Обнаружение персональных данных в фотоальбоме на основе кластеризации лиц и классификации текста сканированных документов

Л.Н. Копейкина[1], А.В. Савченко[1]

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

Аннотация. Исследуется задача обнаружения персональных данных в галерее фотографий пользователя. Предложен новый двухэтапный подход, на первом этапе которого обрабатывается текст в сканированных документах на основе эффективного детектора текста EAST, распознавания с помощью Tesseract и нейросетевого классификатора извлеченного текста. На втором этапе для оставшихся фотографий применяются методы кластеризации лиц для выделения достаточно больших групп близких людей (друзья, родственники), чьи фотографии также относятся к персональным данным и должны обрабатываться непосредственно на мобильном устройстве. Остальные изображения могут быть отправлены на удаленный сервер для высокоточной обработки. Представлены экспериментальные результаты сравнительного анализа известных методов классификации текста и кластеризации для векторов признаков лиц, извлечённых с помощью различных свёрточных сетей.

1. Введение

В эпоху цифровых технологий в связи с развитием облачных вычислений и мобильных устройств объемы мультимедийных данных непрерывно растут. Галерея фотографий содержит уникальную информацию о пользователе, которая потенциально может широко использоваться для моделирования его предпочтений при помощи основанных на технологиях глубокого обучения методов обработки изображений [1] с целью построения рекомендательных систем [2]. Такие методы обычно требуют значительных вычислительных ресурсов и реализуются на удаленном сервере. В таком случае возникает острая необходимость ограничивать обработку фотографий, содержащих персональную информацию о пользователе. В результате все большое внимание уделяется предотвращению вторжения в личную жизнь пользователей с помощью автоматического обнаружения конфиденциальных фотографий [3].

Заметим, что большинство персональных изображений в основном содержат такие общие характеристики, как человеческие лица, текстовые данные (идентификационные данные и номера кредитных карт) и иные объекты (частные автомобили и недвижимость) [4]. Поэтому в настоящей работе предлагается унифицированный подход к обнаружению персональных данных в фотоальбоме с использованием известных методов классификации лиц [5], распознавания текстов (optical character recognition, OCR) [6] и детектирования объектов [1, 7]. В частности, для обнаружения сканированных персональных документов предложено последовательно воспользоваться детектором текстов EAST [8], библиотекой Tesseract OCR и

нейросетевой классификацией обнаруженного на фотографии текста. Для обнаружения персональных изображений, содержащих лица самого пользователя, его близких друзей и родственников, используются известные методы кластеризации лиц на основе свёрточных нейронных сетей (СНС) [8, 9]. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области распознавания изображений и рекомендательных систем.

2. Предложенный подход

Для решения задачи бинарной классификации каждого изображения в галерее пользователя и его отнесения к одному из двух классов: персональное или общедоступное изображение, в работе предлагается следующий подход (рис. 1). Рассмотрим его основные элементы более подробно.

2.1. Обнаружение сканированных персональных документов

Для обнаружения изображений отсканированных личных документов предлагается воспользоваться методами распознавания текста. Вначале детектируются области с текстом, например, с помощью алгоритма EAST [8]. Далее для распознавания текста в каждой области используется Tesseract OCR в режиме image_to_string с использованием рекуррентной модели LSTM (Long-Short Term Memory) Engine. После этого для классификации персональных данных в распознанном тексте предлагается использовать нейронную сеть, которая обучается на основе входной последовательности слов, распознанных в обучающей выборке сканированных документов [10]. Для представления входных данных в виде вектора признаков применяется one-hot кодирование. А именно, создается словарь из V наиболее часто встречающихся в обучающем множестве слов, и каждый текст представляется в виде V-мерного бинарного вектора, при этом v-й компонент вектора равен 1, только если v-е слово из словаря представлено во входном тексте (т.н. модель bag-of-words). Для решения задачи бинарной классификации предлагается использовать вычислительно эффективную реализацию полносвязной нейронной сети, которая показала высокую точность в аналогичной задаче анализа тональности [11]. Как будет показано далее, такой подход не уступает по точности более сложным методам на основе СНС и LSTM, а также с традиционным методом обнаружения персональных данных с помощью поиска ключевых слов.

2.2. Обнаружение персональных фотографий с изображениями лиц

Сканированные документы не являются единственным вариантом персональных данных в фотоальбоме. В частности, персональными обычно считаются изображения, содержащие лица самого пользователя, его близких друзей и родственников [5, 12].

