УДК 004.932

DOI: 10.17587/it.25.538-544

И. С. Гречихин, аспирант, ст. преп., e-mail: igrechikhin@hse.ru,
А. В. Савченко, д-р техн. наук, проф., e-mail: avsavchenko@hse.ru,
Национальный исследовательский университет Высшая школа экономики, Нижний Новгород

Метод анализа предпочтений пользователя по фото- и видеоизображениям на мобильном устройстве на основе нейросетевых детекторов объектов на изображениях[1]

Предложен метод извлечения предпочтений пользователей в результате анализа галереи их мобильных устройств. На первом этапе выделяются публичные фото- и видеоизображения, не содержащие лиц из предварительно выделенных кластеров. На втором этапе такие изображения обрабатываются на сервере с помощью высокоточных детекторов объектов. Объекты на остальных (персональных) фото- и видеоизбражениях детектируются непосредственно на устройстве. Представлены экспериментальные результаты сравнительного анализа нескольких предварительно обученных нейросетевых детекторов.

Ключевые слова: *обработка изображений, детектирование объектов, мобильные системы, анализ предпочтений пользователя, кластеризация лиц*

Введение

В настоящее время в связи с одновременным развитием социальных сетей и мобильных устройств [1] наблюдается взрывной рост объема мультимедийных данных, которые создаются пользователями мобильных платформ. При этом такие данные нередко содержат уникальную информацию о пользователе, которая может использоваться, например, для повышения полезности разнообразных рекомендательных систем. Как известно, в последнее время для обработки изображений большинство исследователей и практиков применяют методы, основанные на технологиях глубокого обучения [2]. В контексте задачи анализа предпочтений пользователя по его фотографиям и видеоизображениям наибольший интерес представляют алгоритмы детектирования объектов на изображениях (предметы интерьера, виды еды, транспорт, спортивные принадлежности, музыкальные инструменты и т.п.) [3].

Стоит отметить, что, поскольку указанные пользовательские данные могут содержать персональную информацию, не всегда приемлемой является их передача на удаленный сервер для анализа с помощью современных высокоточных методов [2]. В связи с этим в настоящее время наблюдается заметная тенденция к разработке эффективных архитектур сверточных нейронных сетей (СНС) [3, 4]. В частности, для детектирования объектов на изображениях могут использоваться известные нейросетевые алгоритмы, обеспечивающие баланс между точностью и вычислительной эффективностью [5], такие как SSDLite [6], Faster R-CNN [4], YOLO [7, 8], в которых в качестве базовой СНС используются различные модификации MobileNet [4, 9] и т.п.

К сожалению, точность таких детекторов обычно оказывается намного ниже точности наилучших методов, использующих нейросетевые архитектуры с очень большим числом слоев, такие как ResNet или InceptionResNet [10]. Кроме того, заметим, что далеко не все фотографии и видеоизображения пользователя содержат персональные данные. Например,

[1]Статья подготовлена в ходе проведения исследования (№ 19-04-004) в рамках Программы "Научный фонд Национального исследовательского университета "Высшая школа экономики" (НИУ ВШЭ)" в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

обработка на удаленном сервере вполне приемлема для панорамных снимков достопримечательностей, еды в ресторанах, интерьеров музеев, театров, спортивных сооружений и т.п. Вместе с тем, именно такие изображения содержат наиболее важную информацию о предпочтениях пользователя. Поэтому в настоящей статье предлагается автоматически находить публичные фото- и видеоизображения для их последующей обработки на удаленном сервере с помощью высокоточных детекторов объектов. Предполагается, что персональными являются данные, содержащие лица самого пользователя, его близких друзей и знакомых, выделенных автоматически с помощью известных методов распознавания [11, 12] и кластеризации лиц [13, 14]. При этом объекты во всех остальных данных в галерее можно детектировать с помощью более простых методов непосредственно на мобильном устройстве пользователя. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области распознавания образов.

