

ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ
И ТЕЛЕКОММУНИКАЦИИ

INFORMATION
TECHNOLOGIES
AND TELECOMMUNICATIONS

УДК 004.8

DOI: 10.17212/2782-2001-2025-1-105-120

Интеллектуальный метод автоматического выявления подмены лица на изображении*

М.Д. ХАЛЕЕВ

РФ, 199178, Санкт-Петербургский федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 14-я линия В. О., 39

Haleev.M@iiias.spb.su

Подмена лиц на изображениях является серьезной проблемой в современном цифровом пространстве. Такие изображения обычно создаются с использованием современных методов машинного обучения. Эти поддельные изображения могут вводить пользователей в заблуждение, создавая риски для конфиденциальности, безопасности и доверия. В настоящей статье представлен инновационный подход к формированию целевого набора признаков для выявления подменных изображений и последующему обнаружению подмены лиц, сформирован набор данных размеченных изображений подмены на основе публичного набора данных с использованием генератора подмены лиц; вычислены и проанализированы ключевые точки лица в объемном формате для изображений набора данных и вычислены соответствующие признаки, выявленные экспертным способом; обучены несколько моделей классификаторов на основе трех наборов признаков, включающих в себя ключевые точки лица, экспертные признаки и их смешение; выполнена оценка работы классификаторов; построены матрицы спутанности и прослежены закономерности в зависимости от целевого набора признаков; достигнута точность 0.682 и F1-мера 0.680 на объединенном наборе признаков. При использовании коэффициентов полученных моделей была оценена важность признаков и выявлены наиболее значимые из них для более глубокого подхода к формированию наборов данных при обучении моделей машинного обучения. Для вычисленных экспертных признаков наиболее важными оказались черты лица, характеризующие его геометрическую форму, а в случае с ключевыми точками лица наиболее значимыми оказались границы объектов лица: контур лица, области вокруг глаз и уголки губ.

Ключевые слова: подмена лиц, компьютерное зрение, искусственный интеллект, генеративный контент, маркировка контента, дипфейк, информационная безопасность, лицевые ориентиры

* Статья получена 19 сентября 2024 г.

ВВЕДЕНИЕ

Изображения подмены лиц представляют собой сложную форму синтетических медиа, использующих современные алгоритмы подмены лиц, включающие глубокое обучение и нейронные сети, для изменения или полного замещения сегментов оригинального изображения или видео. Эти изменения могут быть настолько качественно выполнены, что полученный контент часто неотличим от подлинных медиа для человеческого глаза. Распространение изображений подмены создает значительные проблемы, так как их реалистичная природа может быть использована для распространения дезинформации, манипуляции общественным мнением и причинения различных форм вреда. Следовательно, разработка и внедрение надежных механизмов обнаружения являются необходимыми для защиты от потенциального злоупотребления этой технологией и для поддержания целостности информации в цифровую эпоху.

В статьях [1, 2] говорится, что подмена лиц представляет собой значительную угрозу, включая мошенничество через обманные видео, подделку доказательств в судебных разбирательствах, злоупотребление личными данными и распространение дезинформации для влияния на общественное мнение и политические процессы. Эти опасности требуют разработки современных технологий обнаружения и предотвращения со стороны правоохранительных органов.

Предложенный метод сосредоточен на анализе лицевых ориентиров, которые имеют решающее значение для различения подлинного контента и подделок. Лицевые ориентиры относятся к конкретным, идентифицируемым точкам на лице, включая уголки глаз, кончик носа и уголки рта. Эти контрольные точки играют важную роль в передаче важной информации о выражении лица, геометрической структуре и индивидуальной идентичности. Тщательно оценивая позиционное и относительное значения этих точек, наш метод может обнаруживать тонкие аномалии и несоответствия, введенные манипуляциями глубоких подделок. Этот подход повышает точность и надежность обнаружения подмены лиц, тем самым способствуя более широким усилиям по борьбе с дезинформацией и защите целостности визуальных медиа.

Также были выявлены экспертные особенности, имеющие наибольшее значение для идентификации изображений подмены лиц, которые были вычислены на основе контрольных точек лица. Подход заключается в анализе контрольных точек лица и экспертных особенностей для повышения точности обнаружения подмены лиц.

Были вычислены следующие экспертные признаки на основе этих контрольных точек.

- Геометрические расстояния. Измеряя расстояния между ключевыми точками лица, были выявлены неестественные изменения, полученные в результате создания изображения подмены лица. Например, несоответствия в расстоянии между глазами или расстоянии от рта до носа могут сигнализировать о потенциальном вмешательстве.

- Объем и площадь лица. Вычисление объема и площади, заключенной контрольными точками лица, предоставляет дополнительные сведения. Отклонения от ожидаемых значений могут указывать на модификации глубоких подделок.

Предложен классификатор для использования при обнаружении подменных изображений и последующем анализе важности признаков. Классифика-

тор обучался отдельно на каждом наборе признаков, и на основе их весов была рассчитана важность признаков в каждом наборе.

- Важность контрольных точек. Назначение весов каждой точке лица на основе ее важности позволяет нам приоритизировать определенные области. Например, уголки рта и глаз более критичны для точного обнаружения.

- Важность экспертных признаков. Оценка важности экспертных признаков позволит оценить ключевые признаки для обнаружения подмены лица, чтобы в будущем использовать только их.

В последующих разделах данной работы сначала проводится обзор текущего состояния исследований, который служит основой для исследования. Затем представлено подробное описание предложенного метода с детализацией по каждому шагу процесса. Этот метод используется для генерации набора данных, который затем используется для обучения модели классификатора. Рассчитаны и предоставлены качественные метрики для оценки производительности модели. Далее оценена значимость различных особенностей в наборе данных, сделаны выводы о их влиянии на точность модели. Подведены итоги исследований, сформированы результаты работы и намечены потенциальные направления для дальнейшего развития и улучшения подхода.

