УДК 004.855.5

РАСПОЗНАВАНИЕ ВОЗРАСТА ПО ИЗОБРАЖЕНИЮ ЛИЦА С ИСПОЛЬЗОВАНИЕМ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

© **Д. В. Пакулич**[1,2], **С. А. Якимов**[2], **С. А. Алямкин**[2]

[1] *Новосибирский государственный университет,*
630090, г. Новосибирск, ул. Пирогова, 2
[2] *ООО «Экспасофт»,*
630090, г. Новосибирск, ул. Николаева, 11
E-mail: d.pakulich@expasoft.ru

Задача определения возраста человека по изображению лица получила развитие с ростом популярности свёрточных нейронных сетей. Благодаря им появилась возможность выделять особенности лиц, которые незаметны взгляду человека, и интерпретировать эти особенности и их совокупность как возрастные характеристики. Проведён анализ существующих подходов к распознаванию возраста. Взяты данные из существующих наборов для обучения с последующей корректировкой в целях уменьшения ошибок, приобретённых при разметке алгоритмами сбора. Обучены и протестированы нейронные сети на полученных данных. Выявлена проблема с поворотом головы, для решения которой при обучении были предоставлены изображения лиц, повёрнутые с помощью нейронной сети PRNet.

Ключевые слова: свёрточные нейронные сети, распознавание возраста, глубокие нейронные сети, компьютерное зрение.

DOI: 10.15372/AUT20190307

Введение. В конце XX века определение возраста человека по фотографии стало одной из популярных задач машинного обучения. Существует много признаков во внешности человека, по которым можно приблизительно оценить возраст, причём не только по лицу, но и другим частям тела, в том числе по рукам. Решение данной задачи имеет множество практических приложений. К примеру, появляется возможность быстрого принятия решения о допуске лиц на мероприятие, имеющее возрастной ценз, или на продажу им некоторых товаров. В области маркетинга и анализа статистики сегментирование посещений по социально-демографическому признаку позволяет лучше понять состояние рынка и предпочтения пользователей.

Очень часто в этих целях требуется обрабатывать видеопоток со множеством лиц в реальном времени, что вызывает дополнительные трудности. По таким данным возможно не только уметь определять возраст человека, но и устанавливать его местоположение на кадре. В реальном времени возникает необходимость либо в больших вычислительных мощностях, либо в быстродействующей модели.

Поскольку задача определения возраста является популярной в течение длительного промежутка времени, в настоящий момент существует множество различных подходов к её решению.

В [1] впервые проведена работа по распознаванию возраста людей по фотографиям. Представлен метод, который по анализу первичных признаков лица, таких как положение его ключевых точек (нос, уголки губ и глаза), а также вторичных признаков (в частности, морщин) разделяет людей на три возрастные группы: детей, молодых людей и старшее поколение.

В последнее время для решения такой задачи, как и для многих задач машинного зрения, используют нейронные сети. Весь процесс можно разделить на три этапа: поиск лиц на изображении, выравнивание лица для дальнейшей обработки и предсказание возраста по лицу. Так как первый и второй этапы используются не только при определении возраста, но и, например, пола, при идентификации людей и т. д., был разработан ряд подходов к их реализации. Задача распознавания возраста в основном связана с реализацией третьего этапа и/или подготовкой набора данных.

Нейронные сети состоят из множества слоёв, для обучения которых используются функции потерь, которые получают результат работы нейронной сети, сравнивают его с ожидаемым результатом и возвращают коэффициент для корректировки параметров сети. Часто для решения различных задач компьютерного зрения применяются одинаковые архитектуры сетей: AlexNet [2], ResNet [3], MobileNetV2 [4] и т. д. Сложность модели определяет скорость её работы.

Задача установления возраста хотя и является задачей регрессии, однако чаще её решают как задачу классификации, где классом служит возраст человека с группировкой по различным возрастным диапазонам. Это происходит из-за того, что признаки взросления для каждой возрастной группы свои и нелинейно изменяются с возрастом.

