УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2021 Том 21 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2021 Vol. 21 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

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

doi: 10.17586/2226-1494-2021-21-5-755-766
УДК 004.932.2

Защита изображений лиц от распознавания в социальных сетях: способы решения и их перспективы

Георгий Александрович Кухарев[1], Калыбек Сапарович Мауленов[2], Надежда Львовна Щеголева[3]✉

[1] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация
[2] Некоммерческое акционерное общество «Костанайский региональный университет имени А. Байтурсынова», Костанай, 110000, Республика Казахстан
[3] Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация

[1] gakukharev@etu.ru; kuga41@mail.ru, https://orcid.org/0000-0003-2188-2172
[2] k_maulenov@inbox.ru, https://orcid.org/0000-0003-4147-3843
[3] stil_hope@mail.ru✉, https://orcid.org/0000-0003-1087-2833

Аннотация

Предмет исследования. Исследована проблема несанкционированного использования в глубоком обучении изображений лиц из социальных сетей. Рассмотрены методы защиты таких изображений от их использования и распознавания на базе процедур де-идентификации и новейшей из них — процедуре Fawkes. **Метод.** Задача решена путем сравнительного анализа изображений, подвергнутых процедуре Fawkes-преобразования, представления и описания текстурных изменений и особенностей структурных разрушений в изображениях лиц. Применены многоуровневые параметрические оценки разрушений для их формальной и численной оценки. **Основные результаты.** Объяснены причины невозможности использования изображений лиц, разрушенных в процессе выполнения процедуры Fawkes, в задачах глубокого обучения. Теоретически доказано и экспериментально показано, что изображения лиц, подвергнутые процедуре Fawkes, хорошо распознаются вне методов глубоко обучения. **Практическая значимость.** Утверждается, что использование простых способов предобработки изображений лиц, подвергнутых процедуре Fawkes, на входе сверточной нейронной сети может привести к их распознаванию с высокой результативностью, что разрушает существующее представление о значимости защиты изображений лиц процедурой Fawkes.

Ключевые слова

социальные сети, несанкционированный доступ, глубокое обучение, защита изображений лиц, де-идентификация, процедура Fawkes, детерминированные методы распознавания

Ссылка для цитирования: Кухарев Г.А., Мауленов К.С., Щеголева Н.Л. Защита изображений лиц от распознавания в социальных сетях: способы решения и их перспективы // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 5. С. 755–766. doi: 10.17586/2226-1494-2021-21-5-755-766

Protecting facial images from recognition on social media: solution methods and their perspective

Georgy A. Kukharev[1], Kalybek S. Maulenov[2], Nadezhda L. Shchegoleva[3]✉

[1] Saint Petersburg State Electrotechnical University (LETI), Saint Petersburg, 197376, Russian Federation
[2] Non-profit limited company “A. Baitursynov Kostanay Regional University”, Kostanay, 110000, Republic of Kazakhstan
[3] Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation

[1] gakukharev@etu.ru; kuga41@mail.ru, https://orcid.org/0000-0003-2188-2172
[2] k_maulenov@inbox.ru, https://orcid.org/0000-0003-4147-3843
[3] stil_hope@mail.ru✉, https://orcid.org/0000-0003-1087-2833

Abstract
The paper deals with the problem of unauthorized use in deep learning of facial images from social networks and analyses methods of protecting such images from their use and recognition based on de-identification procedures and the newest of them — the "Fawkes" procedure. The proposed solution uses a comparative analysis of images subjected to the Fawkes-transformation procedure, representation and description of textural changes and features of structural damage in facial images. Multilevel parametric estimates of these damages were applied for their formal and numerical assessment. The reasons for the impossibility of using images of faces destroyed by the Fawkes procedure in deep learning tasks are explained. It has been theoretically proven and experimentally shown that facial images subjected to the Fawkes procedure are well recognized outside of deep learning methods. It is argued that the use of simple preprocessing methods for facial images (subjected to the Fawkes procedure) at the entrance to convolutional neural networks can lead to their recognition with high efficiency, which destroys the myth about the importance of protecting facial images with the Fawkes-procedure.

Keywords
social networks, unauthorized access, deep learning, face image protection, de-identification, Fawkes procedure, deterministic recognition methods

For citation: Kukharev G.A., Maulenov K.S., Shchegoleva N.L. Protecting facial images from recognition on social media: solution methods and their perspective. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 755–766 (in Russian). doi: 10.17586/2226-1494-2021-21-5-755-766

Введение

В конце XX — начале XXI веков отмечено появление различных платформ, сервисов и сайтов для хранения и обмена фотографиями, что послужило началом создания социальных сетей во всем мире. Первоначально социальные сети охватывали локальные группы школьников и студентов, любителей музыки и людей по общим профессиям, а в последствии они объединяли в общие социальные группы различных людей, ищущих друзей и собеседников. С развитием социальных сетей люди стали не только обмениваться сообщениями друг с другом, но и практически стали жить в социальных сетях, представляя себя и своих друзей в бесконечных фото- и видеофильмах. Так, например, социальная сеть «Facebook»[1] поддерживает на сегодня почти три миллиарда аккаунтов активных пользователей. Очень быстро общие объемы фотоизображений лиц и фотопортретов пользователей во всех социальных сетях, платформах и хостингах обмена фотографиями перевалили за десятки миллиардов. И все эти фото стали доступными, чем воспользовались компании, разрабатывающие технологии распознавания лиц.

Например, российская компания NTech.Lab с технологией FindFace реализовала поиск фото людей в социальной сети «ВКонтакте», а американская компания Clearview AI собрала более трех миллиардов фотопортретов из социальных сетей Facebook и Venmo, видео хостинга YouTube и других подобных платформ, и сайтов[2].

На базе собранных фото и таких, например, как база MegaFace[3], IT компании научили нейронные сети не только распознавать людей по лицам, но и собирать все фото о людях с различных источников, связывая их с аккаунтами в социальных сетях, что можно было рассматривать, как слежка за людьми. В результате этого были созданы два негативных прецедента. Первый — несанкционированный сбор фото граждан, а второй — создание систем распознавания, обученных на подобном сборе данных, и продажа этих систем частным компаниям. Именно это и привело к возможности распознавания людей с различных источников без ведома и согласия этих людей, что фактически привело к вторжению в личное пространство граждан во всем мире, т. е. к нарушению их конфиденциальности. На ежегодной конференции Global Media Forum[4], открывшейся 14.06.2021 года, Хоан Тон-Тат — генеральный директор компании Clearview AI, заявил: «Вся информация собирается на законных основаниях и является общедоступной».

А вот уже и совершенно немыслимый пример. В ноябре 2017 года профессор Калифорнийского университета в Беркли (UCB) Стюарт Расселл (Stuart Russell) представил видео о рое крошечных автономных боевых дронов[5]. Такие дроны способны самостоятельно прокладывать маршрут к цели, узнавать человека по лицу, а затем направленным микровзрывом пробивать ему голову, причиняя травмы мозга, несовместимые с жизнью. При этом данные о цели боевые дроны также получают из социальных сетей.

Естественно, что эти прецеденты вызвали широкий общественный резонанс и поиски решений в защите личных фото от распознавания. Начало этому было положено в технологии де-идентификации изображений лиц «Face De-identification» [1–4] — математической процедуры искажения формы или текстуры исходного

[1] Top 15 Most Popular Social Networking Sites and Apps [Электронный ресурс]. Режим доступа: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/ (дата обращения: 06.07.2021).

[2] Hill Kashmir. The secretive company that might end privacy as we know it [Электронный ресурс]. Режим доступа: www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (дата обращения: 06.07.2021).

[3] MegaFace and MF2: Million-Scale Face Recognition [Электронный ресурс]. Режим доступа: http://megaface.cs.washington.edu/ (дата обращения: 06.07.2021).

[4] Как работает система распознавания лиц Clearview AI и что с ней не так [Электронный ресурс]. Режим доступа: www.dw.com/ru/kak-rabotaet-sistema-raspoznavanija-lic-clearview-ai-i-chto-s-nej-ne-tak/a-57905656 (дата обращения: 06.07.2021).

[5] Крошечные дроны-убийцы могут совершать массовые атаки на людей [Электронный ресурс]. Режим доступа: www.youtube.com/watch?v=4W3s0xh32Ag (дата обращения: 06.07.2021).