1. АНАЛИЗ СОВРЕМЕННОГО СОСТОЯНИЯ ИССЛЕДОВАНИЯ

В этом разделе рассмотрены методы выявления изображений подмены лица.

В статье [3] авторы разделяют подменные изображения на четыре подгруппы: генерация лиц, замена лиц, модификация лиц и изменение выражения лица. Каждая подгруппа соответствует специфическим наборам данных, состоящим из изображений, сгенерированных исключительно архитектурами GAN. Размеры наборов данных варьируются от 100 000 до 330 000 изображений. Примечательно, что в основном обучение происходит на видеоданных.

Работа [4] предлагает единую архитектуру для обнаружения всех типов подмены лиц. Авторы предоставляют краткий анализ и сравнение существующих моделей, адаптированных к специфическим типам подмены. Кроме того, они представляют единую архитектуру, достигающую точности 98 % с использованием общедоступного набора данных «FaceForensic++», который включает 1000 оригинальных и 4000 измененных видеоклипов.

Исследователи [5] используют глубокие сверточные нейронные сети (DCNN) для вычисления коэффициентов схожести лиц. Модель достигает точности 97,5 % на открытых наборах данных.

Авторы [6] сосредотачиваются на обнаружении изображений с замененными лицами. Они представляют набор данных и проводят сравнительный анализ различных архитектур. Кроме того, в исследовании предоставляется исходный код их подхода.

В методе, изложенном в [7], используется совместная некурируемая реконструкция и курируемая классификация для обнаружения подмены лиц. В отличие от предыдущих подходов, метод обращает внимание не только на явные ошибки нейронной сети, но и на менее очевидные признаки общие для генераций. Метод был протестирован на пяти различных наборах данных

(UADFV, FaceForensics++, Celeb-DF (версия 2), DeepFake Detection (DFD) и DeepfakeTIMIT) и превзошел существующие решения на одном из наборов данных.

Авторы [8] предлагают модель, у которой есть два основных модуля: GraphNet, который использует сверточные графовые слои для агрегации и обновления информации о графе, и FFN, который имеет линейные слои для трансформации признаков узлов. Эффективность метода оценивается с использованием разнообразного набора данных Deepfake Detection Challenge (DFDC), FaceForensics++ (FF++), World Leaders (WLRD) и Celeb-DF и достигает 76 % точности.

Используя признаки, извлеченные из трех Vision Transformers (DaViT, iFormer и GPViT), метод [9] анализирует видеоданные. Набор данных состоит из девяти настоящих изображений и 51 поддельного. Он был протестирован на FaceForensics++ с использованием четырех различных техник и достиг точности 97,72 %.

Авторы [10] предлагают фреймворк ART-AVDF, который использует артикуляционное представление для точного аудиовизуального обнаружения замены лиц. В модуле ART аудиокодер и кодер губ разработаны для выполнения обучения артикуляционного представления аудиовизуальных данных с использованием стратегии самообучения. В модуле AVDF они используют замороженные кодеры из модуля ART для получения артикуляционных эмбедингов и комбинируют их с унимодальными признаками, что приводит к лучшему аудиовизуальному анализу для обнаружения подмены лиц, а также используют DFDC и FakeAVCele, DefakeAVMiT и достигают точности до 96 %.

Подход [11] разбивает лицевые изображения на основные элементы и вычисляет характеристики на их основе, такие как текстура лица и степень естественности. Используемая модель распознавания – HDDM, она состоит из двух моделей: конструкции текстуры лица (FTC) и распознавания степени естественности (NDR), была протестирована на различных наборах данных (Flickr-Faces-HQ, StyleGan2, FaceForensics++, Celeb-DFv2) и достигла точности 99,7 %.

При использовании ResNet-Swish-BiLSTM метод [12] маркирует поддельные видео. Шаблоны идентифицированных лиц имеют размеры 224×224 пикселей. Метод был протестирован на наборах данных DFDC и FF++ и достиг точности 99,13 %. Также был использован алгоритм кросс-валидации для оценки эффективности метода.

Метод [13] обнаруживает подмену лиц для фотографий и видео, выявляя незначимые артефакты. Авторы статьи получают два разных типа признаков: один представляет собой глобальные несоответствия между слоями, а другой – пространственные глобальные признаки изображения. Модель была протестирована на FaceForensics++, Celeb-DF (V2) и наборе изображений Deepfakes (DFID) с точностью 95,42 %.

Авторы [14] предлагают сеть, состоящую из двух частей (вспомогательной и основной сетей). Авторы сосредоточены на несоответствии освещения между кадрами. Чтобы улучшить выявления несоответствия освещения на уровне признаков, предлагается многоуровневая доработка признаков, которая состоит из блоков FRRC и слияния признаков, чтобы заново собрать признаки на

различных уровнях. Модель была протестирована на FaceForensics++, Celeb-DF и DFDC и достигла точности 99,11 %.

Авторы [15] предложили метод ансамблирования на основе стеков, в котором признаки, сгенерированные двумя моделями CNN, укладываются в стек, после чего осуществляется выбор оптимальных признаков и исключение несоответствующих. Метод был протестирован и валидирован с использованием пакетов FaceForensics++ и Celeb-DF. Для извлечения признаков были использованы две модели: Xception и EfficientNet-B7. В качестве основных моделей были использованы XG Boost Regressor и Random Forest classifier. Метод достиг точности 98 %.

Метод [16] обнаруживает подмену лица с использованием нейронной сети сверточного внимания. Авторы статьи определили значимость преобразования Фурье как представления в частотной области и закодировали лицевые видео в матрическую визуализацию тепловой карты (MVHM) для ввода в сети классификации изображений. Авторы обнаружили, что механизм пространственного внимания увеличивает производительность сети VGG19 на 9,38 процентных пункта. Метод был протестирован на DeepFakeTIMIT и достиг точности 99,2 %.

