УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2022 Том 22 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2022 Vol. 22 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ОБЗОРНАЯ СТАТЬЯ

REVIEW PAPER

doi: 10.17586/2226-1494-2022-22-3-415-432

УДК 004.932.2

Методы аудиовизуального распознавания людей в масках

Кирилл Эдгарович Косулин[1]✉, Алексей Анатольевич Карпов[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[1] Филиал ООО «Люксофт Профеншнл» в городе Санкт-Петербурге, Санкт-Петербург, 195027, Российская Федерация
[2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] leliclalelic@yandex.ru✉, https://orcid.org/0000-0002-1324-2813
[2] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Аннотация

Предмет исследования. В современном мире очень распространены случаи ношения людьми различных масок, респираторов и одежды на лице. Начавшаяся в 2019 году пандемия новой коронавирусной инфекции существенно увеличила применимость масок в общественных местах. Наиболее эффективными способами бесконтактного распознания личности являются методы идентификации и верификации человека по изображению лица и по записи голоса. Автоматические системы распознавания личности столкнулись с новыми проблемами из-за перекрытия большей части лица маской. Наличие данной проблемы определяет актуальность исследований в области распознания лиц в масках. Предмет исследования работы — системы и корпусы данных для распознавания личности людей в масках. **Метод.** Рассмотрены и проанализированы основные современные подходы и методы распознавания личности людей в масках, использующие изображения лица, записи голоса человека и аудиовизуальные методы. Приведен сравнительный анализ существующих корпусов данных, содержащих изображения и записи голосов людей, необходимые для создания систем распознавания личности. **Основные результаты.** Результаты анализа показали, что среди методов, использующих изображения лиц, наиболее эффективными являются методы, построенные на основе сверточных нейронных сетей, которые применяют область маски для извлечения признаков о геометрии лица. Популярные методы на основе x-векторов показали незначительное падение эффективности, что позволяет сделать вывод об их применимости в задачах распознавания личности говорящего в маске. **Практическая значимость.** На основании полученных выводов сформулированы требования к перспективным системам распознавания личности и определены актуальные направления для дальнейших исследований в данной области.

Ключевые слова

распознавание личности, лицевая биометрия, голосовая биометрия, медицинские маски, средства индивидуальной защиты, аудиовизуальные характеристики, объединение информации

Благодарности

Работа выполнена при поддержке фонда РФФИ (проект № 20-04-60529), Совета по грантам Президента РФ (грант № НШ-17.2022.1.6), а также в рамках бюджетной темы (№ 0073-2019-0005).

Ссылка для цитирования: Косулин К.Э., Карпов А.А. Методы аудиовизуального распознавания людей в масках // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 3. С. 415–432. doi: 10.17586/2226-1494-2022-22-3-415-432

Methods for audiovisual recognition of people in masks

Kirill E. Kosulin[1]✉, Alexey A. Karpov[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation
[1] Luxoft Branch, Saint Petersburg, 195027, Russian Federation
[2] Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation

[1] leliclalelic@yandex.ru✉, https://orcid.org/0000-0002-1324-2813
[2] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Abstract
In the modern world, wearing masks, respirators and facial clothes is very popular. The novel coronavirus pandemic that began in 2019 has also significantly increased the applicability of masks in public places. The most affective person recognition methods are identification by face image and voice recording. However, person recognition systems are facing new challenges due to masks covering most of the subject's face. Existence of new problems for intelligent systems determines the relevance of masked person recognition systems research, therefore the subject of the study is the systems and datasets for masked people recognition. The article discusses analysis of the main approaches to masked people identity recognition: masked face recognition, masked voice recognition and audiovisual methods. In addition, this article includes comparative analysis of images and recordings datasets required for person recognition systems. The results of the study showed that among the methods that use face images the most effective are methods based on convolutional neural networks and the mask area feature extraction. The methods of x-vector analysis showed a slight drop in efficiency which allows us to conclude that they are applicable in the tasks of recognizing the identity of a speaker in a mask. Results of this study help with formulation of requirements for perspective masked person recognition systems and determining directions for further research.

Keywords
person recognition, facial biometrics, voice biometrics, medical masks, personal protective equipment, audiovisual features, information fusion

Acknowledgements
This work was partially supported by the RFBR (project No. 20-04-60529), by the Council for Grants of the President of Russia (grant No. NSH-17.2022.1.6), as well as by the Russian state research (No. 0073-2019-0005).

For citation: Kosulin K.E., Karpov A.A. Methods for audiovisual recognition of people in masks. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 3, pp. 415–432 (in Russian). doi: 10.17586/2226-1494-2022-22-3-415-432

Введение

Распознавание личности по лицу и голосу является самым популярным и перспективным направлением развития биометрических технологий. По данным J'son & Partners Consulting к концу 2018 года доля технологий распознавания лиц и голоса в общем объеме российского биометрического рынка составила почти 85 %, а в течение четырех лет до этого этот сегмент показывал рост на уровне 106,7 % в год [1]. Идентификация личности по лицу и голосу используется в интеллектуальных системах смартфонов, банков и предприятий. Важной частью поимки преступников в современном мире является идентификация личности с помощью общественных камер видеонаблюдения. Для распознавания преступников могут быть использованы записи голоса, реальные изображения и фотороботы [2].

В реальных условиях лицо человека может быть закрыто маской или различными предметами одежды. Преступники часто намерено скрывают свое лицо, как и во время совершения преступления, так и при нахождении на публике. Ношение респираторов и средств индивидуальной защиты обязательно для определенных рабочих и служащих, что может препятствовать распространению биометрических систем на некоторых предприятиях и затруднить распознавание личности этих людей в общественных местах. Часто в северных регионах в холодное время лицо может быть закрыто шарфами, балаклавой или другой одеждой, а в некоторых восточных странах женщины обязаны носить одежду частично или полностью закрывающее лицо.

В начале 2020 года мир столкнулся с новой пандемией коронавирусной инфекции COVID-19. Факторы отсутствия лекарств и постоянная мутация вируса не оставляют надежд на оптимистичные прогнозы по поводу конца пандемии. Даже по самым оптимистичным прогнозам пандемия продолжится многие годы. Одной из самых важных мер для уменьшения распространения COVID-19 является ношение средств индивидуальной защиты, в частности медицинской маски, в общественных местах. В связи с этим ношение медицинских масок распространилось повсеместно, создавая существенные проблемы для биометрических систем.

Технологии распознавания личности по лицу и голосу оказались практически бесполезны для идентификации людей, лица которых закрыты различными масками. Используемые методы разрабатывались и тестировались в доковидную эру, когда маски не были так распространены. Данное утверждение подтверждает анализ, проведенный экспертами из Американского национального института стандартов и технологий (NIST), которые провели сравнение эффективности 89 алгоритмов распознания лиц на изображениях в масках [3]. Точность алгоритмов с маскированными лицами существенно снизилась по всем направлениям. Самые точные рассмотренные алгоритмы не могут аутентифицировать человека без маски примерно в 0,3 % случаев.

При использовании изображений лиц в масках частота ошибок увеличивается до 50 %.

В большинстве случаев при обработке изображения лица в маске алгоритмы часто не могут корректно обработать область лица. Авторы отчета [2] называют данное явление «отказом регистрации» (Failure to enrol, FTE). Алгоритмы распознавания лиц обычно работают, выделяя признаки частей лица, например, их размер и расстояние друг от друга. Ошибка FTE означает, что алгоритм не может достаточно хорошо выделить черты лица, чтобы провести эффективное сравнение.

Важная задача интеллектуальных систем — анализ голосовых характеристик людей. Помимо лексической информации, речевой сигнал передает информацию о личности говорящего, эмоциональном состоянии, акустической среде, языке и акценте. Распознавание говорящего — важная область современных биометрических систем. Голосовая аутентификация применяется во многих системах безопасности. Судебно-криминалистический анализ голоса использует записи с места преступления для поимки преступников. Данные системы также сталкиваются с новыми проблемами при обработке голоса, искаженного защитной маской. В зависимости от материала маски и уровня контакта с органами артикуляции маска различным образом влияет на речевой сигнал, так как часть звуковой энергии поглощается маской, что выражается в искажении сигнала и уменьшении его энергии.

Постановка задачи

Система аутентификации, использующая один биометрический канал, такой как речь или изображение лица, называется одномодальной системой. Часто полученный для классификации образ может быть низкого качества по следующим причинам: некорректного ракурса, недостаточного освещения, наличия фонового шума или низкого пространственного и временного разрешения видео. Проблема повышения качества решается путем использования бимодальной системы, обрабатывающей несколько биометрических потоков данных.

Аудиовизуальные системы вывели мультимодальную биометрию на новый уровень благодаря использованию дополнительной биометрической информации, извлеченной одновременно из голоса и изображения лица человека. Бимодальные биометрические системы эффективны и удобны в использовании, так как часто возможно записать голос и изображение одновременно. Для этого может быть использована видеокамера смартфона, пропускного пункта или наружного слежения.

В настоящее время в отрытом доступе нет полноценных систематических исследований, посвященных аудиовизуальной идентификации людей в масках. Это можно объяснить тем, что повседневное распространение масок в большинстве стран, включая Россию, началось только два года назад. Однако имеется значительное количество публикаций, посвященных одномодальному распознаванию личности по изображениям и голосу. Также методы бимодальной идентификации хорошо исследованы и показывают высокую эффективность в биометрических задачах. Эти две модальности позволяют использовать бимодальные методы для объединения изученных одномодальных систем для работы с масками.

Цель работы — аналитический обзор актуальных подходов распознавания личности людей в новых условиях, подразумевающих постоянное ношение средств индивидуальной защиты. Для достижения данной цели необходимо выполнить сравнительный анализ:

— корпусов данных видео- и аудиоинформации для решения задач распознавания личности в индивидуальной защите;
— методов распознавания личности в индивидуальной защите по видео- и аудиоинформации;
— методов бимодального аудиовизуального распознания людей и исследование применимости их к задачам идентификации людей в индивидуальной защите.

Системы распознания личности в маске по изображению

Разработки технологий искусственного интеллекта в последние 10–15 лет существенно развились в области биометрических систем. Предложено множество методов, реализованных в практических приложениях. Распознание личности по изображению лица — передовое направление исследований в области биометрической идентификации. Преимущество данных систем — простота применения, что достигается за счет повсеместного размещения камер.

В современном мегаполисе установлено огромное количество камер слежения, часто камеры устанавливаются непосредственно на объект, к которому будет предоставляться доступ посредством лицевой биометрии. В связи с распространением ношения масок в последние два года алгоритмы распознавания лиц столкнулись с новой нерешенной задачей. Методы, используемые для идентификации лиц без масок, плохо работают или совсем не работают для идентификации лиц в масках. Возникает потребность в дополнительных исследованиях и разработке новых эффективных методов и моделей, для обучения которых необходимы расширенные корпусы изображений людей в средствах индивидуальной защиты.

Корпусы изображений для распознания личности в маске

После распространения COVID-19 по всему миру и начала повсеместного ношения масок возникла необходимость в расширенных корпусах изображений людей в масках для решения новых задач компьютерного зрения. Основные две задачи, связанные с анализом изображений людей в средствах индивидуальной защиты: распознавания (детектирование) наличия маски на лице человека и распознавание (идентификация или верификация) личности человека в маске.

