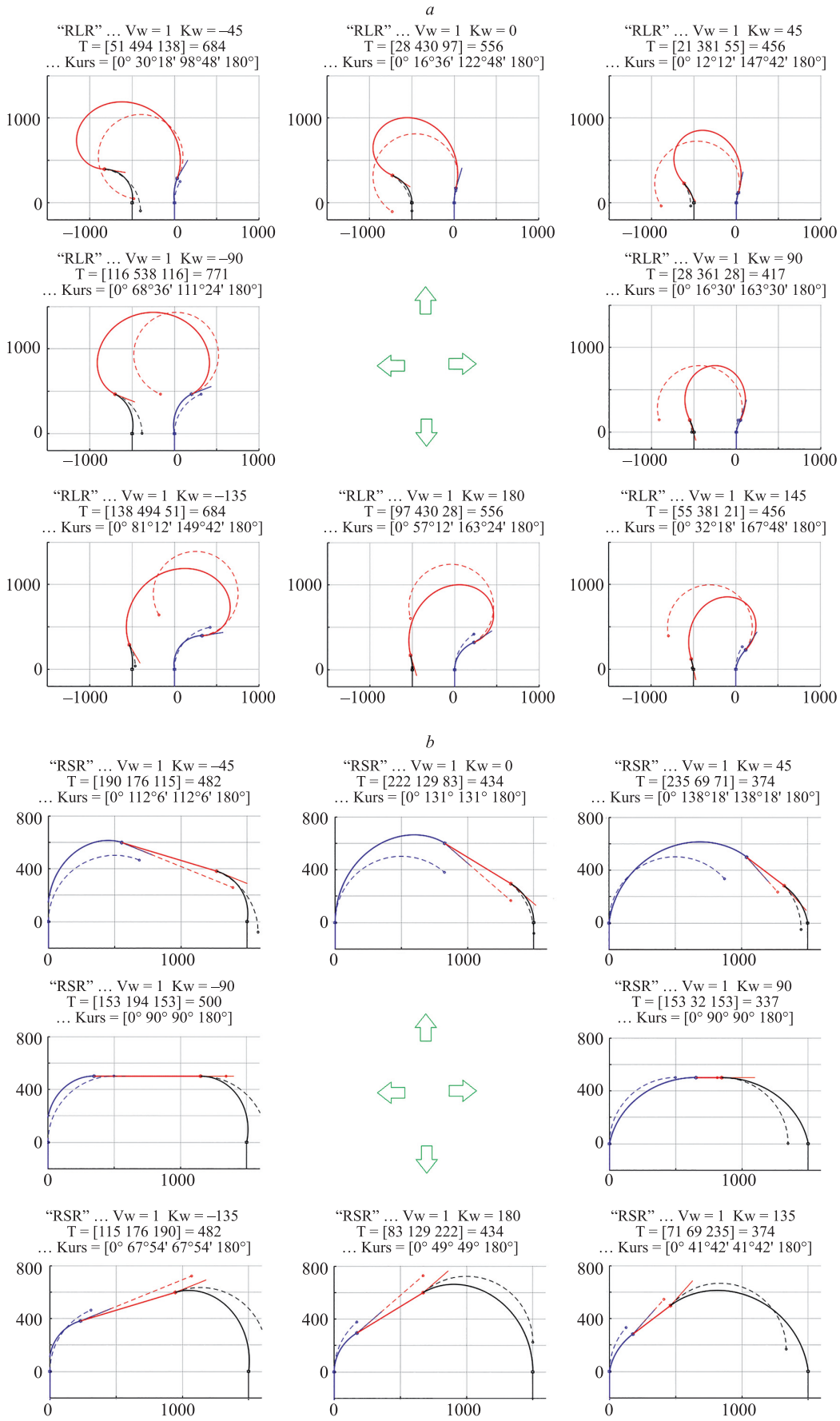


Рис. 5. Изменение траекторий RLR (*a*) и RSR (*b*) движения объекта в зависимости от направления внешнего воздействия *Kw*. Пунктиром показан вид участков траектории при отсутствии внешних воздействий

Fig. 5. Changing the trajectories RLR (*a*) and RSR (*b*) of object movement depending on the direction of external influence *Kw*. The dotted line shows the view of the trajectory sections in the absence of external influences

RLR и RSR по-разному реагируют на формирование программных траекторий при одинаковых типах постоянного возмущения, действующего на подвижный объект, но разных краевых условиях.

Заключение

Рассмотрен подход к формированию программных траекторий движения объектов в горизонтальной плоскости под воздействием постоянных по значению и величине внешних воздействий (ветра, течений) с использованием решения оптимальной по быстродействию задачи для машины Дубинса. Вместо поиска оптимальной траектории на основе принципа максимума Понтрягина использован перебор возможных типов траекторий с целью определения среди них наилучшей по быстродействию. Для каждой траектории расчет точек переключения произведен на основе минимизации разности между заданными координатами конечной точки и точки, в которую траектория приходит в зависимости от выбора параметров двух промежуточных точек переключения управления.

Показано, что при отсутствии внешних воздействий полученные в соответствии с предложенным методом траектории совпадают с траекториями, рассчитанными аналитически в соответствии с известными результатами исследований и построенными с использованием геометрического подхода для кривых Дубинса. Эти данные подтверждают гипотезу о возможности решения классической задачи машины Дубинса в случае действия на объект постоянных внешних воздействий.

Дальнейшие исследования проблемы построения оптимальных по быстродействию траекторий для подвижных объектов, имеющих ограничения по управлению, планируется вести в направлении распространения описанного в работе подхода к движению в горизонтальной плоскости на случай движения объекта в пространстве. Отметим, что система уравнений (1) упрощенно описывает движение судна, и в более общем случае ее уравнения приходится дополнять слагаемыми, учитывающими скорость перекладки руля и угол дрейфа судна, что, в свою очередь, усложняет поиск оптимальной траектории, но позволяет получить наиболее реалистичные решения.

Литература

1. Кумков С.И., Пацко В.С., Пятко С.Г., Федотов А.А. Построение множества разрешимости в задаче проводки самолета при ветровом возмущении // Труды Института математики и механики УрО РАН. 2005. Т. 11. № 1. С. 149–159.
2. Dubins L.E. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents // American Journal of Mathematics. 1957. V. 79. N 3. P. 497–516. https://doi.org/10.2307/2372560
3. Вагизов М.Р., Хабаров С.П. Алгоритм формирования гладких программных траекторий движения БПЛА // Информация и космос. 2021. № 2. С. 122–130.
4. Хабаров С.П., Шилкина М.Л. Геометрический подход к решению задачи для машин Дубинса при формировании программных траекторий движения // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 5. С. 653–663. https://doi.org/10.17586/2226-1494-2021-21-5-653-663
5. Хабаров С.П., Шилкина М.Л. Формирование программных траекторий движения БПЛА с учетом ограничений на их управляемость // Цифровые технологии в лесном секторе: Материалы II Всероссийской научно-технической конференции-вебинара. Санкт-Петербург, 2021. С. 141–143.
6. Пацко В.С., Федотов А.А. Аналитическое описание множества достижимости для машины Дубинса // Труды Института математики и механики УРО РАН. 2020. Т. 26. № 1. С. 182–197. https://doi.org/10.21538/0134-4889-2020-26-1-182-197
7. Patsko V.S., Fedotov A.A. Three-dimensional reachable set at instant for the Dubins car: Properties of extremal motions // Proc. of the 60th Israel Annual Conference on Aerospace Sciences, (IACAS). 2020. P. 1033–1049.
8. Patsko V.S., Fedotov A.A. Reachable set for Dubins car and its application to observation problem with incomplete information // Proc. of the 27th Mediterranean Conference on Control and Automation, (MED 2019). 2019 P. 489–494. https://doi.org/10.1109/MED.2019.8798511
9. Рогачев Г.Н. Использование генетического алгоритма с отсечением по времени в задаче синтеза программного регулятора для машины Дубинса // Вестник Самарского государственного технического университета. Серия: Технические науки. 2011. № 3(31). С. 27–32.
10. Buzikov M.E., Galyaev A.A. Time-optimal interception of a moving target by a Dubins car // Automation and Remote Control. 2021. V. 82. N 5. P. 745–758. https://doi.org/10.1134/S0005117921050015

References

1. Kumkov S.I., Patcko V.S., Piatko S.G., Fedotov A.A. Construction of the solvability set for guiding an aircraft under wind disturbance. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, 2005, vol. 11, no. 1, pp. 149–159. (in Russian)
2. Dubins L.E. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *American Journal of Mathematics*, 1957, vol. 79, no. 3, pp. 497–516. https://doi.org/10.2307/2372560
3. Vagizov M.R., Khabarov S.P. Algorithm for the formation of smooth programmed trajectories of UAV motion. *Information and Space*, 2021, no. 2, pp. 122–130. (in Russian)
4. Khabarov S.P., Shilkina M.L. geometric approach to the solution of the dubins car problem in the formation of program trajectories. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 653–663. (in Russian). https://doi.org/10.17586/2226-1494-2021-21-5-653-663
5. Khabarov S.P., Shilkina M.L. Building programmed motion trajectories for unmanned aircraft with control restrictions. *Proc. of the 2nd All-Russian Scientific and Technical Conference-Webinar "Digital Technologies in the Forestry Sector"*. St. Petersburg, 2021, pp. 141–143. (in Russian)
6. Patsko V.S., Fedotov A.A. Analytic description of a reachable set for the Dubins car. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, 2020, vol. 26, no. 1, pp. 182–197. (in Russian). https://doi.org/10.21538/0134-4889-2020-26-1-182-197
7. Patsko V.S., Fedotov A.A. Three-dimensional reachable set at instant for the Dubins car: Properties of extremal motions. *Proc. 60th Israel Annual Conference on Aerospace Sciences, (IACAS)*, 2020, pp. 1033–1049.
8. Patsko V.S., Fedotov A.A. Reachable set for Dubins car and its application to observation problem with incomplete information. *Proc. of the 27th Mediterranean Conference on Control and Automation, (MED 2019)*, 2019, pp. 489–494. https://doi.org/10.1109/MED.2019.8798511
9. Rogachev G.N. Synthesis of the open-loop regulator for dubins' car by anytime genetic algorithm. *Vestnik of Samara State Technical University. Technical Sciences Series*, 2011, no. 3(31), pp. 27–32. (in Russian)
10. Buzikov M.E., Galyaev A.A. Time-optimal interception of a moving target by a Dubins car. *Automation and Remote Control*, 2021, vol. 82, no. 5, pp. 745–758. https://doi.org/10.1134/S0005117921050015

11. Ветчинкин А.С., Стариченков А.Л. Формирование траектории сближения морских подвижных объектов // Морские интеллектуальные технологии. 2020. № 1-1(47). С. 113–117. https://doi.org/10.37220/MIT.2020.47.1.014
12. Silverberg L., Xu D. Dubins waypoint navigation of small-class unmanned aerial // Open Journal of Optimization. 2019. V. 8. N 2. P. 59–72. https://doi.org/10.4236/ojop.2019.82006
13. Savla K., Frazzoli E., Bullo F. Traveling salesperson problems for the Dubins vehicle // IEEE Transactions on Automatic Control. 2008. V. 53. N 6. P. 1378–1391. https://doi.org/10.1109/TAC.2008.925814
14. Parlangeli G., Indiveri G. Dubins inspired 2D smooth paths with bounded curvature and curvature // IFAC Proceedings Volumes. 2010 V. 43. N 16. P. 252–257. https://doi.org/10.3182/20100906-3-IT-2019.00045
15. Cai W., Zhang M., Zheng Y.R. Task assignment and path planning for multiple autonomous underwater vehicles using 3D Dubins curves // Sensors. 2017. V. 17. N 7. P. 1607. https://doi.org/10.3390/s17071607
16. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Redundant servicing of a flow of heterogeneous requests critical to the total waiting time During the multi-path passage of a sequence of info-communication nodes // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. V. 12563. P. 100–112. https://doi.org/10.1007/978-3-030-66471-8_9
17. Bogatyrev S.V., Bogatyrev A.V., Bogatyrev V.A. Priority maintenance with replication of wait-critical requests // Proc. of the 2021 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2021. P. 9470640. https://doi.org/10.1109/WECONF51603.2021.9470640
18. LaValle S.M. Planning Algorithms. Cambridge University Press, 2006. 1023 p. https://doi.org/10.1017/CBO9780511546877
19. Вагизов М.Р., Хабаров С.П. Построение программных траекторий движения на базе решения задачи «Машина Дубинса» // Информация и космос. 2021. № 3. С. 116–125.

11. Vetchinkin A.S., Starichenkov A.L. Formati on of the trajectory of convergence of marine mobile objects. *Marine Intellectual Technologies*, 2020, no. 1-1 (47), pp. 113–117. (in Russian). https://doi.org/10.37220/MIT.2020.47.1.014
12. Silverberg L., Xu D. Dubins waypoint navigation of small-class unmanned aerial. *Open Journal of Optimization*, 2019, vol. 8, no. 2, pp. 59–72. https://doi.org/10.4236/ojop.2019.82006
13. Savla K., Frazzoli E., Bullo F. Traveling salesperson problems for the Dubins vehicle. *IEEE Transactions on Automatic Control*, 2008, vol. 53, no. 6, pp. 1378–1391. https://doi.org/10.1109/TAC.2008.925814
14. Parlangeli G., Indiveri G. Dubins inspired 2D smooth paths with bounded curvature and curvature. *IFAC Proceedings Volumes*, 2010, vol. 43, no. 16, pp. 252–257. https://doi.org/10.3182/20100906-3-IT-2019.00045
15. Cai W., Zhang M., Zheng Y.R. Task assignment and path planning for multiple autonomous underwater vehicles using 3D Dubins curves. *Sensors*, 2017, vol. 17, no. 7, pp. 1607. https://doi.org/10.3390/s17071607
16. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Redundant servicing of a flow of heterogeneous requests critical to the total waiting time During the multi-path passage of a sequence of info-communication nodes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12563, pp. 100–112. https://doi.org/10.1007/978-3-030-66471-8_9
17. Bogatyrev S.V., Bogatyrev A.V., Bogatyrev V.A. Priority maintenance with replication of wait-critical requests. *Proc. of the 2021 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, 2021, pp. 9470640. https://doi.org/10.1109/WECONF51603.2021.9470640
18. LaValle S.M. *Planning Algorithms*. Cambridge University Press, 2006, 1023 p. https://doi.org/10.1017/CBO9780511546877
19. Vagizov M.R., Khabarov S.P. Construction of software trajectories of motion on the basis of the solution of the problem “Dubins machine”. *Information and Space*, 2021, no. 3, pp. 116–125. (in Russian)

Авторы

Хабаров Сергей Петрович — кандидат технических наук, доцент, доцент, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург, 194021, Российская Федерация, SC 57220744173, https://orcid.org/0000-0003-1337-0150, Serg.Habarov@mail.ru

Шилкина Мария Львовна — кандидат технических наук, доцент, доцент, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург, 194021, Российская Федерация, SC 57246047100, https://orcid.org/0000-0001-6501-5789, mchernobay@inbox.ru

Authors

Sergey P. Khabarov — PhD, Associate Professor, Associate Professor, St. Petersburg State Forest Technical University, Saint Petersburg, 194021, Russian Federation, SC 57220744173, https://orcid.org/0000-0003-1337-0150, Serg.Habarov@mail.ru

Maria L. Shilkina — PhD, Associate Professor, Associate Professor, St. Petersburg State Forest Technical University, Saint Petersburg, 194021, Russian Federation, SC 57246047100, https://orcid.org/0000-0001-6501-5789, mchernobay@inbox.ru

Статья поступила в редакцию 15.11.2021
Одобрена после рецензирования 19.12.2021
Принята к печати 30.01.2022

Received 15.11.2021
Approved after reviewing 19.12.2021
Accepted 30.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-179-186
УДК 621.319.7.001.5

Математическая модель эпидемии с произвольным законом восстановления

Владимир Константинович Сёменов[1], Наталья Борисовна Иванова[2]✉

[1,2] Ивановский государственный энергетический университет им. В.И. Ленина, Иваново, 153003, Российская Федерация

[1] semenov_vk@mail.ru, https://orcid.org/0000-0001-9765-2096
[2] rgr_ivanova@rambler.ru✉, https://orcid.org/0000-0002-9519-6047

Аннотация

Предмет исследования. В работе предложена математическая модель процесса эпидемии, учитывающая зависимость интенсивностей излечения и потери иммунитета от времени. На сегодняшний день большое распространение получили математические модели эпидемий, основанные на базовой модели Кермака–Маккендрика. Среди них наиболее известны модели «Восприимчивые–инфицированные–выздоровевшие» (Susceptible-Infected-Recovered, SIR) и «Восприимчивые–контактные–инфицированные–выздоровевшие» (Susceptible-Exposed-Infected-Recovered, SEIR). Основу данных моделей составляет разбиение населения на отдельные группы, находящиеся в разных эпидемических состояниях. В основу описания моделей положены дифференциальные уравнения, аналогичные уравнениям рождения и гибели в процессе радиоактивных превращений элементов в радиоактивной цепочке. Однако подобный подход не учитывает зависимость вероятностей перехода населения из группы в группу от времени пребывания в процессе лечения и в процессе утрачивания приобретенного иммунитета. Известные модели не предусматривают анализ характера протекания эпидемии для больших промежутков времени, когда процесс может войти в стационарное состояние. **Метод.** В работе предложена математическая модель, которая основана на разбиении населения на отдельные группы. Первую группу составляют здоровые люди, подверженные инфицированию вследствие контакта с членами второй группы, в которую включается инфицированное население. Члены третьей группы находятся на лечении, к четвертой группе относятся переболевшие члены общества с антителами и привитые. Пятую группу составляют умершие члены общества. В отличие от SIR и SEIR моделей в предложенном подходе учтено, что с течением времени иммунитет утрачивается, и выжившие люди снова переходят в группу подверженных инфицированию. Учтены зависимости вероятностей перехода населения из группы в группу от времени пребывания как в процессе лечения, так и в процессе утрачивания приобретенного иммунитета. Таким образом, предложенная математическая модель основана на пяти интегро-дифференциальных уравнениях, два из которых являются уравнениями в частных производных. **Основные результаты.** Сформулирована новая математическая модель, позволяющая учитывать зависимость интенсивности излечения и вероятности перехода из вакцинированного состояния в исходное от времени нахождения в соответствующем состоянии. Показано, что предложенная модель носит автокаталитический характер. При увеличении времени наблюдается состояние бистабильности, когда при одних и тех же граничных условиях сосуществуют два стабильных состояния. Переключение между состояниями определяется найденным в работе управляющим параметром распространения эпидемии. Одно из стабильных состояний — стационарное и приводит к окончанию эпидемии, второе — к вымиранию популяции. Показано, что для стационарного режима вид функции распределения по времени лечения и времени выдержки в вакцинированном состоянии никак не сказывается на окончательном результате. Сформулированы условия подавления эпидемии для управления процессом ее развития. Проведены численные эксперименты по симуляции процесса распространения эпидемии с учетом постоянства всех переходных вероятностей. Интегрирование исходной системы уравнений осуществлено с использованием алгоритма Radaus для жестких дифференциальных уравнений. Результаты численного моделирования подтвердили соответствие экспериментальных данных теории управляющего параметра. **Практическая значимость.** Результаты работы могут применяться для организации управления процессом распространения эпидемии с целью ее скорейшего подавления за счет изменения значения управляющего параметра.

Ключевые слова

эпидемия, математическая модель, плотность распределения по времени восстановления, бистабильное состояние, управляющий параметр, условия подавления

© Сёменов В.К., Иванова Н.Б., 2022

Ссылка для цитирования: Сёменов В.К., Иванова Н.Б. Математическая модель эпидемии с произвольным законом восстановления // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 179–186. doi: 10.17586/2226-1494-2022-22-1-179-186

A mathematical model of an epidemic with an arbitrary law of recovery

Vladimir K. Semenov[1], Natalia B. Ivanova[2]✉

[1,2] Ivanovo State Power University named after V. I. Lenin, Ivanovo, 153003, Russian Federation

[1] semenov_vk@mail.ru, https://orcid.org/0000-0001-9765-2096

[2] rgr_ivanova@rambler.ru✉, https://orcid.org/0000-0002-9519-6047

Abstract

The paper proposes a mathematical model of the epidemic process, taking into account the dependence of the rates of cure and loss of immunity on time. Today, mathematical models of epidemics based on the basic Kermack–McKendrick model have become widespread. The most famous models are Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infected-Recovered (SEIR). These models are based on dividing the population into separate groups that are in different epidemic conditions. The description of the models is based on differential equations similar to the equations of birth and death in the process of radioactive transformations of elements in a radioactive chain. However, this approach does not take into account the dependence of the probabilities of the transition of the population from group to group on time spent in the treatment process and in the process of loss of acquired immunity. The known models do analyze the nature of the course of the epidemic for long periods of time, when the process can enter a stationary state. The paper proposes a mathematical model based on dividing the population into separate groups. The first group consists of healthy people susceptible to infection due to contact with members of the second group, which includes the infected population. Members of the third group are being treated, the fourth group includes members of the society who have recovered with antibodies and are vaccinated. The fifth group consists of deceased members of society. In contrast to the SIR and SEIR models, the proposed approach takes into account that immunity is lost over time, and people who survived again move to the group susceptible to infection. The dependences of the probabilities of transition from group to group on time spent both during the treatment process and in the loss of acquired immunity have been taken into account. Thus, the proposed mathematical model is based on five integro-differential equations, two of which are partial differential equations. A new mathematical model has been formulated that makes it possible to take into account the dependence of the cure rate and the probability of transition from the vaccinated state to the initial state on time spent in the corresponding state. It is shown that the proposed model is autocatalytic. With increasing time, a state of bistability is observed, when, under the same boundary conditions, two stable states coexist. Switching between states is determined by the epidemic spread control parameter found in the work. One of the stable states is stationary and leads to the end of the epidemic, the second one leads to the population's extinction. It has been shown that, for the stationary regime, the form of the distribution function in terms of treatment time and exposure time in the vaccinated state does not affect the final result in any way. The conditions for suppressing the epidemic for managing the process of its development are formulated. Numerical experiments were carried out to simulate the epidemic spreading process, taking into account the constancy of all transition probabilities. Integration of the original system of equations was carried out using the Radaus algorithm for stiff differential equations. The results of numerical simulations have confirmed that the experimental data agree with the theory of the control parameter. The results of the work can be used to organize the management of the epidemic spreading process in order to suppress it as soon as possible by changing the value of the control parameter.

Keywords

epidemic, mathematical model, distribution density over recovery time, bistable state, control parameter, suppression conditions

For citation: Semenov V.K., Ivanova N.B. A mathematical model of an epidemic with an arbitrary law of recovery. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 179–186 (in Russian). doi: 10.17586/2226-1494-2022-22-1-179-186

Введение

Научные работы по созданию математических моделей распространения эпидемий и пандемий появились еще в тридцатые годы прошлого столетия и особенно актуальны они в настоящее время. В продолжение этих работ можно выделить несколько направлений. Одно из них связано с построением математических моделей, основанных на статистической обработке данных эпидемии, подборе по экспериментальным данным параметров статистических распределений, таких, как нормальное, распределение Пуассона, гамма-распределение и др. [1–6]. Развитием этого направления можно считать работы, посвященные анализу и прогнозированию временных рядов с использованием модели Auto Regressive Integrated Moving Average (ARIMA) [7–10]. Такой подход нацелен на вычислении коэффициента летальности, средней продолжительности лечения, эффективности карантинных мер и касается краткосрочных прогнозов протекания эпидемии, но не позволяет отслеживать и исследовать так называемые «волны» эпидемии.

Второе направление связано с развитием теории дифференциальных игр и основано на модели Лотки–Вольтерра, использующей два обыкновенных нелинейных дифференциальных уравнения, описывающих взаимодействие в системе «паразит–хозяин» и созданных для анализа механизма этого взаимодействия [11].

Основу третьего подхода составляет базовая модель Кермака–Маккендрика — «Восприимчивые–инфи-

цированные–выздоровевшие» (Susceptible-Infected-Recovered, SIR), на основе которой сформулирована модель «Восприимчивые–контактные–инфицированные–выздоровевшие» (Susceptible-Exposed-Infected-Recovered, SEIR) и другие ее модификации [12–17].

Данные модели нашли наиболее широкое применение. Остановимся на описании базовой модели, согласно которой население региона в процессе протекания эпидемии разбивается на три группы: восприимчивые к заболеванию, инфицированные и переболевшие/умершие. Передача инфекции происходит от инфицированных индивидов к восприимчивым. Считается, что переболевшие индивиды приобретают иммунитет и не могут быть заражены вторично. Модель SEIR учитывает дополнительно группу людей, находящихся в инкубационном периоде заболевания. Другие модели, основанные на базовой модели, вводят в рассмотрение новые дополнительные группы, однако это не приводит к качественным изменениям. В основу модели положена система обыкновенных дифференциальных уравнений типа рождения и гибели, записанная для общественной системы, состоящей из отдельных групп людей, переходящих из одной группы в другую в процессе развития эпидемии. Данные уравнения аналогичны уравнениям рождения и гибели, описывающим радиоактивные превращения элементов в радиоактивной цепочке. Недостаток этого подхода в том, что он не учитывает зависимость вероятностей восстановления системы от времени лечения и времени нахождения в защищенном состоянии. В частности считается, что функция распределения времени лечения пациента подчиняется экспоненциальному закону, что равносильно тому, что интенсивность излечения не зависит от его длительности. Это означает, что вероятность выздоровления инфицированного человека в единицу времени в начале процесса лечения равняется вероятности его излечения в любой промежуток времени (при условии, что процесс лечения еще не закончен). Такое допущение никогда не выполняется и вводится по причине больших математических трудностей, которые появляются при решении данной задачи. Кроме того, в этих моделях не учитывается тот факт, что с течением времени приобретенный иммунитет утрачивается, и люди из группы приобретенного иммунитета могут быть снова подвержены инфекции, т. е. считается, что время пребывания в защищенном состоянии значительно превышает время продолжительности эпидемии. Если это условие не выполняется, то процесс распространения эпидемии претерпит качественные изменения.

В настоящей работе данные допущения исключены, и считается, что интенсивность излечения — произвольная функция времени лечения. Также учтено, что вероятность перехода из защищенного состояния в исходное зависит от продолжительности действия антител. Рассмотрен практический вопрос, который выпал из внимания исследователей — определение условий окончания процесса распространения эпидемии и возможность управления ею.

Математическая модель

В работе воспользуемся понятием системы, подразумевая под этим общественную формацию, состоящую из отдельных групп людей, члены которых в процессе развития эпидемии могут переходить из одной группы в другую.

Как известно, вирусная эпидемия распространяется за счет контактов здоровых людей с инфицированными, поскольку вирусы не могут размножаться вне живых клеток, так как у них отсутствует аппарат трансляции, т. е. синтеза белка. В связи с этим для построения математической модели распространения эпидемии в замкнутой системе разобьем ее на пять подсистем, каждая из которых представляется соответствующей группой населения. К группе A относятся все члены общества, которые не инфицированы и не защищены от инфекции и при контакте с инфицированными могут подвергнуться инфекции. Ко второй группе (B) относятся инфицированные люди, не находящиеся в изоляции и не проходящие лечения. Третью группу (C) составляют люди, находящиеся на лечении и изолированные от общества. В четвертую группу (D) входят вылечившиеся или привитые члены общества, а пятую группу (E) составляют умершие члены общества.

Для удобства формирования уравнений, описывающих переходы людей из одной группы в другую, изобразим переходы между группами в виде графа (рис. 1).

Для решения вопроса о критерии затухания эпидемии исследуем условия перехода этого процесса в стационарное состояние при неограниченном возрастании времени. Примем допущение, что система является замкнутой, т. е. число людей во всех группах остается неизменным и равным N_0. При нахождении соответствующих условий стационарности процесса основная задача — определение количества людей в каждой из перечисленных групп. Введем следующие обозначения: $N_1(t)$ — количество людей в группе A; $N_2(t)$ — в группе B; $N_3(t)$ — в группе C; $N_4(t)$ — в группе D; $N_5(t)$ — в группе E; $n_3(\tau, t) = \dfrac{\Delta N_3}{\Delta \tau}$ — плотность распределения людей по времени лечения τ в группе C; $n_4(\tau, t)$ — плотность распределения людей по времени нахождения τ в группе D.

В качестве характеристик системы введем следующие параметры: λ_0 — вероятность прививки, характеризующая переход одного человека за единицу времени из группы A в группу D; λ_1 — вероятность заражения

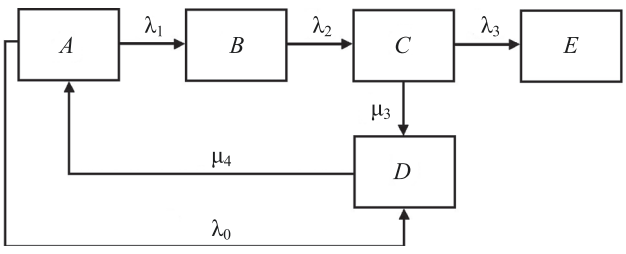


Рис. 1. Структура системы

Fig. 1. System structure

отдельного индивидуума из группы A за единицу времени за счет контактов с зараженными из группы B. Эту вероятность представим в виде:

$$\lambda_1 = \frac{\Delta N_k}{\Delta t_k}\frac{N_2}{N_0}\varepsilon,$$

где $\frac{\Delta N_k}{\Delta t_k}\frac{N_2}{N_0}$ — вероятность контакта одного человека из группы A с инфицированными людьми из группы B, $\frac{\Delta N_k}{\Delta t_k}$ — среднее число контактов одного человека в обществе за единицу времени, $\frac{N_2}{N_0}$ — доля инфицированных в обществе; ε — прозрачность защитного барьера, определяемая применением защитных средств, $0 \le \varepsilon \le 1$. Обозначим $w_1 = \frac{\Delta N_k}{\Delta t_k}\varepsilon$, тогда $\lambda_1 = w_1\frac{N_2}{N_0}$; вероятности перехода человека: λ_2 — за единицу времени из группы B в группу C (на лечение), $\mu_3(\tau)$ — из группы C в группу D (выздоровление), $\mu_4(\tau)$ — из группы D в группу A (после ослабления и потери иммунитета); λ_3 — вероятность летального исхода за единицу времени.

Исходные дифференциальные уравнения имеют вид:

$$\frac{dN_1(t)}{dt} = -\lambda_0 N_1 - w_1 N_1\frac{N_2}{N_0} + \int_0^t \mu_4(\tau) n_4(t, \tau)d\tau, \quad (1)$$

$$\frac{dN_2(t)}{dt} = w_1 N_1\frac{N_2}{N_0} - \lambda_2 N_2, \quad (2)$$

$$\frac{\partial n_3(t, \tau)}{\partial t} = -\frac{\partial n_3(t, \tau)}{\partial \tau} - \mu_3(\tau) n_3(t, \tau) - \lambda_3 n_3(t, \tau), \quad (3)$$

$$\frac{\partial n_4(t, \tau)}{\partial t} = -\frac{\partial n_4(t, \tau)}{\partial \tau} - \mu_4(\tau) n_4(t, \tau), \quad (4)$$

$$\frac{dN_5}{dt} = \lambda_3\int_0^t n_3(t, \tau)d\tau. \quad (5)$$

К уравнениям (1)–(5) необходимо добавить следующие граничные и начальные условия:

$$n_4(t, \tau = 0) = \lambda_0 N_1 + \int_0^t \mu_3(\tau) n_3(t, \tau)d\tau, \quad (6)$$

$$n_3(t, \tau = 0) = \lambda_2 N_2, \quad (7)$$

$$N_2(0) = N_{20},\ N_1(0) = N_0 - N_{20},$$

где N_{20} — число носителей вируса на момент начала эпидемии.

Бистабильное состояние. Исследуем возможность стационарного решения при $t \to \infty$, когда все производные по времени от искомых величин равны нулю и количество людей во всех группах не меняется:

$$-\lambda_0 N_1 - w_1 N_1\frac{N_2}{N_0} + \int_0^\infty \mu_4(\tau) n_4(t, \tau)d\tau = 0, \quad (8)$$

$$w_1 N_1\frac{N_2}{N_0} - \lambda_2 N_2 = 0, \quad (9)$$

$$\frac{\partial n_3(\tau)}{\partial \tau} = -\mu_3(\tau) n_3(t, \tau) - \lambda_3 n_3(t, \tau), \quad (10)$$

$$\frac{\partial n_4(\tau)}{\partial \tau} = -\mu_4(\tau) n_4(\tau), \quad (11)$$

$$\frac{dN_5}{dt} = 0. \quad (12)$$

Решение уравнений (10) и (11) получим в виде:

$$n_3(\tau) = C_1\exp(-\int_0^\tau [\lambda_3 + \mu_3(\tau')]d\tau'),$$

$$n_4(\tau) = C_2\exp(-\int_0^\tau \mu_4(\tau')d\tau'). \quad (13)$$

Используя условия (6) и (7), находим

$$C_1 = \lambda_2 N_2,\ C_2 = C_1 + \lambda_0 N_1.$$

Из уравнений (5) и (12) видно, что $N_3 = 0$, т. е. количество людей в группе E перестанет меняться, когда в группе C никого не останется. В этом случае необходимо, учитывая условие (7), чтобы и N_2 также было равно нулю.