Авторы [17] предлагают детектор текстур и артефактов (TAD) для обнаружения подмены лиц, целью которого является отделение взаимно исключающих текстурных несоответствий и артефактной информации, ослабление их взаимного влияния и улучшение обобщающей способности модели. Авторы использовали пакеты FaceForensics++, WildDF, DFDC, CelebDF в качестве наборов данных и выбрали набор разнообразных базовых методов (XceptionNet, EfficientNetB4, EfficientNetB4Att, DSP-FWA, MCX-API) для сравнения с предложенным методом. Метод достигает точности 99 %.

Авторы [18] интегрировали обнаружение подмены лиц с 3D-оценкой взгляда; предложили стратегию интеграции биометрических признаков методом среднеквадратичной ошибки (MSE) и слиянием утечек признаков, чтобы регуляризовать модель обнаружения подмены; использовали FaceForensics++, Celeb-DF и WildDeepfake и достигли точности до 86 %. Эта интеграция наделяет модель способностью различать подделанные видео, выявляя пространственные несоответствия в области глаз с точки зрения взгляда между заданными кадрами.

Модуль многоуровневого слияния (MSF), разработанный в [19], может получать поддельные лицевые признаки, а интерактивный двуканальный модуль (IDS) может лучше интегрировать информацию о признаках в частотной и пространственной областях. Для решения проблем низкокачественных наборов данных и плохой производительности обнаружения на различных наборах данных это исследование предлагает многомасштабную интерактивную двуканальную сеть. Метод был протестирован на FaceForensics++ и Celeb-DF и достиг точности 99 %.

Статья [20] предлагает конечную преобразовательную пространственно-временную модель SFormer, использующую архитектуру трансформеров для выявления глобальных связей между несколькими локальными точками кадров и временных отношений среди пространственных признаков последовательных кадров. Модель SFormer использует Swin-трансформер для извлечения пространственных признаков, за которым следуют блоки трансформеров для временного анализа. Испытания проводятся на пяти наборах данных,

а именно DFD, Celeb-DF, FF++, DFDC и Deeper-Forensics, достигая точности 97,81 %, 99,1 %, 100 %, 93,67 % и 99,67 % соответственно.

Авторы [21] предлагают модуль слияния пространственных и частотных характеристик (SFFB), структура которого проста и легко реализуется. Для этого используют пространственные признаки, признаки частотной области и логиты для передачи информации, а также FaceForensics++ и Celeb-Deerfake, в результате чего достигается точность 97 %.

Авторы [22] предлагают сеть высокого частотного усиления (HiFE) для обработки низкокачественных данных. Адаптивные частотные признаки реализуются с помощью блочных DCT, механизма внимания по каналам, модуля узкого места по каналам и обратного блочного DCT. Более того, разрабатываются многократные уровни слоев декомпозиции DWT и каскадные стратегии многомасштабного слияния для реализации адаптивного глобального усиления высоких частот. Метод был протестирован на FaceForensics++, Celeb-DF и OpenForensics и достиг точности 99 %.

Хотя большинство существующих методов сосредоточены на обнаружении подмены лиц в видеоданных, этот подход рассматривает замену лиц исключительно на изображениях. Предложен новый метод оценки существующих алгоритмов подмены лиц, в котором подчеркиваются их эффективность и ограничения. Метод не только обнаруживает подмену лиц, но также концентрируется на признаках, которые служили основой для их обнаружения.

2. МЕТОДОЛОГИЯ

В настоящей работе предлагается углубиться в сложную задачу понимания важности признаков для классификации лиц (рис. 1). Предложенный метод позволяет не только обнаружить подмену лиц, но и вычислить важность признаков для оценки алгоритма. Ниже представлено описание каждого шага.

Предложено сформировать набор данных из нескольких существующих наборов данных (например, CelebA, Labeled Faces in the Wild), содержащих изображения лиц. Набор данных включает изображения реальных людей (несгенерированные подмены лиц) и разделяется на базовые, вторичные и оригинальные лица для обмена. Выбираются подмножества лиц для базовых и вторичных лиц в пропорции 1:1. Лица из первой группы выступают в качестве холста, на который будут заменены признаки из второй группы. Оригинальные лица представляют собой неизменное состояние и участвуют в формировании результирующего набора данных.

Было предложено использовать несколько различных генераторов подмены лиц, чтобы каждый из них обрабатывал одну и ту же часть лиц из первых двух наборов. Результатом генераторов станет набор замененных лиц. Оригинальные лица вместе с замененными берутся в равных пропорциях и формируют новый набор данных. Этот набор используется для дальнейшего анализа и формирования набора признаков.

Использование пакета Google Media Pipe Face Mesh позволяет оценивать 3D-ориентиры лиц. Он использует машинное обучение (ML) для определения 3D-структуры лица и требует только изображения без необходимости в глубинной карте. Предлагается использовать его для захвата точек на лице с целью формирования набора лицевых ориентиров и затем вычисления характеристик.