В [5] предложен метод обучения моделей, названный Mean-Variance Loss, который предлагает рассматривать целевую переменную как распределение вероятностей по возрастным классам. При этом используется функция потерь, состоящая из комбинации перекрёстной энтропии, квадрата отклонения среднего значения распределения от целевой переменной и стандартного отклонения полученного распределения вероятности возраста. Этот подход даёт возможность уменьшить разброс предсказания возраста и повысить точность среднего показателя возраста.

В [6] предложен более глубокий подход, который применяет нейронную сеть, предобученную на наборе данных, для распознавания изображения на снимке. После обучения распознаванию возраста нейронная сеть делит изображение лица на части, предсказывает по ним возраст, а затем определяет наиболее репрезентативные части с помощью рекуррентных слоёв.

В [7] представлен метод Soft Stagewise Regression Network (SSR-Net), который благодаря иерархической структуре позволяет использовать более простые и быстрые нейронные сети.

Помимо нейронных сетей, для определения возраста применяются и другие методы машинного обучения [8], описывающие подход, который использует характеристики, полученные свёрточной нейронной сетью, в модели Random Forest для установления возрастной группы человека по изображению.

Задача распознавания возраста человека на видеозаписи отличается от задачи распознавания на статичном снимке. В этом случае на каждого человека приходится последовательность кадров различного качества, особенно если они были получены с помощью камер наблюдения. Поскольку человек редко смотрит прямо на камеру, качество кадров видеозаписи хуже, чем фотографий. В связи с этим точность определения возраста на видеозаписи в общем случае меньше, чем на фотографии. Так как на видеозаписи дана последовательность кадров с одним человеком, то из них можно выбрать те, на которых распознавание работает лучше. В [9] предложен подход, который сначала обучает часть нейронной сети предсказывать возраст на статичных изображениях, а затем, используя результаты этой части, обучает рекуррентные слои, определяющие кадры, подходящие для распознавания возраста.

Актуальность задачи привлекла внимание большого числа исследователей, и сегодня существуют различные методы и наборы данных для работы с ней. Цель представленного исследования — проанализировать различные методы, выделить проблемы и на основе

лучших решений составить и обучить собственную модель для определения возраста по фотографии человека. Для обучения использовались существующие наборы данных, из которых была предварительно исключена ошибочная разметка, и добавлены вариации лиц, в том числе с различных ракурсов съёмки.

Экспериментальная часть. В качестве класса моделей машинного обучения были выбраны нейронные сети со свёрточными слоями. В последние годы нейронные сети обрели большую популярность в различных областях машинного обучения, таких как анализ больших массивов данных, анализ фотографий, видеозаписей и других. Недостатком такого класса моделей зачастую является их вычислительная сложность и, как следствие, временны́е затраты на выполнение всех операций. Также основным требованием к нейронным сетям является наличие размеченного набора данных, содержащего минимальное количество ошибок.

С учётом этих проблем рассмотрены модели MobileNetV2, ResNet50 и SSR-Net.

В качестве метрик для сравнения результатов были использованы MAE (Mean Absolute Error, средняя абсолютная ошибка) и бинарная классификация с порогом в 18 лет, которая находит количество верно определённых лиц до 18 лет и выше.

Были использованы различные функции потерь, такие как категориальная перекрёстная энтропия, Mean-Variance Loss, Angular Softmax (A-Softmax) [10], Large Margin Cosine Loss (LMCL) [11], Additive Angular Margin Loss [12].

Данные для обучения модели. Необходимы правильно размеченные данные, содержащие большое количество примеров. В свободном доступе существует некоторое количество таких наборов, но их разнородность и наличие ошибочных меток могут сильно влиять на результаты работы моделей. В связи с этим были проанализированы следующие наборы данных.

IMDB + Wiki. Набор состоит из двух частей, содержащих 460 723 фотографии с вебсайта о кинематографе IMDB, 62328 фотографий из Википедии, в том числе с большим количеством ошибок, так как размечен автоматически по метаданным изображений. Ошибки разметки появились из-за неправильного выделения интересующего человека на фотографии с изображением нескольких человек. В наборе данных также встречаются ошибочные метки из-за того, что некоторые изображения являются переснятыми со старых кинокартин и неверно датируются автоматическим разметчиком.