На первом шаге области лиц на всех фотографиях выделяются с помощью известных методов детектирования лиц как на основе каскада классификаторов [13], так и СНС [14, 15]. Так как в фотогалерее пользователя отсутствуют метки людей (обучение без учителя), задача сводится к кластеризации лиц [9, 12]. Для этого векторы признаков извлекаются для каждого из $N>0$ выделенных изображений лиц, которые подаются на вход СНС [1, 12], предварительно обученной для идентификации лиц из большого (внешнего) набора данных, например, VGGFace-2, MS-Celeb и т.п. Выходы одного из последних слоев СНС сохраняются в виде вектора характерных признаков лица, для которых и используются методы кластеризации.

Процедура объединения выделенных лиц в кластеры заключается в отнесении каждого i-го изображения ($i=1,\ldots,N$) к одной из $C \geq 1$ заранее неизвестных групп. В общем случае число различных людей C, присутствующих на изображениях пользователя, неизвестно. Поэтому здесь могут использоваться как стандартные алгоритмы, в которых не задано число кластеров (например, иерархическая агломеративная кластеризация), так и методы на основе специальных мер различия между лицами [16, 17].

Авторы предполагают, что персональными могут считаться изображения с лицами из достаточно больших кластеров – лиц, присутствующих на не менее чем M различных фотографиях, где M – конфигурируемый параметр. Такая процедура проводится в

предположении о том, что подобные кластеры включают лица пользователя, его друзей и родственников, поэтому все содержащие их изображения содержат персональные данные.

