Сверточные нейронные сети в задаче распознавания пола и возраста по видеоизображению

А.С. Харчевникова[1], А.В. Савченко[1]

[1]Национальный исследовательский университет Высшая школа экономики, Большая Печерская 25/12, Нижний Новгород, Россия, 603155

Аннотация. Исследуется задача определения пола и возраста по видеоизображению лица с использованием глубоких сверточных нейронных сетей. Проведен сравнительный анализ существующих методов агрегации решений, полученных для отдельных кадров. В целях повышения точности идентификации пола и возраста разработана информационная система, в которой реализованы несколько алгоритмов построения коллективов решающих правил. Проведено экспериментальное исследование для баз видеоданных IJB-A, Indian Movies и Kinect. Показано, что наиболее точные решения для распознавания пола и идентификации возраста достигаются, соответственно, с помощью среднего геометрического и математического ожидания оценок апостериорных вероятностей, полученных со слоя softmax сверточных нейронных сетей.

1. Введение

В настоящее время в связи с бурным ростом интереса к обработке видео современные информационные технологии видеоидентификации по изображению лица ориентированы на выявление различных характеристик наблюдаемых объектов. Например, в коммерческих приложениях контекстной рекламы в целях выявления определенной целевой аудитории применяется автоматическое определение пола и возраста [1]. Несмотря на то, что за последние несколько лет появилось большое число разнообразных алгоритмов распознавания пола и возраста [2], надежность существующих решений остается недостаточной для практического применения [3].

В отличие от традиционных технологий идентификации по одному изображению, системы видеонаблюдения могут опираться на дополнительную информацию, связанную с тем, что в рамках продолжения видеосъемки можно получить более ста кадров классифицируемого объекта в динамике [1]. Для принятия решения достаточно, чтобы система гарантировала принадлежность хотя бы нескольких изображений к одному классу из базы эталонов [1]. Поэтому в данной статье для распознавания возраста и пола по видеоизображению используется выбор наиболее точного решения на основе синтеза коллективов решающих правил (КРП, комитетов классификаторов, алгоритмических композиций) [4, 5, 6]. В работе проводится сравнение КРП, полученных с помощью традиционного усреднения индивидуальных решающих правил [7], с КРП, построенными на основе принципа максимума апостериорной вероятности [8, 9, 10]. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области обработки и распознавания изображений.

2. Постановка задачи

Задача классификации видеоизображения лица состоит в следующем. Требуется отнести вновь поступающую на вход последовательность кадров $\{X(t)\},\ X(t) = \|x_{uv}(t)\|,\ t = \overline{1,T}$ с изображением одного объекта к одному из L классов [3]. Для упрощения мы предполагаем, что видео содержит только одного классифицируемого человека с предварительно выделенной областью лица на кадре, так на каждом изображении $\{X(t)\}$ оставлена только область лица. В таком виде задача распознавания пола является типичным примером бинарной классификации [8]. Несмотря на то, что распознавание возраста является примером задачи регрессии, на практике наибольшая точность достигается при ее сведении к задаче классификации с определением нескольких возрастных категорий (L = 8 в работе [11]).

Ранние методы оценки возраста основаны на вычислении соотношений между различными измерениями ключевых точек лица (нос, рот, глаза) [12]. Такие алгоритмы требуют точного вычисления расположения ключевых точек, что является самостоятельной достаточно сложной задачей, поэтому они обычно неприменимы для видеокадров. Авторы работы [13] предлагают решение для автоматического распознавания возраста - AGing pattErn Subspace (AGES), идея которого заключается в создании шаблона старения. Однако требования к фронтальному выравниванию входных данных накладывает существенные ограничения на набор входных параметров. Комбинация биологических особенностей лица на изображении исследуется в [14] (BIF - Biologically Inspired Features). Классификация возраста также возможна с помощью традиционных линейных фильтров Габора и машины опорных векторов (SVM) [15].

Распознавание пола является более простой задачей [16], так как она включает только L = 2 класса. Поэтому применимы традиционные бинарные классификаторы. Среди них широкое распространение получили такие методы, как SVM [17], алгоритмы, основанные на бустинге и нейронные сети.