Предположение о том, что можно использовать одни и те же корпусы данных для обеих задач не вполне верно. Для детекции наличия маски на лице человека

подойдет любое изображение человека в маске, однако, для идентификации человека по его лицу, корпус должен быть размечен определенным образом. Для обучения алгоритмов распознания лиц субъект на изображении должен быть идентифицирован, и он должен встречаться достаточное количество раз в различных ракурсах, условиях освещенности и одежде для корректного обучения, валидации и тестирования модели.

В [4] представлен аналитический обзор аудиовизуальных систем для определения средств индивидуальной защиты на лице человека, приведен обширный анализ и сопоставление 13 известных корпусов людей в защитных масках. Результаты данного анализа представлены в работе [3]. Однако из данных корпусов изображений для задачи распознания лиц подходит только Real-World Masked Face Recognition Dataset (RMFRD) [4], в остальных базах данных изображений субъекты не аннотированы. Авторы работы [5] представили три корпуса изображений людей в индивидуальной защите: The Masked Face Detection Dataset (MFDD), RMFRD и Masked Face Recognition Dataset (SMFRD). Корпус MFDD включает 24 771 изображение лиц в масках для алгоритмов определения наличия маски. RMFRD содержит 5000 изображений 525 человек в масках, и 90 000 изображений тех же людей без масок. RMFRD — самый большой корпус реальных данных для задач распознавания лиц в масках. Создание большого корпуса для распознавания лиц очень трудоемкая и ресурсозатратная задача. Распознание лиц — популярная область компьютерного зрения, и за время существования этой проблемы возникло достаточное количество корпусов изображений для ее решения. Однако ношение масок стало массово практиковаться только в последние годы, и сейчас существует недостаток качественных баз данных изображений для идентификации людей в масках. Многие исследователи применяют методы синтеза искусственных изображений, в работе [5] рассмотрен данный метод. Для достижения большего разнообразия изображений авторы создали дополнительный корпус SMFRD, который состоит из 500 000 изображений лиц в синтезированных масках.

Корпус данных Masked Face Segmentation and Recognition (MFSR) [6] состоит из двух частей. Первая часть включает 9742 изображения лиц в масках, которые собраны из Интернета. Вторая часть включает 11 615 изображений 1004 личностей, которые сделаны в реальных условиях.

Корпус данных The Synthetic CelebFaces Attributes (Synthetic CelebA) состоит из 10 000 общедоступных синтетических изображений. Для создания данного корпуса в работе [7] синтезированы изображения и добавлены маски из базового корпуса CelebA, который содержит 200 000 изображений знаменитостей. Корпус сгенерирован с использованием 50 типов синтетических масок различных размеров, форм, цветов и структур. Для создания синтетических изображений лицо выровнено по координатам глаз, а затем маска наложена вручную с помощью Adobe Photoshop.

Корпус данных Synthetic Face-Occluded Dataset [8] также создан с использованием общедоступных корпусов CelebA и CelebA-HQ, последний состоит из 30 000 изображений знаменитостей. Отличие CelebA-HD от CelebA состоит в том, что на первом представлены изображения в разрешении HD 1024 × 1024 пикселов. Каждое изображение лица выровнено по положению глаз. Окклюзии синтезированы добавлением одного из пяти объектов, обычно закрывающих лицо: рук, маски, солнцезащитных очков, микрофона. Более 40 различных видов каждого объекта разных размеров, форм, цветов и структур нанесены на лица случайным образом.

Корпус данных MS1MV2 [9] представляет собой модифицированную версию корпуса данных MS-Celeb1M. MS1MV2 включает 58 млн изображений 85 000 различных личностей. Авторы работы [10] произвели добавление масок на изображения из корпуса MS1MV2 и создали версию MS1MV2-Masked. Тип и цвет маски выбраны случайным образом для каждого изображения, чтобы обеспечить большую вариативность обучающих данных. Подмножество из 5000 изображений случайным образом выбрано из MS1MV2-Masked для верификации модели. На этапе тестирования авторы использовали два корпуса данных реальных лиц в масках: MFR2 [11] и EMFR [12]. MFR2 состоит из 269 изображений 53 личностей, взятых из Интернета, корпус включает в себя в среднем по пять изображений на человека в средствах индивидуальной защиты и без них.

Корпус EMFR собран с помощью 48 участников, которые использовали личные веб-камеры для создания изображений [13]. Запись происходила в течение трех сеансов. Авторами создано несколько вариаций корпусов данных. На изображениях участники были в масках. Данная часть корпуса получила название – проба в масках (mask probe).

В работе [13] предложен корпус данных The Labeled Faces in the Wild (LFW), который состоит из 50 000 изображений реальных людей без масок и предназначен для алгоритмов распознания лиц. Авторы работы [11] создали модификацию LFW-SM, которая содержит 13 233 изображения 5749 человек с добавлением смоделированных масок.

Аналогичными синтетическими способами созданы корпусы MFV и MFI [14], BUAA-VisNir [15], VGG-Face2_m [16], CASIA-FaceV5_m [16] и Webface [17]. Полный список рассмотренных корпусов, а также их характеристики представлены в табл. 1.

Сравнительный анализ методов распознания лиц людей в маске

После введения обязательного масочного режима в большинстве стран мира системы распознания лиц, разработанные в доковидную эру, столкнулись с новыми проблемами. Маска закрывает часть лица и препятствует дальнейшей классификации. Область носа очень важна в задаче распознавания лиц, так как она используется для нормализации лица и коррекции позы и поворота головы. Однако маска практически всегда закрывает нос, существенно усложняя задачу распознания лица. Для решения этих проблем необходимы новые модели и методы.

В работе [19] использован самый очевидный подход к распознанию лица в условиях частичного перекры-

Таблица 1. Корпусы изображений людей в защитных масках
Table 1. Masked face datasets

Корпус	Количество изображений	Количество людей	Количество типов масок	Способ наложения масок
RMFRD [5]	95 000	525	1	Реальный
SMFRD [5]	500 000	10 000	1	Синтетический
MFSR [6]	11 615	1004	1	Реальный/синтетический
LFW-SM [11]	13 233	5749	4	Синтетический
MFR2 [11]	269	53	5	Синтетический
MFV [14]	400	200	1	Синтетический
MFI [14]	4916	669	1	Синтетический
MFD [5]	990	45	5	Синтетический
MFW-mini [18]	3000	300	14	Синтетический
SFOD [8]	30 000	307	40	Синтетический
MSIMV2-Masked [9]	5,8 млн	85 000	1	Синтетический
CASIA NIR-VIS 2.0 [16]	17 580	725	5	Синтетический

тия — извлечено максимальное количество признаков из областей лица, не закрытых маской, игнорируя остальные. Применены предварительно обученные сверточные нейронные сети (Convolutional Neural Network, CNN). В основном из исходных изображений использованы только области глаз и лба. Затем для квантования представления на последнем сверточном слое применен алгоритм «мешок признаков» (bag-of-features) для обучения, тестирования и валидации использован корпус данных RMFRD. В [20] также применен корпус RMFRD, каскад Хаара и нейронная сеть MobileNet для обнаружения графического региона маски и дальнейшего его удаления. Для классификации оставшейся области лица использованы нейросети VGG16 и Triplet Loss FaceNet в многопоточном режиме.

В [21] представлена модель глубокого обучения, основанная на функции ArcFace. Изменения внесены в функции извлечения признаков и потерь. Из исходного корпуса данных распознавания лиц сгенерирована версия с масками, а для классификации применена модель ResNet-50. В итоге получена комбинированная модель, состоящая из модифицированной функции ArcFace и нейросети ResNet-50. Данный метод назван мультизадачным ArcFace (Multi-Task ArcFace). В результате ресурсозатратность конечной модели оказалась приемлемой, что достигнуто за счет использования облегченной версии ResNet-50.

В работе [22] разработана сеть FaceMaskNet-21, обученная с помощью 4-мерных кортежей с глубоким метрическим обучением. В сети использованы 128-мерные кодировки, сгенерированные для каждого лица в корпусе данных, а для распознания — признаки, доступные в открытых частях лица, глаз, лба и контура лица. Применены Histogram of Oriented Gradients (HOG) функции для более активного распознания и корпусы данных: RMFRD, MFDD и SMFRD.

В [11] предложен программный инструментарий MaskTheFace для генерации синтетических корпусов изображений лиц в масках. Инструментарий использует ориентиры на лицах, чтобы определить ключевые особенности и наклон лица для дальнейшего наложения маски. Создана модель MaskTheFace для распознавания лиц в маске с помощью системы FaceNet, создающей эмбэддинги лиц. Для обучения FaceNet использованы изображения лица из корпуса VGGFace2 и 42 случайно сгенерированных изображения этого лица в различных масках. Полученный корпус назван VGGFace2-mini-SM и является достаточно разнообразным, а маски, наложенные синтетическим образом, хорошо совпадают с лицом. Однако точность распознания лиц с помощью модели MaskTheFace составляет 86 %, что является не самым высоким показателем.

CNN использованы в работе [20], где предложена конфигурация многозадачной каскадной сверточной нейронной сети (MTCNN) для обнаружения частей лица, не закрытых маской. Соответствующие части лица после некоторой предобработки проанализированы моделью FaceNet, которая необходима для вычисления высокоуровневых признаков. Высокоуровневые признаки классифицированы с помощью метода опорных векторов (Support Vector Machine, SVM) для получения конечного результата идентификации личности. Данная система использована в двух сценариях. В первом случае в качестве входных данных для обучения применены лица без масок, а для тестирования — в масках, во втором — оба варианта изображений лиц при обучении и тестировании.

Кроме корпусов MFR, MFV и MFI в работе [13] предложен метод триангуляции, применяемый для разделения изображений на маленькие треугольники. Каждому треугольнику изображения лица соответствует треугольник изображения маски. Такой подход применен для создания искусственных корпусов изображений лиц в масках. Для распознавания лиц предложена модель обнаружения скрытых частей Latent Part Detection, которая основана на предположении, что человек при распознавании лица анализирует видимые скрытые части лица, закрытые маской. Однако факт того, что маска всегда отражает признаки лица, скрытого под ней, справедлив не для всех типов масок.

Есть сомнения, что предложенный метод достаточно универсален для всех средств индивидуальной защиты.

Одним из способов решения проблемы перекрытия части лица при его распознании может быть автоматическое дополнение его открытой части. В [7] использован этот принцип и предложен метод автоматического удаления объектов с лица и синтеза поврежденных областей с сохранением исходной структуры. Авторы сохранили структурное и форменное постоянство полученного лица с использованием двух дискриминаторов, обученных для восстановления строения закрытой области. Для этой задачи использован синтетический корпус данных на основе CelebA, так как не было возможности применить реальные данные из-за отсутствия пары изображений людей в масках и без нее в открытом доступе. Полученная объединенная модель с прямой связью производит структурно правдоподобные изображения лиц. Исследователи провели тестирование своего метода восстановления в связке с четырьмя современными моделями распознания лиц: VGGFace, FaceNet, OpenFace и DeepFace. Полученные системы продемонстрировали высокую точностью в задаче распознания лиц в масках.

В [20] представлен метод дополнения области маски. Разработанная система улучшает разрешение изображений с помощью расширенной свертки и уменьшает потерю информации с помощью механизма внимания (attention). Для распознания дополненных лиц использована модель ResNet, для обучения — корпусы RMFRD и SMFRD.