LAP. Набор данных LAP использовался в 2016 году в конкурсе по распознаванию возраста и содержал 4113 примеров для обучения, 1500 примеров для валидации и 1978 примеров для тестирования. Все данные размечались разными людьми и представляли собой среднее значение этой разметки со стандартным отклонением.

Unfiltered faces for gender and age classification. Данные базы Adience содержат 26580 фотографий лиц 2284 человек, полученных с сайта Flickr. Эти фотографии не подвергались дальнейшей обработке, что позволяет считать такой набор данных подходящим для обучения моделей, предназначенных для работы в реальном времени. Возрасты в наборе данных разбиты на группы: 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, от 60 и выше.

Morph Album II. Morph является одним из популярных наборов данных для обучения определению возраста. Он содержит 55 134 фотографии лиц 13 617 человек. Для использования данных требуется разрешение создателей набора.

FG-NET. Набор содержит 1002 фотографии 82 человек, сделанные в разном возрасте (до 69 лет). Вариативность возраста одного человека даёт возможность использовать эти данные не только для его предсказания, но и для значительного улучшения обратных методов, позволяющих представить внешность человека через определённый промежуток времени.

Рис. 1. Поворот головы с помощью алгоритма PRNet

MegaAge и MegaAgeAsian. Наборы данных, содержащие 41941 и 40000 фотографий (из наборов данных MegaFace и YFCC100M), аннотированы распределениями по возрасту на основе разметки, выполненной людьми. Потенциально можно использовать распределения для получения метки возраста.

Cross-Age Reference Coding (CARC). Этот набор данных содержит 160000 фотографий 2000 знаменитостей в возрасте 16–62 лет, взятых из списков IMDB. Отличается от IMDB-Wiki методикой получения данных и количеством фотографий одного человека.

Обработка данных.

Так как разметка данных может изначально содержать ошибки, особенно для наборов данных с автоматически сгенерированной разметкой, это усложняет задачу оптимизации моделей и ухудшает их качество. Для улучшения качества разметки была проведена предварительная обработка некоторых из используемых наборов данных.

В автоматически размеченных наборах данных, где есть несколько фотографий одного человека, проведена фильтрация с помощью функции распознавания лиц, предоставленной библиотекой Dlib. Для каждого поднабора фотографий, принадлежащих одному человеку по разметке, с помощью нейронной сети выделялись характеристики, идентифицирующие его. С помощью кластеризации характеристик были установлены различные люди, из которых выбран наиболее вероятный владелец профиля по наибольшему количеству фотографий.

В указанных данных большинство лиц имеет фронтальное положение, т. е. смотрящих прямо на камеру. В общей постановке проблемы получить такие же данные — трудновыполнимая задача, поскольку человек должен позировать перед камерой. В связи с этим появилась необходимость в добавлении к данным повёрнутых лиц, для чего использовалась нейронная сеть PRNet [13], которая строит трёхмерную модель головы с помощью одного снимка. Вращая голову, можно получить снимки с различных ракурсов (рис. 1). Так как фон не применяется в нейронной сети, то не требуется дополнительной обработки повёрнутого лица.

Таблица 1

Структура MobileNetV2

Входные данные	Слой	t	c	n	s
$224 \times 224 \times 3$	conv2d	—	32	1	2
$112 \times 112 \times 32$	Свёрточный блок	1	16	1	1
$112 \times 112 \times 16$	Свёрточный блок	6	24	2	2
$56 \times 56 \times 24$	Свёрточный блок	6	32	3	2
$28 \times 28 \times 32$	Свёрточный блок	6	64	4	2
$14 \times 14 \times 64$	Свёрточный блок	6	96	3	1
$14 \times 14 \times 96$	Свёрточный блок	6	160	3	2
$7 \times 7 \times 160$	Свёрточный блок	6	320	1	1
$7 \times 7 \times 320$	conv2d 1×1	—	1280	1	1
$7 \times 7 \times 1280$	avgpool 7×7	—	—	1	—
$1 \times 1 \times 1280$	conv2d 1×1	—	k	—	—