изображения так, чтобы человеку это лицо было понятным, а компьютеру нет. При этом использовались как сложные методы создания популяций людей (например, моделирование изменения лиц на основе процедур триангуляции и методов двумерной проекции в собственные подпространства), так и простые — сглаживающие или шумовые фильтры. Однако эти решения подходили для небольших корпоративных баз изображений лиц, но не для социальных сетей с большими объемами изображений. В дальнейшем в технологии де-идентификации появились и новые предложения защиты изображений лиц, основанные на глубоком обучении [5, 6]. Де-идентификация изображений лиц состояла здесь в том, чтобы изменить лицо, но сохранить его основные визуальные атрибуты, такие как пол, поза, эмоции и возраст, тем самым сохраняя естественность изображения. Иногда вместе с измененным изображением сохранялись и все параметры изменений, которые позволяли реконструировать исходный оригинал.

Новым революционным решением де-идентификации изображений лиц на базе искусственного интеллекта стала процедура Fawkes[1], реализующая преобразование изображений лиц, которое делает их непригодными для использования в технологии глубокого обучения (Deep Learning, DL) [7]. Разработчики этой процедуры утверждают, что при Fawkes-преобразовании, изображения лиц не претерпевают значимых искажений, но изменяются (маскируются или разрушаются) так, чтобы стать бесполезными в задаче обучения CNN (Convolutional Neural Network), а, стало быть, и не будут ими распознаны. Подробнее эти характеристики можно найти по ссылке[1]. Там же есть переход на видео доклада на конференции «USENIX Security 2020», где была представлена идея Fawkes. Кроме того, авторы процедуры Fawkes разместили адреса доступа к своим программам и описаниям по параметрам управления процессом Fawkes-преобразования, а также предложили всем желающим пользоваться этими программами для защиты своих изображений, перед тем как размещать их в открытых социальных сетях.

Изменения в изображении-оригинале после процедуры Fawkes

Попытаемся ответить на вопрос, что изменяется в изображении-оригинале после процедуры Fawkes, как обнаружить эти изменения и оценить их численно.

Прежде всего отметим, что если у нас есть только результат Fawkes-преобразования, но нет изображения-оригинала, то, действительно, как отмечают авторы процедуры Fawkes, мы ничего не сможем увидеть и/или оценить. При этом использовать соответствующие пары изображений лиц из статьи[1] и статьи [7] также не представляется возможным, поскольку представленные в этих материалах твердые копии изображений (оригинала и результата его Fawkes-преобразования) либо малоразмерны, либо содержат искажения как по их фактическому размеру, так и по форме (смотри, например, соответствующие пары фотопортретов на стр. 2 статьи[1]).

Представленная в настоящей статье методика будет основана на сравнении цифровых изображений-оригиналов с результатами их Fawkes-преобразования при самом высоком параметре маскировки — mode = «high_cloaked». В качестве исходных данных (изображений-оригиналов) мы используем цифровые фото студентов СПбГЭТУ «ЛЭТИ» (Saint Petersburg Electrotechnical University «LETI») — участников представленных далее экспериментов, а также изображения лиц из базы CUFS[2].

На рис. 1, *a* представлены два исходных изображения — цветное изображение лица из базы CUFS, которое принимаем за оригинал (Original), и результат его Fawkes-преобразования (Fawkes). На лица нанесены вычисленные по ним координаты ключевых (антропометрических) точек. Здесь же приведены параметрические оценки различий текстур этих изображений: индекс структурного подобия (**I**ndex **S**tructural **SIM**ilarity) [8] Index SSIM = 0,99, и максимум фазовой корреляции [9] — max Phase Correlation = 0,96. И две эти оценки свидетельствуют о практически полном подобии двух изображений лиц — оригинала и результата его Fawkes-преобразования, хотя фазовая корреляция (0,96) отмечает какие-то изменения в текстурах.

На рис. 1, *b* представлены координаты ключевых точек обоих лиц, а также оценки среднеквадратического отклонения координат их ключевых точек. Эти отклонения составляют около половины пиксела, что можно отнести как к Fawkes-изменениям области лица, так и к погрешностям измерений ключевых точек. С учетом этих замечаний отметим на рис. 1, *a*, *b* отсутствие видимых (или значимых) изменений между текстурой и формой изображения-оригинала и изображения, прошедшего процедуру Fawkes-преобразования. Именно об этом важном факте сообщали авторы-разработчики процедуры Fawkes.

На рис. 1, *c* представлены разность текстур (Difference) в 100 кратном увеличении их значений и матрица структурного подобия (SSIM MAP [9]) между исходными изображениями. Наконец, на рис. 1, *d* представлено цветное изображение «Delta», полученное как разность между цветными битовыми слоями (Color Least Significant Bit, CLSB) оригинала CLSB(O) и результата его Fawkes-преобразования CLSB(F). При этом CLSB вычисляются с помощью операции по модулю 2 по исходным матрицам, которые представлены цветными изображениями O — Original и F — Fawkes так, что

$$\text{Delta} = \text{CLSB(O)} - \text{CLSB(F)}, \qquad (1)$$

где CLSB(O) = mod(Original, 2); CLSB(F) = = mod(Fawkes, 2).

Запись выражений (1) представлена в языке пакета Matlab, в котором были выполнены численные эксперименты.

[1] Image «Cloaking» for Personal Privacy [Электронный ресурс]. Режим доступа: https://sandlab.cs.uchicago.edu/fawkes/ (дата обращения: 06.07.2021).

[2] CUHK Face Sketch Database (CUFS). [Электронный ресурс]. http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html (дата обращения: 06.07.2021).

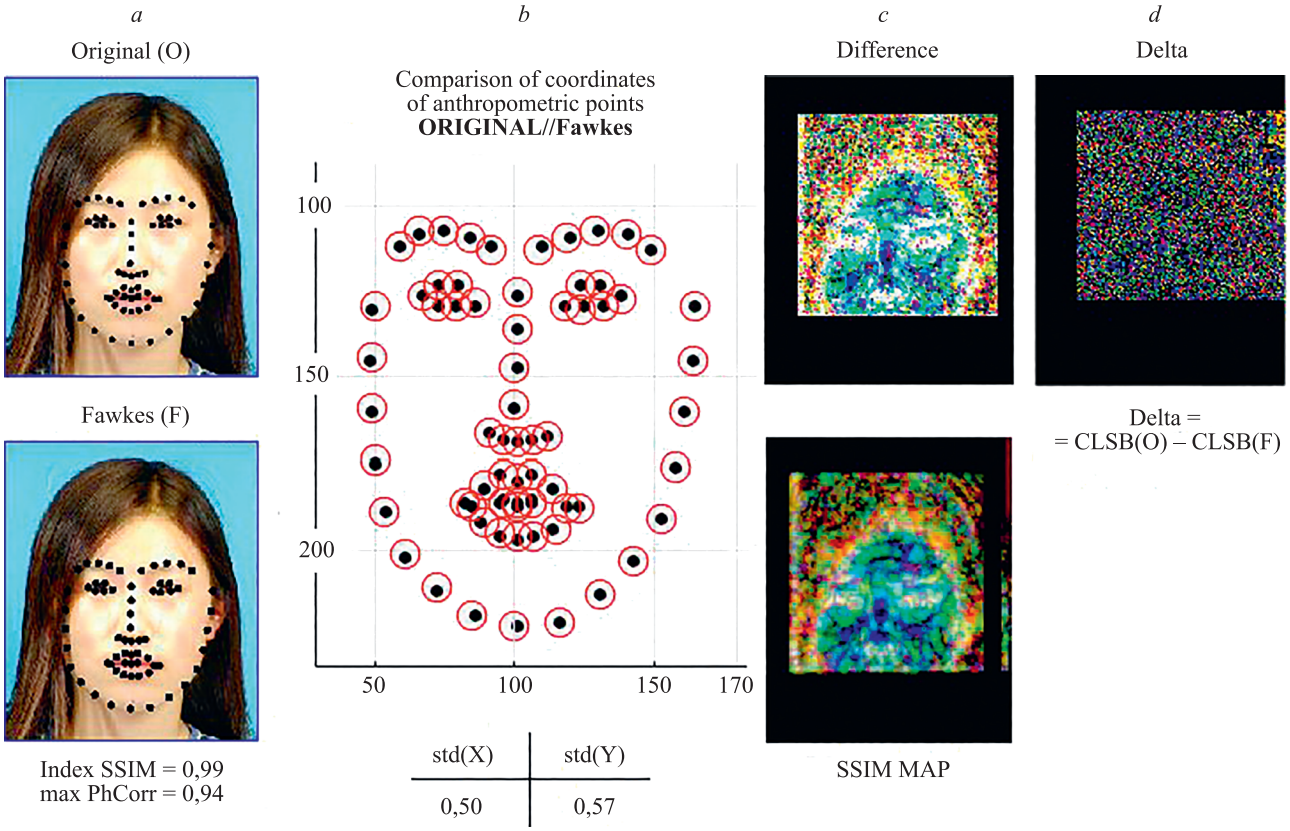

Рис. 1. Исходные данные: оригинал и результат его Fawkes-преобразования (*a*), а также формы представления их взаимных характеристик (*b*, *c*, *d*)