Наличие нелинейного слагаемого в уравнении (2), с физической точки зрения, говорит о том, что процесс носит автокаталитический характер — скорость рождения числа инфицированных пропорциональна их числу. В таких процессах наблюдается явление бистабильности, когда при одних и тех же граничных условиях одновременно сосуществуют два устойчивых стационарных состояния. Выбор системой одного из них зависит от соотношения между характеризующими процесс параметрами. В этом случае введем управляющий параметр, изменяя значения которого можно производить переключения между этими состояниями при достижении параметром состояния некоторого порогового значения.

Действительно, выполнение условия (9) возможно в двух случаях: в случае $N_2 = 0$ при произвольном N_1 и в случае $N_1 = \frac{\lambda_2}{w_1}N_0$ при произвольном N_2.

Вначале разберем случай $N_2 = 0$. С учетом того, что $C_1 = 0$, а, стало быть, и $N_3 = 0$, запишем условие нормировки

$$N_1 + \int_0^\infty n_4(\tau)d\tau + N_5 = N_0, \quad (14)$$

причем $\int_0^\infty n_4(\tau)d\tau = N_4$.

Интеграл, входящий в уравнение (8), представим в виде

$$N_4\int_0^\infty \mu_4(\tau)\frac{n_4(t, \tau)}{N_4}d\tau = N_4\cdot\langle\mu_4\rangle, \quad (15)$$

где $\langle\mu_4\rangle$ — среднее значение вероятности возвращения одного человека из группы D в группу A. Легко показать, что эта вероятность определяется средним временем пребывания людей в группе D. Для этого найдем N_4 прямым вычислением. Учитывая (13), получим

$$N_4 = \int_0^\infty n_4(\tau)d\tau = \lambda_0 N_1\int_0^\infty \exp(-\int_0^\tau \mu_4(\tau')d\tau')d\tau,$$

а интеграл возьмем по частям:

$$\int_0^\infty \exp(-\int_0^\tau \mu_4(\tau')d\tau')d\tau = \int_0^\infty \tau\cdot\mu_4\cdot \exp(-\int_0^\tau \mu_4(\tau')d\tau')d\tau =$$
$$= \int_0^\infty \tau dF = <\tau_4>,$$

где $F = 1 - \exp(-\int_0^\tau \mu_4(\tau')d\tau')$ — функция распределения времени пребывания в защищенном от вируса состоянии; $<\tau_4>$ — среднее время пребывания в защищенном состоянии. Отсюда следует, что

$$N_4 = \lambda_0 N_1 \cdot <\tau_4>.$$

Подставляя соотношение (15) в уравнение (8) и учитывая, что $N_2 = 0$, получим, что $\lambda_0 N_1 = N_4 \cdot <\mu_4>$. Поскольку $N_4 = \lambda_0 N_1 \cdot <\tau_4>$, то получим, что средняя вероятность потери иммунитета обратно пропорциональна среднему времени, прошедшему с момента вакцинации или выздоровления после болезни:

$$<\mu_4> = \frac{1}{<\tau_4>}.$$

Подставив найденное выражение для N_4 в уравнение (14), получим

$$N_1(1 + \frac{\lambda_0}{<\mu_4>}) + N_5 = N_0.$$

Учитывая, что в стационарном случае выполняется неравенство $N_5 << N_0$, найдем долю **неинфицированных** членов общества $\alpha = N_1/N_0$:

$$\alpha = \frac{1}{1 + \dfrac{\lambda_0}{<\mu_4>}}.$$

Итак, для этого случая $N_2 = 0$, $N_3 = 0$, а N_1, N_4 и N_5 — постоянны, как того и требует стационарное состояние.

Вернемся к соотношению (9) и сравним только что полученный результат $N_1 = \alpha N_0$ со второй возможностью $N_1 = \frac{\lambda_2}{w_1} N_0$, означающей, что количество зараженных людей сопоставимо с количеством восприимчивых к заражению людей. В этом случае рассмотрим так называемое переходное равновесие, когда все производные $\frac{dN}{dt}$ равны нулю, за исключением $\frac{dN_5}{dt}$. При этом ни N_2, ни N_3 не равны нулю, т. е. число летальных случаев с течением времени будет расти. В зависимости от соотношения между параметрами процесса осуществляться будет возможность с меньшим N_1.

Управляющий параметр. При выполнении условий

$$\alpha \le \frac{\lambda_2}{w_1} \text{ или } \frac{\lambda_0}{<\mu_4>} \ge \frac{w_1}{\lambda_2} - 1, \qquad (16)$$

процесс — стационарный, эпидемия прекратится, и в этом случае люди будут находиться только в трех группах A, D и E, причем число умерших — минимально. Управляющим параметром является $\frac{\lambda_0}{<\mu_4>}$, а его пороговое значение — $\frac{w_1}{\lambda_2} - 1$. Если значение управляющего параметра меньше порогового значения, то реализуется переходное равновесие с ростом числа летальных исходов. С ростом управляющего параметра достигается его пороговое значение, и происходит переключение на другое решение с прекращением эпидемии. Введенный критерий (16) определяет отличие одного процесса от другого и является условием прекращения эпидемии.

Численный эксперимент

Полученные результаты положены в основу проверки расчета динамики процесса на базе решения обыкновенных дифференциальных уравнений типа рождения и гибели с учетом постоянства всех вероятностей перехода между отдельными группами:

$$\begin{cases} \dfrac{dN_1(t)}{dt} = -\lambda_0 N_1 - w_1 N_1 \dfrac{N_2}{N_0} + \mu_4 N_4, \\ \dfrac{dN_2(t)}{dt} = w_1 N_1 \dfrac{N_2}{N_0} - \lambda_2 N_2, \\ \dfrac{dN_3}{dt} = \lambda_2 N_2 - \mu_3 N_3 - \lambda_3 N_3, \\ \dfrac{dN_4}{dt} = \lambda_0 N_1 + \mu_3 N_3 - \mu_4 N_4, \\ \dfrac{dN_5}{dt} = \lambda_3 N_3. \end{cases}$$

Начальные условия (7) остаются прежними. Численное моделирование осуществлено в пакете Mathcad с применением алгоритма Radaus для интегрирования жестких дифференциальных уравнений. Результаты расчета представлены в безразмерном виде. В качестве масштаба численности групп использована численность всего региона N_0, а в качестве масштаба времени $-1/w_1$. Начальные значения $N_1 = 0{,}999$ и $N_2 = 0{,}001$. Численные эксперименты проведены в широком диапазоне изменения параметров вероятностей перехода, за исключением λ_2, значения которого ограничены условием $N_1 = \frac{\lambda_2}{w_1} N_0$. Заметим, что с целью широкого варьирования параметров процесса их значения не связывались с данными процессов распространения конкретных эпидемий.

Расчеты показали, что при выполнении условия (16) эпидемия заканчивается с совпадением численности групп, предсказываемых приведенной теорией в разделе «Бистабильное состояние».

В примере №1 (рис. 2) выполнен расчет при следующих значениях характеристик системы: $w_1 = 21$ — контакт с зараженными в сутки; $\lambda_0 = 0{,}01\ w_1$, $\mu_3 = 0{,}1\ w_1$, $\lambda_2 = 0{,}7\ w_1$, $\mu_4 = 4{,}67 \times 10^{-3}\ w_1$.

В примере №2 (рис. 3) изменены только значения параметров: $\lambda_2 = 0{,}2\ w_1$, $\mu_4 = 0{,}01\ w_1$, остальные параметры имеют прежние значения. Это привело к тому, что неравенство (16) нарушено, и получено условие $N_1 = \frac{\lambda_2}{w_1} N_0$. В группах A, B, C, D ненулевое количество людей, но в группе E количество летальных исходов с течением времени растет, что подтверждается из-

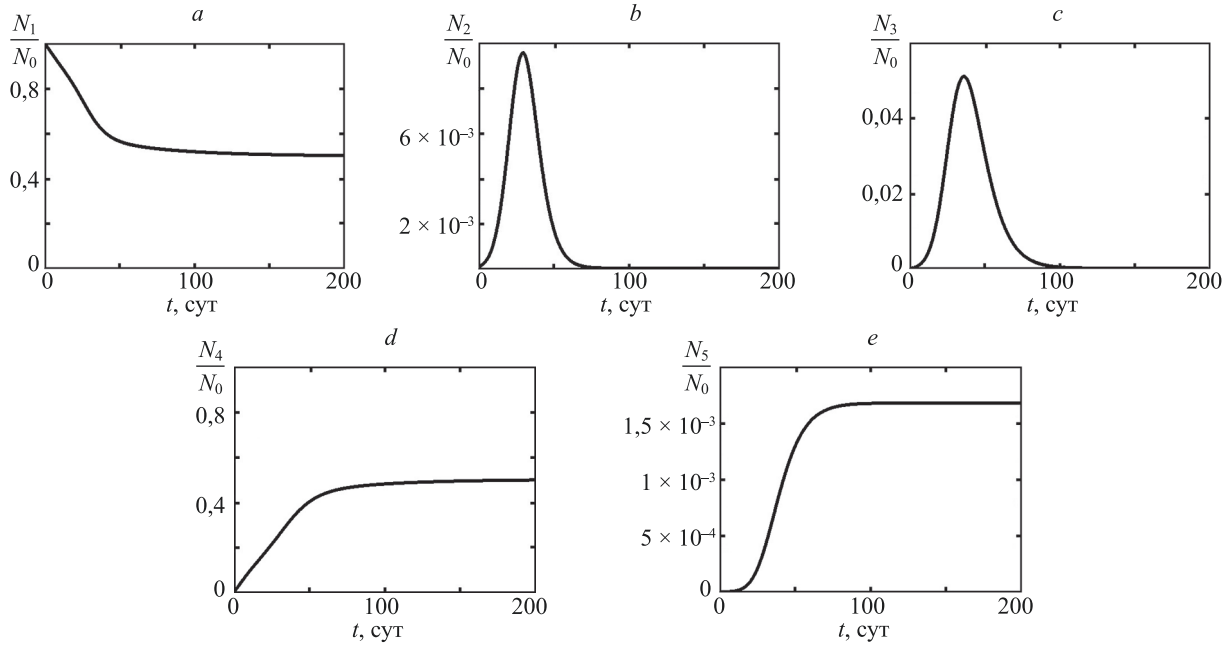


Рис. 2. Результаты примера №1 численного эксперимента для доли людей: восприимчивых к заражению (*a*); инфицированных (*b*); находящихся на лечении (*c*); имеющих антитела (*d*); умерших (*e*); *t* — время в сутках

Fig. 2. Results of the example No. 1 of a numerical experiment for the proportion of people: susceptible to infection (*a*); infected (*b*); being treated (*c*); having antibodies (*d*); deceased (*e*); *t* — time in days

ложенной теорией в разделе «Бистабильное состояние». Этому соответствует аналогия с радиоактивным семейством. В цепочке радиоактивных превращений возможно подвижное равновесие, но превращения будут происходить до тех пор, пока семейство не вымрет.

Из проведенного исследования видно, что для стационарного режима вид функций распределения по времени лечения и выдержки в вакцинированном состоянии на окончательном результате никак не сказывается. Обычное допущение о постоянстве величин μ_3 и μ_4 приведет к правильным результатам при условии, что средние значения времен восстановления $\langle\tau_3\rangle = 1/\mu_3$ и $\langle\tau_4\rangle = 1/\mu_4$ определяются экспериментально. Вопреки распространенному мнению, что коронавирусная пан-

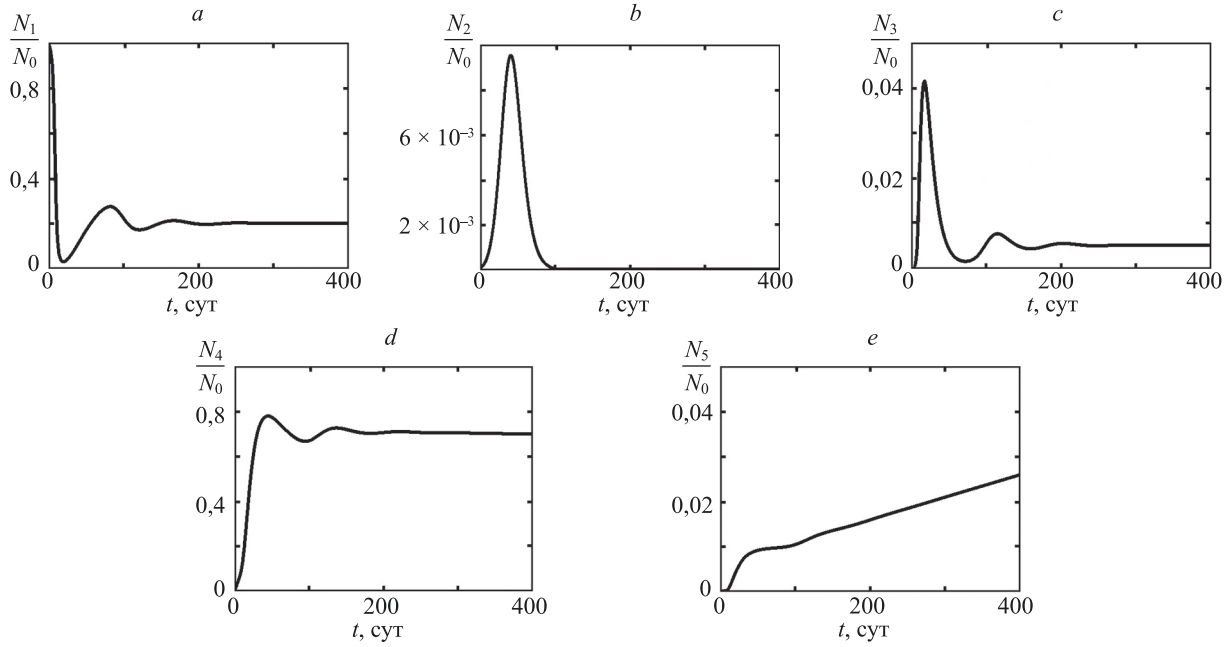


Рис. 3. Результаты примера №2 численного эксперимента для доли людей: восприимчивых к заражению (*a*); инфицированных (*b*); находящихся на лечении (*c*); имеющих антитела (*d*); умерших (*e*); *t* — время в сутках

Fig. 3. The results of the example No. 2 of a numerical experiment for the proportion of people: susceptible to infection (*a*); infected (*b*); being treated (*c*); having antibodies (*d*); deceased (*e*); *t* — time in days

демия сойдет на нет при количестве привитых 80 % от всего населения, из изложенного видно, что прекращение эпидемии, также как и численность отдельных групп, определяется значением управляющего параметра, т. е. выполнением критерия (16).

Выводы

При расчете динамики развития эпидемии зависимость вероятностей восстановления от времени можно не учитывать, определяя их по среднему времени пребывания в данной группе.

Развитие эпидемии носит бистабильный характер и может идти по двум сценариям. При выполнении критерия (16) эпидемия закончится с минимальным количеством летальных исходов. В противном случае процесс развития эпидемии переходит в стадию переходного равновесия между отдельными группами, за исключением последней, в которой число летальных исходов непрерывно растет.

Заключение

В работе предложена математическая модель процесса эпидемии, учитывающая зависимость интенсивности излечения и интенсивности потери иммунитета от времени. Установлено, что для больших промежутков времени развитие эпидемии приводит к бистабильному состоянию, когда при одних и тех же граничных условиях сосуществуют два стабильных состояния, одно из которых говорит об окончании эпидемии. Достичь его можно за счет изменения значения управляющего параметра, который является характеристикой такого рода состояний. Это условие следует применять для практической борьбы с эпидемией. В дальнейшем можно исследовать динамические возможности предложенной модели в полном объеме на основе численного решения исходной системы интегро-дифференциальных уравнений.

Литература

1. Кольцова Э.М., Куркина Е.С., Васецкий А.М. Математическое моделирование распространения эпидемии коронавируса COVID-19 в Москве // Computational nanotechnology. 2020. Т. 7. № 1. С. 99–105. https://doi.org/10.33693/2313-223X-2020-7-1-99-105
2. Getz W.M., Dougherty E.R. Discrete stochastic analogs of Erlang epidemic models // Journal of Biological Dynamics. 2018. V. 12. N 1. P. 16–38. https://doi.org/10.1080/17513758.2017.1401677
3. Verity R., Okell L.C., Dorigatti I., Winskill P., Whittaker C., Imai N., Cuomo-Dannenburg G., Thompson H., Walker P.G.T., Fu H., Dighe A., Griffin J.T., Baguelin M., Bhatia S., Boonyasiri A., Cori A., Cucunubá Z., FitzJohn R., Gaythorpe K., Green W., Hamlet A., Laydon D., Nedjati-Gilani G., Riley S., van Elsland S., Volz E., Wang H., Wang Y., Xi X., Donnelly C.A., Ghani A.C., Ferguson N.M. Estimates of the severity of coronavirus disease 2019: a model-based analysis // The Lancet Infectious Diseases. 2020. V. 20. N 6. P. 669–677. https://doi.org/10.1016/S1473-3099(20)30243-7
4. Kucharski A.J., Russell T.W., Diamond C., Liu Y., Edmunds J., Funk S., Eggo R.M., Sun F., Jit M., Munday J.D., Davies N., Gimma A., van Zandvoort K., Gibbs H., Hellewell J., Jarvis C.I., Clifford S., Quilty B.J., Bosse N.I., Abbott S., Klepac P., Flasche S. Early dynamics of transmission and control of COVID-19: a mathematical modelling study // The Lancet Infectious Diseases. 2020. V. 20. N 5. P. 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4
5. Li R., Pei S., Chen B., Song Y., Zhang T., Yang W., Shaman J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2) // Science. 2020. V. 368. N 6490. P. 489–493. https://doi.org/10.1126/science.abb3221
6. Liu Y., Gayle A.A., Wilder-Smith A., Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus // Journal of Travel Medicine. 2020. V. 27. N 2. P. taaa021. https://doi.org/10.1093/jtm/taaa021
7. Singh S., Parmar K.S., Kumar J., Makkhan S.J.S. Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19 // Chaos, Solitons & Fractals. 2020. V. 135. P. 109866. https://doi.org/10.1016/j.chaos.2020.109866
8. Chakraborty T., Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis // Chaos, Solitons & Fractals. 2020. V. 135. P. 109850. https://doi.org/10.1016/j.chaos.2020.109850
9. Srivastava A., Prasanna V.K. Learning to forecast and forecasting to learn from the COVID-19 pandemic. 2020 [Электронный ресурс].

References

1. Koltsova E.M., Kurkina E.S., Vasetsky A.M. Mathematical modeling of the spread of COVID-19 in Moscow. *Computational nanotechnology*, 2020, vol. 7, no. 1, pp. 99–105. (in Russian). https://doi.org/10.33693/2313-223X-2020-7-1-99-105
2. Getz W.M., Dougherty E.R. Discrete stochastic analogs of Erlang epidemic models. *Journal of Biological Dynamics*, 2018, vol. 12, no. 1, pp. 16–38. https://doi.org/10.1080/17513758.2017.1401677
3. Verity R., Okell L.C., Dorigatti I., Winskill P., Whittaker C., Imai N., Cuomo-Dannenburg G., Thompson H., Walker P.G.T., Fu H., Dighe A., Griffin J.T., Baguelin M., Bhatia S., Boonyasiri A., Cori A., Cucunubá Z., FitzJohn R., Gaythorpe K., Green W., Hamlet A., Laydon D., Nedjati-Gilani G., Riley S., van Elsland S., Volz E., Wang H., Wang Y., Xi X., Donnelly C.A., Ghani A.C., Ferguson N.M. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 2020, vol. 20, no. 6, pp. 669–677. https://doi.org/10.1016/S1473-3099(20)30243-7
4. Kucharski A.J., Russell T.W., Diamond C., Liu Y., Edmunds J., Funk S., Eggo R.M., Sun F., Jit M., Munday J.D., Davies N., Gimma A., van Zandvoort K., Gibbs H., Hellewell J., Jarvis C.I., Clifford S., Quilty B.J., Bosse N.I., Abbott S., Klepac P., Flasche S. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020, vol. 20, no. 5, pp. 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4
5. Li R., Pei S., Chen B., Song Y., Zhang T., Yang W., Shaman J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 2020, vol. 368, no. 6490, pp. 489–493. https://doi.org/10.1126/science.abb3221
6. Liu Y., Gayle A.A., Wilder-Smith A., Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 2020, vol. 27, no. 2, pp. taaa021. https://doi.org/10.1093/jtm/taaa021
7. Singh S., Parmar K.S., Kumar J., Makkhan S.J.S. Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. *Chaos, Solitons & Fractals*, 2020, vol. 135, pp. 109866. https://doi.org/10.1016/j.chaos.2020.109866
8. Chakraborty T., Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, 2020, vol. 135, pp. 109850. https://doi.org/10.1016/j.chaos.2020.109850
9. Srivastava A., Prasanna V.K. *Learning to forecast and forecasting to learn from the COVID-19 PANDEMIC*. 2020. Available at: https://arxiv.org/pdf/2004.11372.pdf (accessed: 01.12.2021).

URL: https://arxiv.org/pdf/2004.11372.pdf, свободный. Яз. англ. (дата обращения: 01.12.2021).
10. Perone G. ARIMA forecasting of COVID-19 incidence in Italy, Russia, and the USA. 2020 [Электронный ресурс]. URL: https://arxiv.org/pdf/2006.01754.pdf, свободный. Яз. англ. (дата обращения: 01.12.2021).
11. Трубецков Д.И. Феномен математической модели Лотки-Вольтерры и сходных с ней // Известия вузов. Прикладная нелинейная динамика. 2011. Т. 19. № 2. С. 69–88. https://doi.org/10.18500/0869-6632-2011-19-2-69-88
12. Томчин Д.А., Фрадков А.Л. Прогнозирование распространения вируса COVID-19 в России на основе простых математических моделей эпидемий. 2020. [Электронный ресурс]. URL: http://onr-russia.ru/sites/default/files/tf20-5.pdf, свободный. Яз. рус. (дата обращения: 01.11.2021).
13. Криворотько О.И., Кабанихин С.И., Зятьков Н.Ю., Приходько А., Прохошин Н., Шишленин М.А. Математическое моделирование и прогнозирование COVID-19 в Москве и Новосибирской области // Сибирский журнал вычислительной математики. 2020. Т. 23. № 4. С. 395–414. https://doi.org/10.15372/SJNM20200404
14. Ed Fontes. Моделирование в Comsol Multiphisics ® распространения вируса Covid-19. 2020 [Электронный ресурс]. URL: https://www.comsol.ru/blogs/modeling-the-spread-of-covid-19-with-comsol-multiphysics/, свободный. (дата обращения: 01.11.2021).
15. Weiss H. The SIR model and the Foundations of Public Health // MATerials MATemàtics. 2013. V. 2013. N 3. P. 17 [Электронный ресурс]. URL: https://mat.uab.cat/matmat_antiga/PDFv2013/v2013n03.pdf, свободный. Яз. англ. (дата обращения: 01.11.2021).
16. Barmparis G.D., Tsironis G.P. Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach // Chaos, Solitons & Fractals. 2020. V. 135. P. 109842. https://doi.org/10.1016/j.chaos.2020.109842
17. Teles P. A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal. 2020 [Электронный ресурс]. URL: https://arxiv.org/pdf/2004.04735.pdf, свободный. Яз. англ. (дата обращения: 01.12.2021).

10. Perone G. *ARIMA forecasting of COVID-19 incidence in Italy, Russia, and the USA*. 2020. Available at: https://arxiv.org/pdf/2006.01754.pdf (accessed: 01.12.2021).
11. Trubetskov D.I. Phenomenon of Lotka-Volterra mathematical model and similar models. *Izvestiya Vysshikh Uchebnykh Zavedeniy. Prikladnaya Nelineynaya Dinamika*, 2011, vol. 19, no. 2, pp. 69–88. (in Russian). https://doi.org/10.18500/0869-6632-2011-19-2-69-88
12. Tomchin D., Fradkov A. *Prediction of the virus COVID-19 spread in Russia based on simple mathematical models of epidemics*. 2020. Available at: http://onr-russia.ru/sites/default/files/tf20-5.pdf (accessed: 01.11.2021). (in Russian)
13. Krivorot'ko O.I., Kabanikhin S.I., Zyat'kov N.Y., Prikhod'ko A.Y., Shishlenin M.A., Prokhoshin N.M. Mathematical modeling and forecasting of COVID-19 in Moscow and Novosibirsk Region. *Numerical Analysis and Applications*, 2020, vol. 13, no. 4, pp. 332–348. https://doi.org/10.1134/S1995423920040047
14. Ed Fontes. *Modeling the Spread of COVID-19 with COMSOL Multiphysics®*. Available at: https://www.comsol.ru/blogs/modeling-the-spread-of-covid-19-with-comsol-multiphysics/ (accessed: 01.11.2021).
15. Weiss H. The SIR model and the Foundations of Public Health. *MATerials MATemàtics*, 2013, vol. 2013, no. 3, pp. 17. Available at: https://mat.uab.cat/matmat_antiga/PDFv2013/v2013n03.pdf (accessed: 01.11.2021).
16. Barmparis G.D., Tsironis G.P. Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. *Chaos, Solitons & Fractals*, 2020, vol. 135, pp. 109842. https://doi.org/10.1016/j.chaos.2020.109842
17. Teles P. *A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal*. 2020. Available at: https://arxiv.org/pdf/2004.04735.pdf (accessed: 01.12.2021).

Авторы

Семёнов Владимир Константинович — доктор технических наук, профессор, профессор, Ивановский государственный энергетический университет им. В.И. Ленина, Иваново, 153003, Российская Федерация, https://orcid.org/0000-0001-9765-2096, semenov_vk@mail.ru
Иванова Наталья Борисовна — кандидат физико-математических наук, доцент, доцент, Ивановский государственный энергетический университет им. В.И. Ленина, Иваново, 153003, Российская Федерация, SC 57210732707, https://orcid.org/0000-0002-9519-6047, rgr_ivanova@rambler.ru

Authors

Vladimir K. Semenov — D.Sc., Full Professor, Ivanovo State Power University named after V. I. Lenin, Ivanovo, 153003, Russian Federation, https://orcid.org/0000-0001-9765-2096, semenov_vk@mail.ru

Natalia B. Ivanova — PhD, Associate Professor, Associate Professor, Ivanovo State Power University named after V. I. Lenin, Ivanovo, 153003, Russian Federation, SC 57210732707, https://orcid.org/0000-0002-9519-6047, rgr_ivanova@rambler.ru

Статья поступила в редакцию 15.11.2021
Одобрена после рецензирования 23.12.2021
Принята к печати 20.01.2022

Received 15.11.2021
Approved after reviewing 23.12.2021
Accepted 20.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-187-192
УДК 532.529

Моделирование импульсного истечения смеси воздуха и мелкодисперсного порошка, частично заполняющего выбросной канал

Дмитрий Викторович Садин

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация
sadin@yandex.ru✉, https://orcid.org/0000-0001-5335-4847

Аннотация

Предмет исследования. Рассмотрены закономерности импульсного истечения смеси воздуха и мелкодисперсного порошка, который частично заполняет выбросной цилиндрический канал в одномерном и двумерном случаях. **Метод.** Динамика газодисперсной среды описана в рамках эйлерового континуального подхода с различными скоростями и температурами газа и частиц порошка. В равновесном приближении построены аналитические автомодельные решения. Для численного решения задачи использован гибридный метод крупных частиц второго порядка точности по пространству и времени. Сопоставление точных автомодельных и численных решений подтвердили достоверность метода. **Основные результаты.** Истечение смеси газа высокого давления и частиц порошка носит ярко выраженный волновой характер, который связан с распадом начального разрыва, движением и преломлением волн на границе раздела сред внутри канала, а также отражением волн от его дна. Установлены характерные временные интервалы волнового процесса и соответствующие им распределения газодинамических величин. В зависимости от обобщенной автомодельной переменной давление, плотность и скорость смеси являются монотонными функциями, а профиль удельного (на единицу поперечного сечения) массового расхода имеет максимум в выходном сечении. Определены безразмерные параметры и удельный массовый расход двухфазной среды в выходном сечении выбросного канала. В случае ограниченного размером камеры высокого давления канала изучена двумерная физическая картина формирования и эволюции газодисперсной смеси. На начальном этапе истечения наблюдается «аномальная» группировка частиц порошка с образованием ударно-волновой структуры на дозвуковом режиме течения несущего газа. После выхода слоя порошка за пределы канала, истекающий из него чистый газ ускоряется до сверхзвуковой скорости, и развивается интенсивное вихревое движение в следе газодисперсной струи. **Практическая значимость.** Расчетные значения параметров позволяют обосновать достижимый уровень технических характеристик (скорости, массового расхода) потока рабочей газодисперсной среды импульсных порошковых устройств. Предложенная методика и полученные результаты являются основой принятия рациональных решений на ранних этапах проектирования и подготовки исходных данных по конструктивным и режимным параметрам для проведения испытаний прототипов импульсных порошковых технических устройств.

Ключевые слова

порошковая струйная технология, импульсное истечение, автомодельные и численные решения

Ссылка для цитирования: Садин Д.В. Моделирование импульсного истечения смеси воздуха и мелкодисперсного порошка, частично заполняющего выбросной канал // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 187–192. doi: 10.17586/2226-1494-2022-22-1-187-192

Simulation of the pulsed outflow of air and fine powder mixture, partially filling the discharge channel

Dmitry V. Sadin

Mozhaisky Military Space Academy, Saint Petersburg, 197198, Russian Federation
sadin@yandex.ru✉, https://orcid.org/0000-0001-5335-4847

Abstract

The paper studies the regularities of the pulsed outflow of air and fine powder mixture, which partially fills the ejection cylindrical channel, in both one-dimensional and two-dimensional formulations. The dynamics of a gas-dispersed

© Садин Д.В., 2022

medium are described in the framework of the Eulerian continuum approach with different velocities and temperatures of gas and powder particles. Analytical self-similar solutions are constructed in the equilibrium approximation. For the numerical solution of the problem, a hybrid large-particle method of the second order of accuracy in space and time is used. Comparison of exact self-similar and numerical solutions confirmed the reliability of the method. The outflow of the mixture of high-pressure gas and powder particles has a pronounced wave character, which is associated with the decomposition of the initial discontinuity, the movement, and refraction of waves at the interface of media inside the channel, as well as the reflection of waves from its bottom. The characteristic time intervals of the wave process and the corresponding distributions of gas-dynamics quantities are established. Depending on the generalized self-similar variable, the pressure, density, and velocity of the mixture are monotonic functions, and the profile of the specific (per unit cross-section) mass flow has a maximum in the critical section. Dimensionless parameters and specific mass flow of a two-phase medium in the outlet section of the discharge channel are determined. In the case of a channel limited by the size of the high-pressure chamber, a two-dimensional physical picture of the formation and evolution of a gas-dispersed mixture was studied. At the initial stage of the outflow, an “anomalous” grouping of powder particles is observed with the formation of a shock-wave structure in the subsonic mode of the carrier gas flow. After the powder layer leaves the channel, the pure gas flowing out of it accelerates to supersonic speed and an intense vortex motion develops in the wake of the gas-dispersed jet. The calculated values of the parameters allow us to justify the achievable level of technical characteristics (speed, mass flow rate) of the flow of the working gas-dispersed medium of pulsed powder devices. The proposed methodology and the results obtained are the basis for making rational decisions at the early stages of design and preparation of initial data on design and operating parameters for testing prototypes of pulsed powder technical devices.

Keywords

powder jet technology, pulse outflow, self-similar and numerical solutions

For citation: Sadin D.V. Simulation of the pulsed outflow of air and fine powder mixture, partially filling the discharge channel. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 187–192 (in Russian). doi: 10.17586/2226-1494-2022-22-1-187-192

Введение

Порошковые струйные технологии находят широкое применение в индустриальных приложениях, в системах пожаротушения, ликвидации аварийных ситуаций, как способ создания заградительных дисперсных образований от воздействия различных поражающих факторов, например ударных волн. Определяющей особенностью применения импульсных газодисперсных систем является волновой характер протекающих процессов.

Математическое моделирование — один из основных инструментов обоснования рациональных технических решений на ранних этапах проектирования и подготовки исходных данных по конструктивным и режимным параметрам для испытаний прототипов устройств. В работах [1–3] рассмотрены основополагающие положения и подходы математического моделирования динамики многофазных сред. Вследствие сложности изучаемых физических явлений основным теоретическим инструментом являются методы численного моделирования. Расчеты взаимодействия ударной волны с облаком частиц изучены в работах [4–6]. Работы [6–12] посвящены задачам истечения и разлета порошкообразных сред.

