Быстрая локализация текстовых полей на изображения документов низкого качества*

Д.А. Ильин[1]

[1] Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия

Аннотация. В данной статье рассматривается проблема точной локализации границ слов в текстовых зонах документа. Обработка документов на мобильном устройстве состоит из этапов локализации документа, коррекции перспективы, локализации отдельных полей, нахождения слов в отдельных зонах, сегментации и распознавания. При захвате изображения с помощью мобильной цифровой камеры в условиях неконтролируемой съемки может возникать цифровой шум, искажения перспективы или блики. Тем не менее, проблема локализации границ слов должна решаться на мобильном процессоре с ограниченными вычислительными возможностями за минимально возможное время. Метод, представленный в данной статье, решает более специализированную проблему, чем задача поиска текста на естественных изображениях. Он использует локальные функции, скользящее окно и легкую нейронную сеть для достижения оптимального соотношения скорости и точности алгоритма. Длительность алгоритма составляет 12 мс за поле, запущенное на ARM-процессоре мобильного устройства. Количество ошибок для локализации границ на тестовом образце из 8000 полей составляет 0,3%.

Ключевые слова: *локализация, изображение, обработка документов, компьютерное зрение.*
DOI: 10.14357/20790279180522

Введение

За последние 50 лет технологии оптического распознавания печатного текста достигли значительного успеха [1, 2]. Для высококачественных изображений, полученных сканерами, современные оптические системы ввода обеспечивают промышленное качество распознавания. Научные исследования в этой области продолжают улучшать качество при работе с низкокачественными изображениями документов, а также сложными и деградированными документами [3, 4].

Основная проблема распознавания изображений идентификационных документов с помощью мобильных устройств – это низкое/среднее качество печати этих документов с большим количеством маркировок и элементов безопасности. Чтобы получить высокое качество финального распознавания, мы используем следующую схему:

После обнаружения документа в кадре [5], изображение проективно исправляется. Затем происходит выделение областей текстовых полей. Несмотря на сложный фон, проекция строк на вертикальную ось довольно информативна, и вместе с анализом шаблона документа она позволяет точно изолировать изображения с отдельными линиями. Более того, вертикальные границы таких зон задаются с большим отрывом, то есть справа и слева от текста зона может содержать большую площадь со структурированным шумом. С точки зрения производительности и надежности результатов, необходимо выполнить точную сегментацию и распознавание символов только в текстовой области с использованием методов распознавания, адаптированных к использованию на мобильных устройствах (например [6, 7]).

Таким образом, процесс определения положения отдельных слов в зоне играет важную роль в процессе оптического распознавания документа.

В этой статье мы предлагаем новый подход, основанный на искусственных нейронных сетях для быстрой локализации границ слов в текстовых областях с изображениями низкого качества. Наш подход иллюстрируется на рис. 1. Метод использует локальные функции, скользящее окно и нейронную сеть с простой архитектурой, что позволяет достичь оптимального соотношения скорости алгоритма с результатами высокого качества.

1. Предыдущие работы

В настоящее время внедрение оптических систем ввода документов промышленного качества на мобильных устройствах, использующих поток

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 17-29-03161 и 17-29-03236).

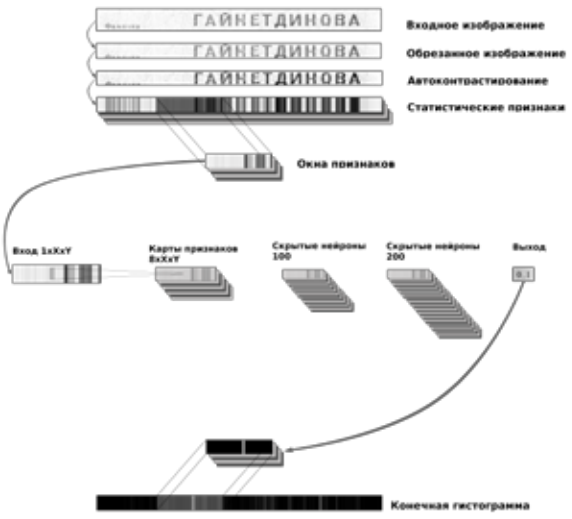

Рис.1. Общая схема алгоритма

Рис.2. Примеры зон текстовых полей

видео, становится актуальным. При внедрении систем этого типа возникает ряд трудностей. Во-первых, неконтролируемые условия съемки часто вызывают необходимость обработки изображений низкого качества с большим количеством искажений. Во-вторых, несмотря на постоянный рост производительности мобильных устройств, их вычислительные возможности ограничены требованиями энергоэффективности. В результате, они несравнимы с возможностями настольных компьютеров и серверного оборудования. Наконец, успех мобильных приложений сильно зависит от опыта пользователя, то есть распознавание должно выполняться во время выполнения программы, когда окончательный ответ предоставляется через 1-2 секунды после обнаружения кадра.