Fig. 1. Input data: the Original and the result of its Fawkes-transformation (*b*), as well as the forms of presentation of their characteristics (*b*, *c*, *d*)

По результатам рис. 1, *c*, *d* мы можем судить, что между изображениями Original и Fawkes существуют отличия, которые охватывают верхнюю часть лиц. И эти изменения происходят в битовых слоях, т. е. «внутри изображений», прошедших процедуру Fawkes-преобразования, что показано ниже на рис. 2 и 3.

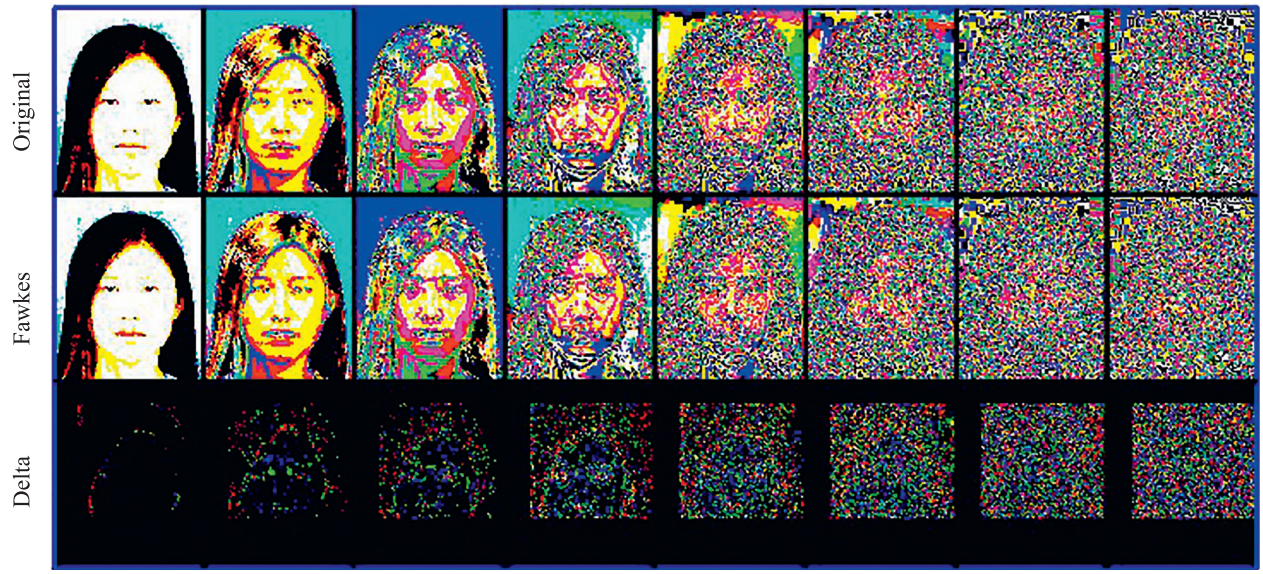

Рис. 2. Цветные битовые слои: изображения-оригинала (Original); результата Fawkes-преобразования оригинала; разница между ними (Delta)

Fig. 2. Colored bit layers: original images (Original); result of the Fawkes-transformation of the original; and the difference between them (Delta)

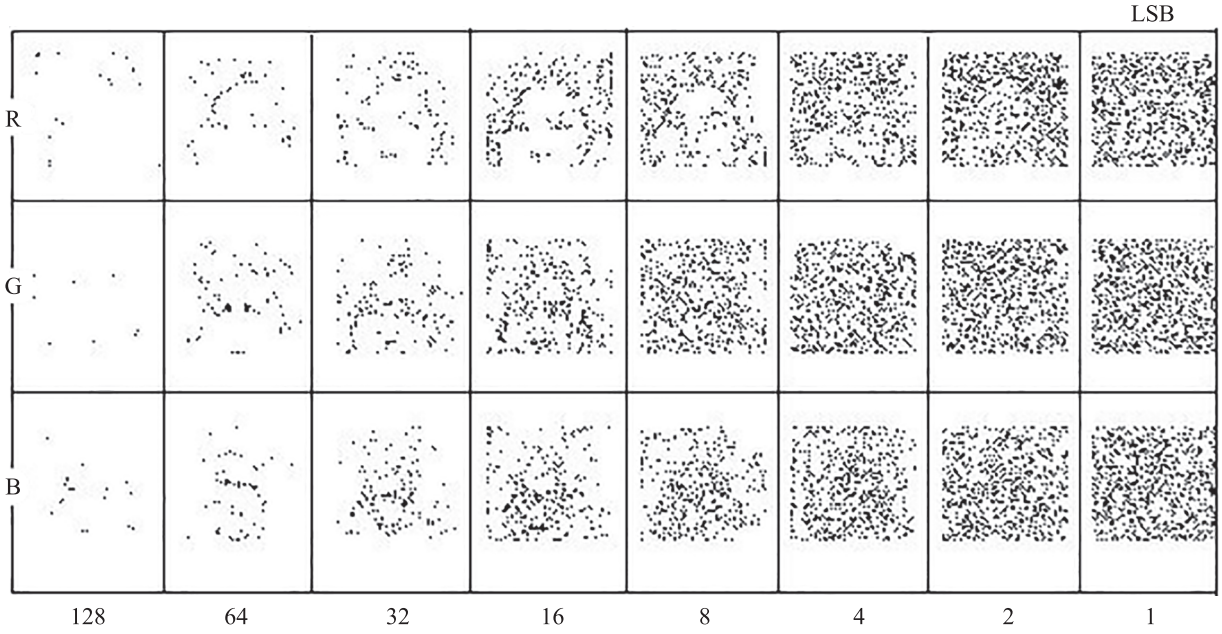

Рис. 3. 24 битовых слоя, представляющие разности между изображением-оригиналом и результатом его Fawkes-преобразования

Fig. 3. 24 bit layers representing the difference between the original image and the result of its Fawkes-transformation

На рис. 2 представлены восемь цветных битовых слоев изображений-оригиналов — CLSB(O), восемь CLSB(F) изображений-результатов Fawkes-преобразования, и восемь CLSB(Delta) разностей между ними. Отметим, что в каждом из восьми CLSB можно заметить два типа областей — черные и цветные.

Черные области в строке «Delta» образованы нулевыми значениями пикселов и определяют факт полного соответствия этих областей в оригинальном изображении и Fawkes-изображении. Цветные области — это отличия Fawkes-изображений от изображений-оригиналов. Более наглядно эти области показаны на рис. 3 в форме 24 битовых слоев, полученных как декомпозиция восьми CLSB на компоненты Red, Green и Blue.

Все битовые слои показаны в инверсной форме: черные точки — это отличия двух изображений, а белые поля — области, не измененные процедурой Fawkes. Как видно, наибольшие изменения (или разрушения) произошли в LSB-слоях (Least Significant Bit) с минимальным весом, равным 1. Изменения в LSB слоях с «1» на «0» и наоборот, будут соответствовать изменению яркости изображения на величину 1/255 (относительно максимума яркости). В следующих слоях слева от LSB слоя эти изменения влияют на яркость с коэффициентом 2/255, в следующем – 4/255 и т. д. В самом левом столбце битовых слоев этот коэффициент составляет 128/255 – или половину диапазона (от 0 до 255) яркости. С учетом данных изменений в процедуре Fawkes размер изменяемых областей в битовых слоях или число изменяемых пикселов в них уменьшается от слоя к слою в направлении влево от LSB-слоя. Этим достигается уменьшение видимых изменений текстуры в результате выполняемого Fawkes-преобразования.