Важную роль в изучении волновых течений двухфазных сред занимают аналитические методы, основанные на автомодельных свойствах задач [13, 14]. Достоинства применения аналитических решений — детальность теоретического анализа и высокая достоверность решений в области их применимости.

Рассмотренные ранее [7–11] схемы импульсного истечения смеси газа повышенного давления и порошка, заполняющего выбросной канал во всем объеме, имеет некоторые недостатки. Первый из них связан с тем, что после окончания процесса истечения часть рабочей дисперсной среды остается в канале. Второй — при реализации схемы канала с полным его заполнением формируется узконаправленная газодисперсная струя. Такой характер течения не является рациональным с точки зрения покрытия площади, например, нейтрализующим порошком при ликвидации аварийных ситуаций.

Цель работы — аналитическое и численное исследования процесса импульсного истечения смеси воздуха и мелкодисперсного порошка, который частично заполняет цилиндрический канал.

Постановка задачи

Основные уравнения. Динамика смеси идеального газа и монодисперсных несжимаемых частиц формулируется в рамках эйлерова подхода в виде законов сохранения [1]:

$$\begin{cases} \dfrac{\partial \rho_i}{\partial t} + \nabla\cdot(\rho_i \mathbf{v}_i) = 0, \dfrac{\partial}{\partial t}(\rho_1 \mathbf{v}_1) + \nabla(\rho_1 \mathbf{v}_1 \mathbf{v}_1) + \alpha_1 \nabla p = -\mathbf{F}_\mu, \\ \dfrac{\partial}{\partial t}(\rho_2 \mathbf{v}_2) + \nabla(\rho_2 \mathbf{v}_2 \mathbf{v}_2) + \alpha_2 \nabla p = \mathbf{F}_\mu, \\ \dfrac{\partial}{\partial t}(\rho_2 e_2) + \nabla\cdot(\rho_2 e_2 \mathbf{v}_2) = Q_T, \\ \dfrac{\partial}{\partial t}(\rho_1 E_1 + \rho_2 E_2) + \nabla\cdot(\rho_1 E_1 \mathbf{v}_1 + \rho_2 E_2 \mathbf{v}_2) + \\ + \nabla\cdot[p(\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2)] = 0, \end{cases} \quad (1)$$

$$\rho_i = \rho_i^\circ \alpha_i \ (i = 1, 2), \ \rho = \rho_1 + \rho_2, \ \alpha_1 + \alpha_2 = 1,$$
$$E_i = e_i + \mathbf{v}_i^2/2,$$

где α_i — объемная доля; ρ_i — приведенная плотность; $\mathbf{v}_i$ — вектор скорости; E_i и e_i — полная и внутренняя

энергии единицы массы i-ой фазы; p — давление газа; $\mathbf{F}_\mu$ — вязкая составляющая силы межфазного взаимодействия; Q_T — мощность теплообмена между газом и частицами в единице объема; t — время; индекс «°» — означает истинные значения плотности газа или материала частиц.

Система уравнений (1) описывает взаимопроникающее континуальное движение несущего газа ($i = 1$) и дисперсной фазы ($i = 2$).

Замыкающие соотношения. Газ считается идеальным и калорически совершенным с уравнением состояния

$$p = (\gamma_1 - 1)\rho_1^\circ e_1,$$

где $e_1 = c_v T_1$; γ_1 — показатель адиабаты; c_v — удельная теплоемкость газа при постоянном объеме.

Частицы — сферические и несжимаемые с линейной зависимостью внутренней энергии от температуры $e_2 = c_2 T_2$, где c_2 — удельная теплоемкость частиц. Величины межфазного трения и теплообмена $\mathbf{F}_\mu$, Q_T рассчитываются на основе эмпирических соотношений в зависимости от чисел Рейнольдса Re_{12} и Нуссельта Nu_1 в широком диапазоне концентраций частиц [8]:

$$\mathbf{F}_\mu = (3/8)(\alpha_2/r)C_\mu(\mathrm{Re}_{12})\rho_1(\mathbf{v}_1 - \mathbf{v}_2)|\mathbf{v}_1 - \mathbf{v}_2|,$$
$$Q_T = (3/2)(\alpha_2/r^2)\lambda_1 \mathrm{Nu}_1(T_1 - T_2),$$

где C_μ — коэффициент межфазного трения; μ_1 — динамическая вязкость; r — радиус частицы.

Начальные и граничные условия. Изучена задача импульсного истечения смеси воздуха с показателем адиабаты $\gamma_1 = 1{,}4$ и мелкодисперсного порошка из цилиндрического канала (рис. 1). Выбросной канал (камера высокого давления) имеет следующие параметры: длина $L = 1$ м; радиус $R = 0{,}1$ м; давление $p_L = 10^6$ Па; температура фаз $T_{1L} = T_{2L} = 293$ К. Канал отделен мембраной в сечении $x = 0$ м от невозмущенной области $x \geq 0$ чистого газа $p_R = 10^5$ Па, при $T_{1R} = T_{2R} = 293$ К.

Порошок в насыпном состоянии с объемной долей $\alpha_2 = 0{,}5$ содержит несжимаемые сферические частицы диаметром d =1 мкм, плотностью $\rho_2^\circ = 2500$ кг/м^3 и занимает в начальный момент времени часть выбросного канала $-x_S < x < 0$, где $x_S = 0{,}25$ м. Краевые условия заданы в виде отражения на стенках и дне канала и свободного вытекания на внешних границах расчетной области при $x \geq 0$.

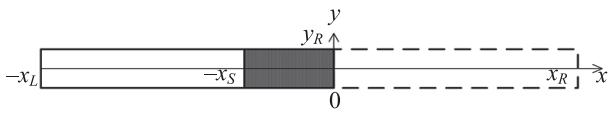


Рис. 1. Расчетная схема задачи. Координата дна канала $-x_L$ с радиусом y_R и выходным сечением $x = 0$, граница порошка $-x_S$, правая граница расчетной области x_R. Штриховой линией показано продолжение канала при одномерном расчете

Fig. 1. Calculation scheme of the problem. The coordinate of the channel bottom is $-x_L$ with the radius y_R and output section $x = 0$, the powder boundary is $-x_S$, the right boundary of the calculated area x_R. The dashed line shows the continuation of the channel in the one-dimensional calculation

Метод расчета. Численное моделирование выполнено гибридным методом крупных частиц с неявной безытерационной схемой расчета источников $\mathbf{F}_\mu$, Q_T [15] с использованием сбалансированного алгоритма второго порядка точности по пространству и времени [16]. Расчеты проведены на равномерной сетке в цилиндрической системе координат с разрешением 200 ячеек на длину L выбросного канала с числом Куранта 0,4.

Результаты моделирования и их обсуждение

После мгновенного удаления мембраны в сечении $x = 0$ происходит распад произвольного разрыва с образованием ударной волны в камере низкого давления при $x > 0$ и центрированной волны разрежения в слое газодисперсной смеси $-x_S < x < 0$. Данная конфигурация имеет точное равновесное решение [13, 15]. В области изменения автомодельной переменной $\xi = x/(a_S\alpha_{1S}t) \in (\xi_S, \xi_R)$, ограниченной левой $\xi_S = -1/\alpha_{1S}$ и правой ξ_R характеристиками, волна разрежения в газовзвеси описывается соотношениями:

$$\left(\frac{\alpha_{2S}\alpha_1}{\alpha_2\alpha_{1S}}\right)^\omega = \frac{\alpha_1 + \omega}{\alpha_1(1 - \omega\xi)},\ \omega = \frac{\gamma_S^* - 1}{2},$$
$$\gamma^* = 1 + (\zeta_1 R_1)/(\zeta_1 c_v + \zeta_2 c_2)\ , \qquad (2)$$

где $\zeta_i = \rho_i/\rho$ — массовые концентрации фаз; R_1 — газовая постоянная; γ^* — эффективный показатель адиабаты двухфазной среды; $M_S = v/a_S$ — безразмерная скорость смеси (число Маха).

На рис. 2 показаны нормированные автомодельные решения: для плотности ρ/ρ_S и удельного (на едини-

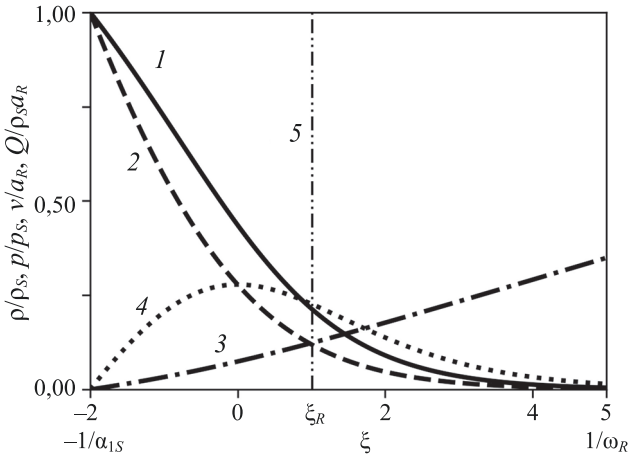


Рис. 2. Распределения относительных величин: плотности ρ/ρ_S (кривая *1*), давления p/p_S (кривая *2*), скорости v/a_R (кривая *3*), массового расхода $Q/(\rho_S a_S)$ (кривая *4*) смеси в зависимости от автомодельной переменной ξ; ξ_R — правая граница волны разрежения (прямая *5*); области: $-1/\alpha_{1S}$ — левая граница волны разрежения, $1/\omega_R$ — газодинамический предел истечения в вакуум

Fig. 2. Distributions of relative values of density ρ/ρ_S (curve *1*), pressure p/p_S (curve *2*), velocity v/a_R (curve *3*), and mass flow $Q/(\rho_S a_S)$ (curve *4*) of the mixture depending on the self-similar variable ξ; ξ_R is the right boundary of the rarefaction wave (line *5*), $-1/\alpha_{1S}$ — the left boundary of the rarefaction wave, $1/\omega_R$ — gas-dynamic limit of the outflow into vacuum

цу поперечного сечения) массового расхода $Q/(\rho_S a_S)$ смеси; давления p/p_S, отнесенные в слое к соответствующим начальным величинам, а также равновесной скорости смеси по отношению к скорости звука в чистом газе v/a_R. Диапазон изменения безразмерной переменной ξ задан от левого фронта волны разрежения $-1/\alpha_{1S} = -2$ до газодинамического предела истечения в вакуум $1/\omega_R = 5$. Область определения поставленной задачи ограничена примыканием правой характеристики к зоне постоянного течения $\xi_R = 1{,}008$ (рис. 2, прямая *5*).

Плотность смеси, давление и скорость являются монотонными функциями от ξ, а удельный массовый расход газодисперсной среды имеет максимум в критическом сечении $\xi = 0$. Это объясняется произведением величин убывающей плотности и возрастающей скорости среды, входящими в выражение массового расхода $Q = \rho v$. Критическое истечение порошкообразной среды для заданных физических свойств фаз достигается при превышении начального отношения давления в камерах высокого и низкого давлений $(p_S/p_R)_* = 3{,}832$, что выполнено в рассматриваемом случае $p_S/p_R = 10$.

Для практического применения технологии выброса порошка имеет большое значение информация о параметрах (давление, скорость, массовый расход) в выходном сечении, которая важна при анализе физической картины развития газодисперсной струи. В равновесном приближении из соотношений (2) при условии $\xi = 0$ (рассчитанные безразмерные параметры в выходном сечении имеют следующие значения: $\rho/\rho_S = 0{,}436$; $p/p_S = 0{,}278$; $v/a_R = 0{,}074$; $Q/(\rho_S a_S) = 0{,}278$).

На рис. 3 приведены профили относительных величин плотности смеси, давления и скорости газа для трех характерных моментов времени (для удобства представления продольная координата отсчитывается от дна канала и нормирована размером расчетной области x/H, где $H = x_L + x_R$). Фронт волны разрежения проходит в слое газодисперсной среды со скоростью — $a_S = \sqrt{\gamma^* p_S/(\rho_S \alpha_{1S})} = 39{,}94$ м/с. Через $t_1 = x_S/a_S = 6{,}259$ мс фронт приходит на границу раздела сред в сечении $x/H = 0{,}4$ (рис. 3, *a*, *d*). Для контроля точности приведены точные автомодельные (референтные) решения.

Затем формируется распад разрыва с двумя волнами разрежения. Через $t_2 = t_1 + (x_L - x_S)/a_L = 8{,}445$ мс левая волна разрежения достигает дна канала $x/H = 0$ и отражается от него (рис. 3, *b*, *e*). В характерный момент времени $t_3 = t_1 + 2(x_L - x_S)/a_L = 10{,}63$ мс отраженная от дна волна разрежения возвращается на границу раздела газа и взвеси. Поскольку для данного этапа решение не является автомодельным, то референтные профили параметров (черные сплошные линии) получены в данной работе численно на подробной сетке $L/2000$.

В случае ограниченного размером камеры высокого давления $-x_S < x < 0$ выбросного канала (рис. 1) формируется осесимметричная двумерная струя газовзвеси. Расчеты численного решения гибридным методом крупных частиц для различных моментов времени приведены на рис. 4. Продольная x/L и радиальная y/L координаты отнесены к длине выбросного канала L. Результаты представлены в виде распределения числа Маха несущего газа (в непрерывной цветовой шкале) и изолиний десятичного логарифма относительной плот-

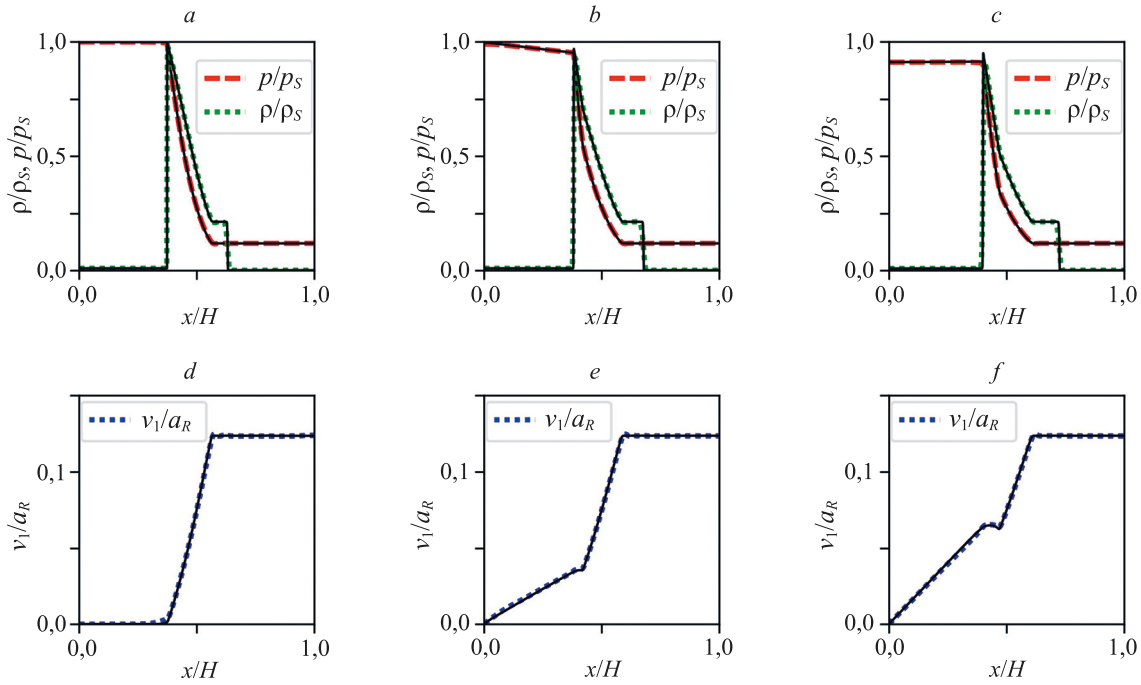


Рис. 3. Распределения расчетных параметров: относительного давления p/p_S (красные штриховые линии), плотности ρ/ρ_S смеси (зеленые пунктирные линии) и скорости v_1/a_R газа (синие пунктирные линии), а также референтные профили (черные сплошные линии) в моменты времени: 6,259 (*a*, *d*); 8,445 (*b*, *e*); 10,63 мс (*c*, *f*)

Fig. 3. Distributions of calculated relative pressures p/p_S (red dashed curves), density ρ/ρ_S of the mixture (green dotted curves) and gas velocity v_1/a_R (blue dotted lines), as well as reference profiles (solid curves) at time points: 6.259 (*a*, *d*); 8.445 (*b*, *e*); 10.63 ms (*c*, *f*)

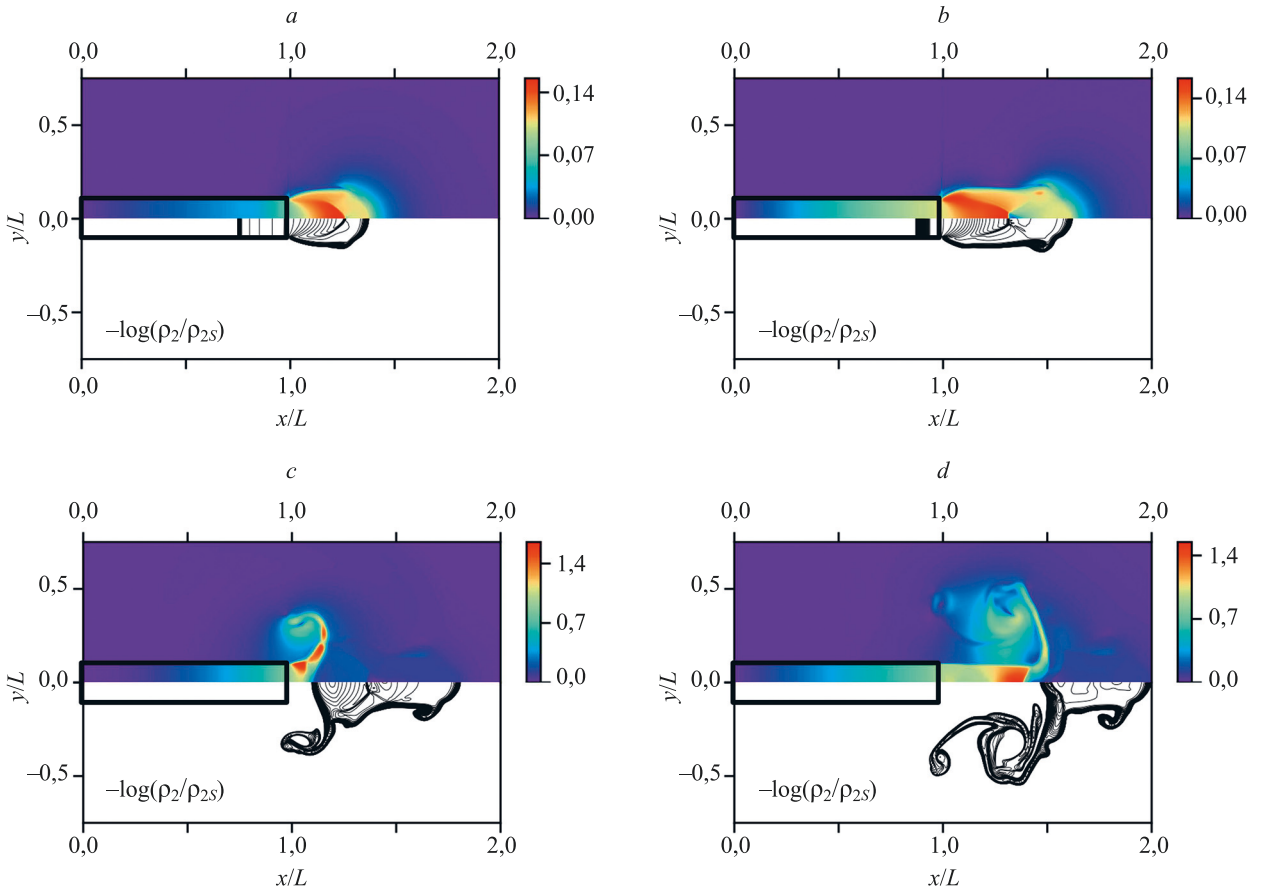


Рис. 4. Расчетные распределения числа Маха несущего газа (в непрерывной цветовой шкале) и относительной плотности дисперсной фазы (линии равных значений $-\log(\rho_2/\rho_{2S})$) в моменты времени: 8,445 (*a*); 15 (*b*); 20 (*c*); 25 мс (*d*)

Fig. 4. Calculated distributions of the Mach number of the carrier gas (in a continuous color scale) and the relative density of the dispersed phase (lines of equal values $-\log(\rho_2/\rho_{2S})$) at time points: 8.445 (*a*); 15 (*b*); 20 (*c*); 25 ms (*d*)

ности дисперсной фазы с обратным знаком $-\log(\rho_2/\rho_{2S})$ в диапазоне от 0 до 4.

Первые два момента времени (рис. 4, *a*, *b*) характеризуют начальный этап истечения. Левая граница раздела чистого газа и взвеси движется внутри канала. А за его срезом начинает формироваться ударно-волновая структура — «бочка». Отметим «аномальное» поведение частиц порошка, которые группируются в скачки уплотнения на дозвуковом по несущему газу режиме течения. Это неочевидное явление было обнаружено и экспериментально подтверждено в работах [9, 11]. Скорость смеси на срезе составляет менее 10 % от скорости звука в газовой фазе, а число Маха (отношение скорости газа к местной скорости звука) во всем поле течения не превышает 0,17 (верхние фрагменты (рис. 4, *a*, *b*)).

После выхода левой границы слоя порошка за пределы канала ударно-волновая структура потока разрушается (рис. 4, *c*). Истекающий из выбросного канала чистый газ в форме искривленного конуса ускоряется до сверхзвуковой скорости с образованием немонотонной структуры (верхний фрагмент (рис. 4, *c*)). Далее формируется интенсивное вихревое движение в следе газодисперсной струи с ее радиальным расширением (рис. 4, *d*).

Заключение

Методами аналитического и численного моделирования изучен процесс импульсного истечения смеси воздуха и мелкодисперсного порошка, частично заполняющего выбросной цилиндрический канал. Получены автомодельные распределения газодинамических параметров и массового расхода двухфазного потока в одномерном случае. Произведено сравнение точных автомодельных и численных решений, подтверждающих достоверность применяемого метода. Исследована двумерная физическая картина формирования импульсной двухфазной струи. На начальном этапе наблюдается «аномальная» группировка частиц порошка с образованием ударно-волновой структуры на дозвуковом режиме течения несущего газа. После выхода слоя порошка за пределы канала развивается интенсивное кольцевое движение в хвосте газодисперсной струи. Приведены параметры в выходном сечении выбросного канала, позволяющие судить о достижимом уровне характеристик (скорости, массовом расходе) потока рабочей газодисперсной среды импульсных порошковых устройств.

Литература

1. Нигматулин Р.И. Динамика многофазных сред. В 2 ч. М.: Наука, 1987.
2. Gidaspow D. Multiphase Flow and Fluidization. Continuum and Kinetic Theory Descriptions. San Diego, USA: Academic Press, 1994. 467 p.
3. Кутушев А.Г. Математическое моделирование волновых процессов в аэродисперсных и порошкообразных средах. СПб.: Недра, 2003. 284 с.
4. Дэвис С.Л., Диттман Т.Б., Якобс Дж.Б., Дон В.С. Дисперсия облака частиц в ударной волне. Влияние формы, угла поворота и геометрических параметров облака на динамику потока и дисперсию // Прикладная механика и техническая физика. 2013. Т. 54. № 6. С. 45–59.
5. Bulat P.V., Volkov K.N., Ilyina T.Y. Interaction of a shock wave with a cloud of particles // IEJME — Mathematics education. 2016. V. 11. N 8. P. 2949–2962.
6. Садин Д.В., Давидчук В.А. Взаимодействие плоской ударной волны с областями различной формы и плотности в мелкодисперсной газовзвеси // Инженерно-физический журнал. 2020. Т. 93. № 2. С. 489–498.
7. Любарский С.Д., Иванов А.С. Движение сжатой двухфазной среды насыпной плотности при внезапном расширении // Физика горения и взрыва. 1989. № 3. С. 78–81.
8. Ивандаев А.И., Кутушев А.Г., Рудаков Д.А. Численное исследование метания слоя порошка сжатым газом // Физика горения и взрыва. 1995. № 4. С. 63–70.
9. Садин Д.В., Гузенков В.О., Любарский С.Д. Численное исследование структуры нестационарной двухфазной тонкодисперсной струи // Прикладная механика и техническая физика. 2005. Т. 46. № 2. С. 91–97.
10. Нигматулин Р.И., Губайдуллин Д.А., Тукмаков Д.А. Ударно-волновой разлет газовзвесей // Доклады Академии наук. 2016. Т. 466. № 4. С. 418–421. https://doi.org/10.7868/S0869565216040101
11. Садин Д.В., Любарский С.Д., Гравченко Ю.А. Особенности недорасширенной импульсной импактной газодисперсной струи с высокой концентрацией частиц // Журнал технической физики. 2017. Т. 87. № 1. С. 22–26. https://doi.org/10.21883/JTF.2017.01.44013.1809
12. Широкова Е.Н. Численное исследование разлета смеси газа и частиц с осевой симметрией // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 4. С. 606–612. https://doi.org/10.17586/2226-1494-2021-21-4-606-612
13. Иванов А.С., Козлов В.В., Садин Д.В. Нестационарное истечение двухфазной дисперсной среды из цилиндрического канала конечных размеров в атмосферу // Известия РАН. Механика жидкости и газа. 1996. № 3. С. 60–66.
14. Садин Д.В. Приложение гибридного метода крупных частиц к расчету взаимодействия ударной волны со слоем газовзвеси // Компьютерные исследования и моделирование. 2020. Т. 12. № 6. С. 1323–1338. https://doi.org/10.20537/2076-7633-2020-12-6-1323-1338
15. Садин Д.В. TVD-схема для жестких задач волновой динамики гетерогенных сред негиперболического неконсервативного типа // Журнал вычислительной математики и математической физики. 2016. Т. 56. № 12. С. 2098–2109. https://doi.org/10.7868/S0044466916120152
16. Садин Д.В. Сбалансированный алгоритм гибридного метода крупных частиц и его проверка на некоторых тестовых задачах // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 5. С. 785–790. https://doi.org/10.17586/2226-1494-2021-21-5-785-790

Автор

Садин Дмитрий Викторович — доктор технических наук, профессор, профессор, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 6602924618, https://orcid.org/0000-0001-5335-4847, sadin@yandex.ru

Статья поступила в редакцию 22.12.2021
Одобрена после рецензирования 30.12.2021
Принята к печати 21.01.2022

References

1. Nigmatulin R.I. *Dynamics of Multiphase Media*. In 2 vol. USA, CRC Press, 1990.
2. Gidaspow D. *Multiphase Flow and Fluidization. Continuum and Kinetic Theory Descriptions*. San Diego, USA Academic Press, 1994, 467 p.
3. Kutushev A.G. *Mathematical Modeling of Wave Processes in aero–dispersed and powdered media*. St. Petersburg, Nedra Publ., 2003, 284 p. (in Russian).
4. Davis S.L., Dittmann T.B., Jacobs G.B., Don W.S. Dispersion of a cloud of particles by a moving shock: Effects of the shape, angle of rotation, and aspect ratio. *Journal of Applied Mechanics and Technical Physics*, 2013, vol. 54, no. 6, pp. 900–912. https://doi.org/10.1134/S0021894413060059
5. Bulat P.V., Volkov K.N., Ilyina T.Y. Interaction of a shock wave with a cloud of particles. *IEJME — Mathematics education*, 2016, vol. 11, no. 8, pp. 2949-2962.
6. Sadin D.V., Davidchuk V.A. Interaction of a plane shock wave with regions of varying shape and density in a finely divided gas suspension. *Journal of Engineering Physics and Thermophysics*, 2020, vol. 93, no. 2, pp. 474–483. https://doi.org/10.1007/s10891-020-02143-2
7. Lyubarskii S.D., Ivanov A.S. Motion of a compressed two-phase medium of bulk density upon sudden expansion. *Combustion, Explosion, and Shock Waves*, 1989, vol. 25, no. 3, pp. 335–337. https://doi.org/10.1007/BF00788810
8. Ivandaev A.I., Kutushev A.G., Rudakov D.A. Numerical investigation of throwing a powder layer by a compressed gas. *Combustion, Explosion, and Shock Waves*, 1995, vol. 31, no. 4, pp. 459–465. https://doi.org/10.1007/BF00789367.
9. Sadin D.V., Guzenkov V.O., Lyubarskii S.D. Numerical study of the structure of a finely disperse unsteady two-phase jet. *Journal of Applied Mechanics and Technical Physics*, 2005, vol. 46, no. 2, pp. 224–229. https://doi.org/10.1007/s10808-005-0040-6
10. Nigmatulin R.I., Gubaidullin D.A., Tukmakov D.A. Shock wave dispersion of gas–particle mixtures. *Doklady Physics*, 2016, vol. 61, no. 2, pp. 70–73. https://doi.org/10.1134/S1028335816020038
11. Sadin D.V., Lyubarskii S.D., Gravchenko Y.A. Features of an underexpanded pulsed impact gas-dispersed jet with a high particle concentration. *Technical Physics*, 2017, vol. 62, no. 1, pp. 18–23. https://doi.org/10.1134/S1063784217010194
12. Shirokova E.N. A numerical study of the expansion of a gas-particles mixture with axial symmetry. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 4, pp. 606–612 (in Russian). https://doi.org/10.17586/2226-1494-2021-21-4-606-612.
13. Ivanov A.S., Kozlov V.V., Sadin D.V. Unsteady flow of a two-phase disperse medium from a cylindrical channel of finite dimensions into the atmosphere. *Fluid Dynamics*, 1996, vol. 31, no. 3, pp. 386–391. https://doi.org/10.1007/BF02030221
14. Sadin D.V. Application of a hybrid large-particle method to the computation of the interaction of a shock wave with a gas suspension layer. *Computer Research and Modeling*, 2020, vol. 12, no. 6, pp. 1323–1338. (in Russian). https://doi.org/10.20537/2076-7633-2020-12-6-1323-1338
15. Sadin D.V. TVD scheme for stiff problems of wave dynamics of heterogeneous media of nonhyperbolic nonconservative type. *Computational Mathematics and Mathematical Physics*, 2016, vol. 56, no. 12, pp. 2068–2078. https://doi.org/10.1134/S0965542516120137
16. Sadin D.V. A balanced algorithm of the hybrid large-particle method and its verification on some test problems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 785–790 (in Russian). https://doi.org/10.17586/2226-1494-2021-21-5-785-790

Author

Dmitry V. Sadin — D.Sc., Full Professor, Mozhaisky Military Space Academy, Saint Petersburg, 197198, Russian Federation, SC 6602924618, https://orcid.org/0000-0001-5335-4847, sadin@yandex.ru

Received 22.12.2021
Approved after reviewing 30.12.2021
Accepted 21.01.2022

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-193-205
УДК 532.529

Использование векторизованных структур данных при реализации вычислительных алгоритмов решения задач механики сплошной среды

Никита Александрович Брыков[1], Константин Николаевич Волков[2]✉, Владислав Николаевич Емельянов[3]

[1,2,3] Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, Санкт-Петербург, 190005, Российская Федерация

[1] brykovna@yandex.ru, https://orcid.org/0000-0002-3568-7522
[2] k.volkov@kingston.ac.uk✉, https://orcid.org/0000-0001-6055-2323
[3] vlademelyanov@gmail.com, https://orcid.org/0000-0002-7039-9948

Аннотация

Предмет исследования. В работе выполнено исследование возможностей, предоставляемых новыми информационными технологиями, средствами объектно-ориентированного программирования и современными операционными системами, для решения краевых задач механики сплошной среды, описываемых дифференциальными уравнениями в частных производных. Для дискретизации основных уравнений применены методы конечных разностей и конечных объемов. Рассмотрена реализация данных методов с векторизованными сеточными структурами, включая обращение к внутренним и граничным ячейкам сетки, а также особенности реализации алгоритмов в особых точках расчетной области. **Метод.** Для решения краевых задач, описываемых дифференциальными уравнениями в частных производных, разработан подход к построению векторизованных алгоритмов и рассмотрены особенности их программной реализации в пакете MATLAB с использованием средств пользовательского программирования. Векторизация в таких задачах, для исключения вложенных циклов, обеспечена за счет организации данных и использования векторизованных операций обращения к сеточным узлам. Разработанные алгоритмы, с одной стороны, широко используют функции MATLAB, предназначенные для обработки векторов и разреженных матриц, а с другой, отличаются высокой эффективностью и скоростью счета. **Основные результаты.** Проведено численное решение ряда задач механики сплошной среды, связанных с расчетами напряжений в твердом теле, а также поля скорости и температуры в потоке вязкой несжимаемой жидкости. Показаны особенности дискретизации основных уравнений и реализации конечно-разностных и конечно-объемных алгоритмов. **Практическая значимость.** Применение системы MATLAB открыло широкие возможности по формализации и реализации конечно-разностных и конечно-объемных методов численного решения краевых задач механики сплошной среды. Несмотря на то, что возможности разработанных алгоритмов иллюстрируются на примере решения достаточно простых задач, они допускают сравнительно простое обобщение на более сложные задачи, например решение уравнений Эйлера и Навье–Стокса. В рамках работы подготовлены расчетные модули, реализованные при помощи средств пользовательского программирования, которые расширяют возможности пакета MATLAB и ориентированы на решение задач механики сплошной среды.