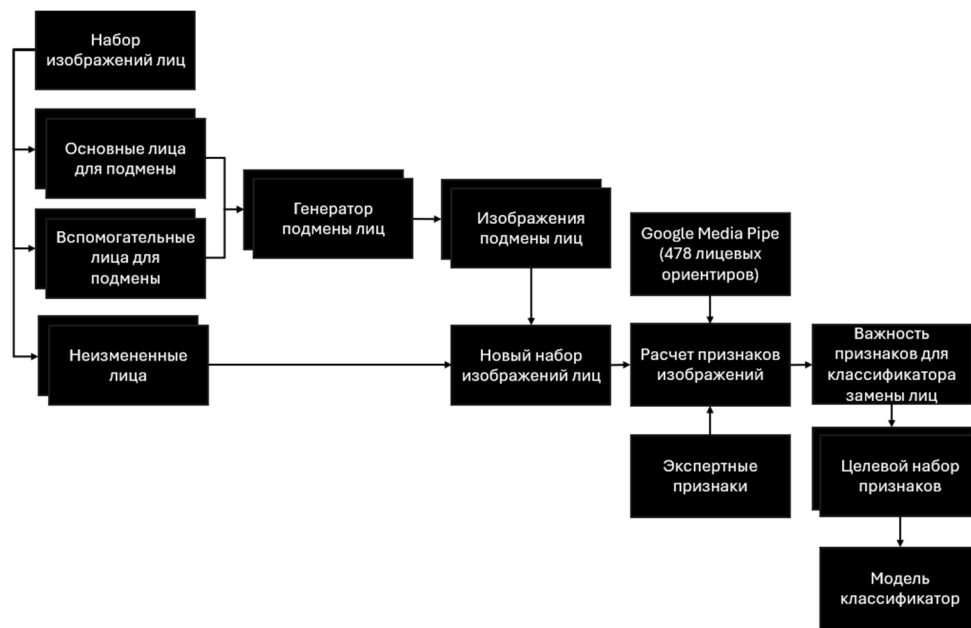


Рис. 1. Подход к определению важности признаков

Fig. 1. An approach to identifying feature importance

Выдвигается предположение, что наиболее значимые признаки для обнаружения изображений замены лиц связываются как с формой лица, так и с отношениями отдельных расстояний на лице. Характеристиками формы лица приняты площадь и объем. Рассматриваются такие признаки, как основные пропорции частей лица (глаза, рот и всё лицо). Для оценки положения частей лица используются отношения, сравнивающие горизонтальные размеры лица с вертикальными. Для этого берутся следующие расстояния: расстояние от глаза до уха, деленное на расстояние от глаза до носа с правой и левой стороны лица; расстояние от глаза до глаза, деленное на расстояние от глаза до носа с левой и правой стороны лица, и расстояние от уха до глаза, деленное на расстояние от уха до рта с левой и правой стороны лица.

Описанные признаки вычисляются для каждого изображения из набора данных, после чего они разделяются на три набора признаков: лицевые ориентиры, экспертные признаки и комбинированный набор признаков. Результирующие наборы признаков определяют окончательные данные, на которых будут обучены модели классификаторов.

После этого обучаются три модели, каждая со своим набором признаков. Выбирается перцептрон в качестве модели классификатора из-за его скорости обучения и возможности получения коэффициентов модели для последующего анализа важности признаков. Перцептрон – это базовый алгоритм в машинном обучении, используемый для бинарной классификации. Он состоит из входных признаков, весов, функции суммирования, функции активации и смещения. Перцептрон вычисляет взвешенную сумму входных данных, применяет функцию активации и выдает бинарный результат. Он обучается, корректируя веса и смещение на основе ошибок предсказания. Несмотря на

свою простоту, персептрон является основой для разработки более сложных нейронных сетей и моделей глубокого обучения (рис. 2).

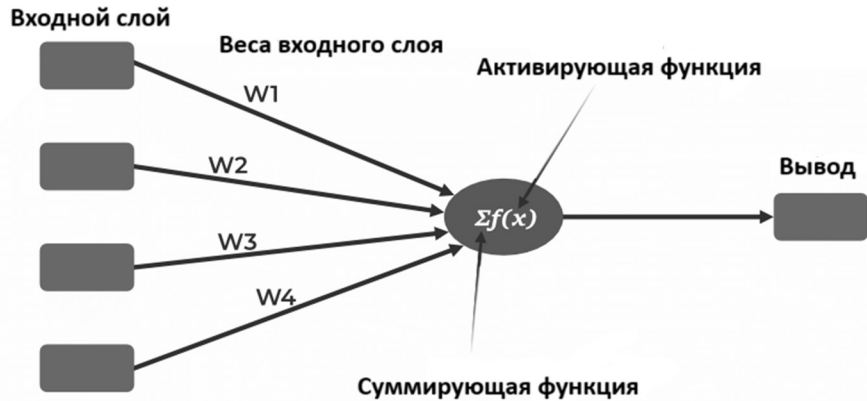


Рис. 2. Архитектура персептрона [23]

Fig. 2. The perceptron architecture [23]

Вычисляются метрики точности и анализируются коэффициенты модели. Используя нормализацию коэффициентов, определяем важность для каждого признака. На основе наиболее важных признаков из полученного набора обучается новая модель-классификатор.

3. РЕЗУЛЬТАТЫ

В этом разделе представлены результаты исследования: важность признаков, обученные модели и выводы. В качестве набора данных для лиц использовался набор данных CelebA. Всего было использовано 35 100 изображений для формирования описанных наборов изображений.

3.1. ОБУЧЕНИЕ МОДЕЛИ КЛАССИФИКАЦИИ

Была обучена модель классификации в соответствии с методологией, представленной в разделе 2, также обучены три разные модели на основе входных данных: экспертные признаки, лицевые ориентиры (см. раздел 2) и их комбинация (комбинированный набор признаков). В качестве тестовой выборки было взято 30 % от общего объема выборки (7800 из 23 400); оставшиеся данные были использованы для обучения. В качестве переменных гиперпараметров было использовано максимальное количество итераций, η_0 (константа, на которую умножаются обновления). Другим гиперпараметром является использование ранней остановки модели во время обучения, если точность не улучшается в течение длительного времени. Также использовались разные типы регуляризации (l2, l1, elasticnet или без регуляризации) с различными коэффициентами. Чтобы определить наиболее подходящий набор гиперпараметров, исследования проводились с каждым набором, и лучший был принят в качестве окончательного результата.

