ПРОГРАММИРОВАНИЕ, 2019, № 3, с. 57–63

КОМПЬЮТЕРНАЯ ГРАФИКА

УДК 004.92

ПОЛУАВТОМАТИЧЕСКИЙ МЕТОД СБОРА ВЫБОРОК ДЛЯ ОБУЧЕНИЯ АЛГОРИТМА ИДЕНТИФИКАЦИИ ЛИЦ

© 2019 г. Н. Ю. Багров[a,*], А. С. Конушин[a,], В. С. Конушин[b,***]**

[a] Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики (ВМК)
119899 Москва, Ленинские горы, д. 1, стр. 8, Россия

[b] ООО "Технологии видеоанализа"
119634 Москва, улица Скульптора Мухиной, д. 7, Россия

** E-mail: nikita.bagrov@graphics.cs.msu.ru*

*** E-mail: ktosh@graphics.cs.msu.ru*

**** E-mail: vadim@tevian.ru*

Поступила в редакцию 14.01.2019 г.
После доработки 18.01.2019 г.
Принята к публикации 18.01.2019 г.

В работе предлагается метод полуавтоматического сбора выборок для обучения алгоритма идентификации лиц. В экспериментальной оценке уделяется внимание работе алгоритма на этнически разнообразных данных. Также проводится оценка работы алгоритма на данных с большой вариацией возрастов. Разработанный метод позволяет проводить дальнейшее увеличение обучающей выборки за счет индексации новых данных.

DOI: 10.1134/S0132347419030026

1. ВВЕДЕНИЕ

Сегодня системы распознавания лиц широко используются как в сфере обеспечения безопасности, так и в финансовой сфере, рекламе и других направлениях. Однако для их применения в системе безопасности городов или крупных транспортных объектах текущей точности лучших алгоритмов часто недостаточно. Среди ключевых проблем можно выделить построение качественных обучающей и тестовой выборок. В научных статьях часто используют обучающие выборки, собранные из открытых данных с неравномерным этническим и возрастным распределением. Тестовые выборки тоже часто похожи по распределению на обучающие, могут иметь пересечения с обучающей выборкой по людям, что может привести к искажению оценки точности. Отдельно стоит рассмотреть проблему низкого качества изображений на практике, что отличается от большинства открытых обучающих выборок. В работе предлагается способ агрегации и кластеризации данных из разных источников, позволяющий создавать большие выборки с качественной разметкой (мало ошибок). Влияние этого оценивается экспериментально, причем используя разнообразные тестовые наборы.

2. ОБЗОР СУЩЕСТВУЮЩИХ КОЛЛЕКЦИЙ

По результатам проведенного обзора можно выделить следующие коллекции изображений, используемых для обучения и тестирования алгоритмов распознавания человека по лицу:

• CASIA-WebFace [1] – 450 тыс. изображений лиц 10575 людей. Собрана автоматически используя поисковые запросы к IBDB (база актеров). Из минусов данной коллекции можно отметить низкую вариативность возрастов и сравнительно высокое качество изображений, при этом число ошибок разметки достаточно небольшое [2].

• CAS-PEAL [3] – 100 тыс. изображений 1040 людей. Выборка собрана в лабораторных условиях, большое разнообразие по углам съемки.

• MS-Celeb-1M [4] – 10 м изображений 100 тыс. людей. Собрана из IBDB и поисковых запросов. Содержит достаточно большое число ошибок, требует применение дополнительной фильтрации.

• NIST IBJ-C [5] – 138 тыс. изображений, 11 тыс. видео. Выборка собрана из публично доступных изображений актеров, политиков, спортсменов. Использовалась в тестировании коммерческих алгоритмов Национальным институтом стандартов и технологий США.