В работе [23] представлена схема для распознания лиц в масках, основанная на деокклюзионной дистилляции (Deoclussion Distillation). Окклюзиями здесь называются маски, закрывающие часть лица, а деокклюзией — процесс отделения маски. Разработанная система включает два модуля: модуль деокклюзии, состоящий из генеративно-состязательной сети (GAN), которая служит для дополнения закрытой части изображения лица и модуля, выполняющего дистилляцию. Предварительно обученная модель распознавания лиц адаптирует свои знания о лицах с помощью дистилляции знаний модели VGGFace2. Дополнительно была обучена разработанная модель для классификации маски на три класса: простые, сложные и гибридные.

Одним из применений систем идентификации лиц в масках — выявление преступников в публичных местах, такой сценарий рассмотрен, например в работе [24], где представлен метод повторной идентификации пешеходов ReID. Данный метод заключается в попытке найти связи между изображениями в маске и без маски одного и того же человека. Метод повторно извлекает локальные и глобальные признаки лица в маске и измеряет сходство между исходным и всеми изображениями из определенной базы данных. При нахождении совпадения производится идентификация с помощью нейросети FaceNet. Сопоставление рассмотренных методов, а также оценки их эффективности представлены в табл. 2.

Таблица 2. Сопоставление методов распознания лиц в масках

Table 2. Masked face recognition methods comparison

Работа	Модель распознавания	Корпус данных	Показатель эффективности	Значение
Din et al. [7] (2020)	GAN	CelebA	Fréchet inception distance	6,102 баллов
Yalavarth et al. [19] (2020)	CNN	RMFRD	Positive Predictive Value	60,17 баллов
Montero et al. [21] (2021)	MTArcFace	LFW, CFP, Agedb	Точность распознавания	99,78 %
Hariri et al. [19] (2021)	VGG-16, AlexNet, ResNet-50	RMFRD, SMFRD	Точность распознавания	91,30 %
Maharani et al. [20] (2020)	VGG-16 И FaceNet	Собственный корпус	Точность распознавания	100 %
Boutros et al. [10] (2021)	ResNet-50, MobileFaceNet	MSIMV2, MFR, MRF2	Equal Error Rate (EER)	7,82 баллов
Golwalkar et al. [22] (2020)	Face MaskNet-21	Собственный корпус	Точность распознавания	88,92 %
Hong et al. [24] (2020)	Сеть с механизмом внимания (Attention-based)	MFDD, RMFRD	Точность распознавания	95,05 %
Anwar et al. [11] (2020)	MaskThe Face	VGGFace2-mini-SM, LFW-SM	Точность распознавания	97,25 %
Mandal et al. [25] (2021)	ResNet-50	RMFRD	Точность распознавания	87,00 %
Ejaz et al. [20] (2019)	MTCNN	MFD	Точность распознавания	98,50 %
Deng et al. [16] (2021)	GAN	MFSR, CASIA-WebFace, VGGFace2	Точность распознавания	86,50 %
Li et al. [23] (2020)	GANS	Celeb-A, LFW, AR	Точность распознавания	95,44 %
Ding et al. [14] (2020)	Two-branch CNN	MFV, MFI, LFW	Точность распознавания (Rank3)	93,70 %
Du et al. [26] (2021)	Siamese сети	Oulu-CASIA NIR-VIS, BUAA-VisNir	Точность распознавания (Rank1)	98,60 %
Wu et al. [27] (2021)	ResNet	RMFRD, SMFRD	Точность распознавания	95,00 %
Li Y et al. [28] (2021)	CBAM	Webface, AR, Yela B, LFW	Точность распознавания	92,61 %

Системы распознания людей в индивидуальной защите по голосу

Устная речь — самый естественный способ общения между людьми. Помимо смысла произнесенных слов, речевой сигнал передает паралингвистическую информацию о личности говорящего, его характеристиках и эмоциональном состоянии, акустической среде, языке, акценте и т. д. Судебно-криминалистический анализ речи и голоса диктора позволяет выявлять правонарушителей, используя записи с места преступления. Однако современные системы идентификации личности по голосу сталкиваются с трудностями при обработке синтезированного, модифицированного, подделанного или естественно измененного сигнала.

Для сокрытия личности преступники стали часто использовать медицинские маски [29]. Ношение маски влияет на произносимую речь как активным, так и пассивным образом. Большинство средств индивидуальной защиты характеризуются звукопоглощающими свойствами, а также затрагивают механизмы речевой артикуляции. В зависимости от типа маски, степени ее контакта и соприкосновения с лицом, маска ограничивает артикуляторные движения в разной степени. Эти ограничения изменяют нормальную артикуляцию некоторых согласных звуков, например, /п/ и /м/.

Влияние маски на идентификацию собеседника исследовано в работе [30], где изучены некоторые способы изменения голоса, в том числе и медицинские маски. Представлена идентификация человека по голосу с помощью системы FASRS, и изучены идентификационные баллы для каждого члена группы целевых дикторов. В результате сделан вывод, что ношение хирургической маски неблагоприятно влияет на распознавание лица.

В работе [31] предпринята попытка измерить влияние масок на частотные характеристики речи. Потери при передаче аудиосигнала измерялись путем воспроизведения речи через громкоговоритель и дальнейшей его записи с помощью микрофона, отделенного от динамика лицевой маской. Потери в передаче мощности аудиосигналов в различных диапазонах слышимых частот измерены для 44 масок из различных тканых материалов. Получен очевидный результат, что потери передачи зависят от веса, толщины и плотности/пористости ткани. Также замечено, что поглощение звуковой энергии в разных тканях приводит к большим потерям энергии в высокочастотных диапазонах, чем в низкочастотных.

Для исследования и решения проблем распознания личности по голосу необходимы обширные речевые базы данных.

Корпусы аудиоданных для распознавания личности в маске

В [30] представлен корпус речевых данных в условиях перекрытия лица диктора различными объектами (Speaking Under Face Cover Corpus, SUFCC). Данная база данных создана для автоматического распознавания личности говорящего. При записи использованы четыре типа одежды на лице, которые люди чаще всего носят в общественных местах: мотоциклетный шлем, резиновая маска, медицинская маска, шарф.

Записи сделаны в звукоизолированной комнате площадью около 5 м^2 с двумя окнами и двойными дверями. Данные изначально записаны с частотой 44,1 кГц, в 16-битном формате моно. Запись велась одновременно с трех микрофонов: гарнитуры, размещенной возле рта говорящего (AKG C444), микрофона (AKG C4000B), прикрепленного к стене с правой стороны динамика, микрофона (AKG C4000B), расположенного за динамиком. Для записи каждый участник произносил заранее подготовленный, фиксированный текст, и некоторый отрывок спонтанной речи. Данные действия записаны для всех типов масок в течение двух сеансов. В эксперименте приняли участие люди в возрасте от 21 до 28 лет. Каждая запись длилась от 60 до 90 с. Контрольная запись без маски записана без усиления аудиосигнала. Речевая база данных включает записи 4 мужчин и 4 женщин, всего 60 аудиофайлов общей длительностью 1,5 ч для каждого из 8 участников. Типы масок, используемых в исследовании, представлены на рис. 1.

Исследование влияния масок на качество распознания личности говорящего рассмотрено в работе [32]. Создан корпус аудиоданных OWR Audio Face Covering Corpus (OWR-AFCC), состоящий из образцов короткой речи 8 участников (6 мужчин и 2 женщин). Записи произведены на личные смартфоны участников: без масок, с тканевым покрытием лица и медицинской маской. Эксперимент выполнен в две сессии: в помещении и на улице.

Audio-Visual Face Cover Corpus (AVFCC) [33] — один из самых представительных корпусов для рас-

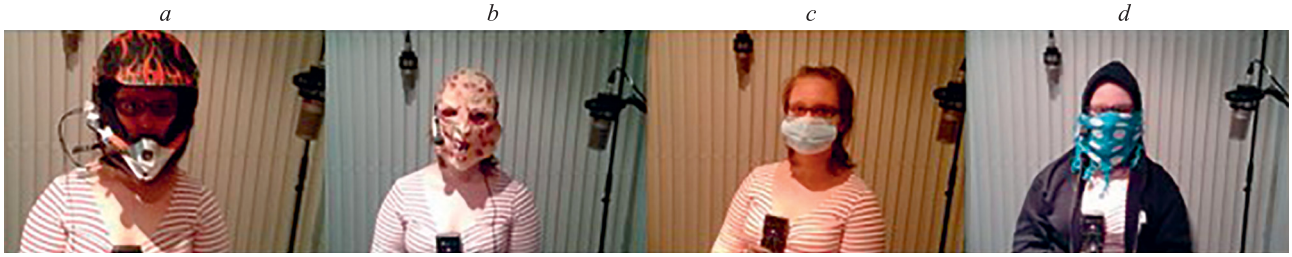

Рис. 1. Типы масок, используемые в исследовании [30]: мотоциклетный шлем (*a*); резиновая маска (*b*); медицинская маска (*c*); шарф (*d*)

Fig 1. Types of masks used in research [30]: motorcycle helmet (*a*), rubber mask (*b*), surgery mask (*c*), scarf (*d*)

познания людей по голосу в средствах индивидуальной защиты. Состоит из высококачественных аудио- и видеозаписей 10 носителей британского английского языка в различных типах лицевых масок. Участники эксперимента произносили вслух корпус из 64 /C1VC2/ слогов (трехбуквенное слово, состоящее из согласной, за которой следует гласная, а затем еще одна согласная), встроенных в несущую фразу. Список слов был повторен 1 + 8 раз: 1 раз в контрольном состоянии и 8 раз в различных видах масок. Аудиозаписи производились с помощью аудиогарнитуры и двух микрофонов, расположенных перед говорящим и позади него. Таким образом, всего на каждое устройство записано 6120 высказываний. Этот корпус создан как для смежных областей исследований голосовых характеристик, так и для тематических исследований, каким и является идентификация людей по голосу.

В 2020 году на международных соревнованиях по компьютерной паралингвистике ComParE-2020 в рамках конференции INTERSPEECH был представлен корпус аудиоданных Mask Augsburg Speech Corpus (MASC) [34]. Корпус предназначен для задачи детектирования наличия маски на лице человека посредством анализа речи. В данном корпусе также идентифицированы сессии для каждого субъекта — участника эксперимента, что позволяет его использовать и для задачи распознания людей по голосу. В эксперименте приняли участие 16 женщин и 16 мужчин, являющихся носителями немецкого языка (средний возраст участников — 25 лет). Записи сделаны с использованием конденсаторного микрофона AKG C4500 BC, исходные данные имели частоту дискретизации 48 кГц, а в дальнейшем были сжаты до 16 кГц. Аудиофайлы разделены на одинаковые короткие фрагменты речи продолжительностью в 1 с.

Собственный речевой корпус Mask Sorbonne Speech Corpus (MSSC) [34] собран одной из команд-участников соревнований ComParE 2020 для расширения обучающих данных корпуса MASC. Для имитации человеческого голоса использована акустическая система — Bose Sound-Link micro. Процесс записи корпуса проходил в два этапа: сначала с помощью динамика проигрывались 1000 высказываний от 30 информантов (15 мужчин и 15 женщин), которые выбраны из речевого корпуса German Distant Speech Data Corpus [35]. Далее на колонку надевалась медицинская маска, и снова проигрывались эти же высказывания. В результате был создан синтетический корпус, содержащий высказывания людей в масках и без них.

Перечень рассмотренных корпусов для распознания людей в средствах индивидуальной защиты представлен в табл. 3. Наличие корпусов данных для распознания личности человека по характеристикам его голоса в маске позволяет проводить исследования и разработки для решения данной задачи.