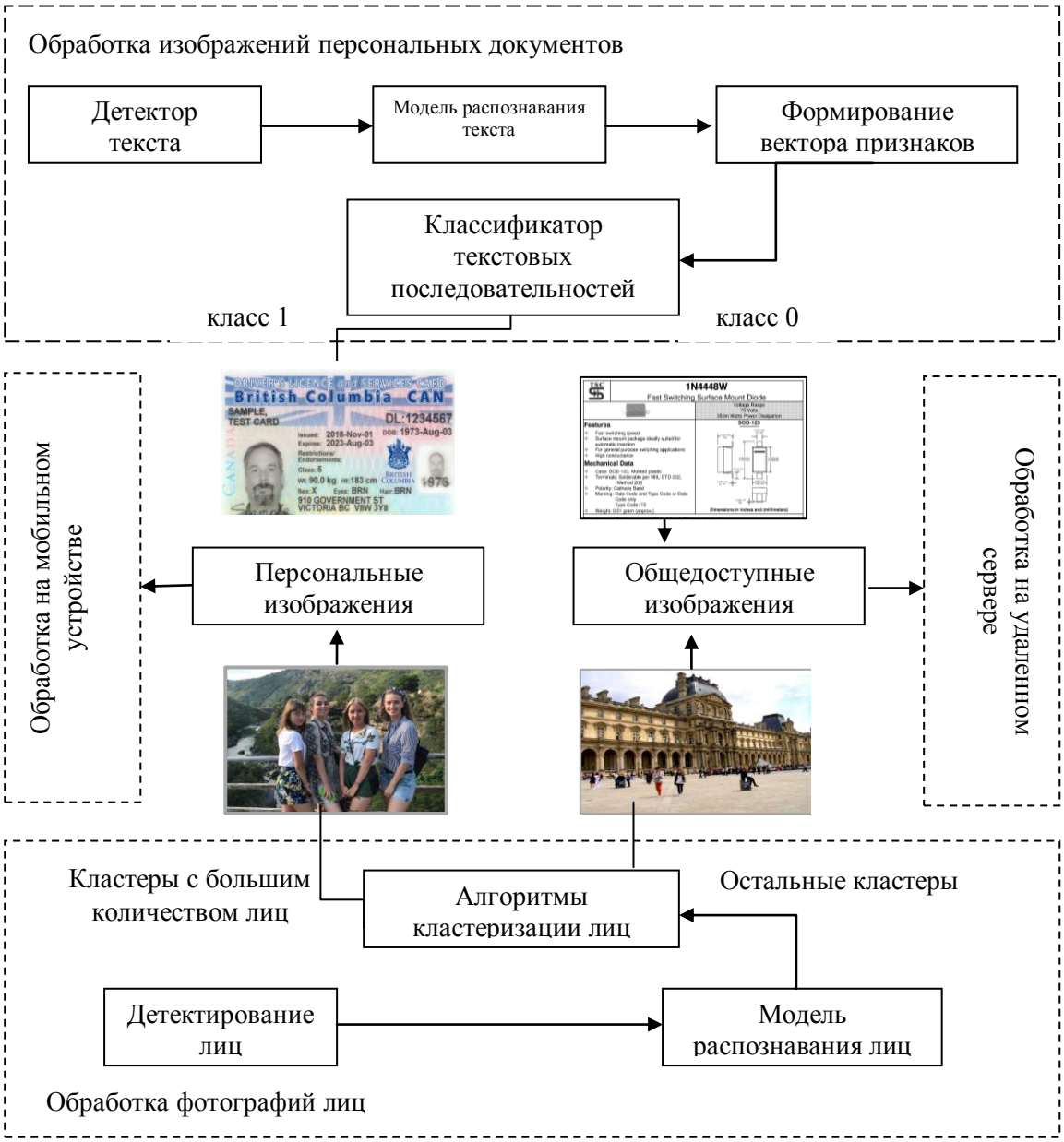

Рисунок 1. Предложенный подход к обнаружению персональных фотографий.

3. Результаты экспериментов

3.1. Обнаружение сканированных персональных документов

В первом эксперименте авторами был создан сбалансированный набор данных из 700 изображений сканированных англоязычных документов. Класс персональных документов представлен 350 изображениями водительских прав и карточек медицинского страхования, паспортов и счетов-фактур из набора данных MIDV [18]. Класс публичных документов состоит из фотографий из общедоступных наборов данных для задач классификации текста DIQA [19] и Ghega [20].

В экспериментальном исследовании использовался традиционный метод поиска специально отобранных ключевых слов (“passport”, “card” и т.п.) [10] при различных вариантах детектирования текста: обнаружение областей текста одновременно с его распознаванием

средствами Tesseract и предварительном детектировании текста с помощью метода EAST [8] с последующим его распознаванием с помощью Tesseract. Дополнительно к традиционному поиску ключевых слов сравнивались три модели нейронных сетей:

- Полносвязная сеть прямого распространения с 2 скрытыми слоями из 16 нейронов с функцией активации гиперболического тангенса. На вход модели поступал V-мерный вектор, закодированный описанным в разделе 2.1 способом (bag-of-words)
- Рекуррентная модель, которая классифицируемую последовательность текста из 400 слов из словаря, состоящего из V = 5000 часто встречаемых слов, подавала на вход слоя векторного представления (embedding) с размером признакового пространства, равным 256. Далее использовался слой LSTM с 128 скрытыми компонентами, слой dropout с частотой выпадения 0.5.
- СНС, состоящей из одного одномерного свёрточного слоя (с 32 нейронами, размером ядра 7 и функцией активации ReLU), слоев подвыборки и dropout (с частотой выпадения 0.5). В качестве первого слоя модели также использовалось векторное представление (embedding) размерности 256.

Последний полносвязный слой каждой модели использовал логистическую функцию для получения оценки апостериорной вероятности принадлежности классу персональных данных. Для обучения классификаторов использовались программные библиотеки TensorFlow и Keras. Все классификаторы обучались в течение 20 эпох с помощью оптимизатора RMSprop.

Таблица 1. Результаты классификации персональных документов.

	Модель	**Точность (precision)**	**Полнота**	**F-мера**	**Вероятность ошибки (error rate)**
Tesseract	Поиск ключевых слов	0.83	0.62	0.70	0.276
	Полносвязная сеть	0.98	0.94	0.95	0.028
	Сеть LSTM	0.97	0.93	0.94	0.043
	СНС	0.88	0.77	0.82	0,161
EAST+ Tesseract	Поиск ключевых слов	0.90	0.75	0.81	0.161
	Полносвязная сеть	1.00	0.97	0.98	0.015
	Сеть LSTM	0.93	0.99	0.95	0.038
	СНС	0.89	0.79	0.83	0.144

Количественное сравнение всех методов с использованием 5-кратной перекрестной проверки представлено в таблице 1. Исходя из полученных результатов, можно отметить, что применение детектора текста EAST для определения областей с текстом было эффективным решением. В то время как вероятность ошибки при использовании только Tesseract на 700 изображениях равна 27.6%, предлагаемое предварительное обнаружение текста с помощью EAST детектора снизило эту ошибку до 16.1%. Кроме того, можно сделать вывод о том, что предлагаемая реализация с детектором текста EAST повышает среднюю точность приблизительно на 2%. Полносвязная сеть достигает наилучших результатов с точностью, превышающей показатель даже традиционной LSTM. Более того, такая реализация на 15% более точно определяет класс изображения документа по сравнению с традиционным поиском ключевых слов.