Однако следует отметить еще одно наблюдение. В каждом CLSB или даже в каждом отдельном битовом слое могут применяться разные способы их изменений (фактически, разрушений):

— циклический сдвиг влево/вправо всего изображения на один столбец и/или вверх и вниз на одну строку (что можно заметить при сравнении парных фотопортретов на стр. 2 статьи[1]);
— циклический сдвиг изменяемых областей битовых слоев влево и/или вправо на один столбец и/или вверх и вниз на одну строку;
— удаление в изменяемых областях битовых слоев одной или двух строк и/или одного или двух столбцов с возвратом результата к исходному размеру изменяемой области;
— выполнение зеркального поворота выделенной области битовых слоев, инвертирование значений отдельных пикселов в выделенных областях (всех «0» в «1» и всех «1» в «0») и т. д.

Однако весь арсенал перечисленных выше изменений битовых слоев не совместим с задачами сверточной фильтрации в CNN. В случае распознавания изображения, прошедшего процедуру Fawkes, изменения в битовых слоях откликаются в конечных результатах свертки иными минимаксными решениями и не только по значениям, но и по их положению — вплоть до хаотического их представления на выходе CNN. С таким результатом не может быть решена и задача классификации, поскольку «хаос» несравним с упорядоченными значениями от эталонов.

Хотя перечисленные и подобные им изменения не столь значительны и масштабны на битовом уровне, но все же становятся заметны на лицах после процедуры Fawkes-преобразования. А именно, в области бровей,

[1] Image «Cloaking» for Personal Privacy [Электронный ресурс]. Режим доступа: https://sandlab.cs.uchicago.edu/fawkes/ (дата обращения: 06.07.2021).

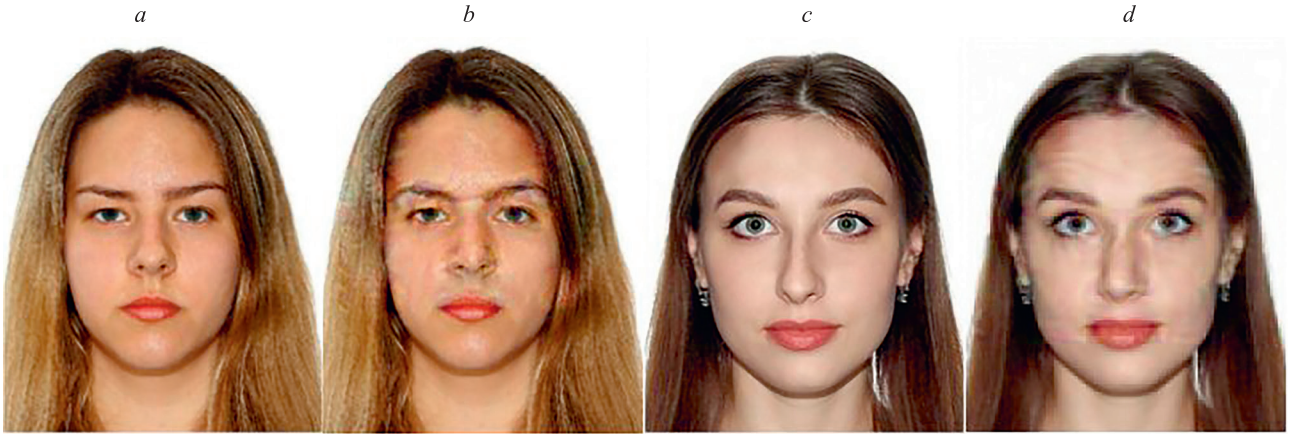

Рис. 4. Оригиналы (*a*, *c*) и результат их Fawkes-преобразования (*b*, *d*)
Fig. 4. Originals (*a*, *c*) and the result of their Fawkes-transformation (*b*, *d*)

переносицы, носа и верхней губы, границ овалов лиц на уровне глаз, а также внешнего контура головы, что можно увидеть на рис. 4 (лица принадлежат студентам СПбГЭТУ «ЛЭТИ» и представлены с их разрешения). Здесь изображения на рис. 4, *a*, *c* являются изображениями-оригиналами, а изображения на рис. 4, *b*, *d* — результатами их Fawkes-преобразования. Если увеличить Fawkes-изображения, то можно увидеть все изъяны такого преобразования.

Kashmir Hill[1], получив фото своей семьи после выполненных по ее заказу специальных процедур Fawkes-преобразования, написала: «*...изменения на фотографиях были заметны невооруженным глазом.... я выглядела омерзительно, у моей 3-летней дочери росли волосы на лице, а у моего мужа синяк под глазом*».

Отметим, что пользователи социальных сетей, помещая в аккаунтах свои наилучшие фото, не хотели бы подобных изменений на своих лицах. Ведь «виртуальный обмен» фотографиями и фотопортретами и их выставка — это одна из важнейших функций социальных сетей. А «испортить эту выставку», значит потерять пользователей и гостей социальных сетей. При этом надо учитывать, что не каждый пользователь социальных сетей сможет применить Fawkes-преобразование самостоятельно... и выполнять его в дальнейшем ежедневно или даже несколько раз в день. В то же время поручить эту процедуру администрации социальных сетей тоже невозможно по причине того, что ваше фото будет проходить через руки администратора. Отсюда фото могут попасть в другие руки и другие базы.

Попробуем подвести итог, насколько процедура Fawkes представляет собой удобный на практике и высоко результативный по факту защиты инструмент для защиты изображений лиц от распознавания. Несомненно, что процедура Fawkes-преобразования позволяет защитить изображения лиц от распознавания их CNN. И, как отмечают авторы технологии Fawkes «… хотя Fawkes не идеален, он предоставляет людям средства, которых у них раньше не было, и может обеспечить некоторую защиту от нежелательного распознавания лиц» [7]. И, если предположить, что все сознательные пользователи социальных сетей отныне и одновременно защитят свои фото Fawkes-преобразованием, то можно надеяться, что на какой-то очень короткий период времени их фото будут в безопасности от натиска существующих CNN, но не от новых, более совершенных нейронных сетей и всех ресурсов развивающегося искусственного интеллекта.

Вот что еще пишет Kashmir Hill в своей статье: «*...генеральный директор компании Clearview AI, Хоан Тон-Тат, ... сказал, что технология (Fawkes) не мешает работе его системы (распознавания) и его компания может использовать изображения, замаскированные Fawkes, чтобы улучшить способность распознавать измененные изображения*»[1].

В рамках выполненных экспериментов нами выяснено, что изображения лиц, подвергнутые процедуре Fawkes, хорошо распознаются вне методов глубоко обучения — широко известными и ранее используемыми детерминированными алгоритмами распознавания. И, хотя эти алгоритмы сегодня находятся за границами нейронных сетей и потока разработок искусственного интеллекта, но это значит, что уже один подход (детерминированный) — распознавания фото из социальных сетей, «укрытых процедурой Fawkes», все же остается доступным и неприкрытым. Неприкрытым потому, что такие системы распознавания легко реализуются, что будет рассмотрено далее.

Модели систем распознавания изображений, подвергнутых процедуре Fawkes

Если использовать «не сверточные» алгоритмы обработки изображений, разрушенных процедурой Fawkes, то можно получить набор признаков, не чувствительный к изменениям в битовых слоях результата Fawkes-преобразования. Тогда на базе этих алгоритмов можно создать простые системы распознавания изобра-

[1] Hill Kashmir. The Secretive Company That Might End Privacy as We Know It and 'Lead to a Dystopian Future [Электронный ресурс]. Режим доступа: https://www.news18.com/news/world/the-secretive-company-that-might-end-privacy-as-we-know-it-and-lead-to-a-dystopian-future-2464201.html (дата обращения: 06.07.2021).