Сравнительный анализ методов распознавания личности людей в маске по голосу

В [30] кроме создания корпуса SUFCC разработана система идентификации человека по его голосу. Текстонезависимая идентификация говорящего стала актуальной задачей в последнее десятилетие. В работе [30] авторы предположили, что самое перспективное направление — подход на основе i-векторов. Архитектура разработанной системы представлена на рис. 2.

В результате использована готовая система RUN, успешно зарекомендовавшая себя в реальных криминалистических условиях. Схема предложенного в работе [30] метода извлечения признаков изображена на рис. 3.

Голосовой сигнал разделен на кадры длительностью 30 мс со сдвигом в 15 мс. Для кадрированного сигнала произведено линейное прогнозирование кратковременного спектра. Далее 19 мел-частотных кепстральных коэффициентов (MFCC) извлечены и дополнены энергиями кадра. Произведена RASTA фильтрация и расчет Δ и $\Delta\Delta$ характеристик. В результате сформированы векторы признаков из 60 компонентов. На последнем этапе

Таблица 3. Сравнение корпусов аудиоданных для распознавания людей в масках

Table 3. Comparison of datasets for masked speaker recognition

Корпусы	Количество дикторов	Язык	Формат данных	Продолжительность, ч	Количество экземпляров в каждом классе
SUFCC [30]	4 мужчины 4 женщины	Английский	44,1 кГц, в 16-битный формат моно	Около 12	96 (речь в шлеме), 96 (речь в резиновой маске), 96 (речь в медицинской маске), 96 (речь в шарфе), 96 (речь без маски)
AVFCC [33]	5 мужчин 5 женщин	Английский	48,0 кГц, 768 кбит/с, 16-бит моно	Около 10	10 (контрольный), 10 × 8 (8 различных масок)
OWR-AFCC [32]	16 мужчин 16 женщин	Немецкий	16 кГц, 16-бит моно	Около 8	16 (в маске в помещении), 16 (без маски в помещении), 16 (без маски на улице)
MSSC [34]	15 мужчин 15 женщин	Немецкий	48,0 кГц, 16-бит моно	Нет данных	1000 (речь без маски), 1000 (речь с маской)
BRAVE-MASKS [36]	15 мужчин 15 женщин	Русский	48,0 кГц, 16-бит моно	60	83 (речь без маски), 83 (речь с маской)

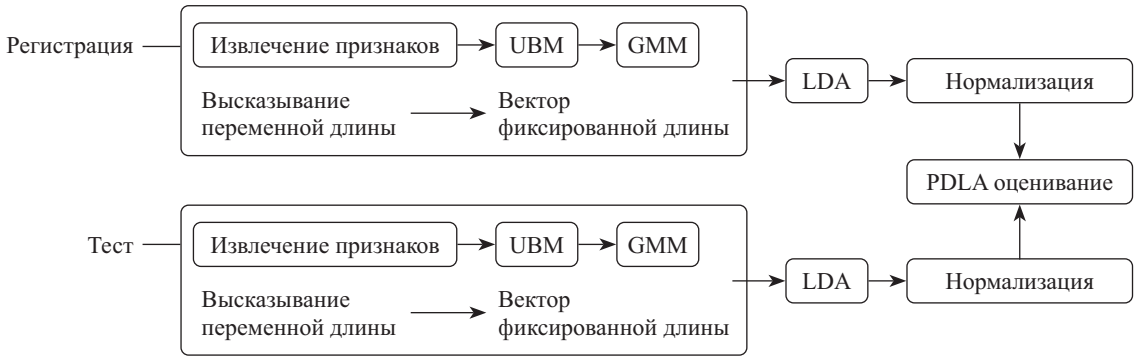

Рис. 2. Архитектура системы SUFCC [30]

GMM — Gaussian Mixture Model; UBM — Unified Background Model; LDA — Linear Discriminant Analysis

Fig. 2. SUFCC system architecture [30]

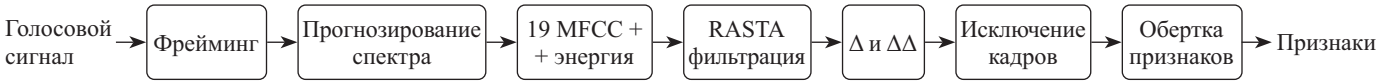

Рис. 3. Схема метода извлечения акустических признаков [30]

Fig. 3. Acoustic features extracting scheme [30]

на основе энергии уровня кадра применена функция обертки признаков. Гендерно-зависимая универсальная фоновая модель UBM с 2048 компонентами обучена с использованием корпуса NIST SRE 2004–2006 [37]. При постобработке i-векторов уровня высказывания для улучшения разделимости классов и уменьшения размерности i-векторов до 200 использована линейная проекция дискриминантного анализа LDA.

Результаты исследований системы идентификации говорящего при использовании различных масок представлены в табл. 4. Для оценки точности использован показатель:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$

где количества срабатываний: TP (True positive) — истинно положительные; TN (True Negative) — истинно отрицательные; FN (False Negative) — ложноотрицательные; FP (False Positive) — ложноположительные.

В столбцах данной таблицы показаны типы масок, использовавшихся в тестовом корпусе данных, а в строках таблицы расположены типы масок, использовавшихся при создании шаблона идентификации говорящего. В ячейке a_(i,j) показана точность (Accuracy, %), полученная при создании шаблона на корпусе i и тестировании на корпусе j. Сравнение выделенных диагональных элементов, находящихся по диагонали таблицы показывает, что при соответствии типа масок, использующихся при обучении и тестировании, система показывает высокие результаты точности. Производительность распознавания снижается, когда в обучающем шаблоне и тестовом сегменте взяты разные маски. Однако даже при таких комбинациях в ситуации, когда шаблон создан без маски, а запись голоса произведена в маске, система показала точность равную 94,2 %. В работе [30] авторами не предложено принципиально нового подхода к распознанию людей в масках по голосу, выполнена только адаптация ранее разработанного метода путем настройки гиперпараметров, и получены высокие показатели эффективности.

В [32] проведен анализ существующих методов распознавания по голосу в условиях, когда лицо говорящего закрыто различными типами масок. Авторами собран корпус данных OWR-AFCC, он содержит телефонные записи образцов речи. 8 участников (6 мужчин и 2 женщины): контрольная запись, запись с тканевой и медицинской масками. Помимо созданного корпу-

Таблица 4. Показатели точности системы SUFCC [30], %

Table 4. Accuracy values of SUFCC evaluation [30], %

Типы масок при создании шаблона идентификации говорящего	Типы масок тестового корпуса данных				
	Без маски	Шлем	Резиновая маска	Медицинская маска	Капюшон + шарф
Без маски	95,2	94,9	88,6	94,2	93,3
Шлем	88,5	97,7	86,0	88,8	88,4
Резиновая маска	90,3	96,5	97,1	94,1	91,4
Медицинская маска	95,1	96,7	90,1	97,9	95,6
Капюшон + шарф	90,3	85,7	82,5	94,9	97,0

са OWR-AFCC для тестирования также использованы аудиоречевые данные из корпуса Native Forensics Audio-Visual Face Cover Corpus (NF-AVFCC). Все тестовые данные объединены в один корпус. Для оценивания эффективности работы системы использован количественный показатель средней равновероятной ошибки:

$$EER = \frac{FAR + FRR}{2},$$

где отношения количества: FAR (False Acceptance Rate) — ложноположительных и FRR (False Rejection Rate) — ложноотрицательных срабатываний к количеству всех срабатываний.

Нулевые значения ошибки EER наблюдались для всех экспериментов на корпусе OWR-AFCC. У тех же участников при наличии тканевой маски оценки ниже, но этого не хватило, чтобы внести ошибки. В ходе экспериментов с маской и без результаты аналогичны. Показатели метрики EER для корпуса NF-AVFCC различались для разных типов масок. Для медицинской маски и шарфов значения EER составили 0 %, для мотоциклетного шлема получена деградация эффективности, при этом EER — 4,00 %. EER для заклеенного лентой рта равен 14,67 %, что является наихудшим показателем. Эксперименты показали, что повседневные предметы одежды и медицинские маски не оказывают негативного воздействия на распознание говорящего с помощью x-векторного анализа.

Можно сделать вывод, что на данный момент наблюдается недостаток исследований по голосовому распознанию личности человека в масках. Однако рассмотренные исследования однозначно подтверждают применимость существующих методов к решению данной задачи, что позволяет сделать предположение о незначительном ухудшении производительности существующих методов при анализе голосовых данных субъектов в маске.

Системы бимодального распознания личности людей

В некоторых реальных задачах недостаточно признаков, которые можно извлечь из изображений лица или речи. Такой задачей является и рассматриваемое в настоящей работе распознавание людей в масках. Для достижения лучших результатов классификации возможно использование объединения информации от двух взаимодополняющих модальностей. Аудиовизуальные биометрические методы используют множество подходов объединения для взаимного дополнения аудио- и видеохарактеристик двух модальностей. Классификация методов объединения представлена на рис. 4.

Методы объединения делятся на три типа: предварительное (Early Fusion), промежуточное (Intermediate Fusion) и позднее (Late Fusion). Рассмотрим различные аудиовизуальные биометрические методы согласно классификации и их показатели эффективности.

Методы раннего объединения

Матрица Лапласа — эффективный метод представления аудио- и видеопризнаков для раннего объединения информации. Метод карт собственных значений лапласиана (Laplacian Eigenmaps) — нелинейный подход к снижению размерности данных, который позволяет сохранить внутренние геометрические взаимосвязи и локальную структуру данных. Данный метод использован в работе [38]. В предложенном подходе фрагменты изображений лиц и MFCC, извлеченные из записей голоса спроецированы на пространство карт собственных значений лапласиана и в дальнейшем объединены в единый вектор признаков для классификации. В ходе экспериментов была достигнута большая точность по сравнению с объединением с помощью линейного дискриминантного анализа (Linear Discriminant Analysis, LDA), а также по сравнению с использованием признаков каждой из модальностей по отдельности.

Многоканальный полуавтоматический дискриминантный анализ (Multi-level Semi-Supervised Discriminant Analysis, MSDA) — расширение полуавтоматического дискриминантного анализа (Semi-supervised Discriminant Analysis, SDA) для осуществления объединения информации на уровне признаков [39]. MSDA основан на методе обучения с полуавтоматическим управлением с несколькими представлениями. Он накладывает локальное ограничение смежности нескольких представлений на целевую функцию традиционного LDA. Ограничение требует, чтобы два соседних образца в первоначальном пространстве признаков первого представления располагались близко друг к другу в конечном пространстве меньшей размерности другого представления. Этот подход лучше использует информацию, полученную из небольших, вручную размеченных корпусов для обучения, и поэтому превосходит традиционный LDA.

В работе [40] для объединения на уровне признаков использованы изображения губ и кривые огибающей спектра речевого сигнала. Извлеченные признаки преобразованы в векторы временных рядов и объединены для дальнейшей обработки. Авторы предложили метод получения расстояния между данными из различных векторов временных рядов, названный динамическим выравниванием времени (Dynamic Time Wrapping, DTW). С помощью DTW в предложенном подходе рассчитано сходство между признаками голоса и черт лица. Векторы сходства далее использованы для классификации.