3.2. Кластеризация лиц

В следующем эксперименте производилось сравнение различных методов кластеризации лиц для набора Gallagher [14], содержащего 589 фотографий с 931 размеченными лицами 32 людей. Поскольку в этом наборе данных доступны только положения глаз, предварительно проводится определение лица с помощью метода MTCNN [15], а далее выбираются объекты с наибольшим пересечением области лица с заданной областью глаз. Если лицо не обнаружено, выделяется квадратная область размером, в 1,5 раза большим заданного расстояния между глазами.

Для извлечения вектора признаков лица рассматривались традиционные предварительно обученные модели, загруженные с официальных сайтов их разработчиков:

- VGGFace (VGGNet-16) [21] извлекает D=4096 признаков;
- VGGFace2 (ResNet-50) [22] извлекает D=2048 признаков;
- MobileNet [12] извлекает D=1024 признака;
- InsightFace (ArcFace) [23] извлекает D=512 признаков;
- FaceNet (Inception ResNet v1) [24] извлекает D=512 признаков.

Таблица 2. Результаты кластеризации лиц для набора Gallagher.

	СНС	K/C	ARI	AMI	Однородность	Полнота	F-мера
Rank-order	VGGFace2	1.25	0.480	0.627	0.794	0.635	0.706
	VGGFace	2.47	0.420	0.506	0.957	0.485	0.552
	MobileNet	2.09	0.674	0.678	0.965	0.611	0.725
	InsightFace	1.78	0.464	0.507	0.906	0.494	0.578
	FaceNet	1.53	0.674	0.681	0.906	0.633	0.760
Single linkage	VGGFace2	3.06	0.267	0.568	0.553	0.752	0.631
	VGGFace	2.75	0.260	0.559	0.531	0.763	0.623
	MobileNet	2.72	0.280	0.586	0.562	0.767	0.636
	InsightFace	2.72	0.109	0.294	0.296	0.607	0.503
	FaceNet	3.09	0.286	0.592	0.579	0.762	0.642
Average linkage	VGGFace2	2.15	0.648	0.771	0.794	0.808	0.802
	VGGFace	1.5	0.662	0.763	0.762	0.819	0.892
	MobileNet	1.75	0.887	0.870	0.932	0.844	0.786
	InsightFace	0.53	0.180	0.395	0.311	0.718	0.497
	FaceNet	2.31	0.886	0.868	0.942	0.835	0.895
Complete linkage	VGGFace2	1.09	0.859	0.867	0.911	0.853	0.888
	VGGFace	1.18	0.616	0.743	0.876	0.690	0.711
	MobileNet	0.41	0.863	0.816	0.798	0.861	0.836
	InsightFace	1.75	0.367	0.576	0.819	0.521	0.512
	FaceNet	0.65	0.710	0.813	0.826	0.830	0.821
Weighted linkage	VGGFace2	1.5	0.891	0.898	0.946	0.876	**0.921**
	VGGFace	1.03	0.599	0.737	0.704	0.830	0.762
	MobileNet	0.59	0.56	0.718	0.629	0.893	0.688
	InsightFace	0.56	0.460	0.556	0.518	0.677	0.625
	FaceNet	1.47	0.884	0.881	0.934	0.857	**0.902**
Centroid linkage	VGGFace2	2.9	0.256	0.559	0.543	0.743	0.632
	VGGFace	2.75	0.093	0.273	0.279	0.600	0.489
	MobileNet	2.34	0.053	0.149	0.173	0.516	0.436
	InsightFace	2.38	0.038	0.082	0.130	0.425	0.422
	FaceNet	2.97	0.252	0.545	0.535	0.726	0.627
Median linkage	VGGFace2	2.63	0.259	0.552	0.523	0.752	0.617
	VGGFace	2.22	0.057	0.191	0.197	0.572	0.447
	MobileNet	2.38	0.048	0.154	0.182	0.498	0.442
	InsightFace	2.59	0.035	0.087	0.143	0.413	0.429
	FaceNet	2.25	0.257	0.580	0.549	0.759	0.627

В ходе эксперимента рассматриваются метод кластеризации на основе рангового расстояния (rank order) [16], а также иерархическая агломеративная кластеризация для расстояния L_2 между нормированными векторами признаков со следующими типами связи: одиночная (single linkage), невзвешенное попарное среднее (average linkage), полная связь (complete linkage),

взвешенное попарное среднее (weighted linkage), центроидная (centroid linkage) и медианная связи (median linkage) из библиотеки SciPy.