Ключевые слова

математическое моделирование, векторизация, численный алгоритм, механика сплошной среды

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 21-19-00657, https://rscf.ru/project/21-19-00657/.

Ссылка для цитирования: Брыков Н.А., Волков К.Н., Емельянов В.Н. Использование векторизованных структур данных при реализации вычислительных алгоритмов решения задач механики сплошной среды // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 193–205. doi: 10.17586/2226-1494-2022-22-1-193-205

© Брыков Н.А., Волков К.Н., Емельянов В.Н., 2022

Vectorized numerical algorithms for the solution of continuum mechanics problems

Nikita A. Brykov[1], Konstantin N. Volkov[2]✉, Vladislav N. Emelyanov[3]

[1,2,3] Baltic State Technical University "VOENMEH" named after D.F. Ustinov, Saint Petersburg, 190005, Russian Federation

[1] brykovna@yandex.ru, https://orcid.org/0000-0002-3568-7522
[2] k.volkov@kingston.ac.uk✉, https://orcid.org/0000-0001-6055-2323
[3] vlademelyanov@gmail.com, https://orcid.org/0000-0002-7039-9948

Abstract

The aim of the work is to study the possibilities provided by new information technologies, object-oriented programming tools and modern operating systems for solving boundary value problems of continuum mechanics described by partial differential equations. To discretize the basic equations, we applied the method of finite differences and finite volumes, which are widely used to solve problems in the mechanics of liquids and gases. The paper considers the implementation of the finite difference methods and the finite volume method with vectorized grid structures, including access to the inner and boundary cells of the grid, as well as the features of the implementation of algorithms at singular points of the computational domain. To solve boundary value problems described by partial differential equations, we developed an approach to the construction of vectorized algorithms and considered the features of their software implementation in the MATLAB package. Vectorization in such tasks, excluding nested loops, is ensured by appropriate data organization and the use of vectorized operations. On the one hand, the developed algorithms widely use MATLAB functions designed for processing vectors and sparse matrices, and on the other hand, they are distinguished by high efficiency and computation speed, comparable to those of programs written in C/C++. The main results imply the numerical solution of a number of problems in continuum mechanics associated with the calculation of stresses in a separate body and the calculation of the field of velocity and temperature in the flow of a viscous incompressible fluid. The features of discretization of the basic equations and the implementation of the corresponding finite-difference and finite-volume algorithms are shown. The use of the MATLAB system opens up new possibilities for the formalization and implementation of finite-difference and finite-volume methods for the numerical solution of boundary value problems in continuum mechanics. Despite the fact that the capabilities of the developed algorithms are illustrated by the example of solving fairly simple problems, they admit a relatively simple generalization to more complex problems, for example, solving the Euler equations and Navier–Stokes equations. As part of the work, computational modules were prepared with user programming tools that expand the capabilities of the package and are focused on solving problems in continuum mechanics.

Keywords

mathematical modeling, vectorization, numerical algorithm, mechanics of continuum medium

Acknowledgements

The research was supported by the Russian Science Foundation (project No. 21-19-00657), https://rscf.ru/en/project/21-19-00657/.

For citation: Brykov N.A., Volkov K.N., Emelyanov V.N. Vectorized numerical algorithms for the solution of continuum mechanics problems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 193–205 (in Russian). doi: 10.17586/2226-1494-2022-22-1-193-205

Введение

Современные системы вычислительного моделирования представлены достаточно сложными программными комплексами, оформленными в виде пакетов прикладных программ. Исследование и моделирование мультифизичных процессов в технологических системах неразрывно связано с использованием подобных технологий. Важными оказываются вопросы создания моделей и алгоритмов для выполнения многовариантного анализа различных конструкций на этапах поисковых исследований, в которых важную роль играет оперативность получения результата. При этом значительная часть вычислительных ресурсов при решении задач, описываемых дифференциальными уравнениями в частных производных, используется на решение систем линейных алгебраических уравнений (СЛАУ). В частности, в отечественном пакете вычислительной газовой динамики ЛОГОС на решение СЛАУ расходуется от 40 % времени расчета задачи [1, 2]. Создание алгоритмов, повышающих эффективность решения СЛАУ — одно из основных направлений повышения производительности пакетов прикладных программ [3–5].

Векторизация вычислений — наиболее перспективный метод оптимизации, наряду с реализацией параллельных вычислительных алгоритмов. Векторизация представляет собой один из способов выполнения параллельных вычислений, при котором программа определенным образом модифицируется для выполнения нескольких однотипных операций одновременно. Такой подход потенциально приводит к значительному ускорению однотипных вычислений над большими массивами данных [6].

Значительные возможности в рамках данного направления реализованы в среде MATLAB и родственных программах, в которых поддерживаются операции векторизации. Векторизация — стандартная часть эффективного использования MATLAB. При этом в MATLAB реализован интерпретируемый язык, в связи с чем циклические конструкции, типичные для реализации многих сеточных методов, характеризуются весьма медленным исполнением. Для ускорения векторно-матричных вычислений в MATLAB реализованы функции для работы с векторами и матрицами, оптимизированные для конкретной архитектуры процессора. Современные языки программирования (например,

C++) также позволяют создавать векторизованные алгоритмы, эффективность которых проявляется при соответствующем аппаратном обеспечении.

Несмотря на многочисленные примеры решения задач механики сплошной среды при помощи средств MATLAB [7–18], включая генерацию сеток, реализацию конечно-элементных подходов, возможности векторизации вычислительных алгоритмов, связанные с решением дифференциальных уравнений в частных производных конечно-разностными и конечно-объемными методами, не получили достаточного внимания в научных работах. Применение неявных разностных схем в существенной степени усложняет реализацию алгоритмов. Для реализации численных подходов применяются в основном циклические конструкции, не позволяющие в полной мере воспользоваться векторизованными функциями низкого уровня, имеющимися в MATLAB. Векторизованные алгоритмы, предназначенные для решения задач сверхзвуковой газовой динамики конечно-объемными методами, основанными на решении задачи распада произвольного разрыва, обсуждаются в работах [6, 7].

В состав MATLAB входит пакет PDE Toolbox (Partial Differential Equation), который содержит средства для решения нестационарных дифференциальных уравнений в частных производных второго порядка при помощи метода конечных элементов. Команды и графический интерфейс пакета PDE Toolbox могут быть использованы для математического моделирования широкого класса инженерных и научных приложений (сопротивление материалов, электротехника, тепломассоперенос, механика жидкости и газа). Однако при решении задач механики жидкости и газа конечно-элементный подход получил менее широкое распространение, чем метод конечных разностей или метод конечного объема. Во многих случаях требуются подходы, учитывающие структуру исходной системы уравнений, особенности формулировки граничных условий, сложную пространственно-временную зависимость правой части уравнения и вид расчетной области.

Применение Simulink, представляющей собой среду динамического междисциплинарного моделирования сложных технических систем и инструмент для модельно-ориентированного проектирования, также не решает проблемы решения задач, описываемых дифференциальными уравнениями в частных производных. Соответствующие численные методы по-прежнему нуждаются в реализации при помощи имеющихся в распоряжении пользователя функций.

В данной работе рассмотрен один из возможных подходов к построению векторизованных алгоритмов для решения задач механики сплошной среды, описываемых дифференциальными уравнениями в частных производных различного вида. Реализация предложенного подхода позволяет заменить циклические конструкции операциями подстановок групп соответствующих сеточных узлов. Возможности разработанного подхода демонстрируются на примере решения ряда задач механики сплошной среды.

Векторизованные сеточные структуры

Тип дискретизации и используемый сеточный шаблон в значительной степени определяют свойства и особенности разностной схемы и вычислительной процедуры в целом (устойчивость, точность, время счета и другие показатели).

Сетка. Разностная сетка, покрывающая расчетную область, имеет структуру, схожую со структурой двумерного массива. Кроме избираемых данных в текущем узле, необходимо обеспечить адресацию к данным в соседних узлах, входящих в разностный шаблон. Традиционный способ адресации к ячейкам или узлам сетки состоит в указании двух индексов i и j, определяющих строку и столбец. Узел сетки с индексами (i, j) имеет соседей, к которым легко адресоваться, давая соответствующие приращения индексов. Указанный принцип адресации работает как для обозначения узлов сетки в методе конечных разностей, так и для обозначения ячеек в методе конечных объемов, причем такая адресация возможна и в физическом, и в вычислительном пространствах при использовании криволинейных сеток.

При программировании вычислительных задач работа с сеточными структурами обычно осуществляется при помощи перебора узлов сетки и реализации алгоритма на основе вложенных циклов. Данный подход подразумевает вычисление индексного выражения и занесение результатов вычислений в двумерный массив.

Векторизация позволяет работать со множеством данных как с единой вычислительной структурой, что не только делает компактным запись вычислительного алгоритма, избегая вложенных циклов, но и повышает эффективность вычислений.

Адресация к внутренним ячейкам. В задачах механики сплошной среды двумерным структурам данных соответствуют поля величин и координаты сеточных узлов при решении задачи в области прямоугольной формы или в криволинейной области, которую можно отобразить на вычислительный прямоугольник при помощи преобразования координат [19, 20].

Рассмотрим адресацию к внутренним ячейкам на примере L-образной расчетной сетки (рис. 1), где для наглядности принято малое число узлов. Адресация к данным в узлах состоит из указания двух индексов, определяющих строку и столбец.

При помощи одномерного массива можно указать определенный набор узлов расчетной области. Обозначим через **C** вектор индексов всех внутренних узлов. При организации вычислительного алгоритма возникает необходимость конечно-разностного представления дифференциальных операторов, фигурирующих в уравнениях и граничных условиях. В этом случае необходимо обеспечить адресацию к данным в соседних узлах, входящих в вычислительный шаблон. С помощью вектора центральных ячеек найдем подмножество узлов, которые лежат: левее центральных ячеек — **L**, правее — **R**, ниже — **D**, выше — **U**, выше левее — **UL**, выше правее — **UR**, ниже левее — **DL** и ниже правее — **DR** (рис. 1). Номера соответствующих ячеек определим по следующим формулам:

$$\mathbf{L} = \mathbf{C} - N,\ \mathbf{R} = \mathbf{C} + N,\ \mathbf{D} = \mathbf{C} - 1,\ \mathbf{U} = \mathbf{C} + 1,$$
$$\mathbf{UL} = \mathbf{C} + 1 - N,\ \mathbf{UR} = \mathbf{C} + 1 + N,\ \mathbf{DL} = \mathbf{C} - 1 - N,$$
$$\mathbf{DR} = \mathbf{C} + 1 + N,$$

где N — количество узлов в столбце расчетной области.

Адресация к граничным ячейкам. В вычислительных алгоритмах, реализованных при помощи метода конечных разностей на разнесенной сетке или метода конечных объемов, обычно выделяются фиктивные граничные ячейки, образующие вместе с внутренними ячейками расширенную расчетную область. Способ адресации к фиктивным граничным ячейкам, вводимым для постановки граничных условий, показан на рис. 2, где p_1–p_6 — значения переменной в угловых точках. Построены индексные вектора **BL**, **BU**, **BR**, **BD** и для рассматриваемого случая **BRU** и **BUR**, соответствующие левой, нижней, верхним и правым границам области. Способ формирования этих векторов такой же, как и индексных векторов внутренних ячеек (угловые ячейки не используются).

Построение разностной схемы. Рассмотрим вопрос о векторизованном вычислении потоков на границах расчетных ячеек для двумерных задач, реализованных на структурированных сетках. В качестве объекта, которому соотносится поток, выбрана грань, разделяющая две соседние ячейки. Для вычисления потоков формируются два множества индексных векторов, адресуемых к ячейкам. Один вектор **pCFL** определяет множество всех ячеек, расположенных слева от боковых граней. Другой — **pCFR** обеспечивает операцию подстановочной адресации для множества всех правых ячеек. Индексный вектор множества правых ячеек определяется как **pCFR** = **pCFL** + m. Векторная операция вычисления потоков задается соотношением

$$\mathbf{F}(\mathbf{pC}) = \mathbf{F}[\mathbf{U}(\mathbf{pCFR}), \mathbf{U}(\mathbf{pCFL})].$$

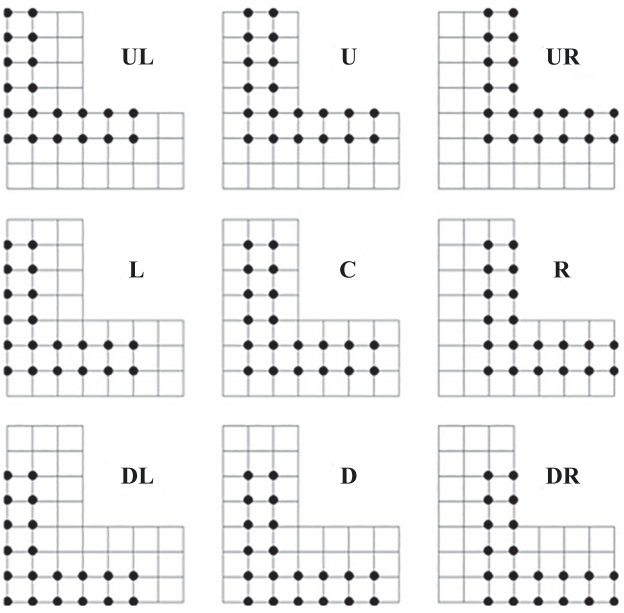


Рис. 1. Адресация к внутренним ячейкам
Fig. 1. Internal cells addressing

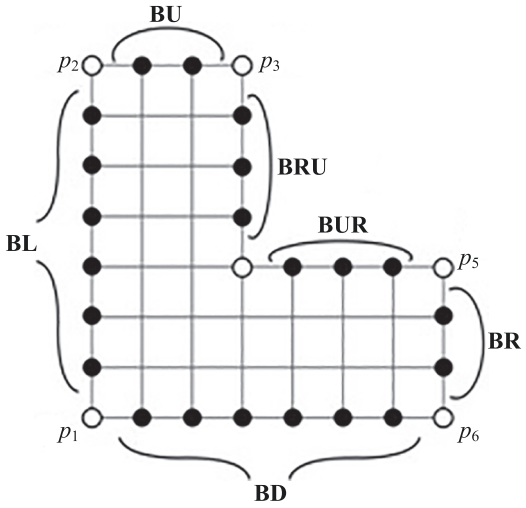


Рис. 2. Адресация к граничным ячейкам
Fig. 2. Boundary cells addressing

При такой организации данных можно записать векторизованные соотношения для каждой ячейки расчетной области. Для всех расчетных ячеек, определяемых индексным вектором **pC**, потоки через их левые грани определяются индексным вектором **pC** – m, а через правые грани — **pC**. Величина m представляет собой количество ячеек в столбце расширенной области ячеек и определяет сдвиг при переходе от двумерной матричной структуры к линейному размещению элементов.

Используя индексные представления сеточных структур, запишем одним векторизованным оператором разностную схему рассматриваемых уравнений. При этом нет необходимости использовать циклы, присущие традиционным разностным схемам, которые заменяются заполнением матрицы коэффициентов разностных уравнений, в которой каждая строка представляет сеточное уравнение и имеет высокую размерность, но при этом является существенно разреженной. Обращение этой матрицы дает решение задачи на следующем временном слое.

Решение системы разностных уравнений. Совокупность конечно-разностных уравнений и граничных условий, записанных в дискретной форме, представляет собой связанную СЛАУ для определения сеточных значений искомой функции в узлах расчетной области:

$$\mathbf{AU} = \mathbf{B}, \qquad (1)$$

где **A** — матрица коэффициентов линейной системы уравнений; **U** — вектор-столбец неизвестных; **B** — вектор-столбец правых частей. Разреженная структура матрицы **A** в (1) позволяет реализовать компактные методы хранения ее компонент и использовать эффективные в вычислительном плане методы обращения.

Последовательно пронумеруем все точки расчетной области и занесем их номера в вектор-столбец **pG**. Точки, не входящие в расчетную область, тоже нумеруются, но их номера не заносятся в вектор **pG**. Для формирования вектора подстановок используем функцию **k** = find(**X**), которая возвращает индексы ненулевых элементов матрицы **X** (если таких элементов нет,

то возвращается пустой вектор). При этом матрицу **X** рассмотрим, как вектор-столбец, образованный объединением столбцов исходной матрицы. Используя вектор подстановок **pG**, в матрицу **G** заносятся порядковые номера всех точек расчетной области. Элементы матрицы **G**, которые соответствуют точкам, не входящим в пределы области заданной формы, имеют нулевые значения.

Применяя аналогичный прием, сформируем вектора, содержащие номера граничных узлов. При этом функция **k** = find(«условие») возвращает индексы элементов, удовлетворяющие заданному условию. Используем сквозную нумерацию граничных узлов, включая точки, находящиеся за пределами расчетной области. В результате при помощи простого слияния получим вектор подстановок **pB**, содержащий номера всех граничных точек.

На следующем шаге построим матрицу **Gp**, содержащую номера только внутренних узлов. Матрица **Gp** строится в два приема. Сначала, используя матрицу **G** и вектор подстановок **pB**, создадим промежуточную матрицу, содержащую номера только граничных узлов, остальные ее элементы считаем равными нулю. Окончательную структуру матрицы **Gp** получим путем простого вычитания и выполнения оператора присваивания.

Сформируем вектор подстановок **pP**, содержащий номера внутренних узлов расчетной области. Для его построения используем функцию **k** = find(**X**). Для построения разностного шаблона, следуя векторному принципу построения вычислительного алгоритма, создаются массивы, содержащие номера узлов, используемые для построения разностного шаблона.

Перед тем как сформировать матрицу коэффициентов системы разностных уравнений, определим ее структуру. Для этого введем два целочисленных вектора **indI** и **indJ**, задающие позиции элементов матрицы коэффициентов, а также вещественный вектор **valS**, определяющий числовые значения элементов этой матрицы.

Построим массивы **indI**, **indJ** и **valS** для граничных узлов, индексы которых хранятся в векторе подстановок **pB**. Используя операцию слияния векторов и алгоритм векторных подстановок, сформируем массивы **indI**, **indJ** и **valS** для внутренних узлов при соблюдении последовательности обхода узлов разностного шаблона. Элементы вектора **valS** находятся в соответствии с используемой разностной схемой.

Вектор правых частей **B** построим отдельно для внутренних и граничных узлов. Элементы **B**(**pP**), связанные с внутренними узлами, приравняем к правой части решаемого уравнения, которую для простоты примем равной единице. Для задания значений элементам **B**(**pB**), относящихся к граничным узлам, используем соответствующие граничные условия.

Сформированные данные применим для построения матрицы СЛАУ и ее решения прямым методом. Функция sparse(**i**, **j**, **s**) сформирует матрицу в соответствии с правилами записи разреженных матриц, принятыми в системе MATLAB. При вызове этой функции строки массива [**i j s**] используются для формирования разреженной матрицы размерности $m \times n$, где m = max(**i**) и n = max(**j**). Вектора **i** и **j** задают позиции элементов и являются целочисленными, а вектор **s** определяет числовое значение элемента матрицы. При формировании разреженной матрицы все строки вида [**i j** 0] из описания удаляются. Длина вектора **s** совпадает с количеством ненулевых элементов разреженной матрицы.

Примеры применения векторизованных алгоритмов

Возможности разработанных средств численного моделирования показаны на примере решения ряда задач механики сплошной среды.

Задача о концентрации напряжений в трещинах конструкции. При эксплуатации различных конструкций появляются дефекты в виде трещин, которые образуются в результате изменения прочностных характеристик, вызванных со старением материала. Появление трещин вследствие действия на их внутреннюю поверхность различных факторов (например, высокого давления) приводит к возникновению концентрации напряжений в материале в области вершины трещины и ее разрастанию. Образование трещин в приповерхностном слое может привести к разрушению конструкции. Моделирование напряженно-деформированного состояния, возникающего в результате действия неравномерно распределенного давления внутри дефекта конструкции, представляет практический интерес для прогнозирования нештатных ситуаций.

Рассмотрим двумерную статическую задачу теории упругости, связанную с определением напряжений в плоской пластине с продольным разрезом. На пластину действуют внешние нагрузки, приложенные к двум противоположным сторонам так, что нормальная компонента напряжения поперек пластины равна нулю. В этом случае пластина находится в плоском напряженном состоянии. Решение задачи о плоском напряженном состоянии в координатах x и y сводится к определению напряжений σ_x, σ_y, (нормальные компоненты) и τ_{xy} (тангенциальные компоненты), которые удовлетворяют условиям равновесия, граничным условиям и условиям совместности для напряжений.

Задача о плоском напряженном состоянии в координатах x и y может быть сведена к решению бигармонического уравнения для функции напряжения Эри [21]

$$\frac{\partial^4 \varphi}{\partial x^4} + 2\frac{\partial^4 \varphi}{\partial x^2 \partial y^2} + \frac{\partial^4 \varphi}{\partial y^4} = 0. \qquad (2)$$

Функция φ вводится таким образом, чтобы $\tau_{xy} = -\partial^2\varphi/\partial x \partial y$.

Уравнение (2) позволяет определить распределение напряжений в сплошном теле. Известно решение классической задачи для трещины в безграничной среде при ее равномерном растяжении. Для плоской задачи теории упругости выражения для полей напряжений рассмотрены в работе [21]. В том случае, когда среда находится в произвольном плоском напряженном состоянии, а берега трещины нагружены напряжениями

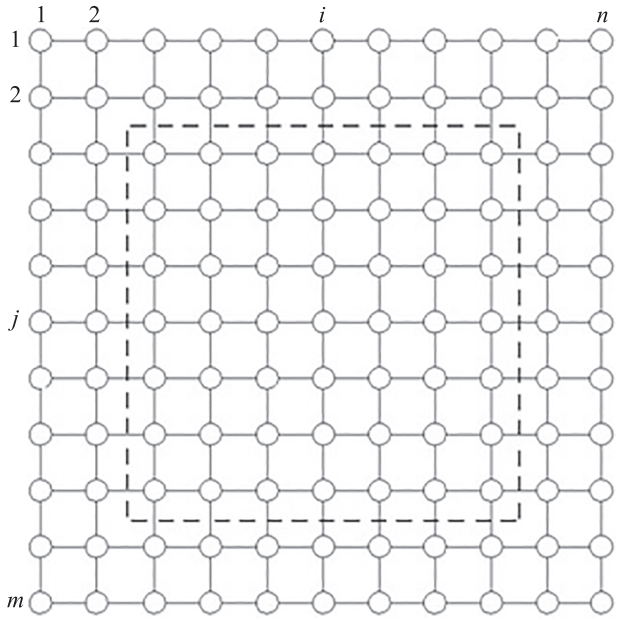


Рис. 3. Расчетная область и сетка для пластины без разреза

Fig. 3. Computational domain and grid for a plate without cut

неравномерного распределения, задача решается численно. Для численного решения уравнения (2) воспользуемся методом конечных разностей.

Рассмотрим пластину без центрального разреза (рис. 3).

Внутренние узлы. Составим разностный шаблон для внутренних узлов сетки, лежащих в области, ограниченной штриховой линией. Выберем любую точку внутри выделенной области и примем ее за центральную. Соответствующим образом обозначим соседние точки (рис. 4, *a*). Тогда запишем в виде конечных разностей производные, входящие в уравнение (2), используя обозначения для узлов *RR*, *R*, *C*, *L*, *LL*, *DD*, *D*, *U*, *UU*, *UL*, *UR*, *DL*, *DR*. В результате получим

$$\frac{\partial^4\varphi}{\partial x^4} = \frac{1}{h^4}(\varphi_{RR} - 4\varphi_R + 6\varphi_C - 4\varphi_L + \varphi_{LL});$$

$$\frac{\partial^4\varphi}{\partial y^4} = \frac{1}{h^4}(\varphi_{DD} - 4\varphi_D + 6\varphi_C - 4\varphi_U + \varphi_{UU});$$

$$\frac{\partial^4\varphi}{\partial x^2\partial y^2} = \frac{1}{h^4}(\varphi_{UL} - 2\varphi_U + \varphi_{UR} - 2\varphi_R + 4\varphi_C - 2\varphi_L + \varphi_{DL} - 2\varphi_D + \varphi_{DR}),$$

где h — шаг сетки.

В разностном виде уравнение (2) имеет вид

$$\frac{\partial^4\varphi}{\partial x^4} + 2\frac{\partial^4\varphi}{\partial x^2\partial y^2} + \frac{\partial^4\varphi}{\partial y^4} =$$

$$= \frac{1}{h^4}(\varphi_{RR} - 8\varphi_R + 20\varphi_C - 8\varphi_L + \varphi_{LL} + \varphi_{DD} - 8\varphi_D - 8\varphi_U + \varphi_{UU} + 2\varphi_{UL} + 2\varphi_{UR} + 2\varphi_{DL} + 2\varphi_{DR}).$$

Разностный шаблон для внутренних узлов сетки показан на рис. 4, *b*.

Граничные узлы. В узлах сетки, лежащих вне обозначенной области на рис. 3, зададим граничные условия. Первое граничное условие на левой и правой границах – равенство нулю нормальных напряжений, которое запишем в виде:

$$\sigma_x = \frac{\partial^2\varphi}{\partial y^2} = \frac{1}{h^2}(\varphi_D - 2\varphi_C + \varphi_U) = 0.$$

На верхней и нижней границах зададим напряжение, растягивающее пластину. Запишем для узлов на этих границах разностное представление производной:

$$\sigma_y = \frac{\partial^2\varphi}{\partial x^2} = \frac{1}{h^2}(\varphi_R - 2\varphi_C + \varphi_L).$$

Второе граничное условие — равенство нулю касательных напряжений имеет вид

$$\tau_{xy} = -\frac{\partial^2\varphi}{\partial x\partial y} = 0.$$

Составим разностный шаблон с учетом данного граничного условия (рис. 5, *a*). Выберем на нижней границе точку и примем ее за центральную. Для пред-

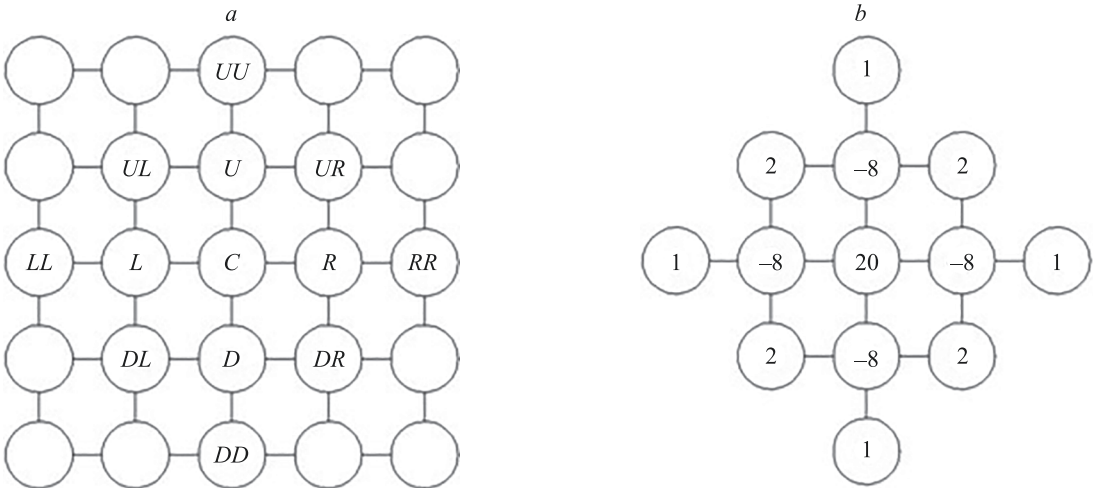


Рис. 4. Внутренние узлы (*a*) и коэффициенты разностного шаблона (*b*) сетки

Fig. 4. Internal grid nodes (*a*) and coefficients of the finite-difference template (*b*)

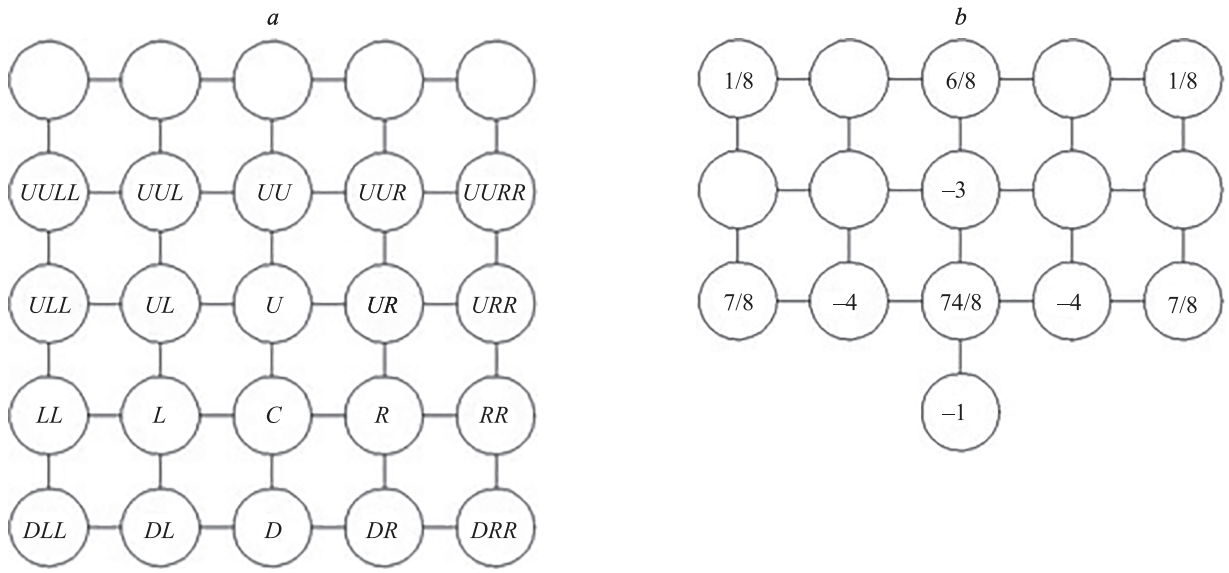


Рис. 5. Граничные узлы (*a*) и коэффициенты разностного шаблона (*b*) сетки

Fig. 5. Boundary grid nodes (*a*) and coefficients of the finite-difference template (*b*)

ставления производных, входящих в уравнение (2), в граничных узлах использованы односторонние конечно-разностные соотношения. Тогда уравнение примет вид

$$\frac{\partial^4\varphi}{\partial x^4}+2\frac{\partial^4\varphi}{\partial x^2\partial y^2}+\frac{\partial^4\varphi}{\partial y^4}=$$
$$=\frac{1}{h^4}(-\varphi_D+\frac{74}{8}\varphi_C-3\varphi_U+\frac{6}{8}\varphi_{UU}+\frac{7}{8}\varphi_{LL}+\frac{1}{8}\varphi_{UULL}+$$
$$+\frac{7}{8}\varphi_{RR}+\frac{1}{8}\varphi_{UURR}-4\varphi_L-4\varphi_R).$$

Разностный шаблон на граничных узлах сетки показан на рис. 5, *b*.

Угловые точки. Разностный шаблон для приугловых точек рассмотрим на примере правой нижней приугловой точки (рис. 6, *a*). Используя односторонние конечно-разностные формулы и граничные условия, запишем уравнение (2) в виде:

$$\frac{\partial^4\varphi}{\partial x^4}+2\frac{\partial^4\varphi}{\partial x^2\partial y^2}+\frac{\partial^4\varphi}{\partial y^4}=$$
$$=\frac{1}{h^4}(-\frac{63}{8}\varphi_C+\varphi_D+6\varphi_U-\frac{33}{8}\varphi_{UU}+\varphi_{UUU}+\varphi_L+6\varphi_R-$$
$$-\frac{33}{8}\varphi_{RR}+\frac{1}{8}\varphi_{UURR}+\varphi_{RRR}).$$

Разностный шаблон для угловых точек приведен на рис. 6, *b*. В крайних угловых точках задано определенное значение функции Эри (в данном случае нулевое). Это значение произвольно, и решение задачи о нахождении распределения напряжений в пластине не зависит от него.