Модели, обученные на первых двух наборах, показали низкую точность, при этом первая модель не продемонстрировала разницы в точности идентификации оригинальных и изображений подмены (рис. 3), в то время как вторая модель более точно идентифицировала оригинальные изображения, чем изображения подмены лица (рис. 4), но всё равно с низкой точностью. Исходя из этого можно судить о том, что закономерность прослеживается, но точность нужно улучшить. Третья модель достигла большей точности, чем ее предшественники, и показала равную эффективность как на оригинальных изображениях, так и на изображениях подмены лица (рис. 5), достигнув 68,2 %.

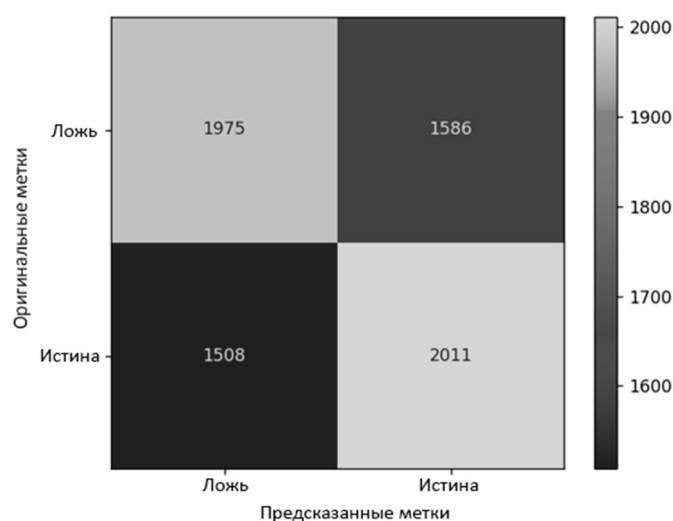


Рис. 3. Матрица спутанности для модели набора экспертных признаков

Fig. 3. The confusion matrix for the expert feature set model

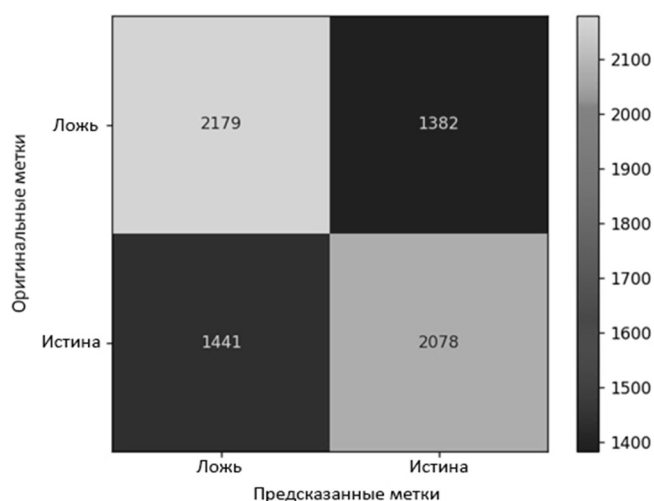


Рис. 4. Матрица спутанности для модели набора лицевых ориентиров

Fig. 4. The confusion matrix for the facial landmark feature set model

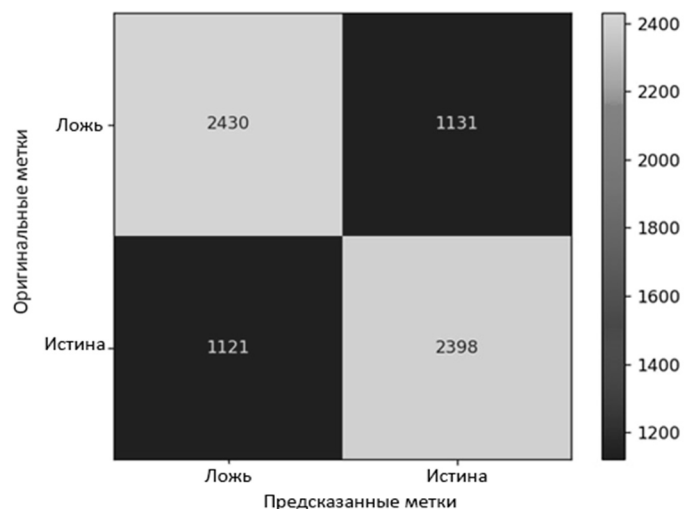


Рис. 5. Матрица спутанности для модели комбинированного набора признаков

Fig. 5. The confusion matrix for the combined feature set model

На основе полученных результатов (табл. 1) можно сделать вывод, что модель предсказывала лучше на лицевых точках, чем на экспертных признаках, и при комбинировании двух наборов признаков она показывала лучшие результаты, чем на каждом из наборов отдельно.

Таблица 1

Table 1

Метрики

Metrics

Набор признаков	Точность	F1-мера
Лицевые ориентиры	0.601	0.596
Экспертные признаки	0.563	0.565
Комбинированный	0.682	0.680

3.2. АНАЛИЗ ВАЖНОСТИ ЭКСПЕРТНЫХ ПРИЗНАКОВ

Важность экспертных признаков показана в табл. 2. Было замечено, что наиболее значимые признаки были связаны с формой лица (объем и площадь лица). Следующие четыре характеристики учитываются при определении подмены лиц: 1) расстояние от глаза до уха, деленное на расстояние от глаза до носа с правой стороны лица; 2) пропорции лица (ширина, деленная на высоту); 3) расстояние от глаза до уха, деленное на расстояние от глаза до носа с левой стороны лица; 4) расстояние от глаза до глаза, деленное на расстояние от глаза

до носа с левой стороны лица. Остальные характеристики имеют незначительное отношение к замене лиц.