1. Анализ предпочтений по изображениям и видеоданным на основе нейросетевых детекторов

Задача анализа предпочтений по фотографиям и видеоданным состоит в том, чтобы по поступившему на вход фотоальбому — множеству фотографий и видеоизображений — выделить наиболее интересные для пользователя категории (виды еды, спортивное оборудование, музыкальные инструменты и т.п.). Предполагается, что для обучения системы для каждой категории задано множество изображений, соответствующих категории объектов, и данные об их местонахождении на изображении (обрамляющие прямоугольники или маска границ). В таком случае результатом анализа предпочтений можно считать частоты встречаемости объектов каждой категории на пользовательских фотографиях и видеоизображениях.

Для детектирования объектов на изображениях и определения их категорий могут использоваться известные высокоточные детекторы, основанные на СНС. В работах [4, 5] предложена архитектура SSDLite и СНС MobileNet v2, которая специально спроектирована для ускорения работы нейронной сети и поэтому удобна для использования в мобильных устройствах. СНС MobileNet извлекает карты признаков входного изображения, используя специальные "разделяемые по глубине" (depthwise-separable) сверточные слои, которые имеют значительно меньшее число параметров и бóльшую скорость обработки данных по сравнению со стандартными сверточными слоями без существенной потери качества. Детектор SSD (Single Shot Detector) использует карту признаков на выходе СНС для предсказания классов и положения объектов за один проход, а его модификация (SSDLite) включает разделяемые по глубине сверточные слои для снижения вычислительной сложности и затрат памяти детектора. В совокупности такая архитектура обнаруживает объекты значительно быстрее, но за счет некоторого уменьшения точности предсказаний.

Faster R-CNN-архитектуры [7] также используют СНС (backbone) для создания карты признаков, но с их помощью определяются несколько (100...200) регионов, в которых могут содержаться потенциально интересные объекты. После этого на основании карты признаков и выделенных регионов предсказывается класс объекта. В качестве СНС в детекторе хорошо зарекомендовали себя архитектуры Inception или InceptionResNet [10], которые считаются одними из самых точных для детектирования объектов на изображениях и их классификации, однако требуют значительных вычислительных ресурсов. Первая СНС (Inception) использует специальные блоки, состоящие из факторизованных, работающих параллельно сверток разного размера, результаты которых соединяются в один слой. СНС состоит из нескольких таких идущих подряд Inception-блоков. InceptionResNet создает более глубокую (и, как следствие, более точную) сеть с помощью добавления к Inception-блокам остаточных (residual) связей.

Таким образом, вычислительная эффективность и сложность по затратам памяти наиболее точных детекторов является недостаточной для их реализации даже на современных мобильных устройствах. При этом использование удаленного сервера для обработки *всех* мультимедийных данных пользователя может оказаться неприемлемым с точки зрения сохранности персональных данных.

2. Предложенный подход

В данной статье предлагается автоматически определять потенциальные публичные изображения на основе известных методов распознавания лиц. Так как в галерее фото- и видео-

файлов отсутствуют идентификаторы (метки) запечатленных на них людей, задача сводится к кластеризации (обучению без учителя). Для ее решения вначале необходимо детектировать лица, например, с помощью описанных в предыдущем разделе методов. Задача группировки состоит в том, чтобы каждому r-му изображению поставить в соответствие одну из $K \geqslant 1$ меток, где общее число различных людей K в общем случае неизвестно. Здесь $r = 1, ..., R$ — номер изображения, а R — общее число обнаруженных в альбоме лиц. Для каждого r-го доступного изображения осуществляется извлечение вектора признаков $\mathbf{x}_r$. В наиболее часто используемых сейчас методах переноса знаний (transfer learning) [2, 15] для предварительного обучения характерных признаков используется внешняя база данных изображений лиц с известными метками классов, с помощью которой происходит обучение глубокой СНС. Далее все изображения лиц приводятся к одному размеру (высота U и ширина V) и подаются на вход СНС [16]. Выходы из $D \gg 1$ значений одного предпоследнего слоя нейронной сети нормируются (в метрике L_2) и формируют вектор признаков $\mathbf{x}_r$ r-го изображения [12]. Для полученных векторов могут использоваться традиционные методы кластеризации, не требующие знания числа кластеров, например иерархическая агломеративная кластеризация [17].