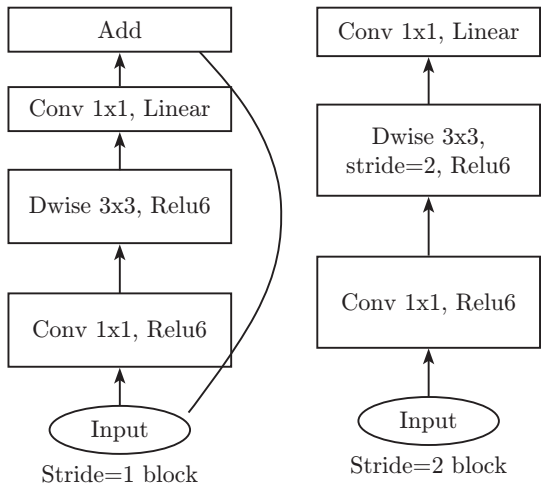

Рис. 2. Структура свёрточного блока [4]

Для увеличения количества данных были использованы различные преобразования, такие как поворот лица на случайный угол, горизонтальное отражение, случайное масштабирование. Это позволяет уменьшить вероятность переобучения моделей и даёт им бóльшую устойчивость.

Архитектура нейронной сети.

MobileNetV2. Первой рассмотренной архитектурой нейронной сети была MobileNetV2. Благодаря отсутствию полносвязных слоёв и использованию depthwise свёрточных слоёв эта архитектура обладает как большой обучаемостью, так и большой скоростью. Depthwise свёрточный слой (Dwise) отличается от обычного тем, что каналы входных данных не комбинируются, т. е. для каждого канала применяется свой фильтр.

В табл. 1 [4] описана структура MobileNetV2. Сначала картинка обрабатывается свёрточным слоем с шагом s (stride) 2. Затем результат обрабатывается серией свёрточных блоков (рис. 2). В зависимости от шага используется один из двух вариантов. Первый слой блока представляет собой набор из t фильтров размерности $1 \times 1 \times C$, где C — количество каналов во входных данных. Количество фильтров в последнем свёрточном слое равно c. Каждый блок повторяется n раз.

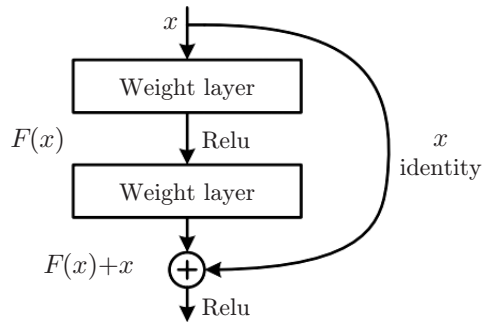

Рис. 3. Структура Residual-блока [3]

ResNet. Следующей рассмотренной архитектурой является ResNet50. Основу сети ResNet составляет Residual-блок (остаточный блок) с shortcut-соединением, через которое данные проходят без изменений. Res-блок представляет собой несколько свёрточных слоёв с активациями, которые преобразуют входной сигнал x в $F(x)$. Shortcut-соединение — это тождественное преобразование $x \to x$.

С помощью комбинации блоков создаётся полная архитектура сети. Число 50 в названии архитектуры указывает на то, что при её создании используется 50 блоков (рис. 3).

SSR-Net. Soft Stagewise Regression Network является архитектурой для оценки возраста по одному изображению с компактным размером модели. Для решения поставленной задачи выполняется мультиклассовая классификация с последующим преобразованием результатов классификации в регрессию и вычислением ожидаемых значений. Архитектура SSR-Net использует стратегию многоклассовой классификации в несколько этапов. Каждый этап отвечает только за уточнение решения предыдущего этапа для более точной оценки возраста. Таким образом, каждый этап выполняет задачу с несколькими классами и требует несколько нейронов, значительно уменьшая размер модели. Для решения проблемы квантования, связанной с группированием возрастов по классам, SSR-Net назначает динамический диапазон каждому возрастному классу, позволяя ему сдвигаться и масштабироваться в соответствии с входным изображением лица.