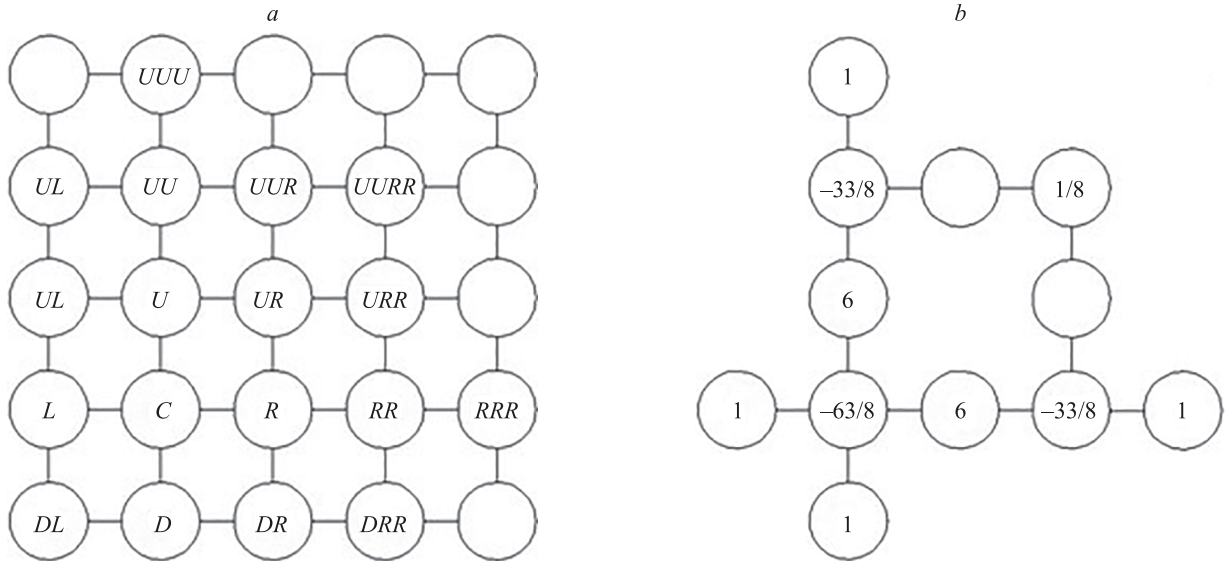


Рис. 6. Угловые узлы (*a*) и коэффициенты разностного шаблона (*b*) сетки

Fig. 6. Corner grid nodes (*a*) and coefficients of the finite-difference template (*b*)

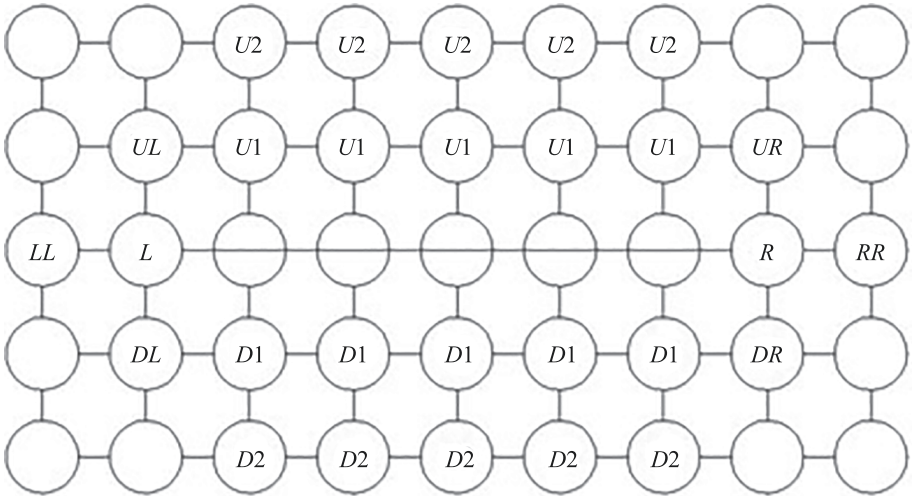


Рис. 7. Расчетная область и сетка для пластины с разрезом
Fig. 7. Computational domain and grid for a plate with cut

Поперечный разрез. Введем в рассмотрение поперечный разрез в пластине, расположенный параллельно нагруженным сторонам (рис. 7). Верхний берег разреза разместим в массиве основных узлов сетки, а для нижнего берега выделим дополнительный массив. Наличие разреза приводит к возникновению точек в его окрестности, для которых необходимо построить особые разностные шаблоны. Выделим следующие точки в окрестности массива: $U1$ и $D1$ — точки, непосредственно прилегающие к верхнему и нижнему берегам разреза; $U2$ и $D2$ — внешние точки разреза; L и R — левые и правые точки разреза; LL и RR — крайние левые и правые точки разреза; UL, UR, DL, DR — угловые точки разреза.

Разностный шаблон для точек $U2$, $D2$, UL, UR, DL, DR аналогичен для внутренних точек при условии, что для точек, относящихся к нижнему берегу разреза, значения для функции Эри выбраны из дополнительного массива. Учтем данное условие при формировании индексных массивов.

Составим разностный шаблон для левых и правых узлов (рис. 8, *a*). Уравнение (2) преобразуем к виду

$$\frac{\partial^4\varphi}{\partial x^4} + 2\frac{\partial^4\varphi}{\partial x^2\partial y^2} + \frac{\partial^4\varphi}{\partial y^4} =$$

$$= \frac{1}{h^4}(\varphi_{DD} + \varphi_{UU} - 8\varphi_D + \frac{25}{12}\varphi_C - 8\varphi_U + 2\varphi_{UL} + 2\varphi_{UR} +$$

$$+ \frac{92}{12}\varphi_R - \frac{42}{12}\varphi_{RR} + 2\varphi_{DL} + 2\varphi_{DR} - \frac{4}{12}\varphi_{RRR} + \frac{1}{12}\varphi_{RRRR}).$$

Приведем разностный шаблон для узлов сетки в зависимости от разреза (рис. 8, *b*).

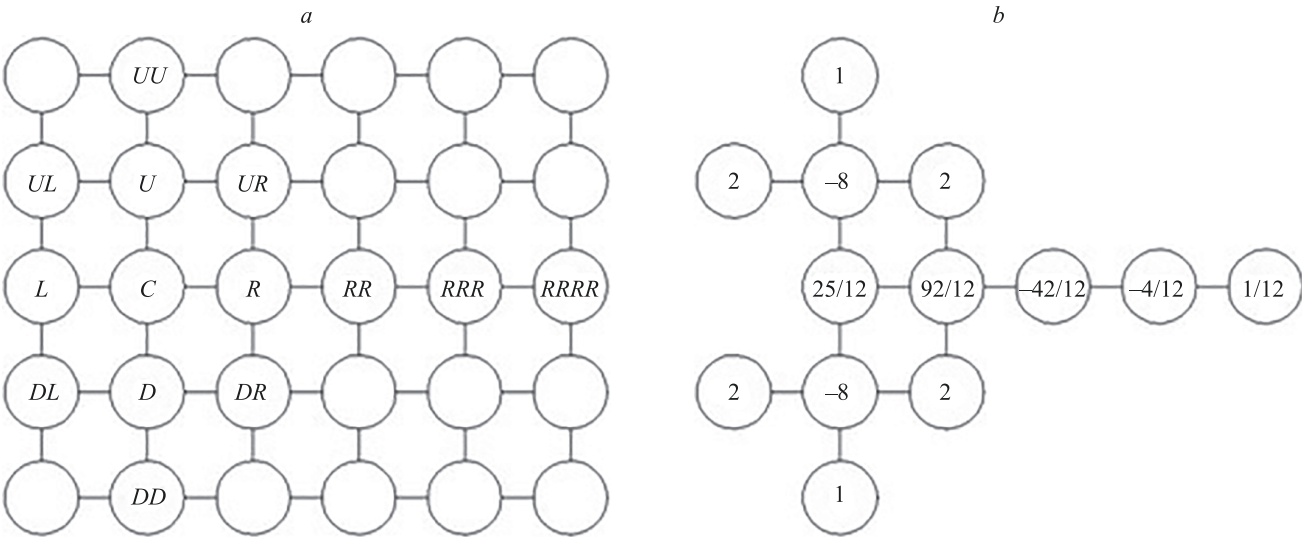


Рис. 8. Правые узлы от разреза (*a*) и коэффициенты разностного шаблона сетки (*b*)
Fig. 8. Right grid nodes from cut (*a*) and coefficients of the finite-difference template (*b*)

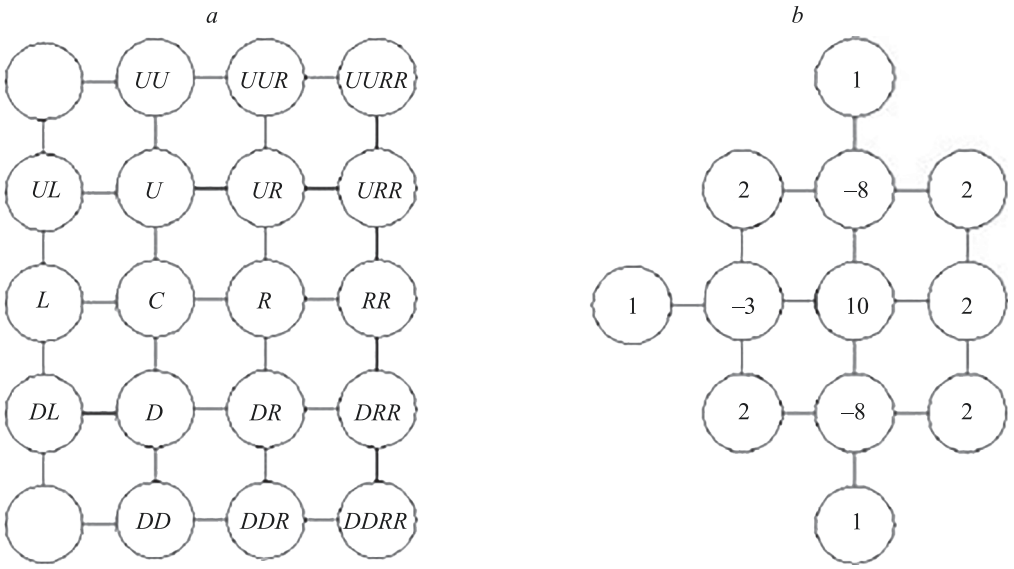


Рис. 9. Крайние правые узлы сетки от поперечного разреза (*a*) и коэффициенты разностного шаблона (*b*)

Fig. 9. Last in a row right grid nodes from transverse cut (*a*) and coefficients of the finite-difference template (*b*)

Рассмотрим разностный шаблон для крайних левых и крайних правых узлов. В частности, в узле 4*R* (рис. 9, *a*), уравнение (2) примет вид

$$\frac{\partial^4\varphi}{\partial x^4} + 2\frac{\partial^4\varphi}{\partial x^2\partial y^2} + \frac{\partial^4\varphi}{\partial y^4} =$$

$$= \frac{1}{h^4}(-4\varphi_{RR} + 2\varphi_R + 10\varphi_C - 3\varphi_L + \varphi_{DD} - 8\varphi_D - 8\varphi_U +$$

$$+ \varphi_{UU} + 2\varphi_{UL} + 2\varphi_{UR} + 2\varphi_{DL} + 2\varphi_{DR}).$$

Разностный шаблон для крайних правых узлов сетки показан на рис. 9, *b*.

Зададим условия равенства нулю: на верхнем и нижнем берегах разреза для нормального напряжения и для разностного шаблона (точки *U*1 и *D*1) для касательных напряжений.

Система разностных уравнений. Запишем разностную аппроксимацию для каждой точки, тогда дифференциальное уравнение можно заменить системой алгебраических уравнений (1). Для нахождения вектора неизвестных умножим (1) справа на матрицу, обратную **A**, и получим $\mathbf{U} = \mathbf{A}^{-1}\mathbf{B}$.

В качестве входных данных используем следующие значения: размер пластины 1 × 1 м^2; длина центрального разреза — 0,2 м. Нагрузим пластину постоянным напряжением по длине краев, а также зададим квадратичный закон распределения напряжения. Применим квадратичное распределение при решении задачи без центрального разреза для проверки работоспособности алгоритма.

На рис. 10, *a* приведено распределение напряжения на краях пластины, которое задано при решении задачи с пластиной без центрального разреза (изображение симметрично относительно оси ординат), поэтому показана половина графика). В результате расчета получим распределение нормальных напряжений в пластине (рис. 10, *b*).

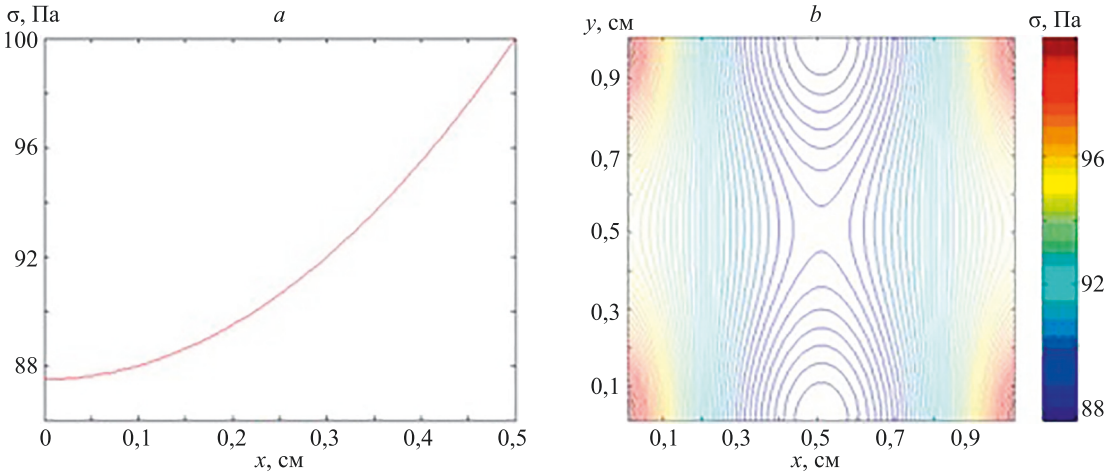


Рис. 10. Распределение напряжения на краях пластины (*a*) и изолинии нормальных напряжений (*b*)

Fig. 10. Stress distributions on plate edges (*a*) and contours of normal stresses (*b*)

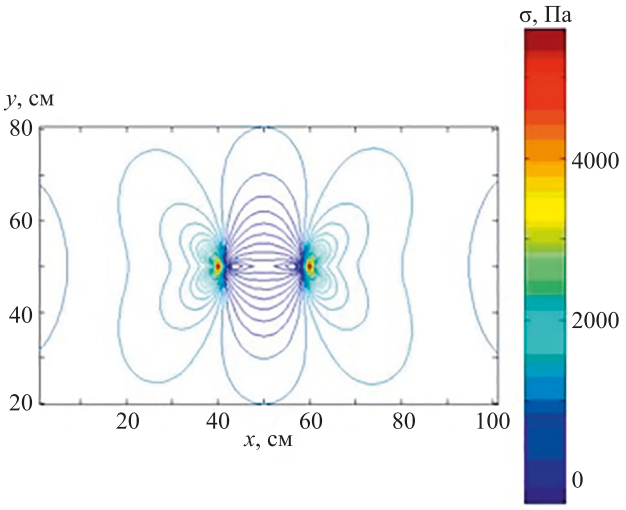


Рис. 11. Изолинии нормальных напряжений
Fig. 11. Contours of normal stresses

Данный метод можно применить к известной задаче о нагружении пластины с разрезом. Учет разреза и определение напряжений в угловых точках удобно реализовать с помощью векторизованного алгоритма. На рис. 11 показан результат решения, где максимум нормальных напряжений приходится на угловые точки разреза. В данном случае к пластине применена постоянная нагрузка.

Задача о свободной конвекции с внутренним источником тепла. Процесс конвекции определяется основными законами сохранения массы, количества движения и энергии. Особенности задачи конвекции проявляются в конкретизации сил и потоков, входящих в основные уравнения системы, а также в особенностях постановки граничных условий. В общем случае для однокомпонентной среды свободная конвекция описывается системой уравнений вязкой несжимаемой жидкости. Используя допущения модели Буссинеска о свободной конвекции, решение задачи свободноконвективного течения в переменных «функция тока ψ — вихрь скорости ω» имеет вид

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}=-\omega; \tag{3}$$

$$\frac{\partial\omega}{\partial t}+\frac{\partial\psi}{\partial y}\frac{\partial\omega}{\partial x}-\frac{\partial\psi}{\partial x}\frac{\partial\omega}{\partial y}=v\left(\frac{\partial^2\omega}{\partial x^2}+\frac{\partial^2\omega}{\partial y^2}\right)-\beta g\frac{\partial^2 T}{\partial x^2}; \tag{4}$$

$$\frac{\partial T}{\partial t}+\frac{\partial\psi}{\partial y}\frac{\partial T}{\partial x}-\frac{\partial\psi}{\partial x}\frac{\partial T}{\partial y}=a\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right)+\frac{q}{\rho c}, \tag{5}$$

где t — время; ρ — плотность; v — кинематическая вязкость; β — коэффициент объемного расширения; g — ускорение свободного падения; T — температура; a — температуропроводность; q — внутренний источник тепла; c — теплоемкость.

Уравнения (3)–(5) записаны для случая несжимаемой жидкости (ρ = const), что является следствием предположения Буссинеска о том, что изменение плотности учитывается только в выражении для гравитационной силы. Особенность системы — наличие внутреннего источника тепла q, влияние которого определяет не только параметры конвективного течения, но и изменение физических характеристик среды (зависимость вязкости от температуры).

Задача решается методом конечных разностей, используя возможности построения векторизованных алгоритмов и работой с разреженными матрицами в среде MATLAB. В качестве примера построим разностную схему для уравнения функции тока (2). Используя центрированные конечные разности 2-го порядка, получим

$$\psi_L+\psi_R+\frac{h_x^2}{h_y^2}(\psi_U+\psi_D)-2\left(1+\frac{h_x^2}{h_y^2}\right)\psi_C-\omega_C=0,$$

где h_x и h_y — шаги сетки в координатных направлениях x и y.

На рис. 12 представлены результаты моделирования свободной конвекции с внутренним источником тепла. На рис. 12, *а* показана структура матрицы коэффициентов размером 20 × 20, которая имеет 325 нулевых элементов, на рис. 12, *b* — поле температуры.

Задача о тепловом состоянии пластины. Рассмотрим задачу оценки нестационарного теплового состояния L-образной пластины при ее локальном нагреве. Часть энергии, которая поступает в определенную область пластины, за счет теплопередачи передается в соседние ячейки, а часть может излучаться от задней стенки области. Математическая модель сводится к уравнению двумерной нестационарной теплопроводности с добавлением в правую часть двух источниковых членов, один из которых учитывает энергоподвод, а другой — потери системой тепла за счет излучения

$$\frac{\partial T}{\partial t}=a\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right)+\frac{q}{\rho c\delta}-\frac{\sigma\varepsilon}{\rho c\delta}T^4, \tag{6}$$

где δ — толщина пластины; σ — постоянная Стефана–Больцмана; ε — степень черноты пластины.

В векторизованном алгоритме установим множества: всех центральных узлов сетки; левых, правых, нижних и верхних узлов; граничных узлов. Для дискретизации уравнения (6) по времени применим неявную схему:

$$\frac{T_C^{n+1}-T_C^n}{\tau}=a\left(\frac{T_L^{n+1}-2T_C^{n+1}+T_R^{n+1}}{h_x^2}\right)+$$

$$+a\left(\frac{T_U^{n+1}-2T_C^{n+1}+T_D^{n+1}}{h_y^2}\right)+\frac{q}{\rho c\delta}-\frac{\sigma\varepsilon}{\rho c\delta}(T^{(n+1)})^4,$$

где n — индекс по времени.

Во многих случаях в нестационарных системах в условиях интенсивного нагрева зависимостью теплопроводности от температуры пренебречь нельзя. В этом случае добавим фактор нелинейности. Разностная схема для уравнения (6) имеет вид

$$(\rho c)_C\frac{T_C^{n+1}-T_C^n}{\tau}=$$

$$=\left[\frac{\lambda_R+\lambda_C}{2}\frac{T_R^{n+1}-T_C^{n+1}}{h_x^2}-\frac{\lambda_C+\lambda_L}{2}\frac{T_C^{n+1}-T_L^{n+1}}{h_x^2}\right]+$$

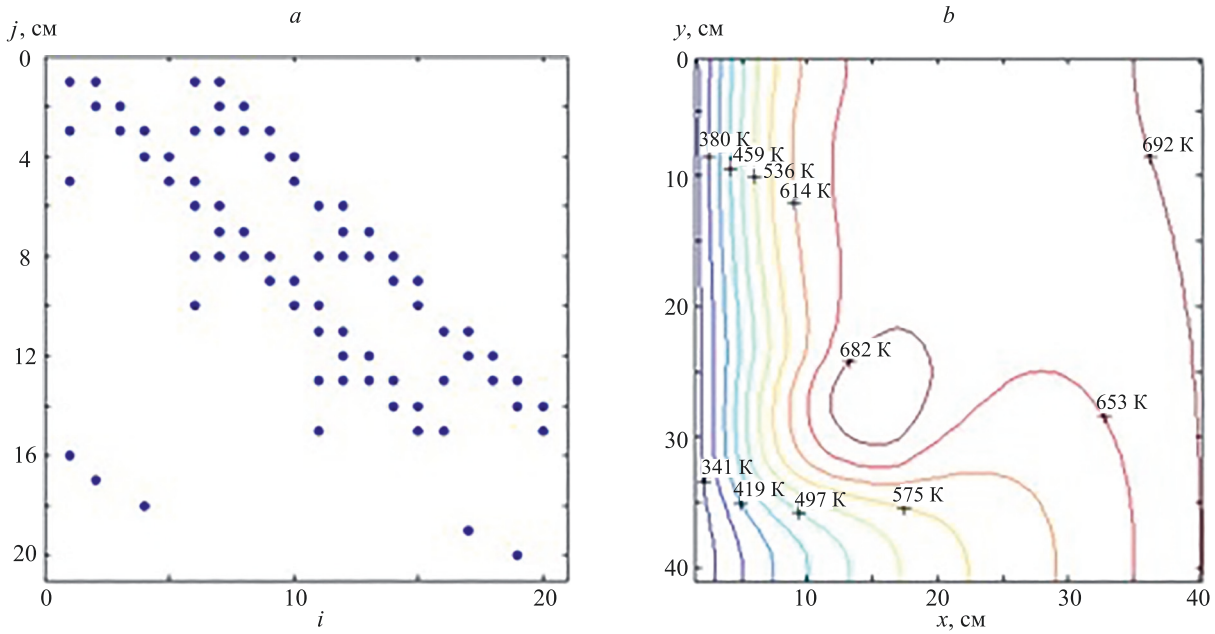


Рис. 12. Матрица коэффициентов (*a*) и поле температуры (*b*)

Fig. 12. Matrix of coefficients (*a*) and contours of temperature (*b*)

$$+\left[\frac{\lambda_U+\lambda_C}{2}\frac{T_U^{n+1}-T_C^{n+1}}{h_y^2}-\frac{\lambda_C+\lambda_D}{2}\frac{T_C^{n+1}-T_D^{n+1}}{h_y^2}\right]+$$
$$+\frac{q}{\delta}-\frac{\sigma\varepsilon}{\delta}(T^{(n+1)})^4,$$

где λ — коэффициент теплопроводности.

В неявную разностную схему добавим излучение членом правой части, имеющим сомножителем температуру в четвертой степени. Неявное включение данного члена вводит еще один фактор существенной нелинейности. Проведем линеаризацию этого члена по Ньютону, получим следующее представление этого соотношения

$$(T^{(n+1)})^4 = (T^n)^4 + \frac{\partial T^4}{\partial t}\tau = (T^n)^4 + \frac{\partial T^4}{\partial t}\frac{\partial T}{\partial t}\tau =$$
$$= (T^n)^4 + 4(T^n)^3(T^{n+1}-T^n) = 4(T^n)^3T^{n+1} - 3(T^n)^4.$$

На рис. 13 представлен результат решения уравнения в фиксированный безразмерный момент времени.

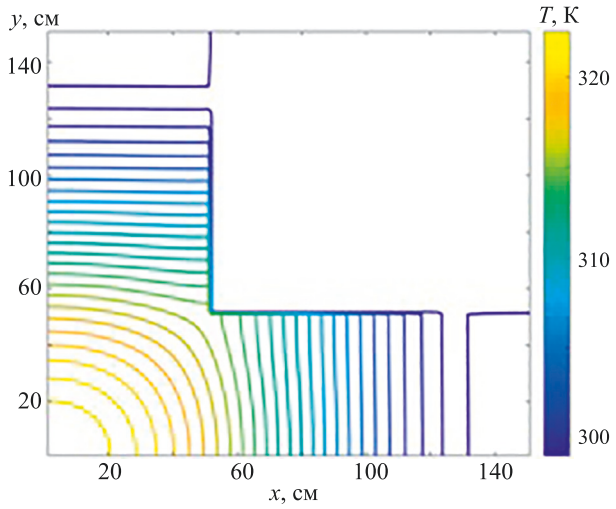


Рис. 13. Температурное поле системы в безразмерный момент времени 180

Fig. 13. Contours of temperature at dimensionless time 180

Заключение

Разработан подход к построению и реализации векторизованных алгоритмов решения задач механики сплошной среды. В качестве конкретных примеров рассмотрены задачи о: свободной конвекции с внутренним источником тепла; концентрации напряжений в трещинах конструкции; тепловом состоянии пластины.

Разработанные алгоритмы применимы для решения более сложных задач механики сплошной среды, к числу которых относятся задачи построения криволинейных сеток для областей сложной геометрической конфигурации, а также механики жидкости и газа, описываемые уравнениями Эйлера и Навье–Стокса.

Созданные программные средства могут быть привлечены для поддержки научных исследований в различных областях, где применяются методы математического моделирования.

Литература

1. Козелков А.С., Дерюгин Ю.Н., Зеленский Д.К., Глазунов В.А., Голубев А.А., Денисова О.В., Лашкин С.В., Жучков Р.Н., Тарасова Н.В., Сизова М.А. Многофункциональный пакет программ ЛОГОС для расчета задач гидродинамики и тепломассопереноса на суперЭВМ. Базовые технологии и алгоритмы // Труды XII Международного семинара «Супервычисления и математическое моделирование», 11–15 октября 2010 г., Саров. Саров, 2010. С. 215–230.
2. Волков К.Н., Дерюгин Ю.Н., Емельянов В.Н., Карпенко А.Г., Козелков А.С., Смирнов П.Г., Тетерина И.В. Реализация параллельных вычислений на графических процессорах в пакете вычислительной газовой динамики ЛОГОС // Вычислительные методы и программирование. 2013. Т. 14. № 1. С. 334–342.
3. Emelyanov V.N., Brykov N.A., Efremov A.V. Application of vectorized algorithms for solving problems of continuum mechanics // Journal of Physics: Conference Series. 2021. V. 2094. P. 022047. https://doi.org/10.1088/1742-6596/2094/2/022047
4. Емельянов В.Н. Механика сплошной среды: теория напряжений и основные модели. М.: Юрайт, 2018. 162 с.
5. Емельянов В.Н. Численные методы: введение в теорию разностных схем. М.: Изд-во Юрайт, 2018. 188 с.
6. Гергель В.П. Теория и практика параллельных вычислений. М.: БИНОМ, 2010. 423 с.
7. Волков К.Н., Емельянов В.Н. Применение пакета MATLAB для решения краевых задач механики жидкости и газа // Труды всероссийской научной конференции «Проектирование научных и инженерных приложений в среде MATLAB» (25–26 мая 2004, Москва, Россия). М.: Изд-во ИПУ РАН, 2004. С. 92–115.
8. Волков К.Н., Емельянов В.Н. Реализация векторизованных конечно-разностных алгоритмов решения краевых задач механики жидкости и газа в пакете MATLAB // Вычислительные методы и программирование. 2004. Т. 5. № 3. С. 13–29.
9. Alberty J., Carstensen C., Funken S.A. Remarks around 50 lines of Matlab: short finite element implementation // Numerical Algorithms. 1999. V. 20. N 1-2. P. 117–137. https://doi.org/10.1023/A:1019155918070
10. Persson P.-O., Strang G. A simple mesh generation in MATLAB // SIAM Review. 2004. V. 46. N 2. P. 329–345. https://doi.org/10.1137/S0036144503429121
11. Koko J. Vectorized MATLAB codes for linear two-dimensional elasticity // Scientific Programming. 2007. V. 15. N 3. P. 157–172. https://doi.org/10.1155/2007/838942
12. Funken S., Praetorius D., Wissgott P. Efficient implementation of adaptive P1-FEM in MATLAB // Computational Methods in Applied Mathematics. 2011. V. 11. N 4. P. 460–490. https://doi.org/10.2478/cmam-2011-0026
13. Rahman T., Valdman J. Fast MATLAB assembly of FEM matrices in 2D and 3D: nodal elements // Applied Mathematics and Computation. 2013. V. 219. N 13. P. 7151–7158. https://doi.org/10.1016/j.amc.2011.08.043
14. Anjam I., Valdman J. Fast MATLAB assembly of FEM matrices in 2D and 3D: edge elements // Applied Mathematics and Computation. 2015. V. 267. P. 252–263. https://doi.org/10.1016/j.amc.2015.03.105
15. Koko J. Fast MATLAB assembly of fem matrices in 2D and 3D using cell-array approach // International Journal of Modeling, Simulation, and Scientific Computing. 2016. V. 7. N 2. P. 1650010. https://doi.org/10.1142/S1793962316500100
16. Walker S.W. FELICITY: a MATLAB/C++ toolbox for developing finite element methods and simulation modeling // SIAM Journal on Scientific Computing. 2018. V. 40. N 2. P. 234–257. https://doi.org/10.1137/17M1128745
17. Čermák M., Sysala S., Valdman J. Efficient and flexible MATLAB implementation of 2D and 3D elastoplastic problems // Applied Mathematics and Computation. 2019. V. 355. P. 595–614. https://doi.org/10.1016/j.amc.2019.02.054
18. Tsega E.G. A finite volume solution of unsteady incompressible Navier–Stokes equations using MATLAB // Numerical and Computational Methods in Sciences and Engineering. 2019. V. 1. N 3. P. 117–131. http://dx.doi.org/10.18576/ncmsel/010302
19. Годунов С.К., Забродин А.В., Иванов М.Я., Крайко А.Н., Прокопов Г.П. Численное решение многомерных задач газовой динамики. М.: Наука, 1976. 400 с.

References

1. Kozelkov A.S., Deriugin Iu.N., Zelenskii D.K., Glazunov V.A., Golubev A.A., Denisova O.V., Lashkin S.V., Zhuchkov R.N., Tarasova N.V., Sizova M.A. Multifunctional software package LOGOS for the problems of hydrodynamics and heat and mass transfer on a supercomputer. *Basic technologies and algorithms. Proc. of the 12th International seminar "Supercomputing and mathematic modeling"*. Sarov, 2010, pp. 215–230. (in Russian)
2. Volkov K.N., Derugin Yu.N., Emelyanov V.N., Karpenko A.G., Kozelkov A.S., Smirnov P.G., Teterina I.V. Implementation of parallel calculations on graphics processor units in the LOGOS computational fluid dynamics package. *Numerical Methods and Programming*, 2013, vol. 14, no. 1, pp. 334–342. (in Russian)
3. Emelyanov V.N., Brykov N.A., Efremov A.V. Application of vectorized algorithms for solving problems of continuum mechanics. *Journal of Physics: Conference Series*, 2021, vol. 2094, pp. 022047. https://doi.org/10.1088/1742-6596/2094/2/022047
4. Emelianov V.N. *Mechanics of Continuous Media: Strain Theory and Main Models*. Moscow, Jurait Publ., 2018, 162 p. (in Russian)
5. Emelianov V.N. *Numerical Methods: An Introduction to the Theory of Difference Schemes*. Moscow, Jurajt Publ., 2018, 188 p. (in Russian)
6. Gergel V.P. *Theory and Practice of Parallel Computations*. Moscow, BINOM Publ., 2010, 432 p. (in Russian)
7. Volkov K.N., Emelianov V.N. Application of MATLAB for the solution of boundary-value problems of liquid and gas mechanics. *Proc. of All-Russian scientific conference "Of Research and Engineering Applications in MATLAB"*. Moscow, ICS RAS Publ., 2004, pp. 92–115. (in Russian)
8. Volkov K.N., Emelyanov V.N. Implementation of vectorized finite-difference algorithms for solving boundary value problems of fluid and gas mechanics with MATLAB package. *Numerical Methods and Programming*, 2004, vol. 5, no. 3, pp. 13–29. (in Russian)
9. Alberty J., Carstensen C., Funken S.A. Remarks around 50 lines of Matlab: short finite element implementation. *Numerical Algorithms*, 1999, vol. 20, no. 1-2, pp. 117–137. https://doi.org/10.1023/A:1019155918070
10. Persson P.-O., Strang G. A simple mesh generation in MATLAB. *SIAM Review*, 2004, vol. 46, no. 2, pp. 329–345. https://doi.org/10.1137/S0036144503429121
11. Koko J. Vectorized MATLAB codes for linear two-dimensional elasticity. *Scientific Programming*, 2007, vol. 15, no. 3, pp. 157–172. https://doi.org/10.1155/2007/838942
12. Funken S., Praetorius D., Wissgott P. Efficient implementation of adaptive P1-FEM in MATLAB. *Computational Methods in Applied Mathematics*, 2011, vol. 11, no. 4, pp. 460–490. https://doi.org/10.2478/cmam-2011-0026
13. Rahman T., Valdman J. Fast MATLAB assembly of FEM matrices in 2D and 3D: nodal elements. *Applied Mathematics and Computation*, 2013, vol. 219, no. 13, pp. 7151–7158. https://doi.org/10.1016/j.amc.2011.08.043
14. Anjam I., Valdman J. Fast MATLAB assembly of FEM matrices in 2D and 3D: edge elements. *Applied Mathematics and Computation*, 2015, vol. 267, pp. 252–263. https://doi.org/10.1016/j.amc.2015.03.105
15. Koko J. Fast MATLAB assembly of fem matrices in 2D and 3D using cell-array approach. *International Journal of Modeling, Simulation, and Scientific Computing*, 2016, vol. 7, no. 2, pp. 1650010. https://doi.org/10.1142/S1793962316500100
16. Walker S.W. FELICITY: a MATLAB/C++ toolbox for developing finite element methods and simulation modeling. *SIAM Journal on Scientific Computing*, 2018, vol. 40, no. 2, pp. 234–257. https://doi.org/10.1137/17M1128745
17. Čermák M., Sysala S., Valdman J. Efficient and flexible MATLAB implementation of 2D and 3D elastoplastic problems. *Applied Mathematics and Computation*, 2019, vol. 355, pp. 595–614. https://doi.org/10.1016/j.amc.2019.02.054
18. Tsega E.G. A finite volume solution of unsteady incompressible Navier–Stokes equations using MATLAB. *Numerical and Computational Methods in Sciences and Engineering*, 2019, vol. 1, no. 3, pp. 117–131. http://dx.doi.org/10.18576/ncmsel/010302

20. Волков К.Н., Емельянов В.Н. Вычислительные технологии в задачах механики жидкости и газа. М.: Физматлит, 2012. 468 с.
21. Лурье А.И. Теория упругости. М.: Наука, 1970. 940 с.