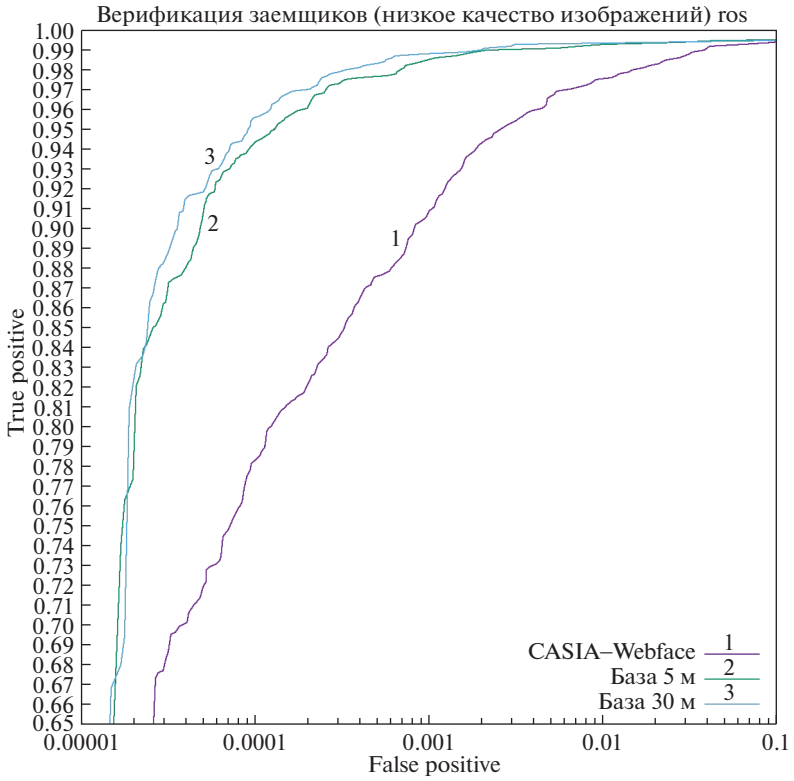

Рис. 1. Датасет 2. Верификация заемщиков в финансовых организациях (сравнение фотографии из паспорта с лицом – большая разница в возрасте между фотографией в паспорте и лицом, низкое качество изображений).

• Celebfaces [6] – 200 тыс. изображений, 10 тыс. различных людей. Собрана с использованием ресурса Google Images, актеры, политики, спортсмены.

• FERET [7] – 2400 изображений 856 различных людей. Одна из самых первых выборок, но при этом похожа на CAS-PEAL по распределению ракурсов, собрана в лабораторных условиях.

Все эти коллекции обладают следующими недостатками:

1) Низкая вариативность данных. Отсутствие данных, полученных в реальных условиях эксплуатации систем распознавания лиц.

2) Недостаточное количество изображений редких классов: мало людей с возрастом больше 50 лет, на фотографиях присутствует макияж и сценический грим (что характерно для этих данных, т.к. большую часть составляют изображения актеров и политиков).

3) Объем выборок недостаточно большой для обучения нейросетевых моделей, что приводит к переобучению.

Для решения этих проблем в работе предложен способ сбора и объединения данных для обучения распознавания лиц из различных источников данных.

3. ОБЗОР НЕЙРОСЕТЕВЫХ АРХИТЕКТУР ДЛЯ ИДЕНТИФИКАЦИИ ЛИЦ

Почти все современные методы для идентификации лиц основаны на сверточных нейронных сетях. На практике они отличаются глубиной и числом параметров в нейронной сети, количеством сетей для принятия решения (комитет нейронных сетей). Авторы используют различные функции потерь, на текущий момент это – открытая задача, и однозначного ответа какая из функций работает лучше на задаче идентификации лиц нет. При обучении нейронных сетей для идентификации лиц часто используют подход с обучением многоклассового классификатора, например, как в работе [6], но иногда функция потерь может быть основана на расстоянии в некотором пространстве [8]. Авторы также используют различные методы предобработки данных, например, в работе [9] используется трехмерное преобразование для нормализации изображений лиц и приведения их к фронтальному ракурсу. В работе [10] применялся ансамбль из 25 глубоких нейронных сетей, что повысило точность метода, но при этом усложнило процедуру обучения.