К сожалению, точность традиционных методов компьютерного зрения и распознавания образов не удовлетворяет требованиям практического применения. В связи с эффективностью осуществления сверточных нейронных сетей (СНС) для разнообразных задач классификации изображений, авторы в [11] предлагают использовать данное решение в задачах определения возраста и пола по фотографии лица. В работе [18] была описана глубокая СНС VGG-16 [19], обученная для распознавания пола и возраста по фотографии, которая характеризуется наиболее высокой точностью по сравнению с известными аналогами. Именно основанный на СНС подход и будет использоваться в настоящей работе для обработки кадров видеопотока.

3. Предлагаемый алгоритм

На первом этапе вновь поступающий кадр с изображением лица необходимо отнести к одному из L классов. Для этого СНС на предварительном этапе обучается с использованием достаточно большого набора данных фотографий лиц с размеченными классами пола и возраста [11]. В процессе распознавания на вход СНС подадим RGB матрицу пикселей t-го кадра. Выход нейросетевой модели обычно получается в слое Softmax, который выдает оценку апостериорной вероятности $P\left(l \middle| X(t)\right)$ принадлежности t кадра к каждому l-му классу [20]:

$$P\left(l \middle| X(t)\right) = \operatorname{softmax} z_l(t) = \frac{\exp(z_l(t))}{\sum_{j=1}^{L} \exp(z_j(t))},\ l = 1,2,\ldots,L \qquad (1)$$

где $z_l(t)$ – выход l нейрона в последнем (чаще всего, полносвязном) слое нейронной сети. Решение для каждого кадра принимается в пользу класса с максимальной апостериорной вероятностью (1).

Вследствие влияния различных внешних факторов, как недостаток освещения, слабое разрешение видеокамеры, быстрая смена ракурса и др., принятие решения в пользу класса с максимальной апостериорной вероятностью для каждого кадра обычно отличается низкой

точностью. Поэтому в настоящей работе предлагается свести задачу распознавания видеоизображений к построению коллективов решающих правил (КРП). Обзор и анализ публикаций в области обработки данных показывает, что синтез КРП является одним из наиболее эффективных подходов к увеличению точности и устойчивости классификации [21, 22, 23]. В КРП для принятия решения о классификации изображения используется не один, а несколько критериев, каждый из которых самостоятельно присваивает метку класса, после чего на основе некоторого принципа [6] формируется общий результат классификации. В задаче распознавания видеоизображений для каждого поступающего кадра $X(t)$ решается обычная задача автоматического распознавания изображений с помощью СНС, а затем все индивидуальные решения комбинируются в одно общее решение для конкретной видеозаписи. Наиболее очевидный подход состоит здесь в использовании более сложных алгоритмов построения КРП, основанных на алгебраическом подходе [6, 22]. Б*о*льшая часть таких алгоритмов (такие как комитет взвешенного большинства, бэггинг и бустинг [8, 24]) требуют достаточной представительной обучающей выборки. К сожалению, во многих задачах распознавания изображений имеющаяся база данных содержит недостаточное число эталонов для каждого класса. В настоящей работе предлагается воспользоваться известными статистическими способами синтеза КРП [6], когда тем или иным способом для каждого члена комитета определяются вероятности принадлежности входного объекта к классам. Для агрегации решений в настоящей работе применяются следующие критерии [25, 9]:

1. *Простое голосование,* в котором окончательное решение принимается в пользу класса [4, 9]:

$$l^* = \underset{l=\overline{1,L}}{\operatorname{argmax}} \sum_{t=1}^{T} \delta(l^*(t) - l) \tag{2}$$

2. *Среднее арифметическое* апостериорной вероятности (1) или правило суммы [4]:

$$l^* = \underset{l=\overline{1,L}}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} P\left(l \middle| X(t)\right) \tag{3}$$

3. Для «наивного» предположения о независимости всех кадров [25] решение принимается по правилу произведения [4] – критерию максимума *среднего геометрического* апостериорной вероятности:

$$l^* = \underset{l=\overline{1,L}}{\operatorname{argmax}} \prod_{t=1}^{T} P\left(l \middle| X(t)\right) = \underset{l=\overline{1,L}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P\left(l \middle| X(t)\right) \tag{4}$$

Кроме того, так как распознавание возраста существенным образом относится к задаче регрессии, для нее возможно вычислять *оценку математического ожидания:*

$$l^* = \sum_{l=1}^{L} P\left(l \middle| X(t)\right) \cdot l \tag{5}$$

Общая схема предлагаемой системы распознавания пола и возраста по видеоизображению представлена на Рисунке 1.