Зоны текстовых полей идентификационных документов содержат одно или несколько слов и могут быть охарактеризованы высокой изменчивостью шрифтов, фоновой заливки и печати артефактов.

Зона поля может содержать разлинеенную бумагу и статический текст (рис.2a). Коррекция проекционных искажений во время обработки изображения документа может быть выполнена недостаточно точно, поэтому отдельные строки или даже символы могут быть искажены. Наклонное положение строки в зоне также может быть результатом неточного ввода текста в пробел. Захваченные цифровой камерой небольшого формата в режиме видео, изображения не имеют высокого разрешения и содержат цифровой шум. Также возможны дефокусировка и размытие, неравномерное освещение или блики (рис.3a, 4a). Элементы защиты документов от подделок могут вносить дополнительные искажения.

Подробный обзор методов сегментации печатных текстов можно найти в [8, 9]. Хорошо зарекомендовавшие себя методы сегментации для высококачественных изображений печатных деловых документов, полученных с помощью планшетных сканеров, основанные на гистограммах и компонентах связности, работают быстро, но не применимы из-за низкого качества исходного изображения. Например, возможная гетерогенность освещения зоны текстового поля, бликов и сложного фона (рис. 3a) предотвращает получение двоичного изображения со всеми символами методом бинаризации по порогу и его производных (рис.3b,c).

Адаптивные методы бинаризации окон страдают от множества мусорных объектов на сложном фоне вне текстовой зоны (рис. 3e). Мы провели исследование более сложных методов обработки изображений и бинаризации и в рамках установленных ограничений сложности вычислений не нашли подходящего надежного метода бинаризации.

Анализ гистограмм проекций может быть выполнен не только для двоичных, но и для серых изображений (например [10]). Кроме того, помимо низкого качества самих изображений, разлинованную бумагу и статические элементы формы документа можно найти в области поля. Если контраст-

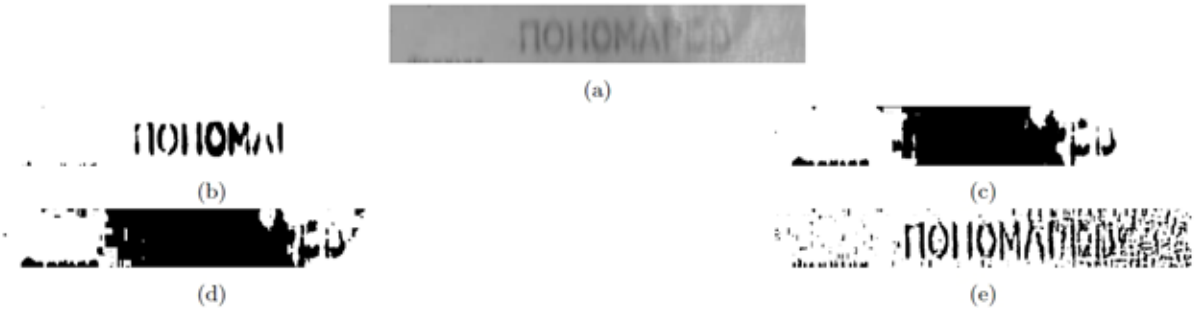

Рис. 3. Проблемы бинаризации: a – исходное изображение зоны текстового поля, b и c – бинаризация по порогу и его производных, d – метод Отсу, e – адаптивные методы бинаризации

ность этих шумовых элементов значительно выше контрастности текстового поля или имеет блики (рис. 4a), тогда проекционный профиль сильно искажен и выбор границ ближайших слов становится сложным (рис.4b).

Методы сегментации текстовой строки на основе символа и методы распознавания текста в естественных сценах [11-13] могут преодолевать низкое качество изображений, но имеют чрезмерно высокую алгоритмическую сложность. Они не подходят для использования на мобильных устройствах как часть системы распознавания документов, которая полностью распознает документ и объединяет результаты распознавания несколько раз в секунду для получения общей высокой точности ввода.

2. Наш метод

Наш метод предназначен для определения границ слов. Основная идея состоит в том, чтобы сочетать преимущества быстрых вычисляемых инженерных характеристик и методов машинного обучения для точного обнаружения границ слов. Такой подход позволяет нам увеличить вычислительную эффективность этого алгоритма и использовать его на устройствах с малой мощностью, таких как мобильные устройства. Разумный выбор технических характеристик позволяет повысить устойчивость алгоритма для различных искажений. Хорошо обученная нейронная сеть поверх технических функций может использоваться на разных полях без переобучения. Полная схема алгоритма показана на рис. 5.