В методе, предложенном в работе [41], записи голоса и изображения лица отдельно обработаны совместной машиной Больцмана (joint-training Deep Boltzman Machine, jDBM) и ограниченной совместной машиной Больцмана (joint-training Restricted Boltzman Machine, jRBM). Входные данные были зашумлены с помощью JPEG-сжатия и добавления случайных звуков множества голосов. В результате авторы сделали вывод, что данный бимодальный метод показывает лучшие результаты в условиях ухудшения качества входных данных.

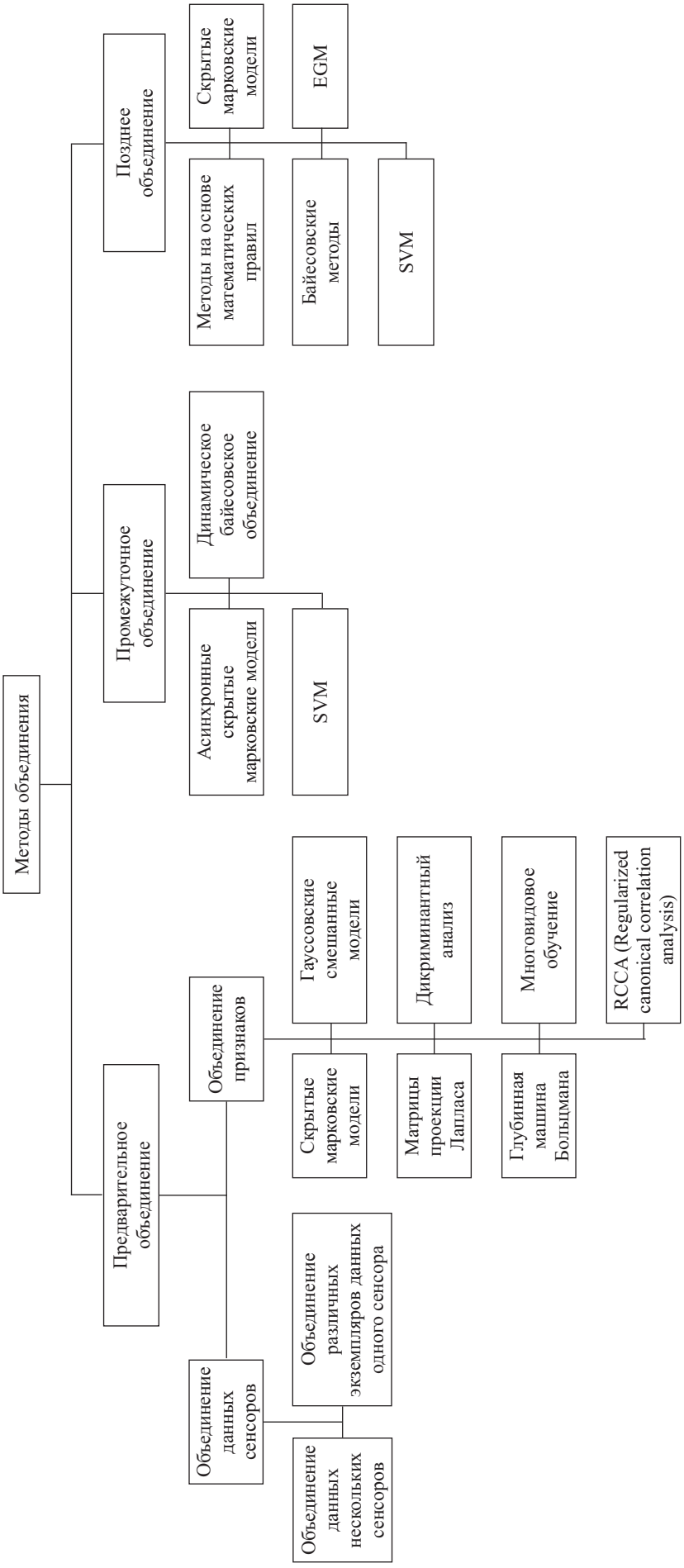

Рис. 4. Классификация методов объединения аудиовизуальной информации
Fig. 4. Audiovisual fusion methods classification

Еще один эффективный метод объединения был предложен в работе [42], где разработан подход, основанный на объединении MFCC и гистограммы направленных градиентов с помощью дискриминантного корреляционного анализа (Discriminant Correlation Analysis, DCA). В ходе экспериментов для классификации вектора объединенных признаков были использованы следующие методы: SVM, LDA, SDA, случайный лес, K-Nearest Neighbors и SVM. Среди протестированных классификаторов наилучшую производительность показал случайный лес, а наихудшую — метод опорных векторов.

Методы промежуточного объединения

Промежуточное объединение информации — более сложный метод по сравнению с ранним объединением. Методы, использующие промежуточное объединение, должны обрабатывать информационные потоки различных модальностей при отображении из пространства признаков в пространство решений. Для промежуточного объединения часто используется временная синхронизация между потоками информации, с помощью которой можно избежать проклятья размерности пространства признаков, возникающего в методах раннего объединения. Примерами моделей, используемых для промежуточного объединения, являются скрытые марковские модели (Hidden Markov Models, HMM), которые могут обрабатывать несколько потоков данных параллельно.

Промежуточное объединение с использованием асинхронных HMM (AHMM) для осуществления текстонезависимой мультимодальной аутентификации описано в работе [43]. Обучение AHMM выполнено с помощью алгоритма максимизации ожидания (Expectation Maximization, EM) на чистых данных. Эксперименты проведены на аудиовизуальных данных с различным уровнем шума (0,5 и 10 дБ). В результате экспериментов значение показателя ошибок EER оказалось в два раза меньше по сравнению с моделью, использующей одну модальность.

В работе [44] рассмотрен подход, названный авторами байесовской аудиовизуальной идентификацией говорящего (Bayessian audio-visual speaker authentication). Для распознавания личности по записи голоса использованы сдвоенные скрытые марковские модели (Coupled HMM, CHMM), а для распознавания личности по изображению — скрытые марковские модели (Enhanced HMM, EHMM). Разработанная модель в качестве входных данных принимает видеофрагменты. На первом этапе происходит разделение аудио- и видеосигналов, а также выделение на отдельных изображениях области лица и губ говорящего. Изображения лица в дальнейшем находят применение для вычисления оценки схожести с предоставляемыми шаблонами с помощью EHMM, а изображения губ совместно с фрагментами записи голоса подаются на CHMM для вычисления оценки схожести по второй модальности. На последнем этапе оценки объединяются по следующему правилу [44]:

$$L(\mathbf{O}^f, \mathbf{O}^a, \mathbf{O}^v|k) = \lambda_f L(\mathbf{O}^f|k) + \lambda_{av} L(\mathbf{O}^a, \mathbf{O}^v|k),$$

где $\mathbf{O}^f$, $\mathbf{O}^a$, $\mathbf{O}^v$ — последовательности данных записей голоса, изображений губ и изображений лица соответственно; $L(*|k)$ — оценка схожести с k-й личностью из базы данных; λ_f и λ_{av} — некоторые веса, полученные в ходе обучения.

Эксперименты проведены на корпусе данных XM2VTS, использование двух модальностей уменьшило количество ошибок на 15 % и 13 % при различных уровнях Гауссова шума соответственно.

Методы позднего объединения

Методы позднего объединения осуществляют объединение результатов, полученных индивидуально от различных классификаторов аудио- и видеомодальностей. Существует большое количество различных подходов к позднему объединению. Простой и распространенный подход — сочетание результатов аудио- и видеомоделей с использованием различных математических правил. Правила взвешенной суммы и взвешенного произведения применяются к модальностям распознавания лиц и голоса отдельно в работе [45]. Эксперименты показали, что производительность мультимодальный системы с поздним объединением на основе взвешенной суммы увеличилась по сравнению с одномодальной системой (значение EER уменьшилось на 14 %) [45].

Объединение с помощью взвешенного суммирования также применено в [46]. Для распознания диктора по голосу использована модель гауссовских смесей (Gaussian Mixture Model, GMM) совместно с универсальной фоновой моделью (Unified Background Model, UBM). Для распознания лиц предложен алгоритм Local Binary Pattern (LBP). Вероятностная оценка по GMM-UBM и метрика взвешенного расстояния на LBP объединены на уровне оценок с помощью взвешенной суммы. Исследователи достигли значения EER, равного 22,7 % при распознавании мужчин и 19,3 % — женщин, что в среднем на 10 % меньше, чем при использовании одной модальности.

В [47] для объединения результатов двух моделей GMM и HMM использована сеть Байеса. Метод байесовского объединения выбирает изображения и аудиоданные из каждой сессии на основе оценок достоверности (confidence scores). Эксперименты показали высокие результаты производительности модели, точность классификации составляет 99,5 % ± 0,3 %.

В работе [48] применено позднее объединение на уровне вероятностных оценок. Для распознавания лиц задействованы четыре варианта моделей CNN: ResNet-50, PyramidNet, ArcFace-50, ArcFace-100. Для распознавания голоса использован метод на основе x-векторного анализа. Конечный результат распознавания личности вычислен с помощью объединения результатов двух моделей на уровне выдаваемых ими оценок (score fusion). Выходными данными моделей каждой из модальностей являются три скалярные величины оценок: s — верификации, q_e — качества шаб-

Таблица 5. Сравнение бимодальных методов распознавания личности
Table 5. Bimodal person recognition methods comparison

Работа	Аудио-признаки	Визуальные признаки	Классификатор	Метод объединения	База данных	Показатель эффективности, %
Alam et al. [38] (2013)	MFCC	Признаки Хаара	Linear Regression + Gaussian Mixture Models + Universal Background Model	Объединение на уровне оценок	Austalk [49]	Accuracy: 31,3
Khoury et al. [50] (2014)	MFCC	DCT-mod2	Gaussian Mixture Models + Inter Session Variability + Total Variability	Объединение данных на уровне LLR-оценок	MOBIO [51]	EER: 6,32
Alam et al. [52] (2013)	MFCC	Признаки Хаара	Linear Regression + Gaussian Mixture Models + Universal Background Model	Правило суммы	Austalk [49]	Accuracy: 100
Tresadren et al. [53] (2013)	MFCC	LBP	Классификатор срезов с усилением	Quadratic Discriminant Analysis (QDA)	MOBIO [51]	EER: 0,5
Islam et al. [54] (2014)	MFCC	ASM	HMM	Back-Propagation Neural Network (BPN)	Собственная	Total Error Rate: 0
Primorac et al. [55] (2016)	MFCC	Признаки Хаара	Joint sparse-классификатор	Объединение на уровне признаков	MOBIO [51]	Accuracy: 94,2
Memon et al. [56] (2017)	MFCC	SVM	Матрица подобия MFCC признаков и оценок SVM	Объединение на уровне признаков	Собственная	Accuracy: 73,8
Gofman et al. [42] (2018)	MFCC	HOG	SVM	Объединение на уровне признаков	CSUF-SG5 [57]	EER: 20,59
Antipov et al. [48] (2019)	MFCC	4 эмбэдинга лиц, извлеченные с помощью 4-х вариаций CNN	Логистическая регрессия (Logistic regression – LogR)	Объединение на уровне оценок	NIST SRE19 [58]	EER: 0,6
Chenxi et al. [59] (2012)	MFCC	Признаки пирамидального фильтра Габора	Probabilistic Neural Network	Объединение на уровне признаков	Собственная	Accuracy: 100
Shi et al. [40] (2012)	MFCC	LBP	DTW	Правило суммы	Собственная	Total Error Rate: 0
Shen et al. [52] (2010)	MFCC	LBP	Gaussian Mixture Models + Universal Background Model + Local Binary Pattern	Объединение на уровне оценок	MOBIO [51]	EER: 21

лона, q_t — качества тестового образца. Таким образом, объединение происходит по следующему правилу:

$$LLR_{spk+face} = \sum_{i \in \{spk,face\}} a_i s_i + b_i q_{ei} + c_i q_{ti} + d,$$

где a_i, b_i, c_i, d — веса, полученные в ходе обучения. Результаты показали, что объединение на уровне баллов дает лучшую производительность на всех тестовых корпусах данных.