На рис. 1 представлена функциональная схема предлагаемой информационной системы анализа предпочтений пользователей мобильных устройств. На предварительном этапе осуществляется обучение детектированию объектов заданных категорий двух нейросетевых моделей: одной — вычислительно эффективной — для реализации непосредственно на мобильном устройстве и другой — высокоточной — для обработки на удаленном сервере. При этом к обучающему множеству добавляется набор фотографий лиц для их детектирования в дополнение к требуемому списку категорий, характеризующих интересы пользователя.

На первом этапе для обнаружения объектов на всех фотографиях и видеоизображениях непосредственно на мобильном устройстве используют первый детектор, который в дополнение к списку интересов выделяет все лица. Далее с помощью специальной вычислительно эффективной СНС для каждого лица извлекается вектор его характерных признаков и выполняется кластеризация векторов признаков всех лиц. Вначале такая процедура проводится для каждого видеофайла, и векторы признаков центров кластеров лиц, выделенных на каждом видеоизображении, добавляются в общее множество векторов признаков лиц, выделенных на фотографиях, после чего осуществляется итоговая кластеризация и выделяются кластеры с достаточно большим числом лиц. Предполагается, что такие кластеры соответствуют самому пользователю и его друзьям и родственникам, поэтому все содержащие их фото- и видеоизображения объявляются содержащими персональную информацию. Среди остальных данных в галерее пользователь может дополнительно указать их приватность.

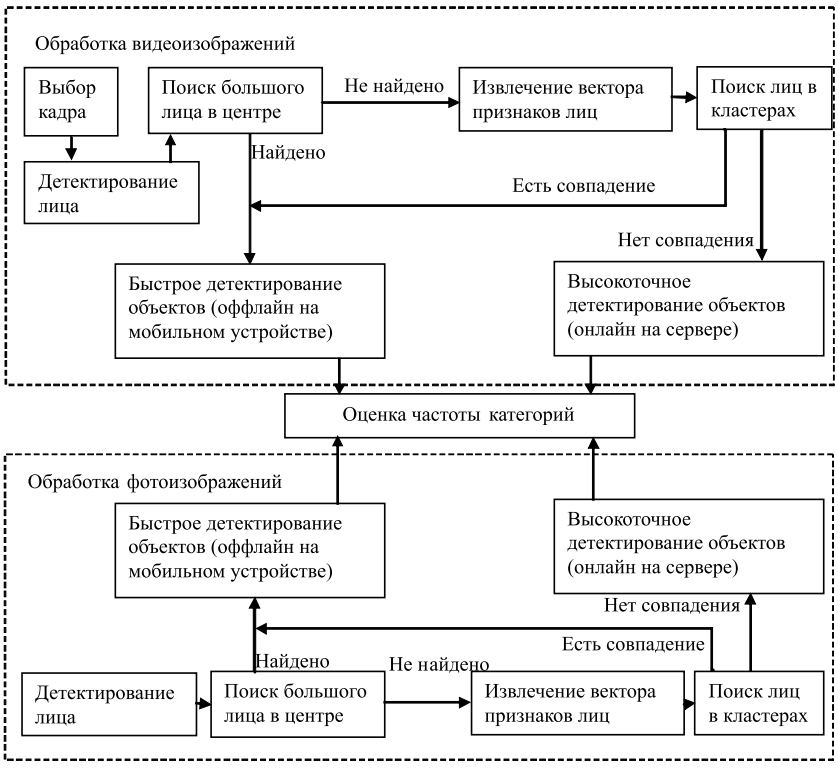

Рис. 1. Схема устройства для анализа предпочтений пользователя

Далее на втором этапе для обработки публичных изображений используется высокоточный нейросетевой детектор, которыей может быть реализован на удаленном сервере. Список обнаруженных объектов возвращается на мобильное устройство и объединяется с результатами первого (эффективного) детектора для подсчета частоты встречаемости каждой категории. Пример экранной формы отображения результатов анализа предпочтений в разработанном Android-приложении приведен на рис. 2.