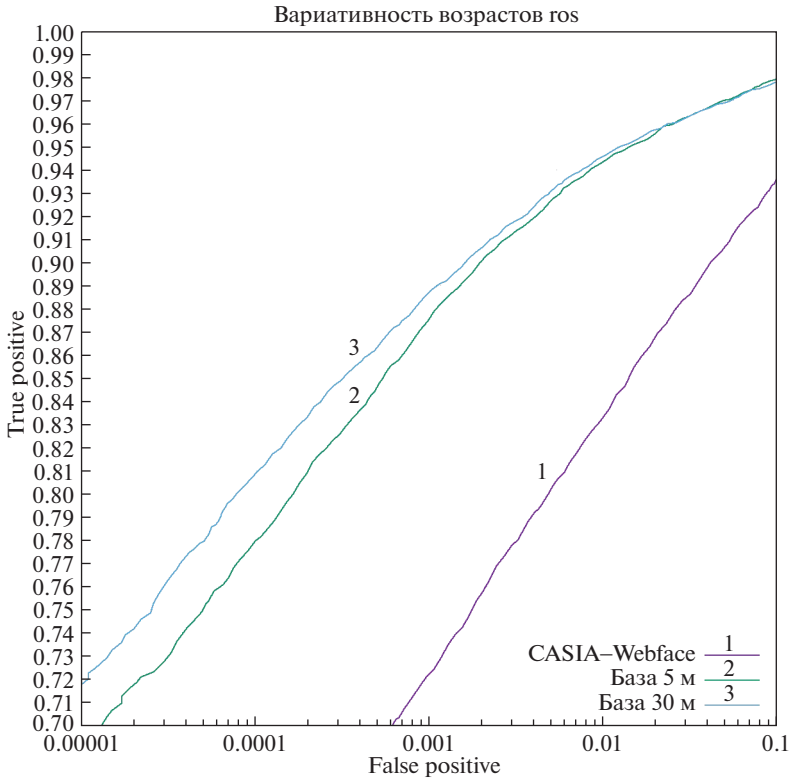

Рис. 2. Датасет 4. Выборка с высокой вариативностью возрастов (равномерное распределение по возрастам от 10 до 70 лет).

4. ОПИСАНИЕ ПРЕДЛОЖЕННОГО МЕТОДА

В работе предлагается новый подход к построению коллекции фотографий с лицами людей, основанный на сборе данных из различных ресурсов и последующем их объединении в одну выборку. На этапе объединения предложен метод кластеризации данных с использованием дополнительных метаданных из источников данных. В качестве базовой нейросетевой модели для построения кластеризации была выбрана архитектура resnet-32 [11], обученная на базе изображений лиц CASIA-WebFace. Нейросети на базе архитектуры ResNet показывают высокое качество работы, при этом более глубокий вариант сети использовать нет смысла, т.к. обучение этой сети ведется на относительно небольшой коллекции.

Для сбора обучающей выборки были использованы следующие источники данных:

1) IMDB (https://www.imdb.com/) – The Internet Movie Database. На ресурсе представлены сведения об актерах и их ролях в фильмах, а также загруженные пользователями и профессиональными фотографами фотографии актеров с различных мероприятий. Данные структурированы, есть текстовые метаданные в виде разметки присутствующих актеров на изображениях. Полная выгрузка данных содержит более 100 тыс. различных людей, количество изображений более 2 млн. На одном изображении могут присутствовать сразу несколько различных людей из базы. Изображения высокого качества, снятые профессиональными фотографами. База поддерживается в актуальном состоянии, данные верифицируются.

2) LISTAL (https://www.listal.com/). Данные плохо структурированы, база наполняется пользователями. Есть пересечения с IMDB, фотографии актеров дополнены как кадрами из фильмов, так и пользовательскими данными. Содержит более 3 млн изображений.