Сравнение всех рассмотренных методов аудиовизуального распознавания личности представлено в табл. 5.

Заключение

Распространение ношения масок существенно изменило область биометрических технологий. Рассмотрены результаты исследований, подтверждающих наличие новых проблем, с которыми столкнулись системы распознавания личности. Актуальность темы распознавания людей в средствах индивидуальной защиты подтверждается большим количеством новых исследований, призванных решить эту проблему.

В работе выполнен анализ основных корпусов данных аудио- и видеоинформации, содержащих изображения лиц и фонограммы голоса людей в масках.

На текущий момент наблюдается недостаток в обоих типах данных для создания автоматических систем. Представительных корпусов реальных изображений людей в масках очень мало, в открытом доступе находятся только два корпуса данных — RMFRD и MFSR. В условиях недостатка изображений предпринимаются попытки синтетического наложения масок на реальные изображения лиц. Однако данный тип корпусов не является универсальным для всех методов распознания лиц, так как синтезированные маски не всегда сохраняют необходимые признаки перекрываемой части лица. Для голосовых данных наблюдаются те же проблемы, в открытом доступе находятся всего 5 корпусов записей голоса общей продолжительностью порядка 60 ч.

Проведен сравнительный анализ современных методов распознания лиц в масках. Данная область хорошо исследована, существует большое количество разработанных систем. Исследованные системы демонстрируют высокую эффективность. Наилучшие показатели точности распознавания лиц 91,30 % и 95,05 % получены на корпусах RMFRD и MFDD. Однако дефицит данных усложняет оценку их производительности. Наблюдается недостаток исследований методов идентификации людей по голосу в масках, при этом выводы исследователей единогласно подтверждают снижение производительности существующих методов в новых условиях.

Сравнительный анализ бимодальных методов аудиовизуального распознавания личности без масок показал, что на данный момент исследования решения задачи распознавания людей в индивидуальной защите не найдены. Результаты анализа подтверждают применимость данных методов для задач, в которых наблюдается неполнота информации от каждой модальности. Это позволяет сделать предположение о применимости бимодальных методов к задаче распознавания людей в индивидуальной защите.

Согласно проанализированным работам, наибольшую эффективность для распознавания лиц в масках продемонстрировали вариации CNN, VGG-16 и FaceNet обеспечили точность распознавания до 100 % на собственном закрытом корпусе данных. Мультизадачная сеть MTCNN продемонстрировала точность распознавания лиц в 98,5 %, а сиамская вариация CNN показала 98,6 %. Большинство рассмотренных методов разработаны для распознавания личности только в медицинской маске, так как в современных условиях медицинская маска — самый распространенный предмет, носимый на лице. Однако, помимо медицинской маски, возможно ношение других средств индивидуальной защиты и предметов одежды. Актуальным направлением дальнейших исследований является расширение области применения текущих методов для случаев ношения всех возможных предметов, перекрывающих лицо. Перспективными направлениями развития существующих методов в области распознания лиц в маске является усовершенствование CNN и создание реальных баз данных изображений людей в масках.

В задаче распознавания личности по голосу методы i-векторного анализа показали незначительную деградацию точности при использовании маски. Данные методы совместно с CNN представляются наиболее перспективными для дальнейших исследований и создания бимодальных систем распознавания личности людей в медицинских масках.

На основе проведенных анализов сформулированы следующие требования, предъявляемые к перспективным системам распознавания людей в масках.

1. Высокая точность распознавания. Биометрические системы распознавания личности используются, в том числе для обеспечения безопасности людей, предоставления доступа к защищенным ресурсам, а также в криминалистической экспертизе. Для данных областей критически важно отсутствие ошибок. Важным направлением является проверка эффективности систем на реальных данных в натурных условиях, а не синтетических.
2. Одновременное использование аудио- и видеомодальностей. Система распознавания личности должна сохранять надежность работы при частичном отсутствии (неполноте) входных данных. Бимодальные системы способны преодолеть данные трудности за счет взаимного дополнения источников информации.
3. Робастность. Разработанные системы должны стабильно и надежно работать с различными типами входных данных и с данными различной степени качества. Изображения и записи голоса могут производиться в различных окружающих условиях, ракурсах, при динамическом освещении и акустической обстановке, так как запись может производиться как с близкого расстояния камерой персонального мобильного устройства, так и дистанционно с помощью камеры видеонаблюдения.

Литература

1. Chernenkova A. Facial recognition technology in Russia: do the citizens of Russia accept it? / University of Twente. BMS Faculty Department of Communication Science University of Twente. 2021. 78 p.
2. Кухарев Г.А., Матвеев Ю.Н., Форчманьски П. Поиск людей по фотороботам: методы, системы и практические решения // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 4. С. 640–653. https://doi.org/10.17586/2226-1494-2015-15-4-640-653
3. Grother P., Ngan M. Face recognition vendor test (FRVT): NIST Interagency Report 8009 / US Department of Commerce, National Institute of Standards and Technology. 2014. https://doi.org/10.6028/NIST.IR.8009

References

1. Chernenkova A. *Facial recognition technology in Russia: do the citizens of Russia accept it?* University of Twente. BMS Faculty Department of Communication Science University of Twente, 2021, 78 p.
2. Kukharev G.A., Matveev Yu.N., Forczmański P. People retrieval by means of composite pictures — methods, systems and practical decisions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 4, pp. 640–653. (in Russian). https://doi.org/10.17586/2226-1494-2015-15-4-640-653
3. Grother P., Ngan M. *Face recognition vendor test (FRVT)*. NIST Interagency Report 8009. US Department of Commerce, National

4. Двойникова А.А., Маркитантов М.В., Рюмина Е.В., Рюмин Д.А., Карпов А.А. Аналитический обзор аудиовизуальных систем для определения средств индивидуальной защиты на лице человека // Информатика и автоматизация. 2021. Т. 20. № 5. С. 1116–1152. https://doi.org/10.15622/20.5.5
5. Wang Z., Wang G., Huang B., Xiong Z., Hong Q., Wu H., Yi P., Jiang K., Wang N., Pei Y., Chen H., Miao Y., Huang Z., Liang J. Masked face recognition dataset and application // arXiv. 2020. arXiv:2003.09093.. https://doi.org/10.48550/arXiv.2003.09093
6. Geng M., Peng P., Huang Y., Tian Y. Masked face recognition with generative data augmentation and domain constrained ranking // Proc. of the 28th ACM International Conference on Multimedia. 2020. P. 2246–2254. https://doi.org/10.1145/3394171.3413723
7. Din N.U., Javed K., Bae S., Yi J. A novel GAN-based network for unmasking of masked face // IEEE Access. 2020. V. 8. P. 44276–44287. https://doi.org/10.1109/ACCESS.2020.2977386
8. Din N.U., Javed K., Bae S., Yi J. Effective removal of user-selected foreground object from facial images using a novel GAN-based network // IEEE Access. 2020. V. 8. P. 109648–109661. https://doi.org/10.1109/ACCESS.2020.3001649
9. Deng J., Guo J., Xue N., Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition // Proc. of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 4685–4694. https://doi.org/10.1109/CVPR.2019.00482
10. Boutros F., Damer N., Kirchbuchner F., Kuijper A. Self-restrained triplet loss for accurate masked face recognition // Pattern Recognition. 2022. V. 124. P. 108473. https://doi.org/10.1016/j.patcog.2021.108473
11. Anwar A., Raychowdhury A. Masked face recognition for secure authentication // arXiv. 2020. arXiv:2008.11104. https://doi.org/10.48550/arXiv.2008.11104
12. Damer N., Grebe J.H., Chen C., Boutros F., Kirchbuchner F., Kuijper A. The effect of wearing a mask on face recognition performance: an exploratory study // Proc. of the 19th International Conference of the Biometrics Special Interest Group (BIOSIG). 2020. P. 9210999.
13. Huang G.B., Ramesh M., Berg T., Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments // Workshop on Faces in'RealLife'Images: Detection, Alignment, and Recognition. 2008.
14. Ding F., Peng P., Huang Y., Geng M., Tian Y. Masked face recognition with latent part detection // Proc. of the 28th ACM international Conference on Multimedia. 2020. P. 2281–2289. https://doi.org/10.1145/3394171.3413731
15. Li S., Yi D., Lei Z., Liao S. The CASIA NIR-VIS 2.0 face database // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2013. P. 348–353. https://doi.org/10.1109/CVPRW.2013.59
16. Deng H., Feng Z., Qian G., Lv X., Li H., Li G. MFCosface: a masked-face recognition algorithm based on large margin cosine loss // Applied Sciences. 2021. V. 11. N 16. P. 7310. https://doi.org/10.3390/app11167310
17. Yi D., Lei Z., Liao S., Li S.Z. Learning face representation from scratch // arXiv. 2014. arXiv:1411.7923. https://doi.org/10.48550/arXiv.1411.7923
18. Hong J.H., Kim H., Kim M., Nam G.P., Cho J., Ko H.-S., Kim I.-J. A 3D model-based approach for fitting masks to faces in the wild // Proc. of the IEEE International Conference on Image Processing (ICIP). 2021. P. 235–239. https://doi.org/10.1109/ICIP42928.2021.9506069
19. Hariri W. Efficient masked face recognition method during the COVID-19 pandemic // Signal, Image and Video Processing. 2022. V. 16. N 3. P. 605–612. https://doi.org/10.1007/s11760-021-02050-w
20. Maharani D.A., MacHbub C., Rusmin P.H., Yulianti L. Improving the capability of real-time face masked recognition using cosine distance // Proc. of the 6th International Conference on Interactive Digital Media (ICIDM). 2020. P. 9339677. https://doi.org/10.1109/ICIDM51048.2020.9339677
21. Montero D., Nieto M., Leskovsky P., Aginako N. Boosting masked face recognition with multi-task arcface // arXiv. 2021. arXiv:2104.09874. https://doi.org/10.48550/arXiv.2104.09874
22. Golwalkar R., Mehendale N. Masked Face Recognition Using Deep Metric Learning and FaceMaskNet21 // SSRN Electronic Journal. 2020. P. 3731223. http://dx.doi.org/10.2139/ssrn.3731223
23. Li C., Ge S., Zhang D., Li J. Look through masks: Towards masked face recognition with de-occlusion distillation // Proc. of the 28th Institute of Standards and Technology, 2014. https://doi.org/10.6028/NIST.IR.8009
4. Dvoynikova A., Markitantov M., Ryumina E., Ryumin D., Karpov A. Analytical review of audiovisual systems for determining personal protective equipment on a person's face. *Informatics and Automation*, 2021, vol. 20, no. 5, pp. 1116–1152. (in Russian). https://doi.org/10.15622/20.5.5
5. Wang Z., Wang G., Huang B., Xiong Z., Hong Q., Wu H., Yi P., Jiang K., Wang N., Pei Y., Chen H., Miao Y., Huang Z., Liang J. Masked face recognition dataset and application. *arXiv*, 2020, arXiv:2003.09093. https://doi.org/10.48550/arXiv.2003.09093
6. Geng M., Peng P., Huang Y., Tian Y. Masked face recognition with generative data augmentation and domain constrained ranking. *Proc. of the 28th ACM International Conference on Multimedia*, 2020, pp. 2246–2254. https://doi.org/10.1145/3394171.3413723
7. Din N.U., Javed K., Bae S., Yi J. A novel GAN-based network for unmasking of masked face. *IEEE Access*, 2020, vol. 8, pp. 44276–44287. https://doi.org/10.1109/ACCESS.2020.2977386
8. Din N.U., Javed K., Bae S., Yi J. Effective removal of user-selected foreground object from facial images using a novel GAN-based network. *IEEE Access*, 2020, vol. 8, pp. 109648–109661. https://doi.org/10.1109/ACCESS.2020.3001649
9. Deng J., Guo J., Xue N., Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. *Proc. of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694. https://doi.org/10.1109/CVPR.2019.00482
10. Boutros F., Damer N., Kirchbuchner F., Kuijper A. Self-restrained triplet loss for accurate masked face recognition. *Pattern Recognition*, 2022, vol. 124, pp. 108473. https://doi.org/10.1016/j.patcog.2021.108473
11. Anwar A., Raychowdhury A. Masked face recognition for secure authentication. *arXiv*, 2020, arXiv:2008.11104. https://doi.org/10.48550/arXiv.2008.11104
12. Damer N., Grebe J.H., Chen C., Boutros F., Kirchbuchner F., Kuijper A. The effect of wearing a mask on face recognition performance: an exploratory study. *Proc. of the 19th International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020, pp. 9210999.
13. Huang G.B., Ramesh M., Berg T., Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Workshop on Faces in'RealLife'Images: Detection, Alignment, and Recognition*, 2008.
14. Ding F., Peng P., Huang Y., Geng M., Tian Y. Masked face recognition with latent part detection. *Proc. of the 28th ACM international Conference on Multimedia*, 2020, pp. 2281–2289. https://doi.org/10.1145/3394171.3413731
15. Li S., Yi D., Lei Z., Liao S. The CASIA NIR-VIS 2.0 face database. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353. https://doi.org/10.1109/CVPRW.2013.59
16. Deng H., Feng Z., Qian G., Lv X., Li H., Li G. MFCosface: a masked-face recognition algorithm based on large margin cosine loss. *Applied Sciences*, 2021, vol. 11, no. 16, pp. 7310. https://doi.org/10.3390/app11167310
17. Yi D., Lei Z., Liao S., Li S.Z. Learning face representation from scratch. *arXiv*, 2014, arXiv:1411.7923. https://doi.org/10.48550/arXiv.1411.7923
18. Hong J.H., Kim H., Kim M., Nam G.P., Cho J., Ko H.-S., Kim I.-J. A 3D model-based approach for fitting masks to faces in the wild. *Proc. of the IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 235–239. https://doi.org/10.1109/ICIP42928.2021.9506069
19. Hariri W. Efficient masked face recognition method during the covid-19 pandemic. *Signal, Image and Video Processing*, 2022, vol. 16, no. 3, pp. 605–612. https://doi.org/10.1007/s11760-021-02050-w
20. Maharani D.A., MacHbub C., Rusmin P.H., Yulianti L. Improving the capability of real-time face masked recognition using cosine distance. *Proc. of the 6th International Conference on Interactive Digital Media (ICIDM)*, 2020, pp. 9339677. https://doi.org/10.1109/ICIDM51048.2020.9339677
21. Montero D., Nieto M., Leskovsky P., Aginako N. Boosting masked face recognition with multi-task arcface. *arXiv*, 2021, arXiv:2104.09874. https://doi.org/10.48550/arXiv.2104.09874
22. Golwalkar R., Mehendale N. Masked Face Recognition Using Deep Metric Learning and FaceMaskNet21. *SSRN Electronic Journal*, 2020, pp. 3731223. http://dx.doi.org/10.2139/ssrn.3731223