19. Godunov S.K., Zabrodin A.V., Ivanov M.Ia., Kraiko A.N., Prokopov G.P. *Numerical Solutions of Multidimensional Problems of Gas Dynamics*. Moscow, Nauka Publ., 1976, 400 p. (in Russian)
20. Volkov K.N., Emelianov V.N. *Computational Technologies in the Problems of Fluid and Gas Mechanics*. Moscow, Fizmatlit Publ., 2012, 468 p. (in Russian)
21. Lure A.I. *Theory of Elasticity.* Moscow, Nauka Publ., 1970, 940 p. (in Russian)

Авторы

Брыков Никита Александрович — кандидат технических наук, доцент, доцент, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, Санкт-Петербург, 190005, Российская Федерация, sc 57209849110, https://orcid.org/0000-0002-3568-7522, brykovna@yandex.ru

Волков Константин Николаевич — доктор физико-математических наук, ведущий научный сотрудник, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, Санкт-Петербург, 190005, Российская Федерация, sc 8663950000, https://orcid.org/0000-0001-6055-2323, k.volkov@kingston.ac.uk

Емельянов Владислав Николаевич — доктор технических наук, профессор, профессор, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, Санкт-Петербург, 190005, Российская Федерация, sc 7004697417, https://orcid.org/0000-0002-7039-9948, vlademelyanov@gmail.com

Authors

Nikita A. Brykov — PhD, Associate Professor, Associate Professor, Baltic State Technical University “VOENMEH” named after D.F. Ustinov, Saint Petersburg, 190005, Russian Federation, sc 57209849110, https://orcid.org/0000-0002-3568-7522, brykovna@yandex.ru

Konstantin N. Volkov — D.Sc., Leading Researcher, Baltic State Technical University “VOENMEH” named after D.F. Ustinov, Saint Petersburg, 190005, Russian Federation, sc 8663950000, https://orcid.org/0000-0001-6055-2323, k.volkov@kingston.ac.uk

Vladislav N. Emelyanov — D.Sc., Full Professor, Baltic State Technical University “VOENMEH” named after D.F. Ustinov, Saint Petersburg, 190005, Russian Federation, sc 7004697417, https://orcid.org/0000-0002-7039-9948, vlademelyanov@gmail.com

Статья поступила в редакцию 28.12.2021
Одобрена после рецензирования 17.01.2022
Принята к печати 30.01.2022

Received 28.12.2021
Approved after reviewing 17.01.2022
Accepted 30.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-1-206-216

A comparative analysis of computational intelligence algorithms for estimation of LTE channels

Siraj Pathan[1]✉, Ajit Noonia[2], Mujib Tamboli[3], Sunil Pathak[4]

[1,2,4] Amity University Rajasthan, Jaipur, 303006, India
[3] Anjuman-I-Islam's Kalsekar Technical Campus, Panvel, 410206, India

[1] sirajpathan404@gmail.com✉, https://orcid.org/0000-0002-1728-3030
[2] ajit009noonia@gmail.com, https://orcid.org/0000-0002-3110-4908
[3] mujibtamboli@yahoo.co.in, https://orcid.org/0000-0002-5891-3713
[4] sunilpath@gmail.com, https://orcid.org/0000-0001-7409-9814

Abstract
Precise modelling and accurate estimation of long-term evolution (LTE) channels are essential for numerous applications like video streaming, efficient use of bandwidth and utilization of power. This deals with the fact that data traffic is increasing continuously with advances in Internet of things. Previous works were focused mainly on designing models to estimate channel using traditional minimum mean square error (MMSE) and least squares (LS) algorithms. The proposed model enhances LTE channel estimation. The designed model combines LS and MMSE methods using Taguchi genetic (GE) and Particle Swarm Intelligence (PSO) algorithms. We consider LTE operating in 5.8 GHz range. Pilot signals are sent randomly along with data to obtain information about the channel. They help to decode a signal in a receiver and estimate LS and MMSE combined with Taguchi GA and PSO, respectively. CI-based model performance was calculated according to the bit error rate (BER), signal-to-noise ratio and mean square error. The proposed model achieved the desired gain of 2.4 dB and 5.4 dB according to BER as compared to MMSE and LS algorithms, respectively.

Keywords
genetic algorithm (GA), particle swarm intelligence (PSO), long term evolution (LTE), minimum mean square error (MMSE), least squares (LS)

Acknowledgements
We thank Amity School of Engineering & Technology of Amity University, Jaipur, India for providing material and resources for our study, we also thank Anjuman-I-Islam's Kalsekar Technical Campus of Mumbai University Navi Mumbai, India for providing laboratory to do research.

For citation: Pathan S., Noonia A., Tamboli M., Pathak S. A comparative analysis of computational intelligence algorithms for estimation of LTE channels. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 206–216. doi: 10.17586/2226-1494-2022-22-1-206-216

УДК 621.396.946

Сравнительный анализ алгоритмов вычислительного интеллекта для оценки канала LTE

Сирадж Патан[1]✉, Аджит Нуния[2], Муджиб Тамболи[3], Сунил Патхак[4]

[1,2,4] Амити Университет Раджастана, Джапур, 303006, Индия
[3] Технический кампус Анджуман-и-Ислам в Калсекаре, Панвел, 410206, Индия

[1] sirajpathan404@gmail.com✉, https://orcid.org/0000-0002-1728-3030
[2] ajit009noonia@gmail.com, https://orcid.org/0000-0002-3110-4908
[3] mujibtamboli@yahoo.co.in, https://orcid.org/0000-0002-5891-3713
[4] sunilpath@gmail.com, https://orcid.org/0000-0001-7409-9814

© Pathan S., Noonia A., Tamboli M., Pathak S., 2022

Аннотация
Выполнение моделирования и точной оценки канала беспроводной связи в стандарте Long-Term Evolution (LTE) необходимо для работы многочисленных приложений, таких как потоковое видео, а также для эффективного использования полосы пропускания и энергии. Это связано с постоянным увеличением трафика данных и развитием интернета вещей. Существующие исследования в основном направлены на изучение моделей оценки канала с использованием традиционных алгоритмов вычисления минимальной среднеквадратичной ошибки Minimum Mean Square Error (MMSE) и наименьших квадратов Least Square (LS). Предложенная модель позволяет улучшить оценку канала в мобильных сетях стандарта LTE. Модель основана на объединении методов наименьших квадратов и наименьшей среднеквадратической ошибки с применением Тагучи-генетического алгоритма и алгоритма оптимизации роя частиц Particle Swarm Intelligence (PSO). Приведен пример сети LTE, работающей в диапазоне 5,8 ГГц. Случайные пилотные сигналы следуют вместе с данными для получения сведений о канале, помогают декодировать сигнал в приемнике и оценивать LS и MMSE за счет сочетания Тагучи-генетического алгоритма и PSO соответственно. Выполнена оценка эффективности модели по частоте битовых ошибок Bit Error Rate (BER), отношению сигнал/шум и среднеквадратической ошибки. С учетом величины BER представленная модель на основе искусственного интеллекта обеспечивает лучшие результаты по сравнению с MMSE на 2,4 дБ и с LS – на 5,4 дБ.

Ключевые слова
генетический алгоритм, GA, алгоритм оптимизации роя частиц, PSO, мобильная сеть LTE, минимальная среднеквадратическая ошибка, MMSE, метод наименьших квадратов, LS

Благодарности
Авторы благодарят Школу инженерии и технологий Университета Амити в Джайпуре, Индия, за предоставление учебных материалов и ресурсов. Также благодарны за предоставление лаборатории для проведения исследований Техническому кампусу Анджуман-и-Ислам в Калсекаре, Университет Мумбаи, Нави Мумбаи, Индия

Ссылка для цитирования: Патан С., Нуния А., Тамболи М., Патхак С. Сравнительный анализ алгоритмов вычислительного интеллекта для оценки канала LTE // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 206–216 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-1-206-216

Introduction

LTE means Long Term Evolution, and it began as a project in early 2005 by the 3rd Generation Partnership Project, a telecommunications organization. The corresponding evolution of 3G packet network is System Architecture Evolution (SAE). LTE is a term that stands for the combination of LTE and SAE [1, 2].

The Universal Mobile Telecommunication System (UMTS) is one of the branches of Global System for Mobile Communications and Evolved UMTS Terrestrial Radio Access network, whereas E-UTRAN is the air interface in an LTE Cellular Network. LTE is the first version.

Fig. 1 shows the evolution of LTE and how different version of LTE added new technologies for better services. The 3rd Generation Partnership Project has been working

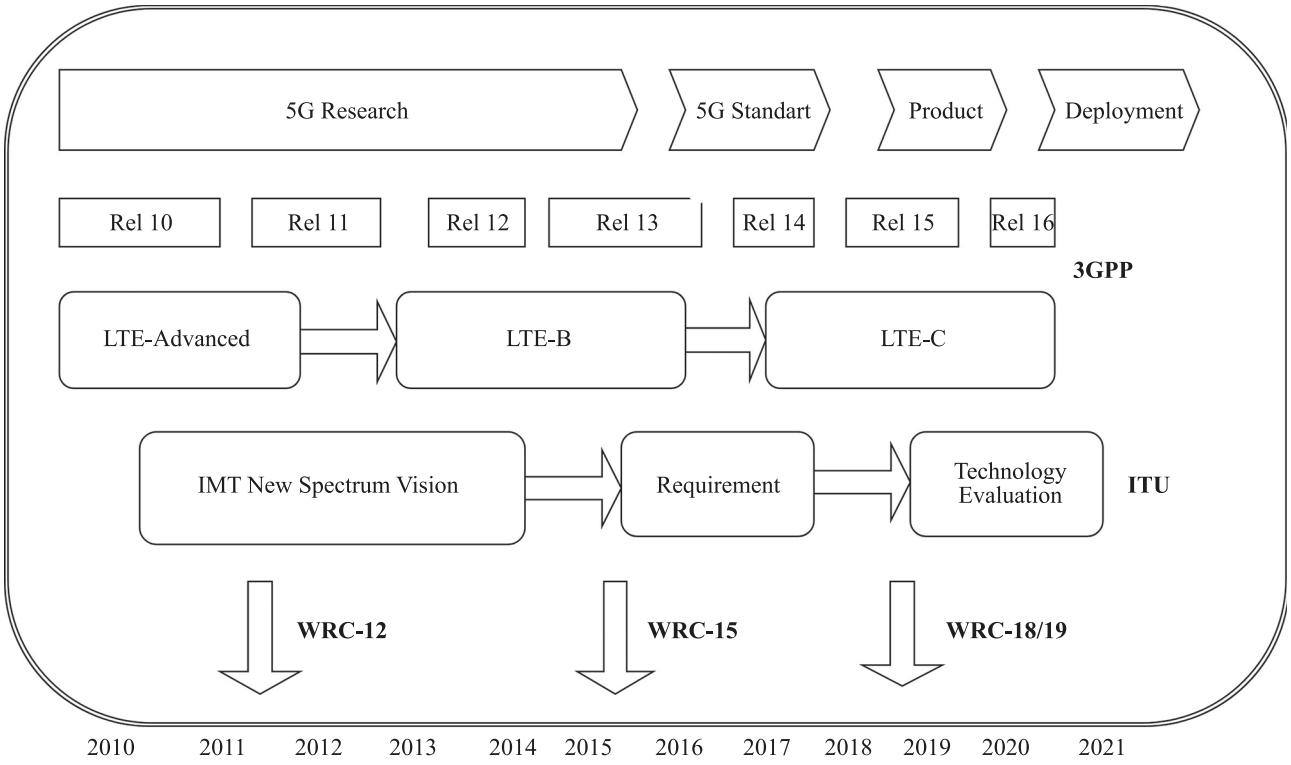


Fig. 1. Evolution of LTE

on LTE on the path to the fourth generation mobile due to the increase in mobile devices data usage and the advent of new applications for Internet of Things devices [3, 4].

LTE's major aim is to form a high-data-rate, low error, packet-optimized wireless radio technology that allows for changing capacity deployments. LTE network architecture was created to perform switched traffic while maintaining high quality of service and mobility.

The most significant parameter in wireless communications is channel state information (CSI). This information aids in understanding signal propagation via the channel, as well as the distortion and delay created by the signal from the transmitter side to the receiver side.

The LTE channel estimation depends on present information that is shared between the sending and LTE receiving side. This approach of channel estimation aids in signal reconstruction at the receiver end. At the receiver, the LTE channel impulse response can be estimated by using inserted pilot symbols, which are familiar to both the sending and receiving side, and which apply various interpolation types to estimate the LTE channel response of the subcarriers between the pilots [5, 6].

Although blind channel estimate has a lower overhead, it requires a lot of signals received from antenna to reconstruct the desired signal. The efficiency of the channel estimation can be improved by utilizing the semi-blind channel estimating system, which uses pilot and data symbols.

This method employs the signal tracked as a feedback mechanism to track the performance of the channel, as well as the tracked signal as reference signal for future data prediction. Although based on symbols, channel prediction delivers the best performance, above the described pilot symbols, which are transmitted with data symbols, reduce transmission efficiencies. When we use training symbols, the channel estimation is done using Least Square (LS) and Minimum Mean Square Error (MMSE) algorithms, which improve system performance by lowering the Bit Error Rate (BER) [7, 8].

Optimization deals with the process of creating a system with the goal of lowering production costs or increasing production efficiency. The procedure for algorithms of optimization is to run them continuously while comparing alternative answers until the optimal one is found.

A genetic algorithm reflects the process of natural selection of fittest particle for reproduction in order to produce offspring of next generation.

Particle Swarm Optimization (PSO) is nothing but finding the best solution in space. It works on the objective function and is not affected by any differential shape of the objective [9, 10].

Using pilot signals, this work approximates the noised mixed, faded signal in the system at the receiver side. The random pilot signals are subjected to the optimization technique, which is then compared to the fixed pilot signals subjected to the LS algorithm and MMSE algorithms [11, 12].

Related Work

One of the applications that received much attention is data-driven channel estimation. In [13], a channel prediction technique for LTE was suggested. To increase the performance of a LTE system that consists of Millimeter wave connected with N Multiple-Input Multiple-Output (MIMO) users, this strategy was integrated with hybrid LS techniques.

The researchers proposed a CI-Model which uses Orthogonal Frequency Division Multiplexing (OFDM) as receiver that applies a Computational Intelligence (CI) algorithm to collect channel state information (CSI) and find out transmitted symbols [14].

Montgomery [15] investigated a MIMO based huge LTE system that estimated the channel matrix using an LS and MMSE. However, a major disadvantage of utilizing LS and MMSE for channel estimation is that it necessitates a significant amount of training channel data, which may be impractical for time-changing channels.

The medium in a Frequency-Division-Duplex system was estimated using a bi-direction recurrent Genetic Algorithm (GA) in conjunction with pilot samples in [16].

In addition, the authors suggested a CI-included channel evaluating idea for MIMO antenna system that consists of a MIMO base interacting with many single-antenna points in [17]. Instead of evaluating the medium coefficients using the Least Square (LS) approach, this work applied computational intelligence to denoise the received signal.

In [18], the author introduced a joint GA and LS-based channel estimation for the transmitter that takes advantage of the receiver's Signal to Noise Ratio (SNR) received from output side to derive MIMO wireless channel coefficients.

CI implementation was suggested to overcome difficulties of predicting a mixed selective wireless channel [19]. In this study, the researcher focused on using a CI to calculate channel coefficients during the training phase.

Contributions

A deep examination of the works discussed in the previous section shows that the vast majority of authors have primarily concentrated on using LS and MMSE to acquire the CSI. Processing intelligence algorithms has significant computational needs and requires a lot of data to train. This feature prompted us to study the feasibility of introducing a low-complexity CI-based method specifically tailored for channel estimation. The contribution of our research is as follows.

We present a powerful CI algorithm for estimating and monitoring difficult time changing and frequency selective multipath channels. The parameters are optimised and the Channel Impulse Response (CIR) is obtained using stochastic gradient descent in this algorithm.

To calculate the impulse response of the channel whose time is changing continuously, the proposed CI algorithm is combined with the OFDM based transmission technique.

The performance of the designed technique is compared to that of the well-known least squares and MMSE channel estimation algorithms for a variety of pilot sample numbers. Even with minimal pilot samples for each symbol of OFDM, the suggested approach performs better than LS and MMSE algorithms, according to the obtained findings. The designed CI-Algorithm based wireless medium evaluating complexity is assessed in terms of drifting

numbers operations. Our approach has a lower complexity than the usual MMSE and LS, according to the complexity evaluation.

Finally, the suggested algorithm's convergence is studied by determining the number of iterations required for the designed approach to meet the best feasible estimation channel.

Organization of the paper

The paper is organised as follows. The first section explains a system model with the help of mathematical expression. The subsections describe the transmitter and the designed channel model and their implementation with LS and MMSE technique. The next section deals with the CI based channel estimation explaining the receiver. The section on results considers all parameters and the paper is concluded with the achieved results.

System model

We analyse a basic LTE communication system shown in Fig. 2 in which Multiple Antenna at sending and receiving sides communicate over a changing frequency channel with a changing time frequency. We use the OFDM bock to reduce ISI and a multiplexing (OFDM) method.

In the LTE system, Downlink and Uplink Transmission are organized in frames. Data for Transmission are modulated using M-ary Quadrature Amplitude Modulation (QAM) scheme to get frame of N-subcarrier symbols, i.e. $X(k)$. OFDM signals, i.e. $x(n)$ to be send, are received by applying the Inverse Discrete Fourier Transform on $X(k)$ such that

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X(k) e^{j2\Pi k \frac{n}{N}}.$$

In practice, N is a power of two, and IDFT is carried out with the help (IFFT) inverse-fast-Fourier transform. Actually, we consider an OFDM data with N sub-carriers that are split into N_P pilots to calculate p (number of pilot subcarrier) and track impulse response, as well as payload, N_D stand for data samples, D is number of Data Samples; j is imaginary unit used to indicate complex component in alternating current circuits; k is wavenumber. $X(k)$ is the Discrete Fourier Transform of $x(n)$.

Transmitter. The data is initially coded and then mapped with QAM on the transmitter side, using the block of Modulation. We shall consider system sends data in T time slots, with the QAM symbols at time slot t, $t = 1, \ldots, T$ are combined to Data Vector $x(t) \in C^N$ as

$$x(t) = [x_1(t), x_2(t), x_3(t), \ldots, x_N(t)],$$

where N is symbol of modulation. Following that, the encoded information is divided into N_T vectors that correspond to the Transmit antenna N_T as given below

$$x_i(t) = [x_i(t), x_{i+N_T}(t), x_{i+2N_T}(t), \ldots],$$

where $i = 1, 2, \ldots, N_T$.

Each antenna will send data in serial to parallel form, and pilot symbols are inserted to acquire knowledge of Transmitter and receiver side as shown in Fig. 3. We define $x_a(t)$ and $a = 1, \ldots, N_T$. $x_a(t)$ is a vector for signal which is inserted between pilots. After that the block of inverse fast Fourier transform is applied to $x_a(t)$, transforming the signals from frequency domain to time domain defined by $\tilde{x}_a(t)$ as follows:

$$\tilde{x}_a(t) = IFFT\{x_a(t)\}.$$

Then, we use cyclic prefix insertion of length N_G as a guard interval to remove the inter-symbol interference.

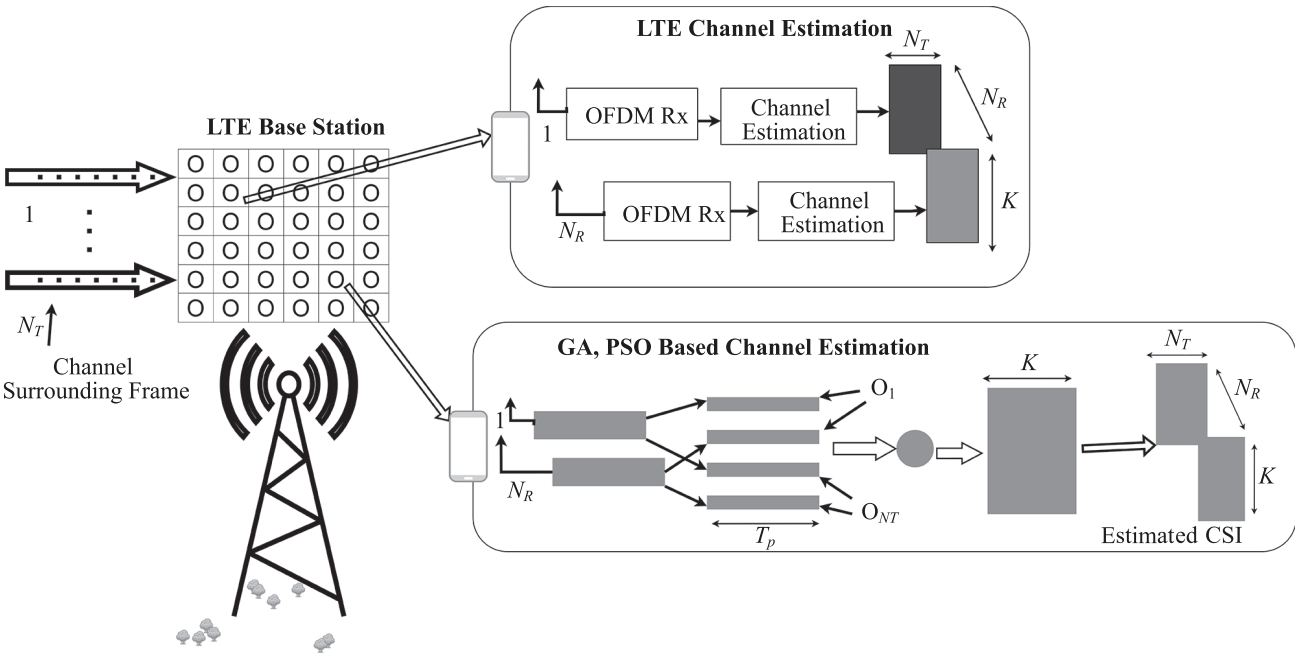


Fig. 2. Overview of the Computational Intelligence algorithm based Channel Estimation for LTE.

K is Number of Subcarriers, N_T is Number of transmit antenna, N_R is Number of receiver antennas, O_{NT} is Orthogonal Sequence of length N_T, T_p is Number of time Domain samples in the channel sounding frame

Fig. 3. Illustration of the considered LTE Model with the proposed CI Based Algorithm Module

Fig. 4. Block Diagram of Computational Intelligence Based LS and MMSE Channel estimation

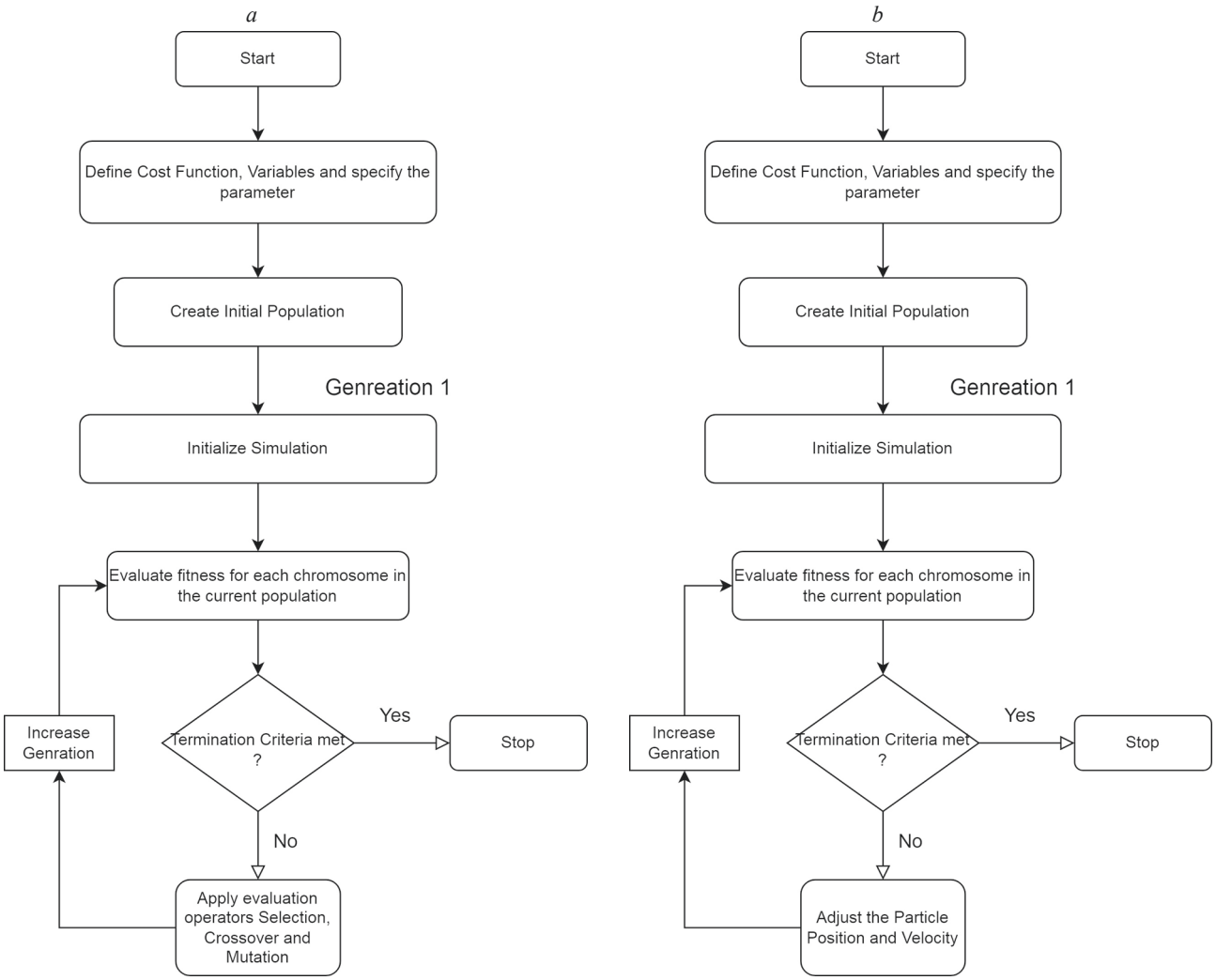


Fig. 5. Flow chart of Computational Intelligence based GA (*a*) and PSO (*b*) Algorithm

Fig. 4 represents CI-based and MMSE channel estimation. The GA is an automatic search algorithm for determining the lowest cost function and maximizing fitness. This algorithm is defined by parameters as mutation, cross over, and process of selection. If number of variables is more used in MIMO-OFDM based LTE system, GA provides the best solution. This method, which is one of the best, provides the best answer to the problem and is utilized in a variety of applications, including mobile communication systems. GA consists of three components: reproduction, mutation, and crossover. These components are used to calculate the Fitness value to select the best channel. The fitness formula is as follows:

$$fitness = \left(\frac{H - H_{GA}}{H}\right)^2.$$

Each particle is referred to as an agent in PSO, and it is kept in the search area to locate the objective function from the present location (Fig. 5). Every particle in the solution search space adjusts its velocity to find the optimum spot.

$$fitness = \left(\frac{H - H_{PSO}}{H}\right)^2.$$

The fitness function helps to optimize the problem and identify the best answer.

Designed Channel Model. Received Output Signal is defined as

$$y = xh + n,$$

where n is the zero mean complex Gaussian noise with variance, x defines the frame of modulated N-length *IFFT*, h shows CIR, which is designed and considered as Length **L** vector, * represents convolution operation.

Noise with variance is given by the following formula:

$$N_0 = 2\Sigma_w.$$

The time varying model which uses Jakes's model is described as follows:

$$h_{jake}(t) = \sum_l A_l e^{j[\Theta_L + 2\Pi f_{D\max}\cos(\Phi_L)t]}.$$

Arrival angle, and phase of the lth path are represented by A_l, Θ_L, and Φ_L, respectively. It is worth noting that Θ_l and Φ_L are mutually independent and distributed uniformly between [–Π, Π]. Furthermore, $f_{D\max}$ specifies the highest possible Doppler frequency.

CI based channel estimation

Least Squares Estimation

The LS channel estimation technique assumes a block-type pilot configuration with pilots located at all subcarriers of the OFDM signal.

The received signal after FFT demodulation is as follows:

$$Y_p = H_P X_P + W_P,$$

when H_P denotes the Frequency Selective Channel's Frequency Response at pilot places, X_P and W_P defines pilot symbols of the transmitter and samples of additive white Gaussian noise, respectively.

To secure channel estimation after implementing LS, the cost function should be minimised

$$J(\hat{H}_{LS}) = \|Y_P - X_P \hat{H}_{LS}\|. \quad (1)$$

Furthermore, the equation (1) should be calculated by considering $\hat{H}_{LS}$ and the output is set to zero to achieve the cost function's minimal value.

This expression will be further simplified to give channel estimation due to LS:

$$\hat{H}_{LS} = \frac{Y_P(k)}{X_P(k)}.$$

Minimum Mean Square Error Estimation

Minimum Means square estimation was first introduced in [20] as shown in Fig. 6.

By close analysis of the above described figure, the estimated channel obtained is

$$\hat{H}_{MMSE} = G\check{H}.$$

At pilot position $\check{H}$ is given as:

$$\check{H} = X_P^{-1} Y_P.$$

By minimising the below given cost function, MMSE estimate of the channel can be calculated as follows:

$$J\left(\check{H}_{MMSE}\right) = E\{\|e\|^2\} = E\left\{\left\|H - \check{H}_{MMSE}\right\|^2\right\}.$$

Receiver

To acquire $\tilde{y}_b(t)$ vector of length N_{FFT} on side of receiver, the cyclic prefix is detached from the collected signal $\tilde{y}_{gb}(t)$ on every received antenna by the module for removing cyclic prefix. The FFT block then converts the signal to parallel form and transforms it into the frequency domain, yielding a frequency domain signal $\tilde{y}_b(t)$:

$$y_b(t) = FFT\{\tilde{y}_b(t)\}.$$

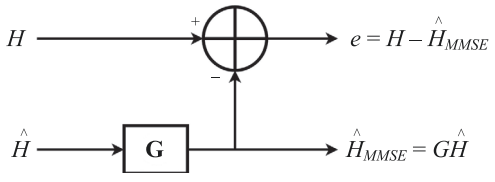


Fig. 6. MMSE System

For channel estimation, the pilot signal is extracted from the domain of the frequency signal. The layer demapping module equalizes and congregates the received signal $y_b(t)$ from all the reception antennas into a serial sequence after calculating the channel.

The signal is subsequently demodulated using the demodulation scheme, which is the same as the transmitter's approach. The final binary data sequence is obtained as the output of the LTE system model.

Results

This section presents the results as well as their discussion. The simulation results in this part were achieved using Spyder (anaconda3) and Google Colab on a core 3 GHz processor with 20 GB RAM and 1053 GPU.

Furthermore, the suggested Computational Intelligence-based channel estimation technique was implemented on a node that receives information through a time changing frequency channel model using Jake's model. In addition, Table 1 shows the simulation and channel model settings, while Table 2 shows the algorithm parameters.

The frame period T_{fp} is set to become shorter than the coherence time of channel T_{Ch} because the channel is believed to be twice selective and frequency-selective. Nonetheless, the channel response varying from frame to frame, the system must collect a correct CIR for every frame. To successfully estimate the CIR and keep track of the CIR variation, we designed the CI-based estimation algorithm at the output side.