На первом шаге в поступающем на вход видеопотоке с фиксированной частотой (10-20 раз в секунду) извлекаются отдельные кадры и происходит их предварительная обработка (нормировка, эквализация гистограмм). Далее в каждом кадре с помощью каскадного классификатора Виолы-Джонса и признаков Хаара [27] средствами библиотеки OpenCV обнаруживаются области лица. Для ускорения работы могут применяться известные процедуры слежения за лицом, выделенном на предыдущих кадрах [26]. На следующем этапе все полученные изображения лиц на одном кадре приводятся к единому масштабу. Кроме того, достаточно часто [11] применяется вычитание среднего изображения (Image mean subtraction) из каждого изображения лица. Наконец, все преобразованные изображения классифицируются с помощью СНС. Результатом работы блока являются оценки апостериорной вероятности со слоя softmax (1). На основе данных, которые определяются выходом блока КРП типа (2)-(4) осуществляется конечное распознавание в пользу соответствующего класса.

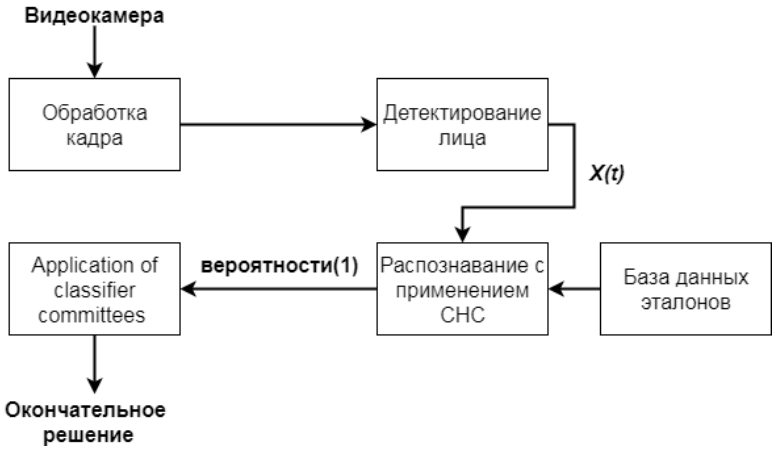

Рисунок 1. Схема предлагаемой системы распознавания.

4. Результаты экспериментов

Реализация описанного подхода (**Рисунок 1**) осуществлялась в MS Visual Studio 2015 средствами C++. Был применен функционал модуля DNN библиотеки OpenCV, а также фреймворка Caffe [28] для сравнения производительности и точности распознавания. Графический интерфейс предлагаемой системы представлен на **Рисунке 2.**

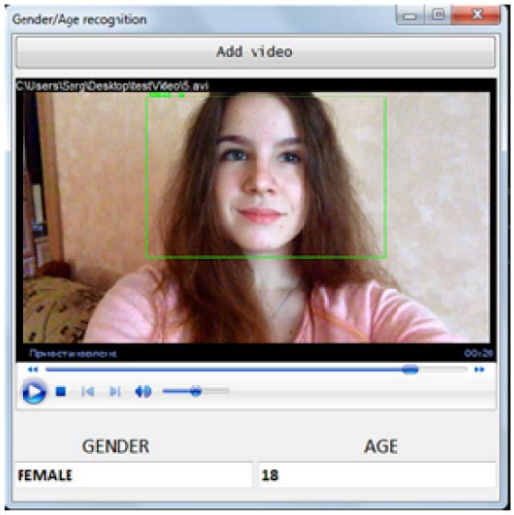

Рисунок 2. Пример экранной формы разработанного приложения.