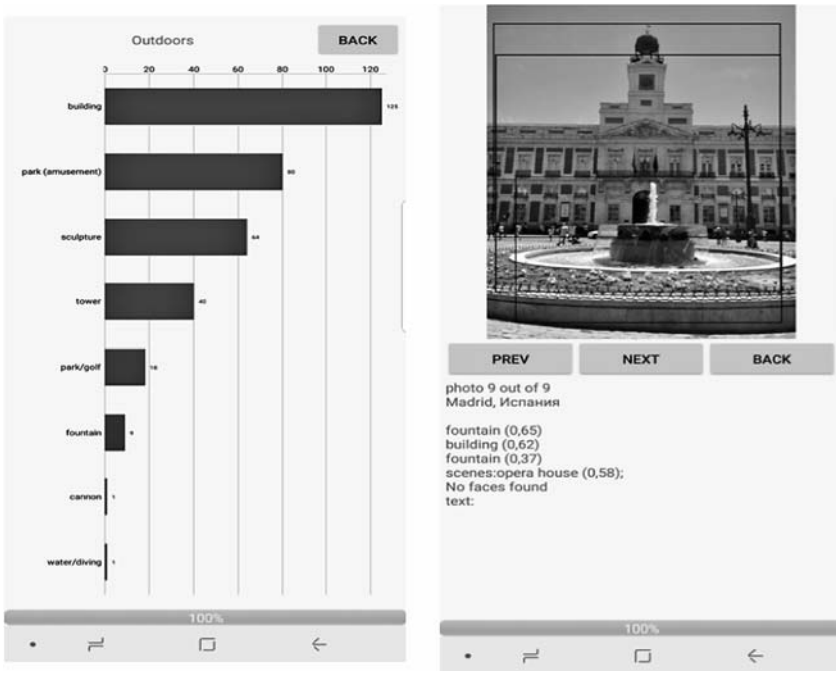

Рис. 2. Экранная форма приложения, реализующего предложенный подход

Таблица 1

Оценки эффективности детекторов категорий

Детектор	Архитектура СНС (backbone)	Размер модели, Мбайт	Время детектирования, с	
			Ноутбук	Смартфон
SSDLite	MobileNet v2, 300 × 300	31,83	0,16	0,30
	MobileNet v2, 512 × 512	31,83	0,21	0,52
Faster R-CNN	Inception v2	64,91	0,4	1,25
	InceptionResNet v2	204,34	1,01	2,39

Таблица 2

Оценки точности и полноты детекторов категорий

Детектор	Архитектура СНС (backbone)	Полнота	mAP	Полнота (родственные категории)	mAP (родственные категории)
Faster R-CNN	InceptionResNet v2	0,425	0,477	0,448	0,534
	InceptionResNet v2 (квантованная)	0,425	0,471	0,448	0,528
	Inception v3	0,393	0,537	0,414	0,593
	ResNet-50	0,332	0,583	0,35	0,636
	ResNet-101	0,465	0,562	0,485	0,618
SSDLite	MobileNet v2, 512 × 512	0,149	0,465	0,166	0,525
	MobileNet v2, 512 × 512 (квантованная)	0,149	0,463	0,166	0,524

3. Вычислительный эксперимент

Для проведения экспериментов и получения детекторов с необходимой точностью создана обучающая выборка из 146 классов (145 категорий интересов пользователя и 1 класс для детектирования лиц) из наборов данных MS COCO, Open Image Dataset и ImageNet. Обучающая выборка была сбалансирована: для каждой категории использовалось не более 5000 изображений. Для обучения применяли библиотеку TensorFlow.

В экспериментах использовали следующие архитектуры нейросетевых детекторов: SSDLite + MobileNet, Faster R-CNN + Inception v2, Faster R-CNN + InceptionResNet v2. Для модели SSDLite исследовались два варианта, соответствующие входным изображениям размера 300 × 300 и 512 × 512 пикселей. Для детекторов Faster R-CNN все изображения масштабировались так, чтобы размер наименьшей стороны был равен 600. Размеры моделей и среднее время предсказаний на ноутбуке (четырехъядерный процессор 4x2.2 ГГц, 16 ГБ ОЗУ) и смартфоне (восьмиядерный процессор: 2x2.2 ГГц, 6x1.6 ГГц, 4 ГБ ОЗУ) представлены в табл. 1.