Результаты. Эксперименты проводились на языке Python. Для реализации нейронных сетей использовалась библиотека TensorFlow.

С помощью идентификации по профилю были проверены наборы данных IMDB + Wiki и CARC и исключены многие примеры. Кроме того, были исключены примеры, в которых исходные размеры лица на фотографии были слишком малы. В результате в IMDB осталось 197470 примеров, в Wiki — 35278, а в CARC — 143030.

Отфильтрованные наборы данных были разделены на три части: тренировочную, валидационную и тестовую в соотношении 80, 10 и 10 % соответственно. При разделении проверялось, чтобы лица одного профиля попали только в один поднабор. Так как распределение по возрастам в наборах данных не равномерно (рис. 4), то это может влиять на результат тестирования.

Сначала были сравнены функции потерь Softmax и Mean-Variance Loss. Для обучения на поднаборе данных была выбрана сеть MobileNetV2. Все сравнения проходили на тестовых поднаборах данных FG-NET, LAP, Adience, IMDB, MegaAge, MegaAgeAsian, Wiki. Как видно из табл. 2, Mean-Variance Loss показывает себя значительно лучше на наборах данных, где больше лиц моложе 20 лет. В остальных случаях она почти не уступает функции Softmax.

Затем сравнивались другие модели (MobileNetV2, ResNet, SSR-Net) и их результаты. Архитектуры MobileNetV2 и ResNet дают приблизительно одинаковые результаты

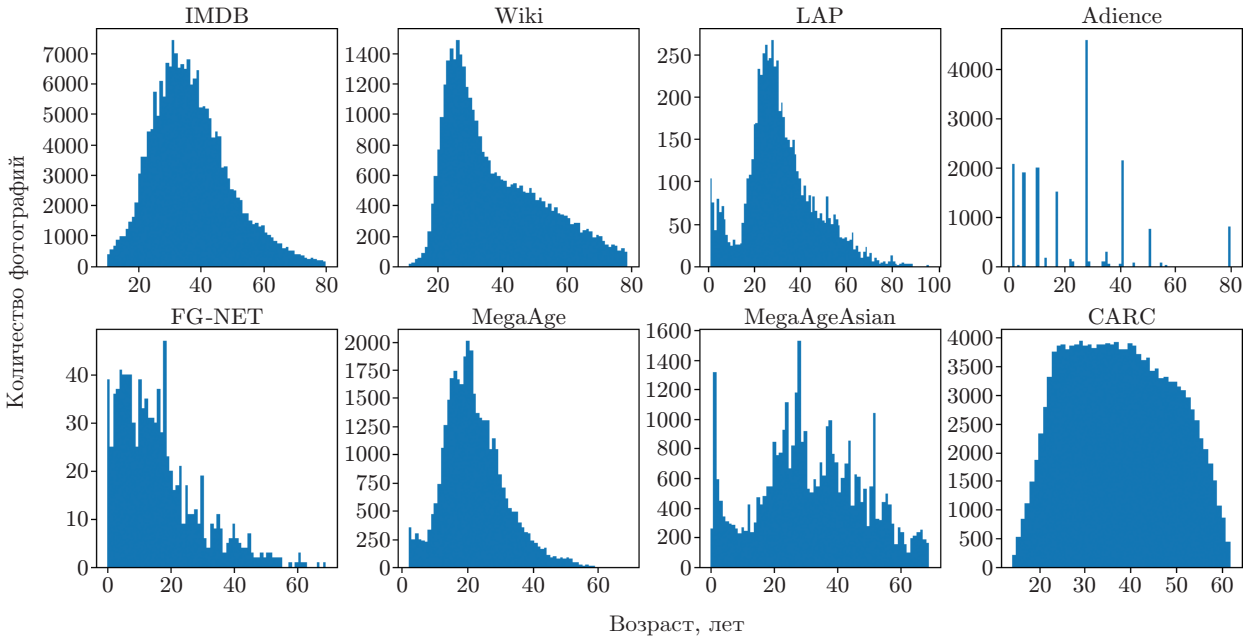

Рис. 4. Распределение изображений по метке возраста в различных наборах данных

Таблица 2

Сравнение метрик MAE тестовых подвыборок различных наборов данных для двух функций потерь