Тестирование точности распознавания и реализации алгоритмов агрегации производится с помощью наборов данных IARPA Janus Benchmark A (IJB-A), Indian Movie и EURECOM Kinect, которые состоят из видео кадров с известной информацией пола и возраста. Первый датасет состоит из 2043 видео. Здесь присутствует информация только о поле человека на кадре [29]. База данных Indian Movie – это коллекция видео фреймов, собранных из индийских фильмов. Всего в базе около 332 различных видео и 34512 фреймов ста индийских актеров, возраст которых разбит на четыре категории: «Child», «Young», «Middle» и «Old» [30]. В данном примере словесное описание возраста было заменено на конкретные возрастные промежутки согласно примерной оценке людей на кадрах: 0-15, 16-35, 36-60, 60+. Далее оцениваются пересечения результатов распознавания с заданными интервалами. Датасет Kinect содержит в себе 104 видео с участием 52 людей (14 женщин и 38 мужчин) [31]. В базе представлена информация о поле человека на кадре, его год рождения, что упрощает оценку возраста. При реализации алгоритма возраст рассматривается в диапазоне с прибавлением и вычетом 5 лет, так как необходимо выявить точность пересечения с распознанным возрастным интервалом (СНС [11]).

В настоящем разделе сравниваются две находящиеся в открытом доступе СНС: модели Age_net и Gender_net [11] и глубокую VGG-16 [19], обученную для распознавания пола и возраста [18]. В связи с тем, что при видеосъемке значительное влияние на точность результатов оказывают внешние факторы, необходима нормализация входных данных. Для этого в СНС можно добавить дополнительный первый слой Mean-Variance Normalization (MVN), который позволяет нормализовать интенсивность значений пикселей изображения.

Среднее время работы алгоритмов на машине Intel Core i5-2400 CPU, 64-bit с NVIDIA GeForce GT 440 представлено в **Таблице 1.** Наилучшие результаты выделены полужирным шрифтом.

Таблица 1. Среднее время распознавания (сек.).

	Gender/Age net	VGG-16
OpenCV DNN	4.805	28.981
OpenCV DNN + MVN layer	8.734	34.984
Caffe + mean image subtraction	**0.867**	**3.395**
Caffe + mean image subtraction + MVN layer	1.064	9.012

Согласно полученным результатам, распознавание с помощью библиотеки Caffe оказывается в несколько раз быстрее по сравнению с реализацией модуля OpenCV DNN.

В Таблице 2 и Таблице 3 продемонстрированы результаты традиционного распознавания пола и возраста по одному изображению каждого кадра (без применения КРП).

Таблица 2. Точность распознавания пола по каждому кадру (%).

IJB-a		Indian Movie		Kinect	
Gender_net	VGG-16	Gender_net	VGG-16	Gender_net	VGG-16
42	55	61	68	51	62

Таблица 3. Точность распознавания возраста по каждому кадру (%).

		Kinect	
Indian Movie			
Age_net	VGG-16	Age_net	VGG-16
10	32	25	58

Наиболее высокая точность распознавания пола и возраста достигнута с помощью фреймворка Caffe и нормализации входных данных методом вычисления среднего изображения. К примеру, разница в вероятности ошибке в сравнении с OpenCV DNN модулем составила примерно 10% для возраста и 20% для пола. Результаты лучшей реализации СНС с применением КРП для пола и возраста представлены в **Таблицах 4** и **5**, соответственно.

Таблица 4. Точность распознавания пола с применением КРП (%).

Алгоритм	IJB-a		Indian Movie		Kinect	
	Gender_net	VGG-16	Gender_net	VGG-16	Gender_net	VGG-16
Простое голосование (2)	**60**	81	71	81	73	83
Правило суммы (3)	59	81	72	87	75	84
Правило произведения (4)	59	**82**	**75**	**88**	**77**	**84**

Таблица 5. Точность распознавания возраста с применением КРП (%).

Алгоритм	Indian Movie		Kinect	
	Age_net	VGG-16	Age_net	VGG-16
Простое голосование (2)	23	62	41	73
Правило суммы (3)	23	63	43	79
Правило произведения (4)	23	65	45	79
Оценка математического ожидания (5)	**31**	**65**	**50**	**79**

Таким образом, при нормализации входных данных, наибольшую эффективность показал метод вычитания среднего изображения. В заключении, вычисление среднего геометрического (4) оказалось в большинстве случаев несколько точнее по сравнению с остальными алгоритмами агрегации. Оценка математического ожидания (5) показала эффективность в определении возраста для СНС Age_net. Для архитектуры VGG-16 лучшие результаты также достигнуты с помощью Caffe и вычитанием среднего изображения. Следует также заметить, что архитектура VGG-16 опережает модели Gender_net и Age_net по точности распознавания. К сожалению, среднее время (**Таблица 1**) работы глубокой VGG-16 намного выше по сравнению с СНС моделями из [11]. Исходя из этого, получаем компромисс между точностью и производительностью.