На начальном этапе происходит предварительная обработка изображений:

1. Обрезка изображений сверху и снизу.
2. Автоконтрастирование изображения.
3. Масштабирование изображения.

Автоконтрастирование необходимо для нормализации входных данных до подсчета функций и подачи их в нейронную сеть. Без автоконтрастирования нейронная сеть может работать не так эффективно. Необходимо масштабировать изображение, чтобы уменьшить количество последующих операций и устранить избыточную информацию. Поскольку мы подсчитываем функции независимо для каждого столбца, процессы расчета могут быть эффективно реализованы параллельно на современных мобильных устройствах. Таким образом, изображение преобразуется в вектор-функцию / набор векторов признаков.

Входной вектор разрезается скользящим окном на отдельные секции. Шаг скользящего окна можно варьировать и выбирать в соответствии с требованиями оптимального соотношения скорости и качества. Ширина окна составляет около 3 символов. Результирующие окна используются в качестве входа нейронной сети. Нейронная сеть принимает в качестве входного окна вектор-функцию. На выходе нейронной сети значение отлича-

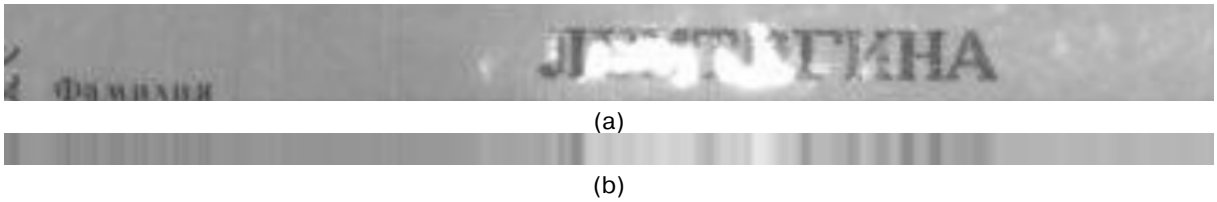

Рис. 4. Проблемы проецирования: a – исходное изображение зоны текстового поля, b – серый проекционный профиль

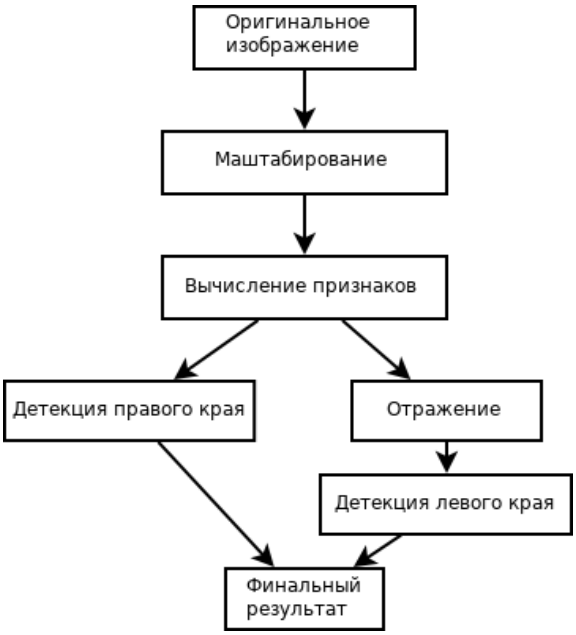

Рис. 5. Схема алгоритма быстрой локализации

ется от 0 до 1. 1 означает, что в окне нет границы слова. Значение от 0 до 1 – это положение границы слова на изображении в окне. Структура нейронной сети изменяется в зависимости от необходимого соотношения качества и скорости, но принципиально она состоит из сверточного слоя и двух полностью связанных слоев. Такой тип архитектуры достаточен для решения проблемы и может быть запущен на различных устройствах, включая мобильные.

После получения ответов нейронной сети для каждого окна ответы суммируются (набирается гистограмма), сглаживаются и выбираются максимальные значения.

3. Метод, основанный на одном признаке

Решение основано на одной функции. Важным выбором, определяющим качественные характеристики системы, является выбор функций, которые учитываются для всего изображения. В качестве единственной функции, которая подсчитывается для каждого столбца изображения, можно выбрать минимум. Его преимущества – низкая вычислительная стоимость, устойчивость к изменениям высоты окна, высоты шрифта и небольшого шума. Недостатки – система с такими характеристиками не устойчива к блику и не соответствует проективным искажениям. Распределение ошибок локализации границы для одного функционального решения можно увидеть на рис. 6.