We consider a period of training to receive the CIR before the start of data transmission. Once data transmission starts, the CI-based system will use define samples of pilot, inserted before each signal OFDM, to keep track of the impulse response change with respect to time. The presented innovative CI-based channel estimation's mean square error (MSE) performance was examined for a

Table 1. Simulation settings

FFT size	512
Number of symbols	100
No of Pilots	4
Carrier Frequency	3.7 GHZ
Mode of Modulation	16-QAM
Sub-carriers	2052
MaxVelocity	85 m/s
Doppler Shift	33–973 Hz
Per user Bandwidth	20 MHz
Channel type	AWGN

Table 2. CI Network Parameters

Input layer Size	2L
Output Layer Size	1
Number of Layers	2
Learning rate	0.001

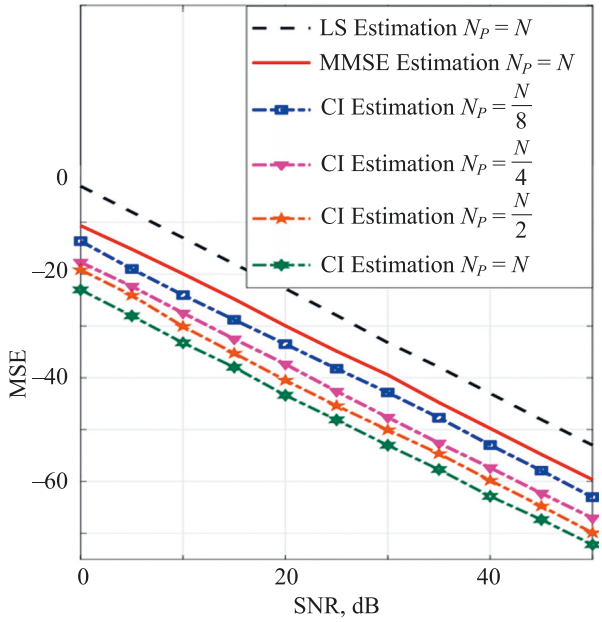


Fig. 7. MSE performance Comparison

different number of blocks of pilot in each block as shown in Fig. 7.

The pilots are considered as follows:

$$N_P = \frac{N}{8}, \frac{N}{4}, \frac{N}{2} \text{ and } N.$$

Mean square error performance of the designed model is compared with LS and MMSE channel estimation algorithm.

After the detailed analysis of Fig. 7, the designed algorithm outperforms MMSE and LS algorithm even when the number of pilots is small.

For $\frac{N}{8}$ data samples, the designed model gives 3.6 dB and 9.9 dB over MMSE and LS channel estimation algorithms. For $\frac{N}{4}$ data samples the designed model has a gain of 8.1 dB and 14.6 dB over MMSE and LS.

For $\frac{N}{2}$ samples, the model attained 11.3 dB and 17.9 dB compared with MMSE and LS.

For a variable number of pilots, we explored how the performance of MSE for CI-based channel estimate algorithm is turned into performance of BER at the receiving point. Fig. 8, *a* depicts the bit error rate performance for different four channel state information scenarios (CSI), including perfect or ideal CSI, CSI derived using the LS method, MMSE algorithm, and the suggested innovative CI-based algorithm.

A close analysis of Fig. 8, *a* reveals that BER performance of the designed model matches with perfect and ideal CSI. This match shows the effectiveness of the designed model.

Then we evaluate the MSE performance of the designed model against the number of pilot samples. Fig. 8, *b* shows the impact of increasing the number of pilot samples when SNR is fixed at 15 dB and 20 dB.

Similarly, BER outcome is evaluated against the number of pilot samples. Fig. 9, *a* shows that the performance of BER is proportional to the number of blocks of pilots. Performance of BER for CI-dependent estimator reaches of the evaluated values for BER of ideal CSI.

Fig. 9, *b* shows the convergence behaviour of the CI-based algorithm. We calculated MSE performance for different iterations to get the lowest MSE when SNR value is 10 dB, 15 dB and 20 dB, respectively. The designed model achieves the lowest MSE in between 17 and 19, respectively.

The analysis of Table 3 (Ô is Complexity constant and *L* is Floating point operations) shows that complexity of CI-model is less than LS and MMSE (Fig. 10, *a*).

Table 3. Complexity analysis

Estimator	Complexity
LS	Ô (*Np*)
MMSE	Ô (N^2p)
CI-based estimator	Ô (L^2Np)

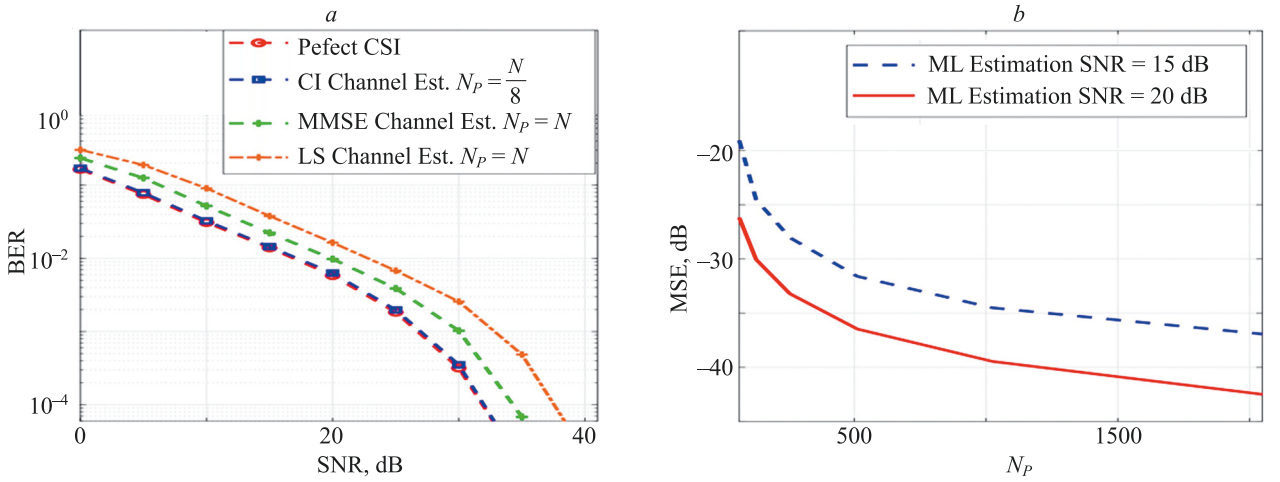


Fig. 8. BER performance comparison (*a*) and MSE performance vs pilots samples (*b*)

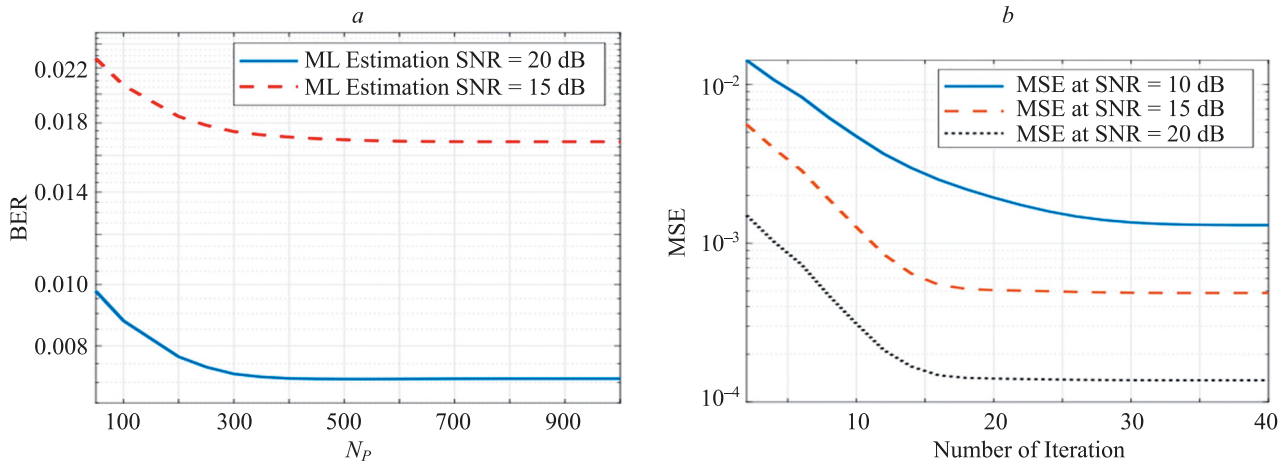


Fig. 9. BER Performance vs pilots samples (*a*) and MSE performance vs number of iterations (*b*)

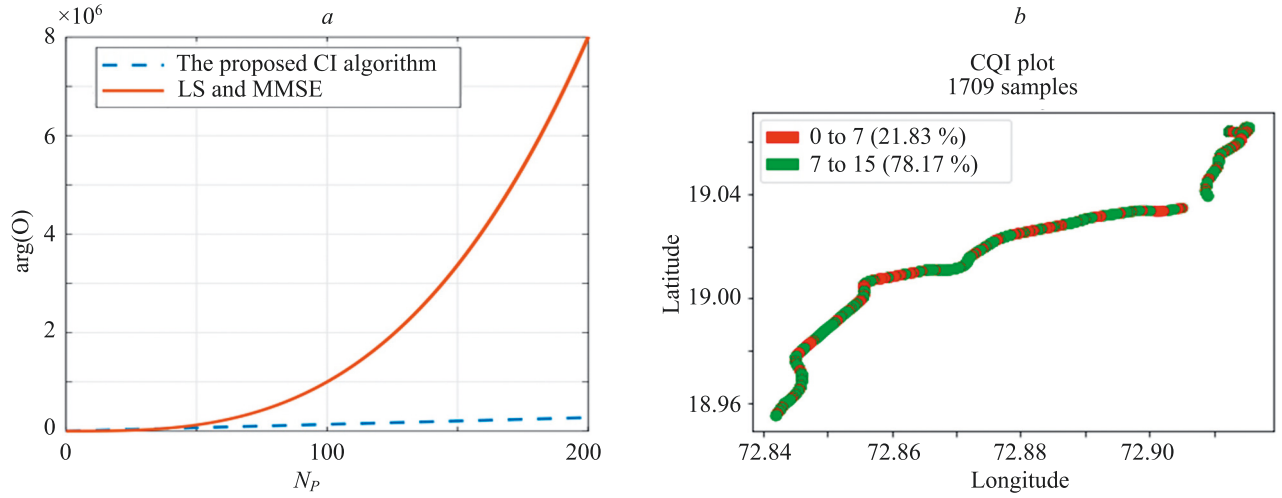


Fig. 10. Complexity comparison (*a*) and CQI analysis (*b*)

The proposed estimator converge the best CIR estimation if L is between 3 to 10 times less than the number of samples N_P.

Finally, Fig. 10, *b* shows the analysis of the channel quality indicator for the designed model taking into account different longitude and latitude for 1,709 drive test data samples.

Conclusion

This paper introduced a straightforward but an effective computational intelligence-depended approach for channel evaluation. We investigated an unmarked OFDM-depended LTE system that receives data across a frequency selective channel that changes over time. The designed algorithm was then applied to this OFDM depended system to evaluate this time changing channel and follows its changes from block to block. First, the suggested algorithm's performance was assessed in terms of MSE channel evaluation performance and differentiated from LS and MMSE channel estimation methods. Even with a modest number of pilot samples used to estimate and track the CIR, the obtained findings showed that the presented approach beat the LS and MMSE algorithms. The suggested CI-based technique performs far better than MMSE and LS channel estimation algorithms by 2.4 dB and 5.4 dB, respectively, in terms of BER. Furthermore, we tested the robustness of the proposed approach against a variety of different maximum Doppler frequency values, and the findings proved the effectiveness of our system in time-changing CIR, as the change in frequency caused only a modest outcome decrease of 0.19 dB.

References

1. Bobde S., Phalnikar R. Software restructuring models for object oriented programming languages using the fuzzy based clustering algorithm. *Scientific and technical journal of information technologies, mechanics and optics*, 2021, vol. 21, no. 5, pp 738–747. https://doi.org/10.17586/2226-1494-2021-21-5-738-747
2. Bothra S.K., Singhal S. Nature-inspired metaheuristic scheduling algorithms in cloud: a systematic review. *Scientific and technical journal of information technologies, mechanics and optics*, 2021, vol. 21, no. 4, pp. 463–472. https://doi.org/10.17586/2226-1494-2021-21-4-463-472
3. Farrokhi F.R., Lozano A., Foschini G.J., Valenzuela R.A. Spectral efficiency of FDMA/TDMA wireless systems with transmit and receive antenna arrays. *IEEE Transactions on Wireless Communications*, 2002, vol. 1, no. 4, pp. 591–599. https://doi.org/10.1109/TWC.2002.804078
4. Balanis C. *Antenna theory: Analysis and design*. 4rd ed. New York, John Wiley & Sons, 2016, pp. 900–908.
5. Durrani S., Bialkowski M.E. Effect of mutual coupling on the interference rejection capabilities of linear and circular arrays in CDMA systems. *IEEE Transactions on Antennas and Propagation*, 2004, vol. 52, no. 4, pp. 1130–1134. https://doi.org/10.1109/TAP.2004.825640
6. Piazza D., Kirsch N.J., Forenza A., Heath R.W., Dandekar K.R. Design and evaluation of a reconfigurable antenna array for MIMO systems. *IEEE Transactions on Antennas and Propagation*, 2008, vol. 56, no. 3, pp. 869–881. https://doi.org/10.1109/TAP.2008.916908
7. Lozano A., Tulino A.M. Capacity of multiple-transmit multiple-receive antenna architectures. *IEEE Transactions on Information Theory*, 2002, vol. 48, no. 12, pp. 3117–3127. https://doi.org/10.1109/TIT.2002.805084
8. Oyman Ö., Nabar R.U., Bölcskei H., Paulraj A.J. Tight lower bounds on the ergodic capacity of Rayleigh fading MIMO channels. *Proc. of the Conference Record / IEEE Global Telecommunications Conference*. V. 2, 2002, pp. 1172–1176. https://doi.org/10.1109/GLOCOM.2002.1188380
9. Du J., Li Y. Optimization of antenna configuration for MIMO systems. *IEEE Transactions on Communications*, 2005, vol. 53, no. 9, pp. 1451–1454. https://doi.org/10.1109/TCOMM.2005.855002
10. Waheed U.A., Kishore D.V. Uplink spatial fading correlation of MIMO channel. *IEEE Vehicular Technology Conference*, 2003, vol. 58, no. 1, pp. 94–98. https://doi.org/10.1109/VETECF.2003.1284985
11. Tsai J.-A., Woerner B.D. The fading correlation function of a circular antenna array in mobile radio environment. *Proc. of the IEEE Global Telecommunications Conference*. V. 5, 2001, pp. 3232–3236. https://doi.org/10.1109/GLOCOM.2001.966023
12. Li X., Nie Z.-P. Spatial fading correlation of circular antenna arrays with Laplacian PAS in MIMO channels. *Proc. of the IEEE Antennas and Propagation Society International Symposium*. V. 4, 2004, pp. 3697–3700. https://doi.org/10.1109/APS.2004.1330149
13. Recioui A., Bentarzi H. Genetic algorithm based MIMO capacity enhancement in spatially correlated channels including mutual coupling. *Wireless Personal Communications*, 2012, vol. 63, no. 3, pp. 689–701. https://doi.org/10.1007/s11277-010-0159-5
14. Recioui A., Azrar A. Use of genetic algorithms in linear and planar array synthesis based on Schelkunoff method. *Microwave and Optical Technology Letters*, 2007, vol. 49, no. 7, pp. 1619–1623. https://doi.org/10.1002/mop.22510
15. Montgomery D. C. *Design and Analysis of Experiments*. New York, Wiley, 1991, pp. 28–39.
16. Ross P.J. *Taguchi Techniques for Quality Engineering*. McGraw-Hill. 2013, pp. 43–52.
17. Weng W.C., Yang F., Elsherbeni A.Z. Linear antenna array synthesis using Taguchi's method: A novel optimization technique in electromagnetic. *IEEE Transactions on Antennas and Propagation*, 2007, vol. 55, no. 3, pp. 723–730. https://doi.org/10.1109/TAP.2007.891548
18. Wu Y. *Taguchi Methods for Robust Design*. ASME Press, 2012.
19. Gen M., Cheng R. *Genetic algorithms and engineering design*. Wiley, 2000, pp. 29–37.
20. Khlifi A., Bouallegue R., Performance analysis of LS and LMMSE channel estimation techniques for LTE downlink systems. *International Journal of Wireless & Mobile Networks (IJWMN)*, 2011, vol. 3, no. 5, pp. 141–149. https://doi.org/10.5121/ijwmn.2011.3511

Литература

1. Bobde S., Phalnikar R. Software restructuring models for object oriented programming languages using the fuzzy based clustering algorithm // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 5. С. 738–747. https://doi.org/10.17586/2226-1494-2021-21-5-738-747
2. Bothra S.K., Singhal S. Nature-inspired metaheuristic scheduling algorithms in cloud: a systematic review // Научно-технический вестник информационных технологий, механики и оптики. 2021. V. 21. № 4. С. 463–472. https://doi.org/10.17586/2226-1494-2021-21-4-463-472
3. Farrokhi F.R., Lozano A., Foschini G.J., Valenzuela R.A. Spectral efficiency of FDMA/TDMA wireless systems with transmit and receive antenna arrays // IEEE Transactions on Wireless Communications. 2002. V. 1. N 4. P. 591–599. https://doi.org/10.1109/TWC.2002.804078
4. Balanis C. Antenna theory: Analysis and design. 4rd ed. New York: John Wiley & Sons, 2018. P. 900–908.
5. Durrani S., Bialkowski M.E. Effect of mutual coupling on the interference rejection capabilities of linear and circular arrays in CDMA systems // IEEE Transactions on Antennas and Propagation. 2004. V. 52. N 4. P. 1130–1134. https://doi.org/10.1109/TAP.2004.825640
6. Piazza D., Kirsch N.J., Forenza A., Heath R.W., Dandekar K.R. Design and evaluation of a reconfigurable antenna array for MIMO systems // IEEE Transactions on Antennas and Propagation. 2008. V. 56. N 3. P. 869–881. https://doi.org/10.1109/TAP.2008.916908
7. Lozano A., Tulino A.M. Capacity of multiple-transmit multiple-receive antenna architectures // IEEE Transactions on Information Theory. 2002. V. 48. N 12. P. 3117–3127. https://doi.org/10.1109/TIT.2002.805084
8. Oyman Ö., Nabar R.U., Bölcskei H., Paulraj A.J. Tight lower bounds on the ergodic capacity of Rayleigh fading MIMO channels // Proc. of the Conference Record / IEEE Global Telecommunications Conference. V. 2. 2002. P. 1172–1176. https://doi.org/10.1109/GLOCOM.2002.1188380
9. Du J., Li Y. Optimization of antenna configuration for MIMO systems // IEEE Transactions on Communications. 2005. V. 53. N 9. P. 1451–1454. https://doi.org/10.1109/TCOMM.2005.855002
10. Waheed U.A., Kishore D.V. Uplink spatial fading correlation of MIMO channel // IEEE Vehicular Technology Conference. 2003. V. 58. N 1. P. 94–98. https://doi.org/10.1109/VETECF.2003.1284985
11. Tsai J.-A., Woerner B.D. The fading correlation function of a circular antenna array in mobile radio environment // Proc. of the IEEE Global Telecommunications Conference. V. 5. 2001. P. 3232–3236. https://doi.org/10.1109/GLOCOM.2001.966023
12. Li X., Nie Z.-P. Spatial fading correlation of circular antenna arrays with Laplacian PAS in MIMO channels // Proc. of the IEEE Antennas and Propagation Society International Symposium. V. 4. 2004. P. 3697–3700. https://doi.org/10.1109/APS.2004.1330149
13. Recioui A., Bentarzi H. Genetic algorithm based MIMO capacity enhancement in spatially correlated channels including mutual coupling // Wireless Personal Communications. 2012. V. 63. N 3. P. 689–701. https://doi.org/10.1007/s11277-010-0159-5
14. Recioui A., Azrar A. Use of genetic algorithms in linear and planar array synthesis based on Schelkunoff method // Microwave and Optical Technology Letters. 2007. V. 49. N 7. P. 1619–1623. https://doi.org/10.1002/mop.22510
15. Montgomery D. C. Design and Analysis of Experiments. New York: Wiley, 1991. P. 28–39.
16. Ross P.J. Taguchi techniques for quality engineering. McGraw-Hill. 2013. P. 43–52.
17. Weng W.C., Yang F., Elsherbeni A.Z. Linear antenna array synthesis using Taguchi's method: A novel optimization technique in electromagnetics // IEEE Transactions on Antennas and Propagation. 2007. V. 55. N 3. P. 723–730. https://doi.org/10.1109/TAP.2007.891548
18. Wu Y. Taguchi Methods for Robust Design. ASME Press, 2012.
19. Gen M., Cheng R. Genetic algorithms and engineering design. Wiley, 2000. P. 29–37.
20. Khlifi A., Bouallegue R., Performance analysis of LS and LMMSE channel estimation techniques for LTE downlink systems // International Journal of Wireless & Mobile Networks (IJWMN). 2011. V. 3. N 5. P. 141–149. https://doi.org/10.5121/ijwmn.2011.3511

Authors

Siraj Pathan — MEng, Research Scholar, Amity University Rajasthan, Jaipur, 303006, India, https://orcid.org/0000-0002-1728-3030, sirajpathan404@gmail.com
Ajit Noonia — PhD, Assistant Professor, Amity University Rajasthan, Jaipur, 303006, India, SC 57221779854, https://orcid.org/0000-0002-3110-4908, ajit009noonia@gmail.com
Mujib Tamboli — PhD, Assistant Professor, Anjuman-I-Islam's Kalsekar Technical Campus, Panvel, 410206, India, https://orcid.org/0000-0002-5891-3713, mujibtamboli@yahoo.co.in
Sunil Pathak — PhD, Associate Professor, Amity University Rajasthan, Jaipur, 303006, India, SC 57194562539, https://orcid.org/0000-0001-7409-9814, sunilpath@gmail.com

Авторы

Патан Сирадж — исследователь, магистр, Амити Университет Раджастана, Джапур, 303006, Индия, https://orcid.org/0000-0002-1728-3030, sirajpathan404@gmail.com
Нуния Аджит — PhD, доцент, доцент, Амити Университет Раджастана, Джапур, 303006, Индия, SC 57221779854, https://orcid.org/0000-0002-3110-4908, ajit009noonia@gmail.com
Тамболи Муджиб — PhD, доцент, доцент, Технический кампус Анджуман-и-Ислам в Калсекаре, Панвел, 410206, Индия, https://orcid.org/0000-0002-5891-3713, mujibtamboli@yahoo.co.in
Патхак Сунил — PhD, доцент, Амити Университет Раджастана, Джапур, 303006, Индия, SC 57194562539, https://orcid.org/0000-0001-7409-9814, sunilpath@gmail.com

Received 21.11.2021
Approved after reviewing 13.01.2022
Accepted 29.01.2022

Статья поступила в редакцию 21.11.2021
Одобрена после рецензирования 13.01.2022
Принята к печати 29.01.2022

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2022-22-1-217-222
УДК 004.58, 004.89

Внедрение системы поддержки принятия решений для повышения качества медицинских данных пациентов с артериальной гипертензией

Михаил Васильевич Ионов[1], Екатерина Владимировна Болгова[2], Надежда Эдвиновна Звартау[3], Наталья Георгиевна Авдонина[4], Марина Андреевна Балахонцева[5]✉, Сергей Валерьевич Ковальчук[6], Александра Олеговна Конради[7]

[1,3,4,7] Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация
[2,5,6] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] Mikeionov90@gmail.com, https://orcid.org/0000-0002-3664-5383
[2] ekaterina_bolgova@itmo.ru, https://orcid.org/0000-0002-6297-1131
[3] n.e.zvartau@gmail.com, https://orcid.org/0000-0001-6533-5950
[4] zybuljak@gmail.com, https://orcid.org/0000-0002-4392-5121
[5] mbalakhontceva@itmo.ru✉, https://orcid.org/0000-0001-7286-3858
[6] kovalchuk@itmo.ru, https://orcid.org/0000-0001-8828-4615
[7] ahleague@mail.ru, https://orcid.org/0000-0001-8169-7812

Аннотация

Цифровизация здравоохранения во многом опирается на аналитику данных, большинство которых содержится в медицинских информационных системах. Такие системы собирают информацию из разнородных источников, в том числе из электронных медицинских карт. Повышение качества данных электронных медицинских карт — современный вызов разработчикам информационных систем для медицинских учреждений. На пути к преодолению проблем человеко-машинного взаимодействия, повышения полноты и достоверности медицинской информации разработана система поддержки принятия решений с расширенным набором вспомогательных функций. В работе исследована применимость существующей системы принятия решений на примере медицинских данных пациентов с артериальной гипертензией. Проведено тестирование системы поддержки принятия решений среди врачей-специалистов. Выполнена оценка влияния предложенной системы на количество ошибок при заполнении электронной медицинской карты. Создан отдельный программный модуль, интегрированный в рабочую версию медицинской информационной системы «qMS» в ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России. Тестовое внедрение системы позволило снизить количество ошибок и повысить удовлетворенность представленной информацией у пациентов с артериальной гипертензией.

Ключевые слова

системы поддержки принятия решений, ценностно-ориентированная медицина, человеко-компьютерное взаимодействие, цифровизация здравоохранения

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 17-15-01177).

Ссылка для цитирования: Ионов М.В., Болгова Е.В., Звартау Н.Э., Авдонина Н.Г., Балахонцева М.А., Ковальчук С.В., Конради А.О. Внедрение системы поддержки принятия решений для повышения качества медицинских данных пациентов с артериальной гипертензией // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 217–222. doi: 10.17586/2226-1494-2022-22-1-217-222

© Ионов М.В., Болгова Е.В., Звартау Н.Э., Авдонина Н.Г., Балахонцева М.А., Ковальчук С.В., Конради А.О., 2022

Implementation of a clinical decision support system to improve the medical data quality for hypertensive patients

Mikhail V. Ionov[1], Ekaterina V. Bolgova[2], Nadezhda E. Zvartau[3], Natalia G. Avdonina[4], Marina A. Balakhontceva[5]✉, Sergey V. Kovalchuk[6], Alexandra O. Konradi[7]

[1,3,4,7] Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation
[2,5,6] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] Mikeionov90@gmail.com, https://orcid.org/0000-0002-3664-5383
[2] ekaterina_bolgova@itmo.ru, https://orcid.org/0000-0002-6297-1131
[3] n.e.zvartau@gmail.com, https://orcid.org/0000-0001-6533-5950
[4] zybuljak@gmail.com, https://orcid.org/0000-0002-4392-5121
[5] mbalakhontceva@itmo.ru✉, https://orcid.org/0000-0001-7286-3858
[6] kovalchuk@itmo.ru, https://orcid.org/0000-0001-8828-4615
[7] ahleague@mail.ru, https://orcid.org/0000-0001-8169-7812

Abstract

The digitalization of healthcare relies heavily on data analytics from medical information systems. Such systems aggregate information from heterogeneous sources, including electronic medical records. Improving the quality of data from electronic medical records is a modern challenge for developers of medical information systems. The authors have designed a decision support system with an expanded set of auxiliary functions to solve the problems of human-computer interaction, increasing the completeness and reliability of medical information. In this paper, the applicability of the existing decision-making system is investigated on the example of medical data of patients with arterial hypertension. The testing of the decision support system among medical specialists was carried out. The impact of the implementation of the system on the number of errors when filling out an electronic medical record was assessed. A software module was created integrated into the working version of the medical information system in the Almazov National Medical Research Centre. Test implementation of the system made it possible to reduce the number of errors and increase satisfaction with the information presented in patients with arterial hypertension.

Keywords

decision support systems, value-based medicine, human-computer interaction, digitalization of healthcare

Acknowledgements

The research was supported by the Russian Science Foundation (project No. 17-15-01177).

For citation: Ionov M.V., Bolgova E.V., Zvartau N.E., Avdonina N.G., Balakhontceva M.A., Kovalchuk S.V., Konradi A.O. Implementation of a clinical decision support system to improve the medical data quality for hypertensive patients. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 217–222 (in Russian). doi: 10.17586/2226-1494-2022-22-1-217-222

В последние десятилетия наблюдаются две взаимосвязанных тенденции, напрямую затрагивающие отрасль здравоохранения: информатизация и автоматизация процессов для повышения эффективности оказания медицинской помощи. Эти тренды отражают глобальный процесс перехода от доказательной к ценностно-ориентированной медицине [1]. Стремление к «ценности», а также персонализация и цифровизация предполагают адекватное человеко-компьютерное взаимодействие, под которым понимается корректное использование медицинскими работниками (в первую очередь, врачами-специалистами) медицинских информационных систем (МИС) [2]. Медицинская аналитика предполагает наличие качественной информации в основной структурной единице МИС, электронной медицинской карте (ЭМК), в которой фиксируется информация медицинского характера из разнородных источников в отдельных временных точках [3, 4]. Сбор массива ЭМК и анализ данных позволяют выработать более персонализированный подход к диагностике и лечению [5–7]. Внедрение ЭМК и анализ их содержимого сопряжены с рядом проблем. В основном они касаются качества данных — их полноты, достоверности, согласованности, точности и единообразия [8]. Ограничения должного анализа данных приводит к снижению темпа реализации новых диагностических и лечебных процедур, ведет к стагнации отрасли, и что более важно — снижению доступности, безопасности и качества получаемой помощи [9, 10]. Важным шагом на пути к преодолению указанных преград стала разработка систем поддержки принятия решений (СППР) с расширенным набором вспомогательных функций. Авторы настоящей работы предполагают, что даже простейшие СППР могут помочь избежать ошибок, опечаток и некорректного ввода данных [11].

На первом аналитическом этапе выполнена оценка полноты и отсутствия противоречий в данных за период 2010–2015 годы (6 лет), извлеченных из ЭМК амбулаторных пациентов с основным диагнозом — артериальная гипертензия. Осуществлен анализ разделов ЭМК с информацией о факторах риска: сердечно-сосудистых осложнений, повреждении органов, обусловленных артериальной гипертензией и метаболических нарушениях (в первую очередь, нарушения углеводного и липидного обменов).

Данные анализировались в соответствии с актуальными клиническими рекомендациями по ведению пациентов с артериальной гипертензией [12]. Выполнен анализ полноты информации (в ЭМК внесена подробная информация о факторах риска сердечно-сосудистых осложнений) и согласованности данных (логика и соответствие рекомендательным документам). Наличие сахарного диабета, гипергликемия натощак, нарушение толерантности к глюкозе и дислипидемия определены

путем автоматизированного интеллектуального анализа данных раздела ЭМК «Диагноз амбулаторный». Информация о назначении дополнительных обследований и лекарственных препаратов была извлечена из раздела «Рекомендации».

В результате аналитического этапа установлено 9 классов основных ошибок (n — число пациентов в каждом классе):

1) дислипидемия была указана в диагнозе в 49,9 % (n = 16 037) случаев, но только в 58,5 % (n = 9389) назначено соответствующее лечение;
2) 67,1 % (n = 10 084) пациентам без дислипидемии не была назначена диагностика, а 14,8 % (n = 1491) — рекомендована терапия без должных показаний;
3) в 2,4 % (n = 787) случаях, согласно лабораторным тестам, диагноз дислипидемии мог быть зафиксирован, но отсутствовал;
4) нарушение толерантности к глюкозе и гипергликемия натощак были установлены в 4,5% (n = 1432) случаях, но не подтверждены результатами лабораторных тестов;
5) в 5,4 % (n = 1551) и 1,7 % (n = 483) случаях рекомендована оценка гликемии натощак и гликированного гемоглобина соответственно;
6) информация об артериальном давлении отсутствовала в 11 % (n = 3283) случаях;
7) не было указаний об повреждении органов, обусловленных артериальной гипертензией (42–99 % в ЭМК в зависимости от сложности показателей), в том числе данных о предыдущих исследованиях. В разделе «Рекомендации» отсутствовали назначения по выполнению соответствующих обследований;
8) в 94,5 % случаев сахарного диабета диагноз не был подтвержден лабораторно;
9) в 12 % (n = 3859) всех обработанных ЭМК фиксировались опечатки и нарушение правил назначения лекарственных препаратов.

На основе установленных классов указанных ошибок выявлены последовательные тренды улучшения качества вносимых данных (формирование кривых обучения, где p — значение уровня значимости, R^2 — коэффициент детерминации). С 2010 по 2012 годы сократились ошибки классов 1 и 2 на 9 % (p = 0,01) и класса 7 — на 34 % (R^2 = 0,93; p = 0,04). Однако, начиная с 2013 года, доля пропусков в указанных областях увеличилась на 42 % (R^2 = 0,99; p = 0,01) и 64 % (R^2 = 0,95; p = 0,14) соответственно. Данная волнообразность обусловлена возросшим количеством первичных визитов и врачебной нагрузкой в последние три анализируемых года, так как средние показатели коэффициентов пропорциональности были практически идентичны для ошибок классов 1, 2 и 7.