Как можно увидеть, детекторы SSDLite быстрее и менее затратны по памяти, чем методы на основе Faster R-CNN. Кроме того, модели Faster R-CNN плохо подходят для использования в режиме "оффлайн" на смартфонах из-за времени, затраченного на детектирование.

С использованием тестового набора из других 5000 изображений каждой из 146 категорий оценены показатели полноты (recall, доля верно определенных объектов класса) и точности mAP (mean average precision, доля верных предсказаний). В дополнение составлен список "родственных" категорий, т. е. таких категорий A и B, что детектирование категории A для объекта из категории B нельзя назвать ошибочным (например, категория

Таблица 3

Оценки точности для надежных категорий (полнота более 0,75)

Детектор	Архитектура СНС (backbone)	Число категорий	mAP (родственные категории)
Faster R-CNN	InceptionResNet v2	78	0,662
	InceptionResNet v2 (квантованная)	79	0,663
	Inception v3	44	0,762
	ResNet-50	30	0,838
	ResNet-101	67	0,76
SSDLite	MobileNet v2, 512 × 512	3	0,773
	MobileNet v2, 512 × 512 (квантованная)	3	0,768

"животное" не является ошибкой для объекта "кошка" или "собака", аналогично "строение" — для "небоскреб" или "дом"). Метрики recall и mAP были посчитаны как для исходных категорий, так и с учетом родственных, результаты усреднены по категориям. Результаты эксперимента представлены в табл. 2.

Для каждой модели были отобраны наиболее надежно определяемые категории, значение полноты для которых превышает 0,75. Оценки точности mAP для них приведены в табл. 3.

В табл. 3 можно выделить две архитектуры с лучшими результатами — это Faster R-CNN с СНС InceptionResNet v2 и СНС ResNet-101. Число отобранных категорий характеризует стабильность моделей, т.е. большое число категорий с высоким значением метрик. В среднем у ResNet-101 значения полноты и точности для отобранных категорий выше, чем у InceptionResNet, однако первая архитектура показывает худшие результаты для некоторых важных категорий (лица, строения), которые были включены в отобранные. Например, полнота для категории "небоскреб" у модели ResNet-101 составляет 0,145, однако в среднем ее mAP выше, а число ложноположительных предсказаний меньше. Обе модели показывают низкий mAP (большое число ложноположительных результатов) для категорий "дом", "машина", "животное" и "лицо".

В заключительном эксперименте исследовалось качество кластеризации лиц для набора данных GFW (Grouping Faces in the Wild) [18], содержащего 60 различных фотоальбомов из одной социальной сети. Число лиц в каждом альбоме варьируется в диапазоне от 120 до 3600, при этом альбомы содержат не более $C = 321$ различных людей. Для извлечения признаков лица используются известные СНС: VGGFace (VGGNet-16) [19] и VGGFace2 (ResNet-50) [20] и обученная нами на наборе данных VGGFace-2 СНС MobileNet [14, 21]. Каждая нейронная сеть извлекает вектор признаков лица (1024 для MobileNet, 4096 для VGGNet-16 и 2048 для ResNet-50). Для группировки лиц использовался метод ранговой кластеризации [22], а также иерархическая агломеративная кластеризация со следующими способами определения расстояния между кластерами: single link (одиночная связь), complete link (полная связь), average link (метод невзвешенного попарного среднего), метод взвешенного попарного среднего (в качестве весового коэффициента используется размер кластеров) и медианное расстояние между элементами кластера. Качество кластеризации оценивалось с использованием следующих метрик: отношение числа полученных кластеров K к исходному числу различных людей C, индекс Ранда (Adjusted Rand Index, ARI), индекс взаимной информации (Adjusted Mutual Information, AMI) и бикубическая F-мера (BCubed F-measure). Результаты разных методов кластеризации представлены в табл. 4.