ACM International Conference on Multimedia. 2020. P. 3016–3024. https://doi.org/10.1145/3394171.3413960
24. Hong Q., Wang Z., He Z., Wang N., Tian X., Lu T. Masked face recognition with identification association // Proc. of the IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). 2020. P. 731–735. https://doi.org/10.1109/ICTAI50040.2020.00116
25. Mandal B., Okeukwu A., Theis Y. Masked face recognition using ResNet-50 // arXiv. 2021. arXiv:2104.08997. https://doi.org/10.48550/arXiv.2104.08997
26. Du H., Shi H., Liu Y., Zeng D., Mei T. Towards NIR-VIS masked face recognition // IEEE Signal Processing Letters. 2021. V. 28. P. 768–772. https://doi.org/10.1109/LSP.2021.3071663
27. Wu G.L. Masked face recognition algorithm for a contactless distribution cabinet // Mathematical Problem in Engineering. 2021. V. 2021. P. 5591020. https://doi.org/10.1155/2021/5591020
28. Li Y., Guo K., Lu Y., Liu L. Cropping and attention based approach for masked face recognition // Applied Intelligence. 2021. V. 51. N 5. P. 3012–3025. https://doi.org/10.1007/s10489-020-02100-9
29. Gover A.R., Harper S.B., Langton L. Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality // American Journal of Criminal Justice. 2020. V. 45. N 4. P. 647–667. https://doi.org/10.1007/s12103-020-09545-1
30. Saeidi R., Niemi T., Karppelin H., Pohjalainen J., Kinnunen T., Alku P. Speaker recognition for speech under face cover // Proc. of the Interspeech. 2015. P. 1012–1016. https://doi.org/10.21437/Interspeech.2015-275
31. Zhang C., Tan T. Voice disguise and automatic speaker recognition // Forensic Science International. 2008. V. 175. N 2-3. P. 118–122. https://doi.org/10.1016/j.forsciint.2007.05.019
32. Fecher N. The “audio-visual face cover corpus”: investigations into audio-visual speech and speaker recognition when the speaker’s face is occluded by facewear // Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH). 2012. P. 2250–2253. https://doi.org/10.21437/Interspeech.2012-133
33. Iszatt T., Malkoc E., Kelly F., Alexander A. Exploring the impact of face coverings on x-vector speaker recognition using VOCALISE // Poc. of the Conference: International Association of Forensic Phonetics and Acoustics. 2021.
34. Schuller B.W., Batliner A., Bergler C., Messner E.-M., Hamilton A., Amiriparian S., Baird A., Rizos G., Schmitt M., Stappen L., Baumeister H., MacIntyre D.A., Hantke S. The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks // Proc. of the Interspeech. 2020. P. 2042–2046. https://doi.org/10.21437/Interspeech.2020-32
35. Montacié C., Caraty M. J. Phonetic, frame clustering and intelligibility analyses for the INTERSPEECH 2020 ComParE Challenge // Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH). 2020. P. 2062–2066. https://doi.org/10.21437/Interspeech.2020-2243
36. Рюмина Е.В., Рюмин Д.А., Маркитантов М.В., Карпов А.А. Метод генерации обучающих данных для компьютерной системы обнаружения защитных масок на лицах людей // Компьютерная оптика. 2022. Т. 46. в печати.
37. Przybocki M.A., Martin A.F., Le A.N. NIST speaker recognition evaluations utilizing the Mixer corpora–2004, 2005, 2006 // IEEE Transactions on Audio, Speech, and Language Processing. 2007. V. 15. N 7. P. 1951–1959. https://doi.org/10.1109/TASL.2007.902489
38. Alam M.R., Bennamoun M., Togneri R., Sohel F. An efficient reliability estimation technique for audio-visual person identification // Proc. of the IEEE 8th Conference on Industrial Electronics and Applications (ICIEA). 2013. P. 1631–1635. https://doi.org/10.1109/ICIEA.2013.6566630
39. Zhao X., Evans N., Dugelay J.-L. Multi-view semi-supervised discriminant analysis: A new approach to audio-visual person recognition // Proc. of the 20th European Signal Processing Conference (EUSIPCO). 2012. P. 31–35.
40. Nishino T., Kajikawa Y., Muneyasu M. Multimodal person authentication system using features of utterance // Proc. of the 20th International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS). 2012. P. 43–47. https://doi.org/10.1109/ISPACS.2012.6473450
41. Alam M.R., Bennamoun M., Togneri R., Sohel F. A joint deep Boltzmann machine (jDBM) model for person identification using mobile phone data // IEEE Transactions on Multimedia. 2017. V. 19. N 2. P. 317–326. https://doi.org/10.1109/TMM.2016.2615524

23. Li C., Ge S., Zhang D., Li J. Look through masks: Towards masked face recognition with de-occlusion distillation. *Proc. of the 28th ACM International Conference on Multimedia*, 2020, pp. 3016–3024. https://doi.org/10.1145/3394171.3413960
24. Hong Q., Wang Z., He Z., Wang N., Tian X., Lu T. Masked face recognition with identification association. *Proc. of the IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 731–735. https://doi.org/10.1109/ICTAI50040.2020.00116
25. Mandal B., Okeukwu A., Theis Y. Masked face recognition using ResNet-50. *arXiv*, 2021, arXiv:2104.08997. https://doi.org/10.48550/arXiv.2104.08997
26. Du H., Shi H., Liu Y., Zeng D., Mei T. Towards NIR-VIS masked face recognition. *IEEE Signal Processing Letters*, 2021, vol. 28, pp. 768–772. https://doi.org/10.1109/LSP.2021.3071663
27. Wu G.L. Masked face recognition algorithm for a contactless distribution cabinet. *Mathematical Problem in Engineering*, 2021, vol. 2021, pp. 5591020. https://doi.org/10.1155/2021/5591020
28. Li Y., Guo K., Lu Y., Liu L. Cropping and attention based approach for masked face recognition. *Applied Intelligence*, 2021, vol. 51, no. 5, pp. 3012–3025. https://doi.org/10.1007/s10489-020-02100-9
29. Gover A.R., Harper S.B., Langton L. Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality. *American Journal of Criminal Justice*, 2020, vol. 45, no. 4, pp. 647–667. https://doi.org/10.1007/s12103-020-09545-1
30. Saeidi R., Niemi T., Karppelin H., Pohjalainen J., Kinnunen T., Alku P. Speaker recognition for speech under face cover. *Proc. of the Interspeech*, 2015, pp. 1012–1016. https://doi.org/10.21437/Interspeech.2015-275
31. Zhang C., Tan T. Voice disguise and automatic speaker recognition. *Forensic Science International*, 2008, vol. 175, no. 2-3, pp. 118–122. https://doi.org/10.1016/j.forsciint.2007.05.019
32. Fecher N. The “audio-visual face cover corpus”: investigations into audio-visual speech and speaker recognition when the speaker’s face is occluded by facewear. *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012, pp. 2250–2253. https://doi.org/10.21437/Interspeech.2012-133
33. Iszatt T., Malkoc E., Kelly F., Alexander A. Exploring the impact of face coverings on x-vector speaker recognition using VOCALISE. *Poc. of the Conference: International Association of Forensic Phonetics and Acoustics*, 2021.
34. Schuller B.W., Batliner A., Bergler C., Messner E.-M., Hamilton A., Amiriparian S., Baird A., Rizos G., Schmitt M., Stappen L., Baumeister H., MacIntyre D.A., Hantke S. The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. *Proc. of the Interspeech*, 2020, pp. 2042–2046. https://doi.org/10.21437/Interspeech.2020-32
35. Montacié C., Caraty M. J. Phonetic, frame clustering and intelligibility analyses for the INTERSPEECH 2020 ComParE Challenge. *Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2062–2066. https://doi.org/10.21437/Interspeech.2020-2243
36. Riumina E.V., Riumin D.A., Markitantov M.V., Karpov A.A. A method for generating training data for a protective face mask detection system. *Computer Optics*, 2022, vol. 46. in press (in Russian).
37. Przybocki M.A., Martin A.F., Le A.N. NIST speaker recognition evaluations utilizing the Mixer corpora–2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, vol. 15, no. 7, pp. 1951–1959. https://doi.org/10.1109/TASL.2007.902489
38. Alam M.R., Bennamoun M., Togneri R., Sohel F. An efficient reliability estimation technique for audio-visual person identification. *Proc. of the IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, 2013, pp. 1631–1635. https://doi.org/10.1109/ICIEA.2013.6566630
39. Zhao X., Evans N., Dugelay J.-L. Multi-view semi-supervised discriminant analysis: A new approach to audio-visual person recognition. *Proc. of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 31–35.
40. Nishino T., Kajikawa Y., Muneyasu M. Multimodal person authentication system using features of utterance. *Proc. of the 20th International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)*, 2012, pp. 43–47. https://doi.org/10.1109/ISPACS.2012.6473450
41. Alam M.R., Bennamoun M., Togneri R., Sohel F. A joint deep Boltzmann machine (jDBM) model for person identification using