Для оценки качества кластеризации в таблице 2 приводятся значения следующих метрик: индекс Ранда (ARI), индекс взаимной информация (AMI), однородность и полнота. Кроме того, вычисляется среднее количество выделенных кластеров K к количеству групп C и традиционная для оценки качества кластеризации лиц бикубическая (bi-cubed) F-мера.

Здесь при извлечении признаков лица с помощью моделей ResNet-50 (VGGFace2) и Inception ResNet v1 (FaceNet) удается достичь более точных результатов кластеризации в большинстве случаев по сравнению с остальными рассматриваемыми моделями. Однако MobileNet уступает незначительно, и при этом ее вычислительная эффективность оказывается в 5-10 раз быстрее по сравнению с VGGFace2. Кроме того, метод взвешенного попарного среднего является лучшим методом по большинству показателей кластерного анализа. Использование рангового расстояния нецелесообразно из-за довольно низких значений по каждой метрике и невысокой производительности (в 3-4 раза выше по сравнению с методами иерархической кластеризации).

В заключительном эксперименте использовался специально созданный набор данных из 400 изображений, 200 из которых содержали фотографии пользователя и его близких друзей, а оставшиеся - общедоступные изображения сцен. Результаты использования предлагаемого подхода (Рис.1) для обнаружения персональных фотографий представлены в таблице 3. Здесь персональными считались изображения, содержащие лица из кластеров, включающих 3 и более объекта. Как видно, все дескрипторы лиц приводят к достаточно высокому качеству обнаружения, но не достигают нулевой вероятности пропуска персональных данных. При этом наилучшие результаты здесь достигаются с использованием «легковесной» мобильной сети, обученной ранее одним из авторов [12].

Таблица 3. Классификации фотографий на основе кластеризации лиц.

Дескриптор лиц	Точность (precision)	Полнота	F1-мера	Вероятность ошибки (error rate)
VGGFace2	0.961	1.0	0.980	0.020
FaceNet	0.956	1.0	0.977	0.023
MobileNet	0.966	1.0	0.982	0.017

4. Заключение

Задача обнаружения персональных фотографий является сложной в плане поиска действенного подхода к решению из-за присущей ей субъективности. В настоящей работе предполагается, что персональные данные содержат конфиденциальную текстовую информацию или же изображения лиц самого пользователя, его близких друзей и родственников. Для выделения таких данных в настоящей работе был предложен новый подход (Рис. 1) для классификации отсканированных документов с использованием детектора текста EAST и распознавания текста в обнаруженной области на основе библиотеки оптического распознавания символов Tesseract. Экспериментально показано, что простая полносвязная нейронная сеть для текста, кодированного с помощью bag-of-words [11], превосходит более сложную сетевую архитектуру, такую как CHC, более чем на 10% и достигает высокой точности обнаружения персональных документов. Кроме того, для набора данных фотоальбома [14] экспериментально показано, что для извлечения групп лиц пользователя, его друзей и знакомых наиболее подходят дескрипторы лиц VGGFace-2 [22] и FaceNet [24] совместно с агломеративной кластеризацией со связью типа взвешенное попарное среднее (weighted linkage).

5. Благодарности

Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа

экономики» (НИУ ВШЭ)» в 2019 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

6. Литература

[1] Гречихин, И.С. Метод анализа предпочтений пользователя по фото- и видеоизображениям на мобильном устройстве на основе нейросетевых детекторов объектов на изображениях / И.С. Гречихин, А.В. Савченко // Информационные технологии. – 2019. – Т. 25, № 9. – С. 538-544. DOI: 10.17587/it.25.538-544.

[2] Demochkin, K.V. Visual product recommendation using neural aggregation network and context gating / K.V. Demochkin, A.V. Savchenko // Journal of Physics: Conference Series. – 2019. – Vol. 1368(3). – P. 032016.

[3] Ren, Z. Learning to anonymize faces for privacy preserving action detection / Z. Ren, Y. Jae Lee, M.S. Ryoo // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 620-636.