3) Google images (https://images.google.com). Загрузка данных из этого источника проводилась по ключевым словам – ФИО актеров, политиков, спортсменов. Список ФИО был взял из IMDB, сайтов различных спортивных федераций (например, IAAF). Проводилась загрузка до 1000 фотографий из поисковой выдачи.

4) Социальные сети. Изображения в этих источниках наиболее близки к реальным данным в видеонаблюдении (высокая вариативность ракурсов, низкое качество изображений, разнообразные условия освещения) и количество данных достаточно для построения репрезентативных выборок по возрастам и этническому составу. Было загружено и обработано более 500 млн фо-

Таблица 1. Сравнение точности алгоритма идентификации в зависимости от обучающей выборки

Точность при фиксированном FAR	CASIA-WebFace	5 млн изображений	30 млн изображений
Датасет 1, far 1e-5	86%	90.5%	92%
Датасет 2, far 1e-4	78%	94%	95.5%
Датасет 3, far 1e-5	74%	93%	95%
Датасет 4, far 1e-4	64%	78%	81%
CAS-PEAL, far 1e-5	75%	94%	97%

тографий 1 млн различных людей. Использовались различные источники данных для этнического разнообразия обучающей выборки. Метаданные, такие как пользовательская разметка людей на фотографиях, также использовались при построении выборки. Проводилась загрузка альбомов только тех людей, для которых есть портретная фотография высокого разрешения.

5) Изображения из фильмов, телетрансляций. Видеоданные позволяют получить практически неограниченное число изображений одного человека в различных условиях съемки. Данные вариативные по ракурсу и уровню освещения. Эти данные дополняют выборку, полученную из обработки ресурса IMDB. Суммарный объем данных около 10 млн изображений из 1000 фильмов и сериалов.

4.1. Использование метаданных источников данных

Использование дополнительной информации может повысить итоговое качество выборки за счет внесения дополнительных весов при кластеризации. Для социальных сетей полезно использовать пользовательскую разметку фотографий: в интерфейсе часто предусмотрена возможность отметить других людей или себя на фотографиях. Это может повышать меру сходства при совпадении работы алгоритма и пользовательской аннотации.

Ресурс IMDB предоставляет информацию об актерском составе фильмов и разметке людей на фотографиях. Разметку актеров на фотографиях можно использовать аналогично сценарию для социальных сетей, тогда как актерский состав фильмов можно использовать для получения дополнительных изображений лиц из видео.

ФИО человека можно использовать для объединения фотографий из разных ресурсов в одну запись в базе. Для актеров в качестве дополнительной информации можно использовать список ролей в фильмах – иногда бывают дубликаты в различных источниках данных по названию роли в фильме.

4.2. Кластеризация

После первичной загрузки и обработки изображений (удаления изображений без людей или поврежденных в процессе загрузки) проводилась следующая процедура:

1) На каждом изображении выполняется поиск всех лиц и вычисляется нейросетевой дескриптор лица (используя базовую модель нейросети).

2) Внутри каждой записи о человеке проводился выбор портретной фотографии. Портретная фотография может быть выбрана из метаданных (если они есть), либо как наиболее похожая на все остальные фотографии этого человека. Метрикой сходства является взвешенная сумма верификационных функций (результатов сравнения двух изображений лиц с использованием нейросети) этой фотографии и всех остальных фотографий человека. Веса суммы подбираются исходя из степени достоверности источника данных (экспертная оценка) и качества изображения (используя дополнительный алгоритм). Для выбранной нейросетевой модели значения степени достоверности источника составляют 0.5 для данных из социальных сетей и 1.2 для IMBD/Listal из-за наличия достоверной информации об актерах. Вес качества изображения суммируется со степенью достоверности источника данных.