Таблица 2

Table 2

Важность экспертных признаков
Expert feature importance

№ п/п	Признак	Важность
1	Объем лица	–0,796
2	Площадь лица	0,414
3	Расстояние от глаза до уха, деленное на расстояние от глаза до носа с правой стороны лица	–0,258
4	Пропорции лица (ширина, деленная на высоту)	–0,224
5	Расстояние от глаза до уха, деленное на расстояние от глаза до носа с левой стороны лица	–0,221
6	Расстояние от глаза до глаза, деленное на расстояние от глаза до носа с левой стороны лица	0,200
7	Пропорции левого глаза (ширина, деленная на высоту)	–0,168
8	Расстояние от уха до глаза, деленное на расстояние от уха до рта с левой стороны лица	0,143
9	Пропорции правого глаза (ширина, деленная на высоту)	–0,120
10	Расстояние от уха до глаза, деленное на расстояние от уха до рта с правой стороны лица	0,101
11	Пропорции рта (ширина, деленная на высоту)	0,099
12	Расстояние от глаза до глаза, деленное на расстояние от глаза до носа с правой стороны лица	0,041

3.3. ВАЖНОСТЬ ЛИЦЕВЫХ ОРИЕНТИРОВ

Важность лицевых ориентиров была вычислена с использованием коэффициентов модели персептрона. Для визуализации было решено нормализовать коэффициенты модели и затем сравнить их с цветом лицевых ориентиров. Конечный цвет лицевого ориентира в RGB был рассчитан как $(k \cdot 255, (1 - k) \cdot 255, 0)$, что дает значения для каждого цвета в пределах 255. Результаты продемонстрированы на рис. 6. Чтобы визуализировать важность лицевых точек, они были увеличены пропорционально их важности. Маленький размер

символизировал низкую важность, в то время как больший символизировал высокую важность. На основе полученных результатов был сделан вывод, что модель классификатора в первую очередь была ориентирована на уголки губ и контур рта, область вокруг глаз и границы лица.



Рис. 6. Значимость лицевых ориентиров (маленькие – низкая важность, большие – высокая важность)

Fig. 6. Facial landmarks importance (small is low importance, big is high importance)

ЗАКЛЮЧЕНИЕ

Проанализировано текущее состояние исследований и на основе выводов и экспертной оценки было сформировано три подхода к извлечению признаков на изображениях для обнаружения подмены лиц. На основе этих наборов обучены три модели классификатора, вычислены их качественные метрики и визуализированы матрицы спутанности. Был найден лучший классификатор, достигнувший точности 0,682.

Проанализированы коэффициенты полученной модели и на их основе вычислена важность каждого признака; определены наиболее значимые параметры через экспертную оценку и выделены наиболее значимые области лица для обнаружения подмены лиц.

В будущем, чтобы повысить точность, планируется использовать разные наборы данных в качестве основы для изображений и разные генераторы для

создания большей разнообразности в наборе данных. Также планируется обучить более сложную архитектуру модели на текущем наборе признаков для более глубокого подхода к обнаружению подмены лица.

СПИСОК ЛИТЕРАТУРЫ

1. Riehle C. Europol report criminal use of deepfake technology // Eu crim. – 2022, 9 May. – URL: <https://eucrim.eu/news/europol-report-criminal-use-of-deepfake-technology/> (accessed: 27.02.2025).
2. Facing reality? Law enforcement and the challenge of deepfakes: an observatory report from the Europol Innovation Lab. – Luxembourg: Publications office of the European Union, 2022. – DOI: 10.2813/158794.
3. Deepfakes and beyond: a survey of face manipulation and fake detection / R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia // Information Fusion. – 2020. – Vol. 64. – P. 131–148. – DOI: 10.1016/J.INFFUS.2020.06.014.
4. Nawaz M. Javed A., Irtaza A. A deep learning model for FaceSwap and face-reenactment deepfakes detection / Applied Soft Computing. – 2024. – Vol. 162. – P. 111854. – 2024. – DOI: 10.1016/J.ASOC.2024.111854.
5. Reis P.M.G.I., Ribeiro R.O. A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores // Forensic Science International. – 2024. – Vol. 358. – P. 111747. – DOI: 10.1016/J.FORSCIINT.2023.111747.
6. Swapped face detection using deep learning and subjective assessment / X. Ding, Z. Raziei, E.C. Larson, E.V. Olinick, P. Krueger, M. Hahsler // EURASIP Journal on Information Security. – 2020. – Art. 6. – DOI: 10.1186/S13635-020-00109-8.
7. Yan B., Li C.T., Lu X. JRC: Deepfake detection via joint reconstruction and classification // Neurocomputing. – 2024. – Vol. 598. – P. 127862. – DOI: 10.1016/J.NEUCOM.2024.127862.
8. DFGNN: An interpretable and generalized graph neural network for deepfakes detection / F. Khalid, A. Javed, Q. ul Ain, H. Ilyas, A. Irtaza // Expert Systems with Applications. – 2023. – Vol. 222. – P. 119843. – DOI: 10.1016/J.ESWA.2023.119843.
9. Essa E. Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection // Neurocomputing. – 2024. – Vol. 598. – P. 128128. – DOI: 10.1016/J.NEUCOM.2024.128128.
10. Wang Y., Huang H. Audio-visual deepfake detection using articulatory representation learning // Computer Vision and Image Understanding. – 2024. – Vol. 248. – P. 104133. – 2024. – DOI: 10.1016/J.CVIU.2024.104133.
11. Preserving manipulated and synthetic Deepfake detection through face texture naturalness / C.J. Chew, Y.C. Lin, Y.C. Chen, Y.Y. Fan, J.S. Lee // Journal of Information Security and Applications. – 2024. – Vol. 83. – P. 103798. – DOI: 10.1016/J.JISA.2024.103798.
12. An efficient deepfake video detection using robust deep learning / A. Qadir, R. Mahum, M.A. El-Meligy, A.E. Ragab, A. AlSalman, M. Awais // Heliyon. – 2024. – Vol. 10 (5). – P. e25757. DOI: 10.1016/J.HELİYON.2024.E25757.
13. ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection / S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, R. Sarkar // Expert Systems with Applications. – 2022. – Vol. 210. – P. 118423. – DOI: 10.1016/J.ESWA.2022.118423.
14. Deepfake detection via inter-frame inconsistency recomposition and enhancement / C. Zhu, B. Zhang, Q. Yin, C. Yin, W. Lu // Pattern Recognition. – 2024. – Vol. 147. – P. 110077. – DOI: 10.1016/J.PATCOG.2023.110077.
15. Deepfake detection using deep feature stacking and meta-learning / G. Naskar, S. Mohiuddin, S. Malakar, E. Cuevas, R. Sarkar // Heliyon. – 2024. – Vol. 10 (4). – P. e25933. – DOI: 10.1016/j.heliyon.2024.e25933.
16. A Deepfake detection algorithm based on fourier transform of biological signal / Y. Ni, W. Zeng, P. Xia, G.S. Yang, R. Tan // Computers, Materials & Continua. – 2024. – Vol. 79 (3). – P. 5295–5312. – DOI: 10.32604/CMC.2024.049911.
17. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection / J. Gao, M. Micheletto, G. Orrù, S. Concas, X. Feng, G.L. Marcialis, F. Roli // Engineering Applications of Artificial Intelligence. – 2024. – Vol. 133. – P. 108450. – DOI: 10.1016/J.ENGAPAI.2024.108450.

18. GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning / Q. He, C. Peng, D. Liu, N. Wang, X. Gao // *Neural Networks*. – 2024. – Vol. 180. – P. 106636. – DOI: 10.1016/J.NEUNET.2024.106636.
19. DeepFake detection method based on multi-scale interactive dual-stream network / Z. Cheng, Y. Wang, Y. Wan, C. Jiang // *Journal of Visual Communication and Image Representation*. – 2024. – Vol. 104. – P. 104263. – DOI: 10.1016/J.JVCIR.2024.104263.
20. Kingra S., Aggarwal N., Kaur N. SFormer: an end-to-end spatio-temporal transformer architecture for deepfake detection // *Forensic Science International: Digital Investigation*. – 2024. – Vol. 51. – P. 301817. – DOI: 10.1016/J.FSIDI.2024.301817.
21. Spatial-frequency feature fusion based deepfake detection through knowledge distillation / B. Wang, X. Wu, F. Wang, Y. Zhang, F. Wei, Z. Song // *Engineering Applications of Artificial Intelligence*. – 2024. – Vol. 133. – P. 108341. – DOI: 10.1016/J.ENGAPPAL.2024.108341.
22. DeepFake detection based on high-frequency enhancement network for highly compressed content / J. Gao, Z. Xia, G.L. Marcialis, C. Dang, J. Dai, X. Feng // *Expert Systems with Applications*. – 2024. – Vol. 249. – P. 123732. – DOI: 10.1016/J.ESWA.2024.123732.
23. Perceptron Algorithm for Classification using Sklearn // *GeeksforGeeks: website*. – 2023. – URL: <https://www.geeksforgeeks.org/sklearn-classification-using-perceptron/?ysclid=m0novhnsyv693925398> (accessed: 04.03.2025).

Халеев Михаил Дмитриевич, младший научный сотрудник лаборатории интегрированных систем автоматизации Федерального государственного бюджетного учреждения науки «Санкт-Петербургский федеральный исследовательский центр Российской академии наук». Основное направление научных исследований – искусственный интеллект, компьютерное зрение. E-mail: Haleev.M@ias.spb.su

Mikhail Dmitrievich Haleev, junior researcher at the Laboratory of Integrated Automation Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). The main area of his scientific research is artificial intelligence and computer vision. E-mail: Haleev.M@ias.spb.su

DOI: 10.17212/2782-2001-2025-1-105-120

An intelligent method for automatic detection of face substitution in an image*

M.D. HALEEV

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 39, 14th line of Vasilyevsky Island, St. Petersburg, 199178, Russian Federation

Haleev.M@ias.spb.su

Abstract

Deepfake images have become a major problem in today's digital landscape. Such images are usually created using advanced machine learning techniques. These fake images can deceive viewers, posing risks to privacy, security, and trust. In this paper we introduce an innovative approach to generating a target feature set for identifying deepfakes and consequently to detect a deepfake, generating a dataset of labeled substitution face images based on a public dataset using a face substitution generator. Facial keypoints are calculated and analyzed in the volumetric format for the images in the dataset. Besides, the corresponding features identified by an expert method are calculated and several classifier models based on three feature sets including facial keypoints, expert features, and their mixture are trained. Also, the performance of the classifiers is evaluated, confusion matrices and tracking patterns depending on the target feature set are

* Received 19 September 2024.

constructed achieving an accuracy of 0.682 and an F1-measure of 0.680 on the combined feature set. Using the coefficients of the obtained models, the importance of features was assessed and the most significant of them were identified for a more in-depth approach to forming data sets when training machine-learning models. For the calculated expert features, the most important were facial features characterizing its geometric shape, and in the case of key points of the face, the most significant areas were the boundaries of facial objects: the contour of the face, the areas around the eyes and the corners of the lips.