42. Gofman M., Sandico N., Mitra S., Suo E., Muhi S., Vu T. Multimodal biometrics via discriminant correlation analysis on mobile devices // Proc. of the International Conference on Security and Management (SAM). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). 2018. P. 174–181.
43. Shenoy R.V. Hidden Markov Models for Analysis of Multimodal Biomedical Images: A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical and Computer Engineering / University of California, Santa Barbara, 2016. 99 p.
44. Chen Y., Yang J., Wang C., Liu N. Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine // Expert Systems with Applications. 2016. V. 64. P. 93–103. https://doi.org/10.1016/j.eswa.2016.07.009
45. Garau M., Fraschini M., Didaci L., Marcialis G.L. Experimental results on multi-modal fusion of EEG-based personal verification algorithms // Proc. of the 9th International Conference on Biometrics (ICB). 2016. P. 7550080. https://doi.org/10.1109/ICB.2016.7550080
46. Shen L., Zheng N., Zheng S., Li W. Secure mobile services by face and speech based personal authentication // Proc. of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS). V. 3. 2010. P. 97–100. https://doi.org/10.1109/ICICISYS.2010.5658534
47. Irfan B., Ortiz M.-G., Lyubova N., Belpaeme T. Multi-modal open-set person identification in HRI // Proc. of the 2018 ACM/IEEE International Conference on Human-Robot Interaction Social Robots in the Wild workshop. 2018.
48. Antipov G., Gengembre N., Le Blouch O., Le Lan G. Automatic quality assessment for audio-visual verification systems. The LOVe submission to NIST SRE challenge 2019 // Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH). 2020. P. 2237–2241. https://doi.org/10.21437/Interspeech.2020-1434
49. Estival D., Cassidy S., Cox F., Burnham D. AusTalk: an audio-visual corpus of Australian English // Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014.
50. Khoury E., El Shafey L., McCool C., Günther M., Marcel S. Bi-modal biometric authentication on mobile phones in challenging conditions // Image and Vision Computing. 2014. V. 32. N 12. P. 1147–1160. https://doi.org/10.1016/j.imavis.2013.10.001
51. McCool C., Marcel S. MOBIO database for the ICPR 2010 face and speech competition: Idiap Communication Report. Idiap Research Institute, 2009.
52. Alam M.R., Togneri R., Sohel F., Bennamoun M., Naseem I. Linear regression-based classifier for audio visual person identification // Proc. of the 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA). 2013. P. 6487281. https://doi.org/10.1109/ICCSPA.2013.6487281
53. Tresadern P., Cootes T.F., Poh N., Matejka P., Hadid A., Lévy C., McCool C., Marcel S. Mobile biometrics: Combined face and voice verification for a mobile platform // IEEE Pervasive Computing. 2013. V. 12. N 1. P. 79–87. https://doi.org/10.1109/MPRV.2012.54
54. Islam R., Sobhan A. BPN based likelihood ratio score fusion for audio-visual speaker identification in response to noise // International Scholarly Research Notices. 2014. V. 2014. P. 737814. https://doi.org/10.1155/2014/737814
55. Primorac R., Togneri R., Bennamoun M., Sohel F. Audio-visual biometric recognition via joint sparse representations // Proc. of the 23rd International Conference on Pattern Recognition (ICPR). 2016. P. 3031–3035. https://doi.org/10.1109/ICPR.2016.7900099
56. Memon Q., AlKassim Z., AlHassan E., Omer M., Alsiddig M. Audio-visual biometric authentication for secured access into personal devices // Proc. of the 6th International Conference on Bioinformatics and Biomedical Science (ICBBS). 2017. P. 85–89. https://doi.org/10.1145/3121138.3121165
57. Gofman M.I., Mitra S., Cheng T.-H.K., Smith N.T. Multimodal biometrics for enhanced mobile device security // Communications of the ACM. 2016. V. 59. N 4. P. 58–65. https://doi.org/10.1145/2818990
58. Sadjadi S.O., Greenberg C., Singer E., Reynolds D., Mason L., Hernandez-Cordero J. The 2019 NIST speaker recognition evaluation CTS challenge // Proc. of the Speaker and Language Recognition Workshop (Odyssey 2020). 2020. P. 266–272. https://doi.org/10.21437/Odyssey.2020-38

mobile phone data. *IEEE Transactions on Multimedia*, 2017, vol. 19, no. 2, pp. 317–326. https://doi.org/10.1109/TMM.2016.2615524
42. Gofman M., Sandico N., Mitra S., Suo E., Muhi S., Vu T. Multimodal biometrics via discriminant correlation analysis on mobile devices. *Proc. of the International Conference on Security and Management (SAM)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2018, P. 174–181.
43. Shenoy R.V. *Hidden Markov Models for Analysis of Multimodal Biomedical Images*: A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical and Computer Engineering. University of California, Santa Barbara, 2016, 99 p.
44. Chen Y., Yang J., Wang C., Liu N. Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine. *Expert Systems with Applications*, 2016, vol. 64, pp. 93–103. https://doi.org/10.1016/j.eswa.2016.07.009
45. Garau M., Fraschini M., Didaci L., Marcialis G.L. Experimental results on multi-modal fusion of EEG-based personal verification algorithms. *Proc. of the 9th International Conference on Biometrics (ICB)*, 2016, pp. 7550080. https://doi.org/10.1109/ICB.2016.7550080
46. Shen L., Zheng N., Zheng S., Li W. Secure mobile services by face and speech based personal authentication. *Proc. of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*. Vol. 3, 2010, pp. 97–100. https://doi.org/10.1109/ICICISYS.2010.5658534
47. Irfan B., Ortiz M.-G., Lyubova N., Belpaeme T. Multi-modal open-set person identification in HRI. *Proc. of the 2018 ACM/IEEE International Conference on Human-Robot Interaction Social Robots in the Wild workshop*, 2018.
48. Antipov G., Gengembre N., Le Blouch O., Le Lan G. Automatic quality assessment for audio-visual verification systems. The LOVe submission to NIST SRE challenge 2019. *Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2237–2241. https://doi.org/10.21437/Interspeech.2020-1434
49. Estival D., Cassidy S., Cox F., Burnham D. AusTalk: an audio-visual corpus of Australian English. *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.
50. Khoury E., El Shafey L., McCool C., Günther M., Marcel S. Bi-modal biometric authentication on mobile phones in challenging conditions. *Image and Vision Computing*, 2014, vol. 32, no. 12, pp. 1147–1160. https://doi.org/10.1016/j.imavis.2013.10.001
51. McCool C., Marcel S. *MOBIO database for the ICPR 2010 face and speech competition*. Idiap Communication Report. Idiap Research Institute, 2009.
52. Alam M.R., Togneri R., Sohel F., Bennamoun M., Naseem I. Linear regression-based classifier for audio visual person identification. *Proc. of the 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013, pp. 6487281. https://doi.org/10.1109/ICCSPA.2013.6487281
53. Tresadern P., Cootes T.F., Poh N., Matejka P., Hadid A., Lévy C., McCool C., Marcel S. Mobile biometrics: Combined face and voice verification for a mobile platform. *IEEE Pervasive Computing*, 2013, vol. 12, no. 1, pp. 79–87. https://doi.org/10.1109/MPRV.2012.54
54. Islam R., Sobhan A. BPN based likelihood ratio score fusion for audio-visual speaker identification in response to noise. *International Scholarly Research Notices*, 2014, vol. 2014, pp. 737814. https://doi.org/10.1155/2014/737814
55. Primorac R., Togneri R., Bennamoun M., Sohel F. Audio-visual biometric recognition via joint sparse representations. *Proc. of the 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3031–3035. https://doi.org/10.1109/ICPR.2016.7900099
56. Memon Q., AlKassim Z., AlHassan E., Omer M., Alsiddig M. Audio-visual biometric authentication for secured access into personal devices. *Proc. of the 6th International Conference on Bioinformatics and Biomedical Science (ICBBS)*, 2017, pp. 85–89. https://doi.org/10.1145/3121138.3121165
57. Gofman M.I., Mitra S., Cheng T.-H.K., Smith N.T. Multimodal biometrics for enhanced mobile device security. *Communications of the ACM*, 2016, vol. 59, no. 4, pp. 58–65. https://doi.org/10.1145/2818990
58. Sadjadi S.O., Greenberg C., Singer E., Reynolds D., Mason L., Hernandez-Cordero J. The 2019 NIST speaker recognition evaluation CTS challenge. *Proc. of the Speaker and Language Recognition*

59. Yu C., Huang L. Biometric recognition by using audio and visual feature fusion // Proc. of the 2012 International Conference on System Science and Engineering (ICSSE). 2012. P. 173178. https://doi.org/10.1109/ICSSE.2012.6257171

Workshop (Odyssey 2020), 2020, pp. 266–272. https://doi.org/10.21437/Odyssey.2020-38
59. Yu C., Huang L. Biometric recognition by using audio and visual feature fusion. *Proc. of the 2012 International Conference on System Science and Engineering (ICSSE)*, 2012, pp. 173178. https://doi.org/10.1109/ICSSE.2012.6257171

Авторы

Косулин Кирилл Эдгарович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; программист, Филиал ООО «Люксофт Профеншнл» в городе Санкт-Петербург, Санкт-Петербург, 195027, Российская Федерация, https://orcid.org/0000-0002-1324-2813, leliclalelic@yandex.ru

Карпов Алексей Анатольевич — доктор технических наук, профессор, Лаборатория речевых и многомодальных интерфейсов, главный научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57195330987, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Authors

Kirill E. Kosulin — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation; Software Developer, Luxoft Branch, Saint Petersburg, 195027, Russian Federation, https://orcid.org/0000-0002-1324-2813, leliclalelic@yandex.ru

Alexey A. Karpov — D. Sc., Professor, Chief Researcher, Multimodal Interfaces Laboratory at the Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57195330987, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Статья поступила в редакцию 18.02.2022
Одобрена после рецензирования 31.03.2022
Принята к печати 31.05.2022

Received 18.02.2022
Approved after reviewing 31.03.2022
Accepted 31.05.2022