3) Проводится построение полного графа зависимостей для каждого из людей в базе, где вершинами являются найденные лица на фотографиях, а весами дуг – мера сходства двух лиц. Затем из графа удаляются дуги с весами меньше заранее определенного числа. Это число подбирается исходя из числа фотографий для данного человека (так как мера сходства соответствует числу ошибок первого рода алгоритма распознавания лиц, возможно регулировать число ошибок в результирующей выборке).

4) Производится удаление мостов в графе. Это позволяет уменьшить число ошибок алгоритма распознавания лиц, который может признавать похожими разные лица. Если не проводить удаление мостов, то к “правильным” лицам человека могут добавиться другие лица через низкокачественную фотографию (экспериментально заме-

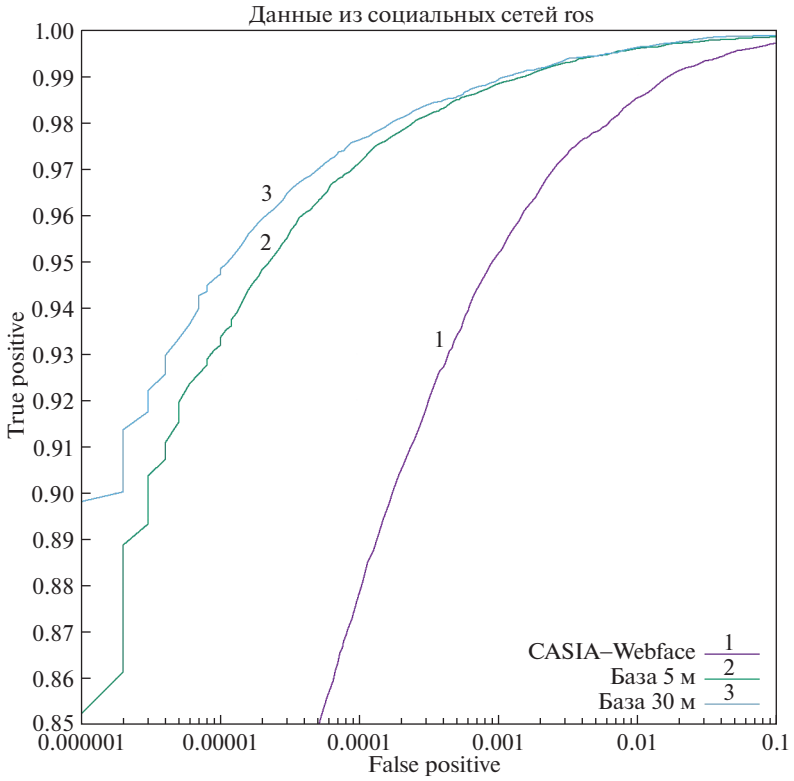

Рис. 3. Датасет 3. Выборка из социальных сетей с большим этническим разнообразием (Средняя Азия, латиноамериканские страны).

чено что алгоритмы распознавания лиц работают с более низкой точностью на некачественных лицах).

5) Производится выбор той компоненты связности графа, в которой находится вершина, соответствующая портретной фотографии. Все лица из этой компоненты помечаются как лица этого человека.

Отдельно следует рассмотреть ситуацию, при которой человек мог продублироваться в базе несколько раз. Это могло произойти при отсутствии метаданных или ошибки в них. Для объединения таких записей для каждого человека (после кластеризации) выбирается его портретная фотография и сравнивается со всеми остальными портретными фотографиями других людей. Экспериментально выбираются два порога:

1) Порог автоматического принятия решения (0.9 для данной нейросетевой модели). В этом случае если мера сходства выше заданного значения, то такие записи будут объединены автоматически.

2) Порог экспертной оценки (0.75 для данной нейросетевой модели). Если значение недостаточно высокое, то решение принимает эксперт.

В результате проведенного эксперимента только 15% случаев потребовали экспертной оценки, при этом процент ошибок-дубликатов составил менее 1%.