Здесь метод иерархической кластеризации с применением межкластерного расстояния на основе среднего расстояния между точками показывает наилучшие результаты. Ожидаемо, что модель VGGFace2 является несколько точ-

Таблица 4

Результаты кластеризации лиц для набора GFW

Метод кластеризации	СНС	K/C	ARI	AMI	F-мера
Одиночная связь	VGGFace	4,10	0,440	0,419	0,616
	VGGFace2	3,21	0,580	0,544	0,707
	MobileNet	4,19	0,492	0,441	0,636
Метод невзвешенного попарного среднего	VGGFace	1,42	0,565	0,632	0,713
	VGGFace2	1,59	0,603	0,663	0,746
	MobileNet	1,59	0,609	0,658	0,751
Полная связь	VGGFace	0,95	0,376	0,553	0,595
	VGGFace2	1,44	0,392	0,570	0,641
	MobileNet	1,28	0,381	0,564	0,626
Метод взвешенного попарного среднего	VGGFace	1,20	0,464	0,597	0,662
	VGGFace2	1,05	0,536	0,656	0,710
	MobileNet	1,57	0,487	0,612	0,697
Медианное расстояние	VGGFace	5,30	0,309	0,307	0,516
	VGGFace2	4,20	0,412	0,422	0,742
	MobileNet	6,86	0,220	0,222	0,411
Ранговое расстояние	VGGFace	0,82	0,319	0,430	0,630
	VGGFace2	1,53	0,367	0,471	0,641
	MobileNet	1,26	0,379	0,483	0,652

нее остальных, однако СНС MobileNet оказывается в 5...10 раз быстрее и занимает в 2...25 раз меньше памяти по сравнению с остальными моделями. Более того, именно для этой модели с помощью метода невзвешенного попарного среднего (average link) получено наибольшее значение (0,751) F-меры, которое превышает наилучший известный результат (0,74) для этого набора данных [18].

Заключение

Многие задачи построения интеллектуальных мобильных систем зачастую содержат противоречивые требования к реализации высокоточных и одновременно вычислительно эффективных процедур распознавания образов. При этом, как показано в настоящей статье, даже несмотря на наличие некоторых ограничений на обработку персональных данных, часто можно автоматически выделить часть "публичных" изображений, которые для повышения точности системы могут быть отправлены на удаленный сервер. Как показано в проведенном эксперименте, такой подход нередко является наиболее приемлемым за счет использования на сервере наиболее современных нейросетевых моделей, более чем на 10...20 % превосходящих по точности алгоритмы, которые могут быть реализованы на современном мобильном устройстве.

Основным ограничением предлагаемого подхода является использование для выделения публичных изображений только информации о распознанных лицах. В результате многие отсканированные персональные документы могут быть ошибочно отправлены на удаленный сервер. Поэтому потенциальная модификация предложенного метода в будущих исследованиях может состоять в его интеграции с алгоритмами распознавания текста и выявлением текстовых фрагментов, характерных для отсканированных документов.

Список литературы

1. **Harrison G.** Next Generation Databases: NoSQL and Big Data. Berlin, Germany, Springer, 2016. 235 p.

2. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series). Cambridge, USA, MIT Press, 2016. 800 p.

3. **Kuznetsova A. et al.** The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale // Cornell University Library, 2018. URL: https://arxiv.org/abs/1811.00982 (date of access 11.02.2019).

4. **Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C.** Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation // Cornell University Library, 2018. URL: https://arxiv.org/abs/1801.04381 (date of access 11.02.2019).

5. **Qin Z., Zhang Z., Chen X., Wang C., Peng Y.** Fd-MobileNet: Improved Mobilenet with a fast downsampling strategy // Proceedings of 25th IEEE International Conference on Image Processing (ICIP). 2018. P. 1363—1367.

6. **Huang J. et al.** Speed accuracy trade-offs for modern convolutional object detectors // Cornell University Library. 2016. URL: https://arxiv.org/abs/1611.10012 (date of access 11.02.2019).

7. **Ren S. et al.** Faster R-CNN towards real-time object detection with region proposal networks // Cornell University Library. 2016. URL: https://arxiv.org/abs/1506.01497 (date of access 11.02.2019).

8. **Redmon J., Farhadi A.** YoloV3: An incremental improvement // Cornell University Library. 2018. URL: https://arxiv.org/abs/1804.02767 (date of access 11.02.2019).