5. Заключение

В работе исследованы простые способы построения КРП для задач идентификации пола и возраста по видеоизображению лица. Проведенное экспериментальное исследование продемонстрировало значительное увеличение точности распознавания при реализации КРП в сравнении с традиционным подходом принятия решения для единичного кадра. Представлен сравнительный результат реализации двух СНС архитектур: моделей Age_net и Gender_net [11] и VGG-16 [19]. Оценка среднего геометрического (правило произведения (4)) с нормализацией входных данных оказался более точным в задаче классификации по полу. В то же время наиболее точный результат распознавания возраста достигается с применением алгоритма оценки математического ожидания (5). Точность архитектуры VGG-16 оказалась на 15% и 20% выше в задачах распознавания пола (**Таблицы 2, 4**) и возраста (**Таблицы 3, 5**), соответственно. Однако среднее время обработки одного изображения для VGG-16 в 4-9 раз выше (**Таблица 1**) вследствие большей глубины. Поэтому одной из задач будущих исследований является применение современных алгоритмов для быстрой классификации [1, 32] и оптимизации глубоких СНС [33].

6. Благодарности

Статья подготовлена в результате проведения исследования (№ 17-05-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2017 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

7. Литература

[1] Savchenko, A.V. Search techniques in intelligent classification systems / A.V. Savchenko. – Springer International Publishing, 2016.

[2] Chao, W.L. Facial age estimation based on label-sensitive learning and age-oriented regression / W.L. Chao, J. Z. Liu, J.J. Ding // Pattern Recognition. – 2013. – Vol. 46(3). – P. 628-641.

[3] Wang, H. Video-based face recognition: A survey / H. Wang, Y. Wang, Y. Cao // World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering. – 2009. – Vol. 3(12). – P. 2809-2818.

[4] Kittler, J. Sum versus vote fusion in multiple classifier systems / J. Kittler, F.M. Alkoot // IEEE transactions on pattern analysis and machine intelligence. – 2003. – Vol. 25(1). – P. 110-115.

[5] Tresp, V. Committee machines / V. Tresp // Handbook for Neural Network Signal Processing. – 2001. – P. 135-151.

[6] Рудаков, К.В. О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания / К.В. Рудаков, К.В. Воронцов // Доклады РАН. – 1999. – Т. 367, № 3. – С. 314-317.

[7] Мазуров, В.Д. Метод комитетов в задачах оптимизации и классификации / В.Д. Мазуров. – М.: Наука, 1990. – 248 с.