В рамках экспериментальной части разработана и внедрена СППР, а также проведена оценка ее влияния на количество ошибок при заполнении ЭМК. Система интегрирована в рабочую версию МИС «qMS» в Национальном медицинском исследовательском центре имени В.А. Алмазова, как отдельный программный модуль. В СППР использован массив данных ЭМК из МИС, на их основании расширен интерфейс пользователя, сгенерированы подсказки и уведомления. СППР включает в себя: модули интеграции с МИС, обработки данных и уведомлений, а также интерфейсы работы для врача-специалиста, эксперта и медицинского ад-

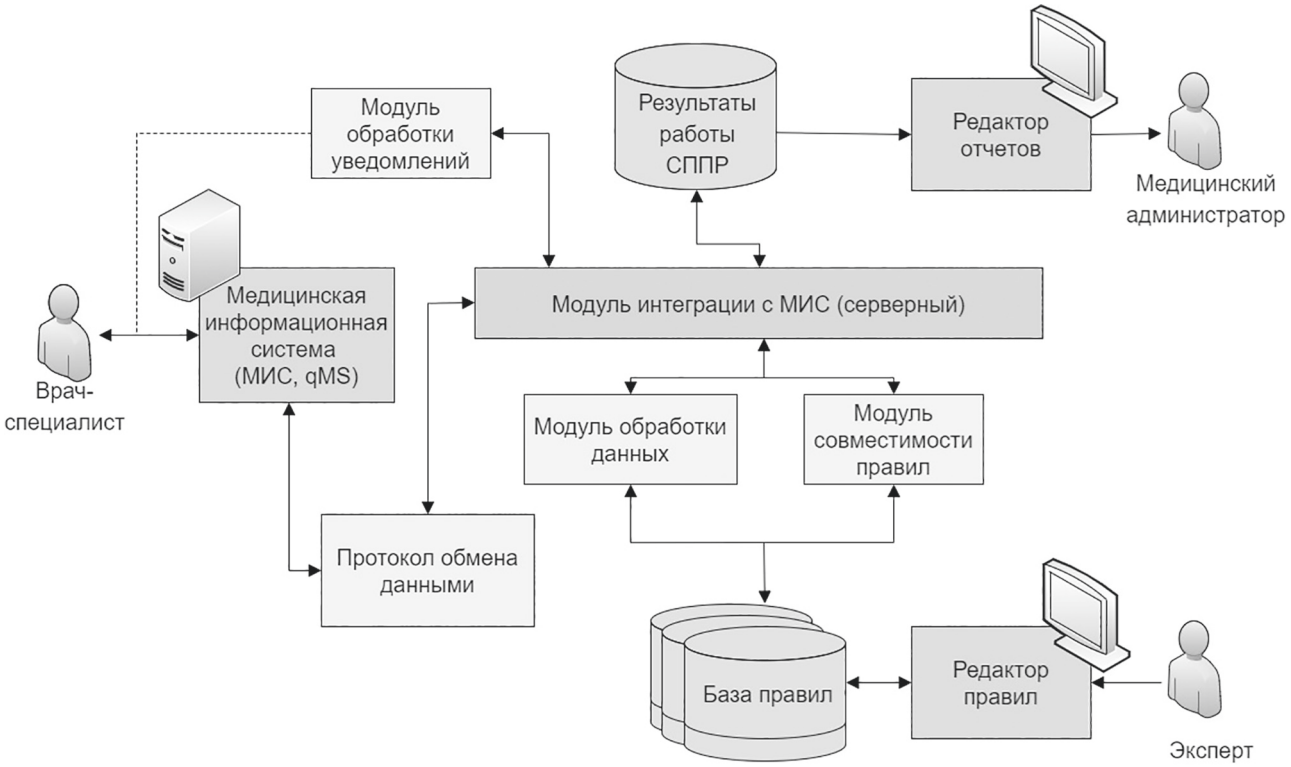


Рис. 1. Общая схема системы поддержки принятия решений, интегрируемой с медицинской информационной системой

Fig. 1. General scheme of the proposed decision support system

министратора. На рис. 1 представлена общая схема архитектуры разработанной СППР.

Введенная вручную в МИС информация выступает основой для базы знаний и правил. На этих данных зациклены процессы контроля знаний, автоматического и полуавтоматического каскадного модельного обучения. Для коррекции ошибок заполнения использован алгоритм формализации естественного языка, а также машинное обучение, включая искусственные нейронные сети, деревья решений и другие классификационные алгоритмы.

В настоящей работе СППР ограничена фиксацией ошибок и уведомлением врачей при погрешностях в заполнении следующих разделов ЭМК: «общий осмотр» (выделение заведомо неправильно вводимых данных); «рекомендации» (анализ электронной библиотеки назначаемых препаратов с контролем их совместимости). В исследовании приняли участие 7 врачей-кардиологов, которые специализируются на ведении пациентов с артериальной гипертензией. Модуль СППР был внедрен на 10 недель. Случайным образом врачи были разделены на две подгруппы: работающие с включенными уведомлениями ($n = 4$) и работающие в режиме «фиксации» без доступа к содержанию уведомлений ($n = 3$). Выполнен анализ 32 158 ЭМК (медиана возраста пациентов — 56 лет; 37 % из них — мужчины). Среди этих пациентов доля лиц с ожирением — 27 % мужчин и 33 % женщин; с дислипидемией 81 %; с нарушениями обмена углеводов — 84 % случаев.

На рис. 2 представлены лепестковые диаграммы результатов анализа для пропусков (_miss) и некорректных значений (_err) в данных для параметров роста, веса, артериального давления (АД), частоты сердечных сокращений (ЧСС) и обхвата талии (ОТ). За 10 недель зарегистрировано 1280 заполненных ЭМК в активной группе (с уведомлениями), 1060 — в контрольной группе, $p = 0{,}33$. При анализе ЭМК в контрольной группе доля ошибок составила 63,7 %, что на 14,6 % больше, чем в группе вмешательства ($p_{\text{межгрупповая}} < 0{,}001$). Из 49,1 % ошибок в активной группе к моменту окончания визитов неисправленными остались 8,7 %. Спустя одну неделю после начала эксперимента отмечено увеличение ошибок (выход количества ошибок на плато), что связано с фактом привыкания к системе и игнорированием уведомлений («событийная усталость»). По результатам интервью с врачами активной группы отмечена полезность и быстрая адаптируемость к работе с СППР. Негативные комментарии касались большого количества всплывающих окон, которые в моменты пиковой нагрузки усложняют работу.

В работе выявлено, что ввод данных в ЭМК без СППР не обеспечивает должного качества и согласованности данных, особенно по отношению к факторам риска сердечно-сосудистых осложнений, повреждений органов, обусловленных артериальной гипертензией и метаболическим нарушениям. Полнота и точность данных, вносимых в ЭМК, формирует основу значительных положительных преимуществ не только здесь и сейчас, но и создает основу для более качественной и «ценной» помощи в будущем. Некорректное сообщение данных о факторах риска и повреждений органов, обусловленных артериальной гипертензией может привести к значительным негативным последствиям для пациента (позднее начало лечение, недостаточно агрессивное лечение вследствие некорректной оценки ри-

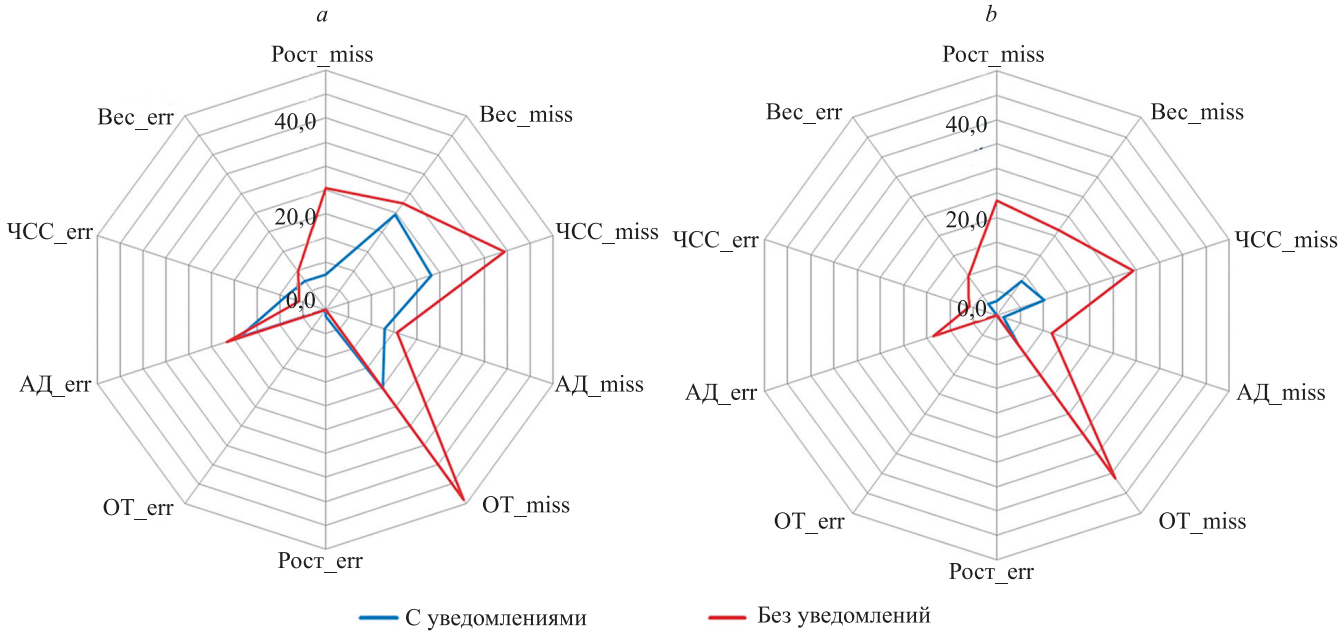


Рис. 2. Доля пользовательских ошибок (в процентах от 0 до 50) за весь период тестирования: общий процент зафиксированных ошибок (*a*); процент ошибок, оставшихся в медицинской информационной системе (после исправлений) (*b*)

Fig. 2. The ratio of user errors for the entire testing period: the total number of errors recorded (*a*); the percentage of errors remaining in the medical information system (after corrections) (*b*)

ска сердечно-сосудистых осложнений) [13]. Обращает на себя внимание, что начало активной эксплуатации МИС (и следовательно, ЭМК) первоначально ассоциируется с прогрессирующим увеличением полноты и содержательности данных, отражая присутствие кривой обучения. Излишняя рабочая нагрузка негативно сказывается на структурированности данных. Этот факт может свидетельствовать о снижении качества оказания услуг в пользу увеличения их объема [14]. Важно подчеркнуть, что качество оказанной помощи может существенно повлиять на результат лечения; в течение последнего десятилетия многие системы здравоохранения в Западном мире перешли на систему оплаты исходя из качества оказанной помощи вместо оплаты исключительно по ее объему [15, 16].

Простая СППР, интегрированная в МИС, помогает связать между собой различные источники данных, применяя приемы семантического описания и объединения гетерогенной информации. В этой связи уместно вновь упомянуть о необходимости наличия исходного некоторого количества качественно заполненных ЭМК для того, чтобы максимизировать опыт от применения СППР и ее эффективность, что может приблизить исследователей к более надежным и валидным данным реальной клинической практики [17]. Использование СППР с минимальным функционалом способствует значительному снижению количества ошибок, но предугадать вероятность сохранения такой положительной тенденции затруднительно. Например, результаты аналитической части допускают возникновение внешних причин, мешающих адекватному ручному вводу данных. Основываясь на обратной связи от врачей активной группы (с уведомлениями), продолжение дальнейшей работы может быть направлено на объективизацию мнений специалистов о работе с цифровыми помощниками; с валидированными опросниками для оценки простоты использования цифровых технологий (к примеру, *Technology Acceptance Model*) [18]. Авторы настоящей работы ориентированы на проведение рандомизированного клинического исследования с оценкой клинической и экономической эффективности внедрения СППР.

Показано низкое качество данных ЭМК пациентов с артериальной гипертензией в части факторов риска сердечно-сосудистых осложнений, повреждений органов, обусловленных артериальной гипертензией и метаболических нарушений: часто данные неполны или противоречивы. Выявлена параболическая динамика качества данных в течение 6 лет: положительный тренд первых трех лет нивелирован к концу наблюдения. Интеграция СППР с минимальным функционалом позволяет существенно повысить и структурированность данных в ЭМК и опыт использования МИС.

Литература

1. Шляхто Е.В., Конради А.О., Звартау Н.Э., Ратова Л.Г. Ценностная медицина, или Value-based medicine. СПб.: ООО «Инфо-ра», 2019. 92 с.
2. Cowie M.R., Bax J., Bruining N., Cleland J.G.F., Koehler F., Malik M., Pinto F., Van Der Velde E., Vardas P. e-Health: a position statement of the European Society of Cardiology // European Heart Journal. 2016. V. 37. N 1. P. 63–66. https://doi.org/10.1093/eurheartj/ehv416
3. Coorevits P., Sundgren M., Klein G.O., Bahr A., Claerhout B., Daniel C., Dugas M., Dupont D., Schmidt A., Singleton P., De Moor G., Kalra D. Electronic health records: new opportunities for clinical research // Journal of Internal Medicine. 2013. V. 274. N 6. P. 547–560. https://doi.org/10.1111/joim.12119
4. Jensen P.B., Jensen L.J., Brunak S. Mining electronic health records: towards better research applications and clinical care // Nature Reviews Genetics. 2012. V. 13. N 6. P. 395–405. https://doi.org/10.1038/nrg3208
5. Dugas M., Lange M., Müller-Tidow C., Kirchhof P., Prokosch H.-U. Routine data from hospital information systems can support patient recruitment for clinical studies // Clinical Trials. 2010. V. 7. N 2. P. 183–189. https://doi.org/10.1177/1740774510363013
6. Prokosch H.U., Ganslandt T. Perspectives for Medical Informatics. Reusing the electronic medical record for clinical research // Methods of Information in Medicine. 2009. V. 48. N 1. P. 38–44. https://doi.org/10.3414/ME9132
7. Turisco F., Keogh D., Stubbs C., Glaser J., Crowley W.F. Current status of integrating information technologies into the clinical research enterprise within US academic health centers: strategic value and opportunities for investment // Journal of Investigative Medicine. 2005. V. 53. N 8. P. 425–433. https://doi.org/10.2310/6650.2005.53806
8. Weiskopf N.G., Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research // Journal of the American Medical Informatics Association. 2013. V. 20. N 1. P. 144–151. https://doi.org/10.1136/amiajnl-2011-000681
9. Birkhead G.S., Klompas M., Shah N.R. Uses of electronic health records for public health surveillance to advance public health // Annual Review of Public Health. 2015. V. 36. P. 345–359. https://doi.org/10.1146/annurev-publhealth-031914-122747

References

1. Shliakhto E.V., Konradi A.O., Zvartau N.E., Ratova L.G. *Value-based medicine*. St. Petersburg, Info-ra Publ., 2019, 92 p. (in Russian)
2. Cowie M.R., Bax J., Bruining N., Cleland J.G.F., Koehler F., Malik M., Pinto F., Van Der Velde E., Vardas P. e-Health: a position statement of the European Society of Cardiology. *European Heart Journal*, 2016, vol. 37, no. 1, pp. 63–66. https://doi.org/10.1093/eurheartj/ehv416
3. Coorevits P., Sundgren M., Klein G.O., Bahr A., Claerhout B., Daniel C., Dugas M., Dupont D., Schmidt A., Singleton P., De Moor G., Kalra D. Electronic health records: new opportunities for clinical research. *Journal of Internal Medicine*, 2013, vol. 274, no. 6, pp. 547–560. https://doi.org/10.1111/joim.12119
4. Jensen P.B., Jensen L.J., Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 2012, vol. 13, no. 6, pp. 395–405. https://doi.org/10.1038/nrg3208
5. Dugas M., Lange M., Müller-Tidow C., Kirchhof P., Prokosch H.-U. Routine data from hospital information systems can support patient recruitment for clinical studies. *Clinical Trials*, 2010, vol. 7, no. 2, pp. 183–189. https://doi.org/10.1177/1740774510363013
6. Prokosch H.U., Ganslandt T. Perspectives for Medical Informatics. Reusing the electronic medical record for clinical research. *Methods of Information in Medicine*, 2009, vol. 48, no. 1, pp. 38–44. https://doi.org/10.3414/ME9132
7. Turisco F., Keogh D., Stubbs C., Glaser J., Crowley W.F. Current status of integrating information technologies into the clinical research enterprise within US academic health centers: strategic value and opportunities for investment. *Journal of Investigative Medicine*, 2005, vol. 53, no. 8, pp. 425–433. https://doi.org/10.2310/6650.2005.53806
8. Weiskopf N.G., Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 2013, vol. 20, no. 1, pp. 144–151. https://doi.org/10.1136/amiajnl-2011-000681
9. Birkhead G.S., Klompas M., Shah N.R. Uses of electronic health records for public health surveillance to advance public health. *Annual Review of Public Health*, 2015, vol. 36, pp. 345–359. https://doi.org/10.1146/annurev-publhealth-031914-122747

10. Zhou L., Soran C.S., Jenter C.A., Volk L.A., Orav E.J., Bates D.W., Simon S.R. The relationship between electronic health record use and quality of care over time // Journal of the American Medical Informatics Association. 2009. V. 16. N 4. P. 457–464. https://doi.org/10.1197/jamia.M3128
11. Старков Е.Ф. Система поддержки принятия решений в медицине // Вестник новых медицинских технологий. 2006. № 2. С. 23–24.
12. Williams B., Mancia G., Spiering W., Agabiti Rosei E., Azizi M., Burnier M. et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension // European Heart Journal. 2018. V. 39. N 33. P. 3021–3104. https://doi.org/10.1093/eurheartj/ehy339
13. Wang C., Yuan Y., Zheng M., Pan A., Wang M., Zhao M., Li Y., Yao S., Chen S., Wu S., Xue H. Association of age of onset of hypertension with cardiovascular diseases and mortality // Journal of the American College of Cardiology. 2020. V. 75. N 23. P. 2921–2930. https://doi.org/10.1016/j.jacc.2020.04.038
14. Kaushal R., Shojania K.G., Bates D.W. Effects of computerized physician order entry and clinical decision support systems on medication safety: A systematic review // Archives of Internal Medicine. 2003. V. 163. N 12. P. 1409–1416. https://doi.org/10.1001/archinte.163.12.1409
15. Ray J.C., Kusumoto F. The transition to value-based care // Journal of Interventional Cardiac Electrophysiology. 2016. V. 47. N 1. P. 61–68. https://doi.org/10.1007/s10840-016-0166-x
16. Christensen T.J. A framework for guiding efforts to reward value instead of volume // International Journal of Health Economics and Management. 2016. V. 16. N 2. P. 175–187. https://doi.org/10.1007/s10754-015-9178-9
17. Sim I., Gorman P., Greenes R.A., Haynes R.B., Kaplan B., Lehmann H., Tang P.C. Clinical decision support systems for the practice of evidence-based medicine // Journal of the American Medical Informatics Association. 2001. V. 8. N 6. P. 527–534. https://doi.org/10.1136/jamia.2001.0080527
18. Sauro J., Dumas J.S. Comparison of three one-question, post-task usability questionnaires // Proc. of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI). 2009. P. 1599–1608. https://doi.org/10.1145/1518701.1518946

10. Zhou L., Soran C.S., Jenter C.A., Volk L.A., Orav E.J., Bates D.W., Simon S.R. The relationship between electronic health record use and quality of care over time. *Journal of the American Medical Informatics Association*, 2009, vol. 16, no. 4, pp. 457–464. https://doi.org/10.1197/jamia.M3128
11. Ctarkov E.F. The support system of the making a decision in medicine. *Journal of New Medical Technologies*, 2006, no. 2, pp. 23–24. (in Russian)
12. Williams B., Mancia G., Spiering W., Agabiti Rosei E., Azizi M., Burnier M. et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. *European Heart Journal*, 2018, vol. 39, no. 33, pp. 3021–3104. https://doi.org/10.1093/eurheartj/ehy339
13. Wang C., Yuan Y., Zheng M., Pan A., Wang M., Zhao M., Li Y., Yao S., Chen S., Wu S., Xue H. Association of age of onset of hypertension with cardiovascular diseases and mortality. *Journal of the American College of Cardiology*, 2020, vol. 75, no. 23, pp. 2921–2930. https://doi.org/10.1016/j.jacc.2020.04.038
14. Kaushal R., Shojania K.G., Bates D.W. Effects of computerized physician order entry and clinical decision support systems on medication safety: A systematic review. *Archives of Internal Medicine*, 2003, vol. 163, no. 12, pp. 1409–1416. https://doi.org/10.1001/archinte.163.12.1409
15. Ray J.C., Kusumoto F. The transition to value-based care. *Journal of Interventional Cardiac Electrophysiology*, 2016, vol. 47, no. 1, pp. 61–68. https://doi.org/10.1007/s10840-016-0166-x
16. Christensen T.J. A framework for guiding efforts to reward value instead of volume. *International Journal of Health Economics and Management*, 2016, vol. 16, no. 2, pp. 175–187. https://doi.org/10.1007/s10754-015-9178-9
17. Sim I., Gorman P., Greenes R.A., Haynes R.B., Kaplan B., Lehmann H., Tang P.C. Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, 2001, vol. 8, no. 6, pp. 527–534. https://doi.org/10.1136/jamia.2001.0080527
18. Sauro J., Dumas J.S. Comparison of three one-question, post-task usability questionnaires. *Proc. of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*, 2009, pp. 1599–1608. https://doi.org/10.1145/1518701.1518946

Авторы

Ионов Михаил Васильевич — кандидат медицинских наук, младший научный сотрудник, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 57200520709, https://orcid.org/0000-0002-3664-5383, Mikeionov90@gmail.com

Болгова Екатерина Владимировна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57211535770, https://orcid.org/0000-0002-6297-1131, ekaterina_bolgova@itmo.ru

Звартау Надежда Эдвиновна — кандидат медицинских наук, доцент, старший научный сотрудник, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 6506439053, https://orcid.org/0000-0001-6533-5950, n.e.zvartau@gmail.com

Авдонина Наталья Георгиевна — научный сотрудник, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 57201382884, https://orcid.org/0000-0002-4392-5121, zybuljak@gmail.com

Балахонцева Марина Андреевна — кандидат технических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57211535770, https://orcid.org/0000-0001-7286-3858, mbalakhontceva@itmo.ru

Ковальчук Сергей Валерьевич — кандидат технических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 55382199400, https://orcid.org/0000-0001-8828-4615, kovalchuk@itmo.ru

Конради Александра Олеговна — доктор медицинских наук, член-корр. РАН, профессор, заведующий научно-исследовательским отделом, Национальный медицинский исследовательский центр имени В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, SC 7004144504, https://orcid.org/0000-0001-8169-7812, ahleague@mail.ru

Authors

Mikhail V. Ionov — PhD, Junior Researcher, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 57200520709, https://orcid.org/0000-0002-3664-5383, Mikeionov90@gmail.com

Ekaterina V. Bolgova — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57211535770, https://orcid.org/0000-0002-6297-1131, ekaterina_bolgova@itmo.ru

Nadezhda E. Zvartau — PhD, Associate Professor, Senior Researcher, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 6506439053, https://orcid.org/0000-0001-6533-5950, n.e.zvartau@gmail.com

Natalia G. Avdonina — Scientific Researcher, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 57201382884, https://orcid.org/0000-0002-4392-5121, zybuljak@gmail.com

Marina A. Balakhontceva — PhD, Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57211535770, https://orcid.org/0000-0001-7286-3858, mbalakhontceva@itmo.ru

Sergey V. Kovalchuk — PhD, Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 55382199400, https://orcid.org/0000-0001-8828-4615, kovalchuk@itmo.ru

Alexandra O. Konradi — D.Sc., Corresponding Member of RAS, Professor, Head of Department, Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation, SC 7004144504, https://orcid.org/0000-0001-8169-7812, ahleague@mail.ru

Статья поступила в редакцию 30.11.2021
Одобрена после рецензирования 20.12.2021
Принята к печати 26.01.2022

Received 30.11.2021
Approved after reviewing 20.12.2021
Accepted 26.01.2022

ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ СТАТЕЙ

Журнал публикует научные обзоры, полнотекстовые статьи и краткие сообщения.

Полнотекстовая статья должна иметь четкую структуру, включающую в себя аннотацию, ключевые слова, введение, несколько содержательных разделов и заключение.

В аннотации, рассчитанной на самый широкий круг читателей, необходимо в объеме 200–400 слов в форме краткого реферата (Abstract) изложить научное содержание статьи: предмет, цель работы, метод или методология проведения работы, краткое описание эксперимента, полученные результаты, рекомендации по их применению.

Ключевые слова должны отражать информацию, представленную в статье. Основное ключевое слово указывается первым.

Во введении необходимо представить содержательную постановку рассматриваемого вопроса, провести краткий анализ известных из научной литературы решений (со ссылками на источники), дать критику их недостатков, показать научную новизну и преимущество (особенности) предлагаемого подхода.

В основном тексте статьи должна быть представлена строгая постановка решаемой задачи, изложены и обстоятельно разъяснены (доказаны) полученные утверждения и выводы, приведены результаты экспериментальных исследований или математического моделирования, иллюстрирующие сделанные утверждения. Основной текст статьи должен быть разбит на содержательные разделы (2–3 раздела).

В заключении необходимо кратко сформулировать основные результаты, прокомментировать их и, если возможно, указать направления дальнейших исследований и области применения.

Пристатейный список литературы (рекомендуется): для обзорной статьи — не менее 50, для полнотекстовой статьи — не менее 15, для краткого сообщения — не менее 8 литературных источников.

Ссылки на стандарты и иные нормативные документы, а также на неперсонифицированные интернет-ресурсы в список литературы не вносятся, а оформляются в виде сносок.

Объем обзорной статьи предварительно согласовывается с редакцией.

Объем полнотекстовой статьи, включая иллюстрации, таблицы и список литературы, не должен превышать 8 страниц машинописного текста (с рисунками и таблицами), шрифт 12 pt, один интервал.

Объем краткого сообщения — до 3 страниц, шрифт 12 pt, один интервал. Рубрикация текста не требуется. Статьи принимаются в электронном виде (ntvitmo@itmo.ru). В распечатанном виде с подписями авторов материалы представляются в редакцию по запросу редакции после прохождения процедуры рецензирования.

Комплект документов должен включать:

- текст статьи с заверстанными рисунками и таблицами;
- формы сведений об авторах (на каждого автора); допускается указание нескольких мест работы автора;
- файлы с рисунками к статье в оригинальном формате (предпочтительно JPEG) с максимальным разрешением; допускается представление цветных рисунков, если в черно-белом варианте теряется полезная информация;
- лицензионное соглашение;
- согласие на обработку персональных данных.

REQUIREMENTS FOR EXECUTION OF PAPERS

Scientific reviews, full-text and brief papers are published.

A full-text paper should have a well-defined structure, including an abstract, keywords, introduction, several substantive chapters and a conclusion.

An abstract is intended for general public. It is necessary to set forth the scientific content of the paper limited by 200–400 words in the form of a brief abstract: the subject, the work objective, the method or methodology of work, a brief description of the experiment, obtained results, recommendations for their application.

Keywords should be connected with the information presented in the paper. The main keyword is given first.

The introduction should contain a meaningful statement of the issue in question, a brief analysis of the solutions known from scientific literature (with references to sources), a criticism of their shortcomings, scientific novelty and advantages (features) of the proposed approach.

The main body of the paper should represent exact statement of the problem being solved, obtained assertions and conclusions are to be set forth and explained in detail (proved), the results of experimental studies or mathematical modeling should be given to illustrate the statements made. The main text of the paper needs to be split up into meaningful sections (2–3 sections).

In conclusion it is necessary to summarize the main results, comment on them and, if possible, to indicate the areas of future research and application.

A list of references is recommended: for a review paper — at least 50 items, for full-text paper — not less than 15, for a brief paper — at least 8 references. References to standards and other regulatory documents, as well as to Internet resources, are not included in a reference list and are documented as footnotes.

The length of a review paper is coordinated with the editorial board beforehand.

The length of a full-text paper, including illustrations, tables and references, should not exceed 8 pages of typewritten text (figures and tables included), font 12 pt, single-spaced.

The length of a brief paper is up to 3 pages, font 12 pt, single-spaced.

No rubrication is required.

Papers are accepted in electronic form (ntvitmo@itmo.ru).

Materials are submitted to the editors on request after the procedure of reviewing in a printed form signed by the authors.

A set of documents should include:

- The text of the paper with make-up figures and tables;
- The forms with information about the authors (for each author), several employers can be specified by the author;
- Files with pictures to the paper in their original format (preferably JPEG) with the maximum resolution; should useful information is lost in the black and white variant, color drawings are permitted;
- Publishers license agreement;
- Consent to the processing of personal data.

SCIENTIFIC AND TECHNICAL JOURNAL OF ITMO UNIVERSITY

INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS

2022, VOLUME 22, NUMBER 1 (JANUARY–FEBRUARY) ISSN 2226-1494 (PRINT), 2500-0373 (ONLINE)

OPTICAL ENGINEERING

A study of a silicone film deposited on quartz glass under laser radiation — *Belikov A.V., Klochkov I.S., Alekseev I.V., Kapralov S.A.* — *1*

Optical composites based on organic polymers and semiconductor pigments — *Volynkin V.M., Evstropiev S.K., Bulyga D.V., Morkovsky A.V., Pashin S.S., Dukelsky K.V., Burdin A.V., Bondarenko I.B.* — *10*

AUTOMATIC CONTROL AND ROBOTICS

A new algorithm for the identification of sinusoidal signal frequency with constant parameters — *Nguyen Kh.T., Vlasov S.M., Pyrkin A.A., Popkov I.V.* — *18*

MATERIAL SCIENCE AND NANOTECHNOLOGIES

A study of silicon *p-n* structures with mono- and multifacial photosensitive surfaces (in English) — *Mirzaalimov A., Gulomov J., Aliev R., Mirzaalimov N., Aliev S.* — *25*

COMPUTER SCIENCE

Detection of yawning in driver behavior based on a convolutional neural network — *Lashkov I.B.* — *33*

A Game Theory approach for communication security and safety assurance in cyber-physical systems with Reputation and Trust-based mechanisms (in English) — *Viksnin I.I., Marinenkov E.D., Chuprov S.S.* — *47*

A study of the influence of human factors on the speed of urban rail transport — *Lyakhovenko Ju.A., Popov I.Yu.* — *60*

An algorithm for detecting RFID duplicates — *Voloshina N.V., Lavrinovich A.A.* — *67*

Reduction of LSB detectors set with definite reliability — *Solodukha R.A., Perminov G.V., Atlasov I.V.* — *74*

Classification of objects in images with distortions based on a two-stage topological analysis — *Eremeev S.V., Abakumov A.V.* — *82*

Dimensionality reduction of the attributes using fuzzy optimized independent component analysis for a Big Data Intrusion Detection System (in English) — *Aswanandini R., Deepa Ch.* — *93*

An optimal swift key generation and distribution for QKD (in English) — *Suma M.R., Madhumathy P.* — *101*

A study of vectorization methods for unstructured text documents in natural language according to their influence on the quality of work of various classifiers — *Shadsky V.V., Sizonenko A.B., Chekmarev M.A., Shishkov A.V., Isakin D.A.* — *114*

Recognition of the emotional state based on a convolutional neural network — *Soma G.M., Kadnova A.M.* — *120*

Intellectualization of personnel development management in high-tech service-oriented companies — *Beresnev A.D.* — *127*

MODELING AND SIMULATION

A study of the efficiency of the magnetic compass correction system — *Gryazin D.G., Paderina T.V.* — *138*

A new analytical model of drain current and small signal parameters for AlGaN-GaN high-electron-mobility transistors (in English) — *Farti A., Touhami A.* — *147*

Imputation and system modeling of acid-base state parameters for different groups of patients (in English) — *Kurapeev D.I., Lushnov M.S., Man T., Zhukova N.A.* — *155*

Construction of movement trajectories for objects based on the Dubins car problem, taking into account constant external influences — *Khabarov S.P., Shilkina M.L.* — *167*

A mathematical model of an epidemic with an arbitrary law of recovery — *Semenov V.K., Ivanova N.B.* — *179*

Simulation of the pulsed outflow of air and fine powder mixture, partially filling the discharge channel — *Sadin D.V.* — *187*

Vectorized numerical algorithms for the solution of continuum mechanics problems — *Brykov N.A., Volkov K.N., Emelyanov V.N.* — *193*

A comparative analysis of computational intelligence algorithms for estimation of LTE channels (in English) — *Pathan S., Noonia A., Tamboli M., Pathak S.* — *206*

BRIEF PAPERS

Implementation of a clinical decision support system to improve the medical data quality for hypertensive patients — *Ionov M.V., Bolgova E.V., Zvartau N.E., Avdonina N.G., Balakhontceva M.A., Kovalchuk S.V., Konradi A.O.* — *217*