4.3. Обучение нейросетевой модели

В качестве базовой архитектуры выбрана resnet-32 с функцией потерь AM-Softmax [2]. Проводится процесс кластеризации с использованием этой обученной модели, затем сеть обучается на полученных данных и используется в дальнейшем в качестве базовой. Процесс кластеризации выборки может повторяться многократно с увеличением точности базовой модели. Каждая модель обучается на 8 видеокартах Tesla P40 в течение фиксированного времени или до стабилизации функции потерь (не более 5 дней, или если loss не будет уменьшаться в течение нескольких эпох обучения). Входное разрешение 112 × 112, все лица проходят предобработку: нормализацию расстояния между зрачками, выравнивание по геометрическому центру середин зрачков и губ. В процессе обучения применяются следующие преобразования:

• случайные вращения изображения на небольшой угол;

• случайные аффинные преобразования для моделирования ошибок детектора ключевых точек лица на этапе выравнивания;

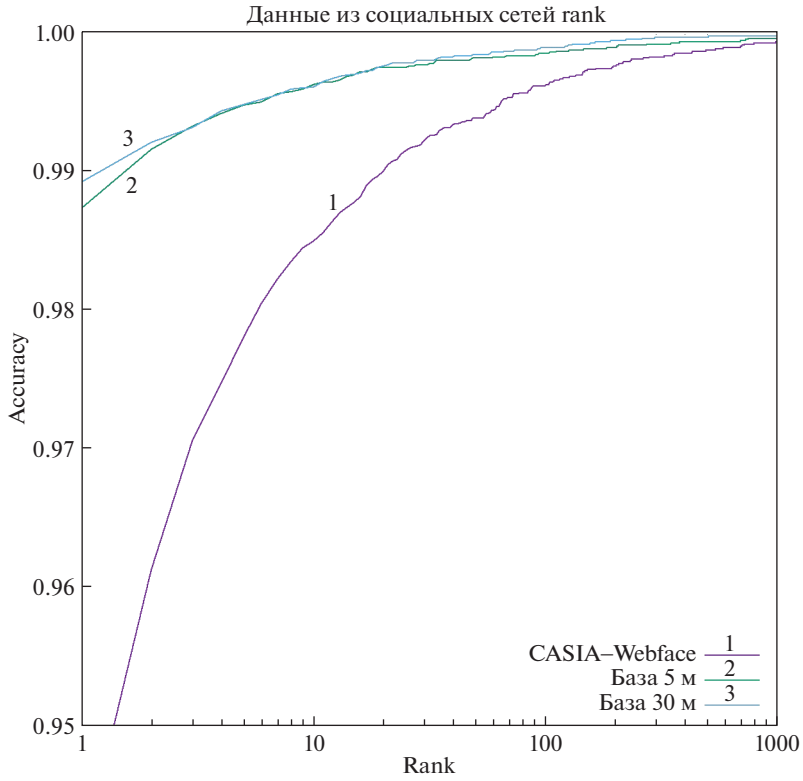

Рис. 4. Датасет 3. RANK-n кривая для задачи идентификации лиц на выборке с этническим разнообразием.

• изменение цветового баланса для моделирования различных условий освещения;

• добавление различных шумов на изображение.

4.4. Экспериментальная оценка

Была реализована система хранения и обработки данных, позволяющая индексировать большие коллекции изображений. Нейросетевые дескрипторы лиц и метаданные хранятся в СУБД MySQL, изображения в распределенной файловой системе cephfs. Суммарно до применения фильтрации было обработано более 700 млн изображений лиц людей, после фильтрации и кластеризации остается более 40 млн.

Экспериментальная оценка алгоритма проводится на фиксированном множестве выборок по следующим протоколам:

1) Замер доли ошибок второго рода при фиксированной доле первого рода. Ошибкой первого рода является тут случай, когда алгоритм разных людей считает одинаковыми, а второго рода когда одинаковых людей разными. Количество ошибок первого рода важно, например, для сценария обеспечения безопасности на транспорте. В таких системах число операторов ограничено и недопустимо их перегружать большим числом ложных обнаружений. На практике требуемый FAR (доля ошибок первого рода) для таких систем составляет от 1e-5 до 1e-7 в зависимости от пассажиропотока (не более 1 ошибки на 100 тыс. сравнений).