жений лиц (Simple **Fa**ce **R**ecognition **S**ystems — Simple FaReS), замаскированные процедурой Fawkes.

Модели Simple FaReS [9, 10] включают:

— базу исходных данных с параметрами изображений лиц;

— три функциональные компоненты, представляющие: алгоритм предобработки изображений лиц; метод экстракции признаков, исходные размеры изображений лиц и размерность признакового пространства, алгоритм выбора признаков и конечное число признаков; тип классификатора (например, классификатор по минимуму расстояния — КМР), используемую метрику и ранг оценки результата распознавания.

Исходные предпосылки для решения задачи распознавания изображений, прошедших процедуру Fawkes-преобразования. Прежде всего отметим, что в собственном подпространстве, реализуемом, например, на базе 2DPCA/2DKLT [9, 11], изображение-оригинал и результат его Fawkes-преобразования находятся в одних и тех же координатах многомерного пространства, т. е. имеют подобные характеристики. Результатом этого подобия является 100 % распознавание изображений Fawkes-преобразования, что подтверждено нашими экспериментами. Однако методы 2DPCA/2DKLT (Two-dimensional Principal Component Analysis/Two-dimensional Karunen-Loeve Transform) не используются с динамическими базами изображений, поскольку любое их изменение даже на одно изображение (добавление или исключение), требует новых вычислений собственного базиса и нового выбора всех главных компонент, а также новой реализации двумерного преобразования Карунена–Лоэва (2DKLT) для всех изображений измененной базы изображений. В социальных сетях изображения исключаются и дополняются постоянно и, стало быть, такие изображения составляют динамические базы. Если принять во внимание, что косинус-преобразование является собственным преобразованием для генеральной совокупности изображений, то вместо метода 2DPCA/2DKLT в рассматриваемой нами задаче, можно также эффективно использовать и Simple FaReS на базе косинус-преобразования 2DDCT (Two-dimensional Discrete Cosine Transform) [9]. Об эффективности этого подхода могут свидетельствовать значения расстояний в косинус-подпространстве между изображениями-оригиналам и результатами их Fawkes-преобразования.

В выполненных нами экспериментах ненормированные расстояния получены классификатором по минимуму расстояния с метрикой L1, представлены тремя параметрами {min, mean, max} для ранга 1 и ранга 2 и показаны на рис. 5. При этом получены следующие тройки параметров: для ранга 1 {0,19; 0,67; 2,0}, а для ранга 2 — {19; 34; 82}. Для 100 разрушенных процедурой Fawkes-изображений лиц базы CUFS, значения расстояний для правильно распознанных лиц (ранг 1) показаны на рис. 5, *a*, а значения расстояний по рангу 2 показаны на рис. 5, *b*. Очевидно, что косинус-преобразование изображений, подвергнутых процедуре Fawkes, обеспечит 100 % их распознавание.

Подтверждением этого являются результаты выполненного нами эксперимента. Так, на рис. 6 в трехмерном подпространстве косинус-преобразования показаны координаты для 100 изображений, прошедших процедуру Fawkes с параметром mode = «high_cloaked» (отмечены знаком «+»), и соответствующие им координаты изображений-оригиналов (отмечены знаком «круг»). Видно, что все знаки «+» находятся в поле знаков «круг», что гарантирует 100 % распознавание изображений лиц, замаскированных процедурой Fawkes.

Действительно, в выполненных нами экспериментах, все изображения, прошедшие Fawkes-преобразования, были правильно распознаны Simple FaReS [10], модель которой показана ниже:

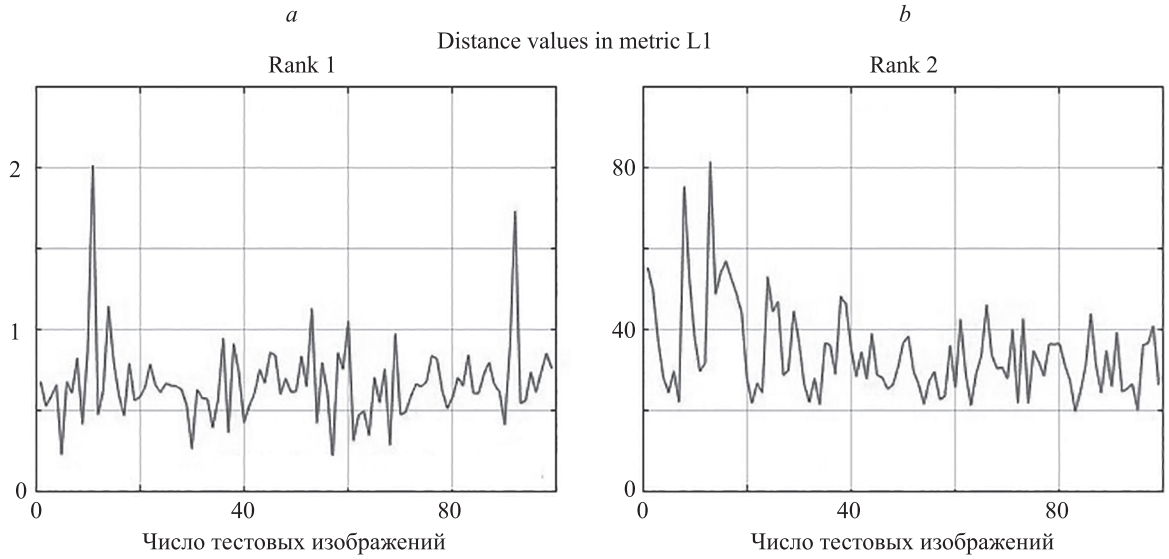

Рис. 5. Расстояния по рангам 1 (*a*) и 2 (*b*) в косинус-подпространстве между изображениями-оригиналами и результатами их Fawkes-преобразования

Fig. 5. Distances in the cos-transform feature space between the original images and the results of their Fawkes-transformation

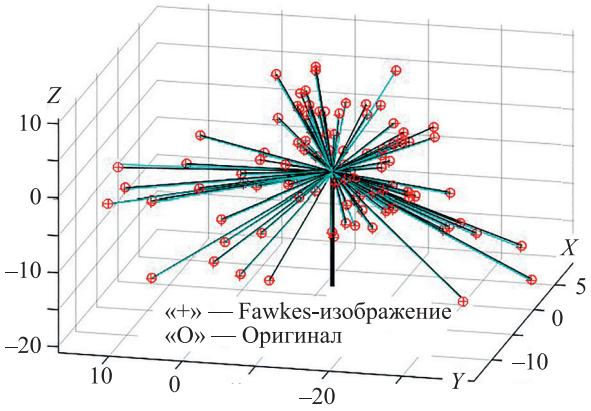

Рис. 6. Отображение изображений лиц, замаскированных процедурой Fawkes, и изображений-оригиналов в признаковом пространстве косинус-преобразования

Fig. 6. Mapping of the face images masked by the Fawkes procedure and original images in the cos-transform feature space

$$\textbf{GUFS}(100/1/100)\ \{\textbf{2DDCT}{:}250\times 200\rightarrow \rightarrow 15\times 15/\textbf{zigzag}(119)\}\ [\textbf{KMP}/\mathrm{L1}/\mathrm{rank}=1], \quad (2)$$

где **GUFS**(100/1/100) — база изображений лиц **GUFS**[1], включающая 100 изображений-оригиналов (100 эталонов, по 1 эталону на класс) и 100 тестовых изображений, прошедших процедуру Fawkes; {**2DDCT**:250 × 200 → 15 × 15/zigzag(119)} — обработка изображений размером 250 × 100 по алгоритму двумерного косинус-преобразования, с выбором спектральных признаков методом Zigzag [9] по 15 диагоналям спектра с общим числом значений, равным 119 (без учета первого значения, соответствующего среднему значению яркости изображения); [**KMP**/L1/rank = 1] — классификатор по минимуму расстояния, с метрикой L1 и рангом 1.

[1] Hill Kashmir. The Secretive Company That Might End Privacy as We Know It and 'Lead to a Dystopian Future [Электронный ресурс]. Режим доступа: https://www.news18.com/news/world/the-secretive-company-that-might-end-privacy-as-we-know-it-and-lead-to-a-dystopian-future-2464201.html (дата обращения: 06.07.2021).

Скриншот работы этой простой системы представлен на рис. 7.