[4] He, J. Puppies: Transformation-supported personalized privacy preserving partial image sharing / J. He, B. Liu, D. Kong, X. Bao, N. Wang, H. Jin, G. Kesidis // 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2016. – P. 359-370.

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

[6] Smith, R. An overview of the Tesseract OCR engine / R. Smith // Ninth International Conference on Document Analysis and Recognition (ICDAR). – 2007. – Vol. 2. – P. 629-633.

[7] Ren, S. Faster R-CNN: Towards real-time object detection with region proposal networks / S. Ren, K. He, R. Girshick, J. Sun // Advances in neural information processing systems. – 2015. – P. 91-99.

[8] Zhou, X. EAST: an efficient and accurate scene text detector / X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017. – P. 5551-5560.

[9] Savchenko, A.V. Efficient statistical face recognition using trigonometric series and cnn features / A.V. Savchenko // 24th International Conference on Pattern Recognition (ICPR), 2018. – P. 3262-3267.

[10] Kopeykina, L. Automatic Privacy Detection in Scanned Document Images Based on Deep Neural Networks / L. Kopeykina, A.V. Savchenko // International Russian Automation Conference (RusAutoCon), 2019. – P. 1-6.

[11] Chollet, F. Deep learning with Python / F. Chollet – Manning Publications, 2017.

[12] Savchenko, A.V. Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet / A.V. Savchenko // J. Computer Science, 2019. – P. 5. e197.

[13] Viola, P. Robust real-time face detection / P. Viola, M.J. Jones // International journal of computer vision. – 2004. – Vol. 57(2). – P. 137-154.

[14] Gallagher, A.C. Clothing cosegmentation for recognizing people / A.C. Gallagher, T. Chen // IEEE Conference on Computer Vision and Pattern Recognition, 2008. – P. 1-8.

[15] Zhang, K. Joint face detection and alignment using multitask cascaded convolutional networks / K. Zhang, Z. Zhang, Z. Li, Y. Qiao // IEEE Signal Processing Letters. – 2016. – Vol. 23(10). – P. 1499-1503.

[16] Zhu, C. A rank-order distance based clustering algorithm for face tagging / C. Zhu, F. Wen, J. Sun // CVPR IEEE, 2011. – P. 481-488.

[17] Shi, Y. Face clustering: representation and pairwise constraints / Y. Shi, C. Otto, A.K. Jain // IEEE Transactions on Information Forensics and Security. – 2018. – Vol. 13(7). – P. 1626-1640.

[18] Arlazarov, V.V. A dataset for identity documents analysis and recognition on mobile devices in video stream / V.V. Arlazarov, K. Bulatov, T. Chernov, V.L. Arlazarov // arXiv, 2018.

[19] Ye, P. Document image quality assessment: A brief survey / P. Ye, D. Doermann // 12th International Conference on Document Analysis and Recognition, 2013. – P. 723-727.
[20] Bartoli, A. Improving features extraction for supervised invoice classification / A. Bartoli, G. Davanzo, E. Medvet, E. Sorio // Proceedings of the 10th IASTED International Conference. – 2010. – Vol. 674(040). – P. 401.
[21] Parkhi, O.M. Deep face recognition / O.M. Parkhi, A. Vedaldi, A. Zisserman // BMVC. – 2015. – Vol. 1(3). – P. 6.
[22] Cao, Q. Vggface2: A dataset for recognising faces across pose and age / Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman // 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018. – P. 67-74.
[23] Deng, J. Arcface: Additive angular margin loss for deep face recognition / J. Deng, J. Guo, N. Xue, S. Zafeiriou // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2019. – P. 4690-4699.
[24] Schroff, F. Facenet: A unified embedding for face recognition and clustering / F. Schroff, D. Kalenichenko, J. Philbin // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – P. 815-823.

Personal data detection in photo album based on face clustering and text classification of scanned documents

L.N. Kopeykina[1], A.V. Savchenko[1]

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

Abstract. In this paper we focus on the task of personal data detection in a photo gallery. A novel two-stage approach is proposed. At first, text of scanned documents is processed based on an effective EAST text detector and then extracted text is recognized using Tesseract and neural network classifier. At the second stage, face clustering is implemented for the remaining photos to identify large groups of people (friends, relatives) whose photos also refer to personal data and must be processed directly on a mobile device. The remaining images can be sent to a remote server for processing with higher accuracy. The experimental results of text recognition and face clustering methods using various convolutional networks for facial features extraction are presented.