2) ROC кривая алгоритма. Данный протокол позволяет оценить в целом качество работы алгоритма.

3) RANK кривая алгоритма. RANK-n – означает то, что правильный результат идентификации был среди первых n результатов распознавания (отсортированных по значению оценивающей функции).

Для тестирования использовались следующие выборки:

1. Датасет 1 – выборка, полученная из данных видеонаблюдения.

2. Датасет 2 – выборка, состоящая из пар фотографий, – фото лица в паспорте и сэлфи. Подобные фотографии используются для верификации заемщиков в банках и микрофинансовых организациях.

3. Датасет 3 – выборка, собранная из социальных сетей, со значительным этническим разнообразием.

4. Датасет 4 – выборка, собранная из социальных сетей, где каждый человек представлен фотографиями, покрывающими большой диапазон возрастов (с 10 до 70 лет).

5. CAS-PEAL – уже упоминавшаяся китайская база лиц, собранная в лабораторных условиях, включает большие повороты лица.

В таблице 1 представлены результаты тестирования 3 версий алгоритмов:

1. Исходный алгоритм, обученный на CASIA-WebFace.

2. База 5 млн изображений – первая итерация алгоритма кластеризации.

3. База 30 млн – вторая итерация кластеризации с применением обученной нейросети на базе в 5 млн изображений.

Стоит отметить, что для баз Датасет 3 и CAS-PEAL прирост точности значительно больше, чем для других баз. Это вызвано тем, что в исходной выборке CASIA-WebFace мало представлены монголоидные лица.

На рисунках 1–4 показаны ROC и Rank кривые для некоторых из тестовых выборок.

5. РЕЗУЛЬТАТЫ

Предложенный алгоритм позволяет постоянно пополнять обучающую выборку как за счет загрузки дополнительных изображений из рассмотренных источников данных, так и за счет добавления новых источников данных. По результатам также видно, что этническое разнообразие выборки позволяет получить значительный прирост на определенных тестовых выборках (CAS-Peal) и специально собранной мультиэтнической выборке. Также алгоритм показал свою эффективность для мультивозрастных данных за счет обработки данных из социальных сетей, в которых широко представлены различные возрастные группы.

СПИСОК ЛИТЕРАТУРЫ

1. *Yi D. et al.* Learning face representation from scratch, 2014. arXiv preprint arXiv:1411.7923.
2. *Wang F. et al.* Additive margin softmax for face verification. IEEE Signal Processing Letters. 2018. V. 25. № 7. P. 926–930.
3. *Gao W. et al.* The CAS-PEAL large-scale Chinese face database and baseline evaluations. /IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2008. V. 38. № 1. P. 149–161.
4. *Guo Y. et al.* Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. European Conference on Computer Vision. Springer, Cham, 2016. P. 87–102.
5. *Maze B. et al.* IARPA Janus Benchmark–C: Face Dataset and Protocol. 11-th IAPR International Conference on Biometrics. 2018.
6. *Sun Y., Wang X., Tang X.* Deep learning face representation from predicting 10,000 classes. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. P. 1891–1898.
7. *Phillips P. J. et al.* The FERET database and evaluation procedure for face-recognition algorithms. Image and vision computing. 1998. V. 16. № 5. P. 295–306.
8. *Schroff F., Kalenichenko D., Philbin J.* Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 815–823.
9. *Taigman Y. et al.* Deepface: Closing the gap to human-level performance in face verification. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. P. 1701–1708.
10. *Sun Y. et al.* Deepid3: Face recognition with very deep neural networks, 2015. arXiv preprint arXiv:1502.00873.
11. *He K. et al.* Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 770–778.