На рис. 7 изображение Delta (см. формулу (1)), представляющее разность между цветными битовыми слоями двух исходных изображений лиц, выполняет роль контроля того, что Query Face прошло процедуру Fawkes-преобразования. Обратим внимание, что на изображении Query Face видны искажения текстуры изображения лица.

Следующие две модели Simple FaReS также строятся на наборах признаков, не чувствительных к изменениям в битовых слоях результата Fawkes-преобразования. Модели этих Simple FaReS имеют вид:

$$\textbf{GUFS}(100/1/100)\ \{\textbf{Face}{:}250\times 200\rightarrow \rightarrow \mathrm{Random}(\mathrm{NUM})\}\ [\mathrm{KMP}/\mathrm{L1}/\mathrm{rank}=1]; \quad (3)$$

$$\textbf{GUFS}(100/1/100)\ \{\ \textbf{Face}{:}250\times 200\rightarrow \rightarrow \mathrm{CAP}(\mathrm{NUM})\}\ [\mathrm{KMP}/\mathrm{L1}/\mathrm{rank}=1], \quad (4)$$

где Random(NUM) — метод генерации координат NUM пикселей, равномерно распределенных на области лица или всего фотопортрета [9, 10]. По этим не изменяемым координатам вычисляется вектор яркостных признаков для каждого оригинала (суть — эталона) и всех распознаваемых Fawkes-изображений. Далее система распознавания выполняет сравнение векторов яркостных признаков и их классификацию; **C**AP(NUM) — метод выбора NUM координат антропометрических точек (**C**oordinates of the **A**nthropometric **P**oints, CAP [12]) лица, по которым вычисляются векторы яркостных признаков. Эти векторы с достаточной точностью представляют каждое отдельное изображение лица (оригинал и результат его Fawkes-преобразования), на чем основан метод их распознавания.

Скриншоты работы системы (3) и (4) представлены на рис. 8. Simple FaReS (3) является демонстрационной и представляет идеи Random [8, 10], как самого простого метода распознавания изображений лиц, прошедших процедуру Fawkes-преобразования. Simple FaReS (4) основана на CAP, более совершенная, поскольку яркостные признаки остаются без изменений при плоском повороте области лица, его увеличении, уменьшении и/или сдвиге. Например, на рис. 8 показан вариант с из-

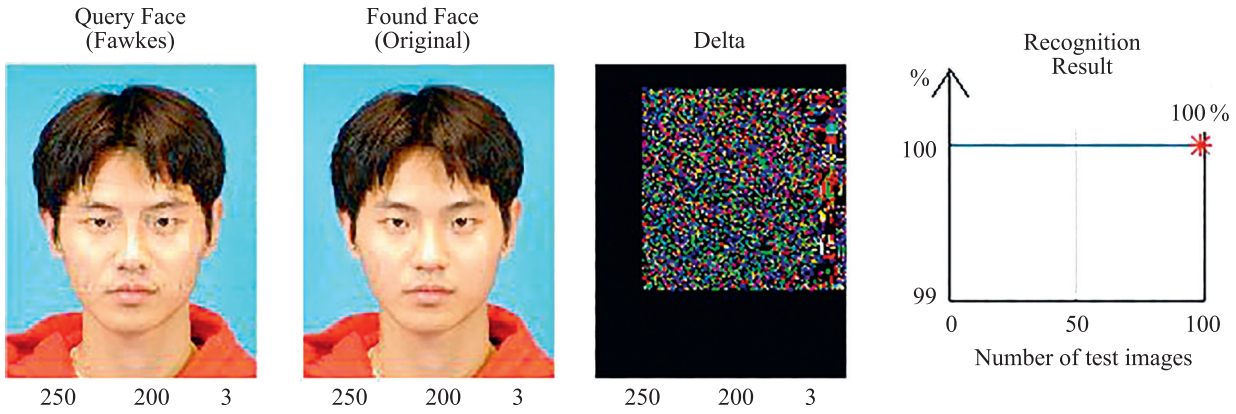

Рис. 7. Скриншот работы системы распознавания модели 2 (формула (2)), замаскированных процедурой Fawkes

Fig. 7. Screenshot of the face recognition system process model (2) masked by the Fawkes procedure

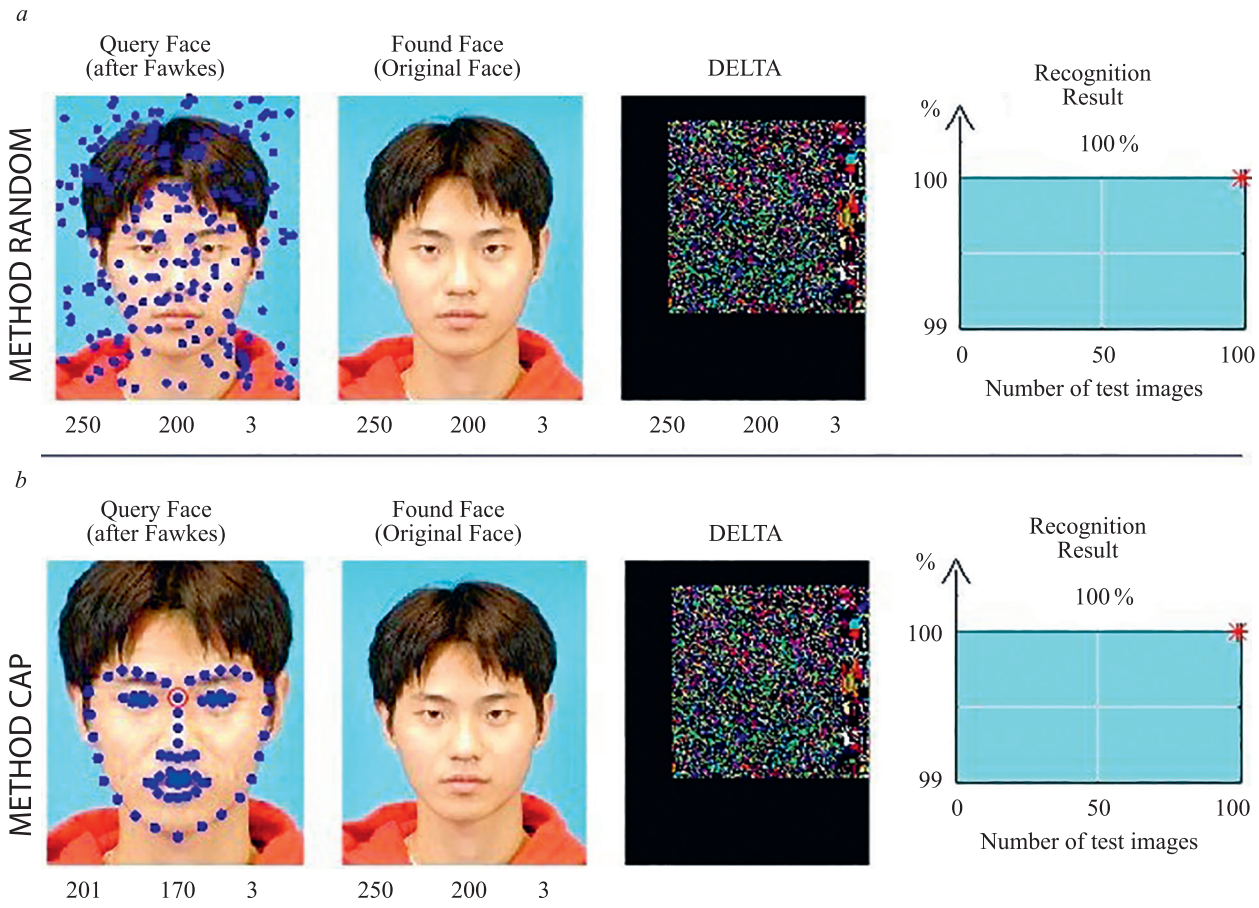

Рис. 8. Скриншот работы систем распознавания по моделям: 3 (формула (3)) (*a*) и 4 (формула (4)) (*b*), замаскированных процедурой Fawkes

Fig. 8. Screenshot of the face recognition system process (models 3 and 4) masked by the Fawkes procedure

менением размера изображения Query Face в сравнении с изображением-эталоном.