Набор данных	Softmax loss	Mean-Variance Loss
FG-NET	11,355	3,652
LAP	6,644	5,171
Adience	11,616	5,985
IMDB	7,269	6,364
MegaAge	7,350	6,845
MegaAgeAsian	8,019	4,887
Wiki	7,816	6,773

(табл. 3), однако время работы MobileNetV2 более чем в 3 раза меньше (MobileNetV2 — 2,5 мс, ResNet — 9,1 мс на видеокарте Titan X).

При повороте головы вверх и вниз модели выдавали наибольшую погрешность, поскольку данные с таким ракурсом практически отсутствовали при обучении, что является большим недостатком. Так как многие камеры находятся выше уровня лица человека, то часто для камеры лицо будет наклонено вниз, поэтому дополнительно были обучены модели, для которых добавлены повёрнутые вниз лица, полученные из набора данных CACD с помощью модели PRNet.

После обучения моделей результаты сравнивались с результатами до обучения. Как видно из табл. 4, при обучении на данных с поворотом головы MobileNet значительно увеличила ошибку, однако ResNet в общем случае улучшила результат даже на фронтальных лицах. Вероятнее всего, это связано с тем, что архитектура первой сети имеет значительно меньше коэффициентов для обучения, чем вторая, следовательно, у неё меньше возможностей выделения дополнительных особенностей изображения.

Результаты работы модели ResNet, обученной распознавать обычные и повёрнутые лица, можно увидеть на рис. 5.

Таблица 3

Сравнение метрик MAE на тестовых подвыборках различных наборов данных
(для разных архитектур нейронных сетей использовалась функция потерь Mean-Variance Loss)

Набор данных	MobileNetV2	ResNet	SSR-Net
FG-NET	3,652	3,025	8,1713
LAP	5,171	5,145	9,894
Adience	5,985	5,425	10,049
IMDB	6,364	6,627	9,353
MegaAge	6,845	6,523	8,702
MegaAgeAsian	4,887	5,337	4,842
Wiki	6,773	6,418	7,521

Таблица 4

Сравнение метрик MAE на тестовых подвыборках наборов данных
для моделей, дообученных на данных с поворотом головы

Набор данных	MobileNetV2	MobileNetV2 с поворотом	ResNet	ResNet с поворотом
FG-NET	3,652	5,191	3,025	3,689
LAP	5,171	5,298	5,145	5,188
Adience	5,985	6,543	5,425	4,645
IMDB	6,364	7,331	6,627	6,234
MegaAge	6,845	7,076	6,523	6,981
MegaAgeAsian	4,887	7,344	5,337	4,776
Wiki	6,773	7,574	6,418	6,319

Рис. 5. Возраст, предсказанный обычной и дообученной моделями: 15 и 22 (a), 35 и 18 (b), 45 и 47 (c) соответственно

Заключение. В представленной работе рассмотрены различные подходы к определению возраста человека на фотографии как к задаче машинного обучения. Были выбраны и протестированы нейронные сети с различной архитектурой, функциями потерь, а также проведена предварительная и последующая обработка наборов данных.

Обнаружено, что широко распространённые наборы данных с автоматически полученными отметками возраста содержат много нерелевантных данных. Предложена методика для фильтрации ошибочных данных и улучшения качества их разметки на основе группирования фотографий.

Была выявлена проблема с определением на фотографиях возраста лиц, имеющих даже небольшой вертикальный поворот. Возможно, это связано с полным отсутствием вертикально повёрнутых лиц в используемых наборах данных. В целях решения этой проблемы в наборы данных были добавлены фотографии с искусственно повёрнутыми лицами, что улучшило результат на тестовых изображениях.