9. **Howard A. G.** et al. MobileNets: Efficient convolutional neural networks for mobile vision applications // Cornell University Library. URL: https://arxiv.org/abs/1704.04861 (date of access 11.02.2019).

10. **Szegedy C. et al.** Inception-v4, Inception-ResNet and the impact of residual connections on learning // Proceedings of the International Conference on Artificial Intelligence (AAAI). 2017. Vol. 4. P. 12.

11. **Prince S. J.** Computer vision: Models, learning, and inference. Cambridge, United Kingdom, Cambridge University Press, 2012. 580 p.

12. **Savchenko A. V., Belova N. S**. Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features // *Expert Systems with Applications,* 2018. Vol. 108. P. 170—182.

13. **Savchenko A. V.** Efficient statistical face recognition using trigonometric series and CNN features // Proceedings of 24th International Conference on Pattern Recognition (ICPR). 2018. P. 3262—3267.

14. **Savchenko A. V.** Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output CNN // Cornell University Library. 2018. URL: https://arxiv.org/abs/1807.07718 (date of access 11.02.2019).

15. **Pan S. J.** A survey on transfer learning // IEEE Transactions on Knowledge and Data Engineering. 2010. Vol. 22, N. 10. P. 1345—1359.

16. **Sharif Razavian A., Azizpour H., Sullivan J., Carlsson S.** CNN features off-the-shelf: an astounding baseline for recognition // Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2014. P. 806—813.

17. **Sokolova A. D., Kharchevnikova A. S., Savchenko A. V.** Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks // Proceedings of International Conference on Analysis of Images, Social Networks and Texts (AIST 2017). Cham, Switzerland, Springer. 2017. P. 223—230

18. **He Y., Cao K., Li C., Loy C. C.** Merge or not? Learning to group faces via imitation learning. Cornell University Library, 2018. URL: https://arxiv.org/abs/1707.03986 (date of access 11.02.2019).

19. **Parkhi O. M., Vedaldi A., Zisserman A.** Deep face recognition // Proceedings of the British Conference on Machine Vision (BMVC). 2015. Vol. 1. P. 6.

20. **Cao Q., Shen L., Xie W., Parkhi O. M., Zisserman A.** VGGFace2: A dataset for recognizing faces across pose and age // Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018). 2018. P. 67—74.

21. **Kharchevnikova A. S., Savchenko A. V.** Neural networks in video-based age and gender recognition on mobile platforms // Optical Memory and Neural Networks (Information Optics). 2018. Vol. 27, N. 4. P. 246—259.

22. **Zhu C., Wen F., Sun J.** A rank-order distance based clustering algorithm for face tagging // Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). 2011. P. 481—488.

I. S. Grechikhin, Postgraduate Student, Senior Lecturer, e-mail: igrechikhin@hse.ru,
A. V. Savchenko, Doctor of Sciences, Professor, e-mail: avsavchenko@hse.ru,
National Research University Higher School of Economics, Nizhny Novgorod

Analysis of User Preferences using Photos and Videos from Mobile Device Based on Object Detection and Neural Networks

In this paper we focus on the problem of user preferences prediction using the gallery of his mobile device. We consider such categories of interests as interior items, food, transport and sport equipment. The novel two-phased method has been proposed. At the first stage, the facial regions are detected on all photos and videos, and the feature vectors are extracted using deep convolutional neural networks. These feature vectors are grouped using known agglomerative clustering techniques. Finally, we select public photos and videos which do not contain faces from the large clusters. At the second stage, these public images are processed on the remote server using high precision Faster R-CNN object detectors. Objects from other images (personal images) are detected on mobile device in offline mode using SSDLite and MobileNet. In the experimental study several neural network-based detectors have been trained using the united training sample from MS Coco, ImageNet and Open Images datasets. Their comparative analysis demonstrated that the Faster R-CNN-based models are characterized with 30 % higher recall when compared to the SSDLite detectors. However, the latter models process each image 3—9-times faster. Finally, we presented the experimental results of facial clustering with GFW (Grouping Faces in the Wild) dataset using either existing feature descriptors (VGGFace, VGGFace2) or the preliminarily trained MobileNet. The latter model with average link hierarchical clustering achieved the highest B-cubed F-measure.