Наконец, представим процедуру предобработки, основанную на предварительном сглаживании изображения QF (Query Face). Процедура сглаживания может быть любой, но ее задачей является «реконструкция сильно искаженной информации» на области лица в изображении, разрушенном процедурой Fawkes. Например, это можно достичь в два шага: сначала уменьшить изображение, а затем вернуть его к исходному размеру.

На первом шаге измененные процедурой Fawkes области сжимаются и, усредняясь, сглаживаются. На втором шаге происходит расширение (через интерполяцию) сжатых областей CLSB, что приводит к сглаживанию искаженной информации на текстуре лица.

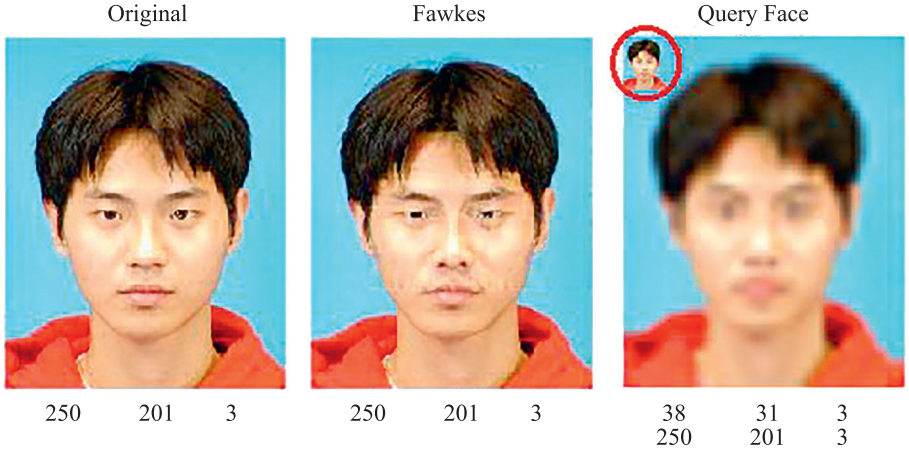

Рис. 9. Этапы предобработки Fawkes-изображения

Fig. 9. Steps of the Fawkes-image preprocessing

Примеры этих изменений приведены на рис. 9. На рисунке показаны: изображение-оригинал; результат Fawkes-преобразования (с видимыми искажениями текстуры лица); уменьшенное изображение (выделенное в круге); результат — сглаженное изображение Query Face. Последнее может быть распознано со 100 % результатом любой системой, представленной выше.

Рецепт для CNN

Для изображений-оригиналов выполним следующие действия: процедуру сглаживания, представленную выше, а затем процедуру 2DDCT, а также проведем выбор набора спектральных признаков по процедуре Zigzag. Полученные признаки нормируем и объединяем (путем конкатенации) с выходными признаками CNN, полученными для изображений-оригиналов. Далее можно с высокой точностью решать задачи классификации результата Fawkes-преобразования.

Представленные эксперименты выполнены нами на популярной CNN ResNet (34, 50, 101)[1], версия ResNet(34). Распознаны все изображения лиц, подвергнутые процедуре Fawkes. Структура смоделированной системы на базе CNN с дополнительной обработкой исходных данных показана на рис. 10.

Блок 1 реализует задачу регистрации (или в терминах глубокого обучения — дообучение) в результате чего создается база векторов признаков эталонных изображений (или база эталонов). Блок 2 — реализует задачу формирования вектора признаков для конкретного изображения QF. Блок 3 — реализует сравнение этого вектора с векторами всех эталонов и определяет тот эталон, с которым найдено минимальное расстояние. Таким образом, блоки 2 и 3 реализуют задачу узнавания изображения QF. В системе одна и также CNN — сначала на этапе дообучения, а затем только на этапе вычисления набора выходных признаков для изображений QF.

При параметре mode = «high_cloaked» в процедуре Fawkes, примененной для всех изображений лиц базы CUFS на оригинальной CNN[1], результат классификации составил 33 % при рекомендуемом пороге (отсечения), равном 0,3. В рамках контролируемой классификации по минимуму расстояния, достижение 100 % результата изображений Fawkes-преобразования стало возможным только на кумулятивной сумме по 10 рангам. В дополненном варианте — с предварительной обработкой результата Fawkes-преобразования (сглаживание + 2DDCT + Zigzag) и фузией с выходом CNN, результат классификации для тех же исходных данных составил 100 % на ранге 1.

Поскольку эксперименты выполнены только для понимания задачи распознавания изображений лиц, прошедших процедуру Fawkes, мы не приводим детали системы на базе CNN, ее компоненты, особенности использованных алгоритмов обработки изображений и их параметров. При этом мы стоим на стороне разработчиков процедуры Fawkes и считаем, что лица людей должны быть защищены всеми возможными и невозможными методами.

Приведем еще одну работу, представляющую новый подход к защите изображений в социальных сетях путем «замусоривания» поисковых баз по лицам [13]. В отличие от технологии Fawkes [7], укрывающей отдельно каждое защищаемое изображение от распозна-

[1] ResNet (34, 50, 101): «остаточные» CNN для классификации изображений [Электронный ресурс]. Режим доступа: https://neurohive.io/ru/vidy-nejrosetej/resnet-34-50-101/ (дата обращения: 06.07.2021).

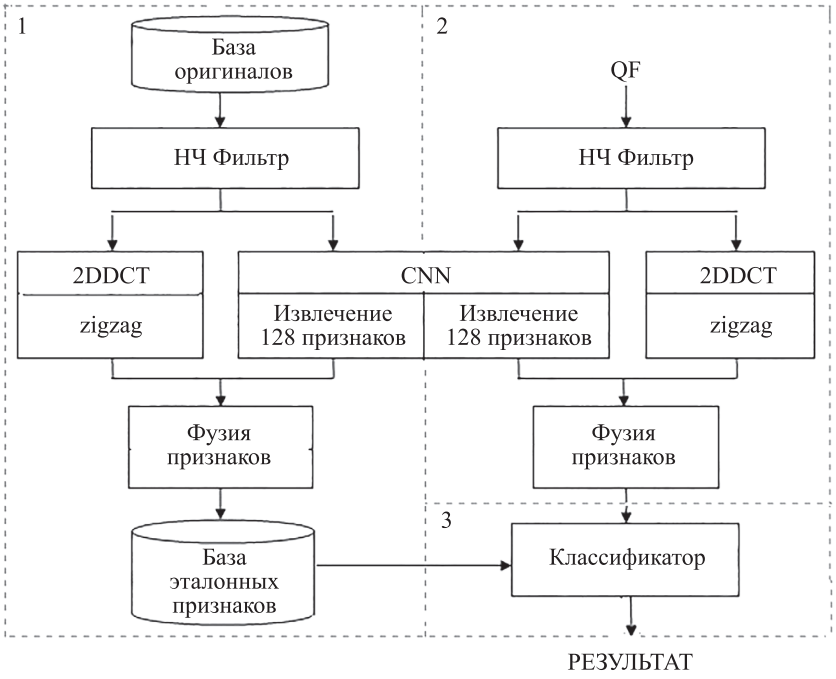

Рис. 10. Структура смоделированной системы на базе CNN с дополнительной обработкой исходных данных

Fig. 10. The structure of a simulated system based on CNN with additional processing of the input data

вания, в [13] предложено расширять наборы изображений в поисковых базах специальной популяцией изображений-приманок. Такие изображения-приманки получаются на основе «состязательных алгоритмов генерации примеров» (adversarial examples-generation algorithms). Стратегии формирования таких популяций могут быть разными — в рамках отдельной социальной сети, групп людей — энтузиастов защиты и даже отдельных пользователей сетей. И, как отмечают авторы [13], огромное число изображений-приманок, загруженных разными людьми, как бы «замещают чистые фотографии» людей в результатах поисковых запросов. И эта стратегия может оказаться также перспективной в ближайшее время.