Keywords: face substitution, computer vision, artificial intelligence, generative content, content labeling, deepfake, information security, facial landmarks

REFERENCES

1. Riehle C. Europol report criminal use of deepfake technology. *Eucrim*, 2022, 9 May. Available: <https://eucrim.eu/news/europol-report-criminal-use-of-deepfake-technology/> (accessed 27.02.2025).
2. *Facing reality? Law enforcement and the challenge of deepfakes: an observatory report from the Europol Innovation Lab*. Luxembourg, Publications office of the European Union, 2022. DOI: 10.2813/158794.
3. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: a survey of face manipulation and fake detection. *Information Fusion*, 2020, vol. 64, pp. 131–148. DOI: 10.1016/J.INFFUS.2020.06.014.
4. Nawaz M., Javed A., Irtaza A. A deep learning model for FaceSwap and face-reenactment deepfakes detection. *Applied Soft Computing*, 2024, vol. 162, p. 111854. DOI: 10.1016/J.ASOC.2024.111854.
5. Reis P.M.G.I., Ribeiro R.O. A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores. *Forensic Science International*, 2024, vol. 358, p. 111747. DOI: 10.1016/J.FORSCIINT.2023.111747.
6. Ding X., Raziei Z., Larson E.C., Olinick E.V., Krueger P., Hahsler M. Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020, art. 6. DOI: 10.1186/S13635-020-00109-8.
7. Yan B., Li C.T., Lu X. JRC: Deepfake detection via joint reconstruction and classification. *Neurocomputing*, 2024, vol. 598, p. 127862. DOI: 10.1016/J.NEUCOM.2024.127862.
8. Khalid F., Javed A., Ain Q., Ilyas H., Irtaza A. DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Systems with Applications*, 2023, vol. 222, p. 119843. DOI: 10.1016/J.ESWA.2023.119843.
9. Essa E. Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection. *Neurocomputing*, 2024, vol. 598, p. 128128. DOI: 10.1016/J.NEUCOM.2024.128128.
10. Wang Y., Huang H. Audio-visual deepfake detection using articulatory representation learning. *Computer Vision and Image Understanding*, 2024, vol. 248, p. 104133. DOI: 10.1016/J.CVIU.2024.104133.
11. Chew C.J., Lin Y.C., Chen Y.C., Fan Y.Y., Lee J.S. Preserving manipulated and synthetic Deepfake detection through face texture naturalness. *Journal of Information Security and Applications*, 2024, vol. 83, p. 103798. DOI: 10.1016/J.JISA.2024.103798.
12. Qadir A., Mahum R., El-Meligy M.A., Ragab A.E., AlSalman A., Awais M. An efficient deepfake video detection using robust deep learning. *Heliyon*, 2024, vol. 10 (5), p. e25757. DOI: 10.1016/J.HELİYON.2024.E25757.
13. Ganguly S., Ganguly A., Mohiuddin S., Malakar S., Sarkar R. ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. *Expert Systems with Applications*, 2022, vol. 210, p. 118423. DOI: 10.1016/J.ESWA.2022.118423.
14. Zhu C., Zhang B., Yin Q., Yin C., Lu W. Deepfake detection via inter-frame inconsistency recomposition and enhancement. *Pattern Recognition*, 2024, vol. 147, p. 110077. DOI: 10.1016/J.PATCOG.2023.110077.
15. Naskar G., Mohiuddin S., Malakar S., Cuevas E., Sarkar R. Deepfake detection using deep feature stacking and meta-learning. *Heliyon*, 2024, vol. 10 (4), p. e25933. DOI: 10.1016/j.heliyon.2024.e25933.

16. Ni Y., Zeng W., Xia P., Yang G.S., Tan R. A deepfake detection algorithm based on fourier transform of biological signal. *Computers, Materials & Continua*, 2024, vol. 79 (3), pp. 5295–5312. DOI: 10.32604/CMC.2024.049911.
17. Gao J., Micheletto M., Orrù G., Concas S., Feng X., Marcialis G.L., Roli F. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Engineering Applications of Artificial Intelligence*, 2024, vol. 133, p. 108450. DOI: 10.1016/J.ENGAPPAI.2024.108450.
18. He Q., Peng C., Liu D., Wang N., Gao X. GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning. *Neural Networks*, 2024, vol. 180, p. 106636. DOI: 10.1016/J.NEUNET.2024.106636.
19. Cheng Z., Wang Y., Wan Y., Jiang C. DeepFake detection method based on multi-scale interactive dual-stream network. *Journal of Visual Communication and Image Representation*, 2024, vol. 104, p. 104263. DOI: 10.1016/J.JVCIR.2024.104263.
20. Kingra S., Aggarwal N., Kaur N. SFormer: an end-to-end spatio-temporal transformer architecture for deepfake detection. *Forensic Science International: Digital Investigation*, 2024, vol. 51, p. 301817. DOI: 10.1016/J.FSIDI.2024.301817.
21. Wang B., Wu X., Wang F., Zhang Y., Wei F., Song Z. Spatial-frequency feature fusion based deepfake detection through knowledge distillation. *Engineering Applications of Artificial Intelligence*, 2024, vol. 133, p. 108341. DOI: 10.1016/J.ENGAPPAI.2024.108341.
22. Gao J., Xia Z., Marcialis G.L., Dang C., Dai J., Feng X. DeepFake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications*, 2024, vol. 249, p. 123732. DOI: 10.1016/J.ESWA.2024.123732.
23. Perceptron Algorithm for Classification using Sklearn. *GeeksforGeeks*. Website. 2023. Available at: <https://www.geeksforgeeks.org/sklearn-classification-using-perceptron/?ysclid=m0novhnsyv693925398> (accessed 04.03.2025).

Для цитирования:

Халеев М.Д. Интеллектуальный метод автоматического выявления подмены лица на изображении // Системы анализа и обработки данных. – 2025. – № 1 (97). – С. 105–120. – DOI: 10.17212/2782-2001-2025-1-105-120.

For citation:

Haleev M.D. Intellektual'nyi metod avtomaticheskogo vyyavleniya podmeny litsa na izobrazhenii [An intelligent method for automatic detection of face substitution in an image]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2025, no. 1 (97), pp. 105–120. DOI: 10.17212/2782-2001-2025-1-105-120.