Планируется провести работу по определению возраста человека на видеозаписи, что является более сложной задачей, так как видеофайлы имеют дополнительные артефакты записи. К тому же необходимы инструмент слежения за движением объекта [14], а также выделение среди кадров тех, которые лучше подходят для распознавания.

СПИСОК ЛИТЕРАТУРЫ

1. **Kwon Y., Lobo N.** Age classification from facial images // Comput. Vis. Image Understanding. 1999. **74**. P. 1–21.
2. **Krizhevsky A., Ilya S., Hinton G. E.** ImageNet classification with deep convolutional neural networks // Proc. of the Conf. on Neural Information Processing Systems. Stateline, USA, 3–8 Dec., 2012. Vol. 25. P. 1097–1105.
3. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition // Proc. of the Conf. on Computer Vision and Pattern Recognition. Las Vegas, USA, 26–30 June, 2016. P. 770–778.
4. **Sandler M., Howard A., Zhu M. et al.** MobileNetV2: Inverted residuals and linear bottlenecks // Comp. Vis. Pattern Recogn. Cornell Univers., 2018. URL: https://arxiv.org/abs/1801.04381 (дата обращения: 20.03.2019).
5. **Pan H., Han H., Shan S., Chen X.** Mean-variance loss for deep age estimation from a face // Proc. of the Conf. on Computer Vision and Pattern Recognition. Salt Lake City, USA, 18–22 June, 2018. P. 5285–5294.
6. **Zhang K., Liu N., Yuan X. et al.** Fine-grained age estimation in the wild with attention LSTM networks // Comp. Vis. Pattern Recogn. Cornell Univers., 2018. URL: https://arxiv.org/abs/1805.10445 (дата обращения: 20.03.2019).
7. **Yang T.-Y., Huang Y.-H., Lin Y.-Y. et al.** SSR-Net: A compact soft stagewise regression network for age estimation // Proc. of the IEEE Intern. Joint Conf. on Artificial Intelligence. Stockholm, Sweden, 13–19 July, 2018. P. 1078–1084.
8. **Shen W., Guo Y., Wang Y. et al.** Deep regression forests for age estimation // Proc. of the Conf. on Computer Vision and Pattern Recognition. Salt Lake City, USA, 18–22 June, 2018. P. 2304–2313.
9. **Pei W., Dibeklioglu H., Baltrusaitis T., Tax D. M. J.** Attended end-to-end architecture for age estimation from facial expression videos // IEEE Trans. Image Process. 2017. **26**, Is. 9. P. 6097–6146.
10. **Liu W., Wen Y., Yu Z. et al.** SphereFace: Deep hypersphere embedding for face recognition // Proc. of the Conf. on Computer Vision and Pattern Recognition. Honolulu, USA, 21–26 July, 2017. P. 6738–6746.

11. **Wang H., Wang Y., Zhou Z. et al.** CosFace: Large margin cosine loss for deep face recognition // Proc. of the Conf. on Computer Vision and Pattern Recognition. Salt Lake City, USA, 18–22 June, 2018. P. 5265–5274.
12. **Deng Z., Guo J., Xue N., Zafeiriou S.** Additive angular margin loss // Comp. Vis. Pattern Recogn. Cornell Univers., 2018. URL: https://arxiv.org/abs/1801.07698 (дата обращения: 20.03.2019).
13. **Feng Y., Wu F., Shao X. et al.** Joint 3D face reconstruction and dense alignment with position map regression network // Proc. of the Eur. Conf. on Computer Vision. Munich, Germany, 8–14 Sept., 2018. Vol. 14. P. 557–574.
14. **Золотухин Ю. Н., Котов К. Ю., Свитова А. М. и др.** Идентификация динамики подвижного объекта с помощью нейронных сетей // Автометрия. 2018. **54**, № 6. С. 107–113.

Поступила в редакцию 20.03.2019
После доработки 01.04.2019
Принята к публикации 02.04.2019