Заключение

В статье рассмотрена процедура Fawkes — как метод защиты от несанкционированного использования и распознавания изображений лиц из социальных сетей. Приведены результаты эксперимента, подтверждающего факт низкого результата распознавания изображений лиц обученной CNN, когда процедура Fawkes применена с наивысшим параметром mode. На основе сравнительного анализа изображений, подвергнутых процедуре Fawkes с исходными изображениями лиц, показаны текстурные изменения и графические особенности структурных разрушений. В дополнение к этому анализу приведены многоуровневые параметрические оценки этих разрушений, и на их основе выяснена причина, затрудняющая использование изображений лиц, подвергнутых процедуре Fawkes, в задачах глубокого обучения и распознавания. В качестве инструментов количественной оценки использован индекс структурного подобия (Index SSIM) и фазовая корреляция изображений (Phase Correlation).

Показано, что изображения лиц, подвергнутые процедуре Fawkes, хорошо распознаются вне методов глубокого обучения. С этой целью исследованы модели систем распознавания изображений лиц, подвергнутых процедуре Fawkes. Приведены результаты выполненных экспериментов. Утверждается, что использование простых способов предобработки изображений лиц (подвергнутых процедуре Fawkes) на входе в CNN, может привести к их распознаванию с высокой результативностью, что разрушает миф о 100 % защите изображений лиц процедурой Fawkes. Однако идеи защиты изображений лиц от несанкционированного распознавания развиваются и дальше, что показано, например, в [13].

Литература

1. Щеголева Н.Л. Модели изображений лиц для решения задач криминалистики // Известия СПбГЭТУ ЛЭТИ. 2016. № 7. С. 37–47.
2. Newton E.M., Sweeney L., Malin B. Preserving privacy by de-identifying face images // IEEE Transactions on Knowledge and Data Engineering. 2005. V. 17. N 2. P. 232–243. https://doi.org/10.1109/TKDE.2005.32
3. Jourabloo A., Yin X., Liu X. Attribute preserved face de-identification // Proc. 8th International Conference on Biometrics (ICB). 2015. P. 278–285. https://doi.org/10.1109/ICB.2015.7139096
4. Kukharev G., Oleinik A. Face photo-sketch transformation and population generation // Lecture Notes in Computer Science. 2016. V. 9972. P. 329–340. https://doi.org/10.1007/978-3-319-46418-3_29
5. Wu Y., Yang F., Xu Y., Ling H. Privacy-protective-GAN for privacy preserving face de-identification // Journal of Computer Science and Technology. 2019. V. 34. N 1. P. 47–60. https://doi.org/10.1007/s11390-019-1898-8
6. Nousi P., Papadopoulos S., Tefas A., Pitas I. Deep autoencoders for attribute preserving face de-identification // Journal Signal Processing: Image Communication. 2020. V. 81. P. 115699. https://doi.org/10.1016/j.image.2019.115699
7. Shan S., Wenger E., Zhang J., Li H., Zheng H., Zhao B.Y. Fawkes: Protecting privacy against unauthorized deep learning models // Proc. 29th USENIX Security Symposium. 2020. P. 1589–1604.
8. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. V. 13. N 4. P. 600–612. https://doi.org/10.1109/TIP.2003.819861
9. Кухарев Г.А., Каменская Е.И., Матвеев Ю.Н., Щеголева Н.Л. Методы обработки и распознавания изображений лиц в задачах биометрии. СПб.: Политехника, 2013. 388 с.
10. De Vel O., Aeberhard S. Line-based face recognition under varying pose // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1999. V. 21. N 10. P. 1081–1088. https://doi.org/10.1109/34.799912
11. Kukharev G.A., Shchegoleva N.L. Algorithms of two-dimensional projection of digital images in eigensubspace: History of development, implementation and application // Pattern Recognition and Image Analysis. 2018. V. 28. N 2. P. 185–206. https://doi.org/10.1134/S1054661818020116

References

1. Schegoleva N.L. Face image models for criminalistics. *Proceedings of Saint Petersburg Electrotechnical University Journal*, 2016, no. 7, pp. 37–47. (in Russian)
2. Newton E.M., Sweeney L., Malin B. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 2005, vol. 17, no. 2, pp. 232–243. https://doi.org/10.1109/TKDE.2005.32
3. Jourabloo A., Yin X., Liu X. Attribute preserved face de-identification. *Proc. 8th International Conference on Biometrics (ICB)*, 2015, pp. 278–285. https://doi.org/10.1109/ICB.2015.7139096
4. Kukharev G., Oleinik A. Face photo-sketch transformation and population generation. *Lecture Notes in Computer Science*, 2016, vol. 9972, pp. 329–340. https://doi.org/10.1007/978-3-319-46418-3_29
5. Wu Y., Yang F., Xu Y., Ling H. Privacy-protective-GAN for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 2019, vol. 34, no. 1, pp. 47–60. https://doi.org/10.1007/s11390-019-1898-8
6. Nousi P., Papadopoulos S., Tefas A., Pitas I. Deep autoencoders for attribute preserving face de-identification. *Journal Signal Processing: Image Communication*, 2020, vol. 81, pp. 115699. https://doi.org/10.1016/j.image.2019.115699
7. Shan S., Wenger E., Zhang J., Li H., Zheng H., Zhao B.Y. Fawkes: Protecting privacy against unauthorized deep learning models. *Proc. 29th USENIX Security Symposium*, 2020, pp. 1589–1604.
8. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, vol. 13, no. 4, pp. 600–612. https://doi.org/10.1109/TIP.2003.819861
9. Kukharev G.A., Kamenskaya E.I., Matveev Y.N., Shchegoleva N.L. *Methods for Face Image Processing and Recognition in Biometric Applications*. St. Petersburg, Politekhnika Publ., 2013, 388 p. (in Russian)
10. De Vel O., Aeberhard S. Line-based face recognition under varying pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, vol. 21, no. 10, pp. 1081–1088. https://doi.org/10.1109/34.799912
11. Kukharev G.A., Shchegoleva N.L. Algorithms of two-dimensional projection of digital images in eigensubspace: History of development, implementation and application. *Pattern Recognition*

12. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees // Proc. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014. P. 1867–1874. https://doi.org/10.1109/CVPR.2014.241
13. Evtimov I., Sturmfels P., Kohno T. FoggySight: A Scheme for facial lookup privacy // Proceedings on Privacy Enhancing Technologies. 2021. V. 3. P. 204–226. https://doi.org/10.2478/popets-2021-0044

and Image Analysis, 2018, vol. 28, no. 2, pp. 185–206. https://doi.org/10.1134/S1054661818020116
12. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees. *Proc. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874. https://doi.org/10.1109/CVPR.2014.241
13. Evtimov I., Sturmfels P., Kohno T. FoggySight: A Scheme for facial lookup privacy. *Proceedings on Privacy Enhancing Technologies*, 2021, vol. 3, pp. 204–226. https://doi.org/10.2478/popets-2021-0044

Авторы

Кухарев Георгий Александрович — доктор технических наук, профессор, профессор, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация, SC 18037842200, https://orcid.org/0000-0003-2188-2172, gakukharev@etu.ru; kuga41@mail.ru

Мауленов Калыбек Сапарович — докторант, Некоммерческое акционерное общество «Костанайский региональный университет имени А. Байтурсынова», Костанай, 110000, Республика Казахстан, https://orcid.org/0000-0003-4147-3843, k_maulenov@inbox.ru

Щеголева Надежда Львовна — доктор технических наук, доцент, профессор, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 36683312100, https://orcid.org/0000-0003-1087-2833, stil_hope@mail.ru

Authors

Georgy A. Kukharev — D.Sc., Full Professor, Saint Petersburg State Electrotechnical University (LETI), Saint Petersburg, 197376, Russian Federation, SC 18037842200, https://orcid.org/0000-0003-2188-2172, gakukharev@etu.ru; kuga41@mail.ru

Kalybek S. Maulenov — Doctoral Student, Non-profit limited company "A. Baitursynov Kostanay Regional University", Kostanay, 110000, Republic of Kazakhstan, https://orcid.org/0000-0003-4147-3843, k_maulenov@inbox.ru

Nadezhda L. Shchegoleva — D.Sc., Associate Professor, Professor, Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation, SC 36683312100, https://orcid.org/0000-0003-1087-2833, stil_hope@mail.ru

Статья поступила в редакцию 18.07.2021
Одобрена после рецензирования 09.08.2021
Принята к печати 04.10.2021

Received 18.07.2021
Approved after reviewing 09.08.2021
Accepted 04.10.2021