Keywords: *image processing, object detection, mobile systems, visual preferences prediction, face clustering, convolutional neural networks (CNN), Faster R-CNN, SSD*

Acknowledgments. The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and by the Russian Academic Excellence Project "5-100".

DOI: 10.17587/it.25.538-544

References

1. **Harrison G.** Next Generation Databases: NoSQL and Big Data, Berlin, Germany, Springer, 2016, 235 p.
2. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series), Cambridge, USA, MIT Press, 2016, 800 p.
3. **Kuznetsova A. et al.** The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale, Cornell University Library, 2018, available at: https://arxiv.org/abs/1811.00982 (date of access 11.02.2019).
4. **Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C.** Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. Cornell University Library, 2018, available at: https://arxiv.org/abs/1801.04381 (date of access 11.02.2019).
5. **Qin Z., Zhang Z., Chen X., Wang C., Peng Y**. Fd-MobileNet: Improved Mobilenet with a fast downsampling strategy, *Proceedings of 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1363—1367.
6. **Huang J. et al.** Speed accuracy trade-offs for modern convolutional object detectors. Cornell University Library, 2016, available at: https://arxiv.org/abs/1611.10012 (date of access 11.02.2019).
7. **Ren S. et al.** Faster R-CNN towards real-time object detection with region proposal networks. Cornell University Library, 2016, available at: https://arxiv.org/abs/1506.01497 (date of access 11.02.2019).
8. **Redmon J., Farhadi A.** YoloV3: An incremental improvement. Cornell University Library, 2018, available at: https://arxiv.org/abs/1804.02767 (date of access 11.02.2019).
9. **Howard A. G. et al.** MobileNets: Efficient convolutional neural networks for mobile vision applications. Cornell University Library, available at: https://arxiv.org/abs/1704.04861 (date of access 11.02.2019).
10. **Szegedy C. et al.** Inception-v4, Inception-ResNet and the impact of residual connections on learning, *Proceedings of the International Conference on Artificial Intelligence (AAAI)*, 2017, vol. 4, pp. 12.
11. **Prince S. J.** Computer vision: Models, learning, and inference, Cambridge, United Kingdom, Cambridge University Press, 2012, 580 p.
12. **Savchenko A. V., Belova N. S**. Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features, *Expert Systems with Applications,* 2018, vol. 108, pp. 170—182.
13. **Savchenko A. V.** Efficient statistical face recognition using trigonometric series and CNN features, *Proceedings of 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3262—3267.
14. **Savchenko A. V.** Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output CNN, Cornell University Library, 2018, available at: https://arxiv.org/abs/1807.07718 (date of access 11.02.2019).
15. **Pan S. J.** A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering,* 2010, vol. 22, no. 10, pp. 1345—1359.
16. **Sharif Razavian A., Azizpour H., Sullivan J., Carlsson S.** CNN features off-the-shelf: an astounding baseline for recognition, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 806—813.
17. **Sokolova A. D., Kharchevnikova A. S., Savchenko A. V.** Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks, *Proceedings of International Conference on Analysis of Images, Social Networks and Texts (AIST 2017),* Cham, Switzerland, Springer, 2017, pp. 223—230.
18. **He Y., Cao K., Li C., Loy C. C.** Merge or not? Learning to group faces via imitation learning, Cornell University Library, 2018, available at: https://arxiv.org/abs/1707.03986 (date of access 11.02.2019).
19. **Parkhi O. M., Vedaldi A., Zisserman A.** Deep face recognition, *Proceedings of the British Conference on Machine Vision (BMVC)*, 2015, vol. 1, pp. 6.
20. **Cao Q., Shen L., Xie W., Parkhi O. M., Zisserman A.** VGGFace2: A dataset for recognizing faces across pose and age, *Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67—74.
21. **Kharchevnikova A. S., Savchenko A. V.** Neural networks in video-based age and gender recognition on mobile platforms. *Optical Memory and Neural Networks (Information Optics),* 2018, vol. 27, no. 4, pp. 246—259.
22. **Zhu C., Wen F., Sun J.** A rank-order distance based clustering algorithm for face tagging, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2011, pp. 481—488.