[8] Theodoridis, S. Pattern Recognition / S. Theodoridis, C. Koutroumbas. – Elsevier Inc (4th Edition), 2009. – 840 p.
[9] Савченко, А.В. Выбор параметров алгоритма распознавания изображений на основе коллектива решающих правил и принципа максимума апостериорной вероятности / А.В. Савченко // Компьютерная оптика.– 2012. – Т. 36, № 1. – С.117-124.
[10] Savchenko, A.V. Adaptive Video Image Recognition System Using a Committee Machine / A.V. Savchenko // Optical Memory and Neural Networks (Information Optics). – 2012. – Vol. 21. – P. 219-226.
[11] Levi, G. Age and gender classification using convolutional neural networks / G. Levi, T. Hassner //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2015. – P. 34-42.
[12] Kwon, Y.H. Age classification from facial images / Y.H. Kwon // Computer Vision and Pattern Recognition. Proceedings CVPR'94. IEEE Computer Society Conference, 1994. – P. 762-767.
[13] Geng, X. Learning from facial aging patterns for automatic age estimation / X. Geng // Proceedings of the 14th ACM international conference on Multimedia. – ACM, 2006. – P. 307-316.
[14] Guo, G. Human age estimation using bio-inspired features / G. Guo, G. Mu, Y. Fu // Computer Vision and Pattern Recognition. CVPR 2009. IEEE Conference, 2009. – P. 112-119.
[15] Choi, S.E. Age estimation using a hierarchical classifier based on global and local facial features // Pattern Recognition. – 2011. – Vol. 44(6). – P. 1262-1281.
[16] Makinen, E. Evaluation of gender classification methods with automatically detected and aligned faces / E. Makinen, R. Raisamo // Trans. Pattern Anal. Mach. Intell. – 2008. – P. 541-547.
[17] Shan, C. Learning local binary patterns for gender classification on real-world face images // Pattern Recognition Letters. - 2012. – Vol. 33(4). P. 431-437.
[18] Rothe, R. Deep expectation of apparent age from a single image / R. Rothe, R. Timofte, L.D. Van Gool // Proceedings of the IEEE International Conference on Computer Vision Workshops. – 2015. – P. 10-15.
[19] Simonyan, K. Very deep convolutional networks for large-scale image recognition / K. Simonyan, A. Zisserman, 2014.
[20] Szegedy, C. Going deeper with convolutions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 1-9.
[21] Krizhevsky, A. Imagenet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. E. Hinton //Advances in neural information processing systems. – 2012. – P. 1097-1105.
[22] Журавлев, Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации / Ю.И. Журавлев // Проблемы кибернетики. – 1978. – Т. 33. – С. 5-68.
[23] Савченко, А.В. Об одном способе повышения надежности решения в задаче распознавания образов / А.В. Савченко // Заводская лаборатория. Диагностика материалов. – 2013. – Т. 79, №1. – С.72-77.
[24] Esmaeili, M. Creating of Multiple Classifier Systems by Fuzzy Decision Making in Human-Computer Interface Systems / M. Esmaeili, M. Rahmati // Conference IEEE Fuzzy Systems. – 2007. – P. 1-7.
[25] Savchenko, A.V. Adaptive Video Image Recognition System Using a Committee Machine / A.V. Savchenko // Optical Memory and Neural Networks (Information Optics). – 2012. – Vol. 21(4). – P. 219-226.
[26] Shan, C. Face recognition and retrieval in video / C. Shan // Video Search and Mining. – Springer Berlin Heidelberg, 2010. – P. 235-260.
[27] Lienhart, R. An extended set of Haar-like features for rapid object detection / R. Lienhart, J. Maydt // Proceedings of the IEEE conference on Image Processing, 2002. – Vol. 1. – P. I.
[28] Jia, Y. Caffe: Convolutional architecture for fast feature embedding / Y. Jia // Proceedings of the 22nd ACM International Conference on Multimedia. – 2014 – P. 675-678.

[29] IJB-a dataset. [Электронный ресурс]. Режим доступа: https://www.nist.gov/itl/iad/image-group/ijba-dataset-request-form.
[30] IMFDB. [Электронный ресурс]. Режим доступа: http://cvit.iiit.ac.in/projects/IMFDB/.
[31] Eurecom Kinect. [Электронный ресурс]. Режим доступа: http://rgb-d.eurecom.fr/.
[32] Savchenko, A.V. Maximum-Likelihood approximate nearest neighbor Method in real-time image recognition / A.V. Savchenko // Pattern Recognition. – 2017. – Vol. 61. – P. 459-469.
[33] Rassadin, A.G. Compressing deep convolutional neural networks in visual emotion recognition / A.G. Rassadin, A.V. Savchenko // Proceedings of the International conference Information Technology and Nanotechnology (ITNT). Session Image Processing, Geoinformation Technology and Information Security Image Processing (IPGTIS), CEUR-WS. – 2017. – Vol. 1901. – P. 207-213. URL: http://ceur-ws.org/Vol-1901/paper33.pdf.

Convolutional Neural Networks in Age and Gender Video-based Recognition

A.S. Kharchevnikova[1], A.V. Savchenko[1]

[1]National Research University Higher School of Economics, Bolshaya Pecherskaya 25/12, Nizhny Novgorod, Russia, 603155

Abstract. In this paper we examine the age and gender video-based recognition problem using deep convolutional neural networks. The comparative analysis of classifier fusion algorithms to aggregate decisions for individual frames is presented. In order to improve the age and gender identification accuracy we implement the video-based recognition system with several aggregation methods. We provide the experimental comparison for IJB-A, Indian Movies and Kinect datasets. It is demonstrated that the most accurate decisions are obtained using the geometric mean and mathematical expectation of the outputs at softmax layers of the convolutional neural networks for gender recognition and age prediction, respectively.

Keywords: Deep learning, gender recognition, age recognition, convolutional neural networks, classifier fusion.