4. Многопризнаковое решение

Чтобы улучшить качество системы и устранить ошибки, возникающие при работе системы с одной функцией, можно использовать несколько функций. Выбор функций определяется ошибками, возникающими во время работы системы. Когда система работает с одной функцией, подавляющее число ошибок является статическим текстом. Добавляя «среднее значение», которое вычисляет-

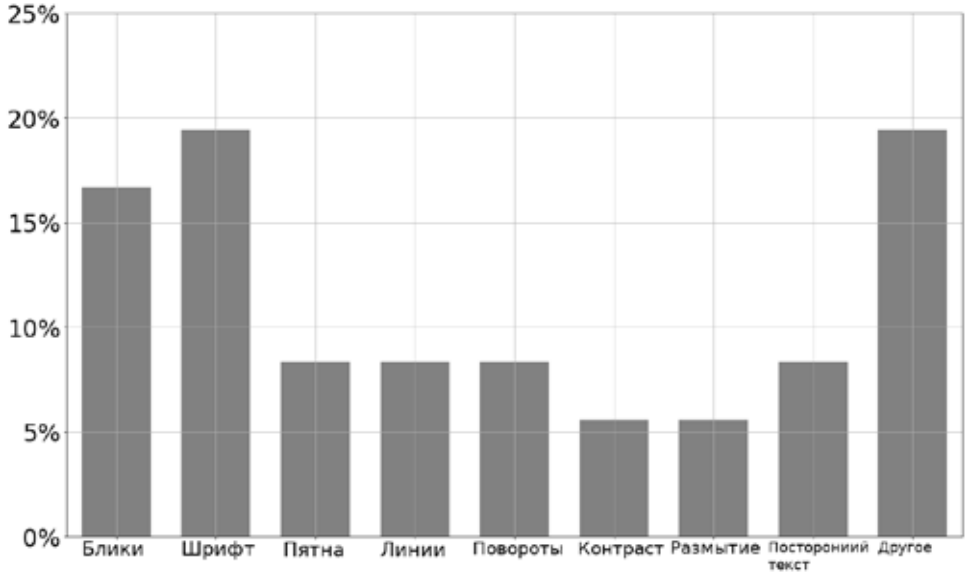

Рис. 6. Гистограмма ошибок локализации границы

ся столбцом, мы получаем прирост качества. Другим решением, более устойчивым к изменениям высоты поля, является вычисление минимума в верхней и нижней половине изображения (2 функции). Пример его применения к изображению со статическим текстом можно увидеть на рис. 7. Мы делаем эту операцию итеративной: рассмотрим ошибки и добавим следующую функцию, чтобы исправить текущую наиболее распространенную ошибку. Аналогичным образом мы можем улучшить качество с некоторым замедлением из-за подсчета дополнительной функции.

Рис. 7. Пример решения с использованием двух функций, примененный к изображению со статическим текстом

5. Описание нейронной сети

Нейронная сеть, обученная полученным функциям, имеет полностью связанную или сверточную структуру:

1. Полностью связанная структура: три полностью связанных слоя с 100, 50 и 1 нейронами.
2. Сверточная структура: 1 сверточный слой вдоль вертикальной оси (через признаки) и 2 полностью связанных слоя.

Нейронная сеть обучалась на образцах изображений со следующими искажениями: сдвигами, поворотами, шумом, проективными искажениями, искажениями яркости, светлыми пятнами.

6. Результаты эксперимента

В наших экспериментах мы использовали 50 000 полей российских национальных паспортов в качестве учебных образцов и 8000 в качестве пробных образцов. Качество полей отличается, поскольку они были получены в разных условиях (например, яркость, наружный / закрытый, время суток и т. д.) и различными мобильными устройствами (например, iPhone 4, 5, 6 поколений и несколько моделей Samsung Galaxy). Мы сделали несколько коротких видеороликов (5-15 секунд) под разными углами к источнику света и к документу, которые в итоге были преобразованы в раскадровки. Таким образом, мы знали условия, в которых было получено каждое полевое изображение. Используя эти данные, мы разделили весь набор примеров изображений полей на обучающие и тестовые множества. Мы не использовали изображения из одного источника (т.е. видео) в разных наборах. Более того, мы старались не использовать изображения, снятые в аналогичных условиях в обучающих и тестовых наборах.

Были получены статистические данные для разных решений (табл. 1):

Максимальное качество было получено для самого большого набора функций. Однако оптимальный коэффициент качества был достигнут для набора признаков (средний, минимальный, сигма).

Более подробная версия статистики для разных методов приведена в табл. 2.

Заключение

В данной статье рассмотрена проблема точной локализации границ слов в текстовых зонах документа. Поскольку локализация является од-

Табл. 1

Результаты эксперимента. Коэффициент качества

Номер	Признаки метода	Качество	Производительность (мс на поле)
1	Минимум	75.6%	11
2	Среднее	83.9%	11
3	Минимум по верхней половине и нижней половине	90.3%	15
4	Минимум, среднее, максимум	94.1%	22
5	Минимум, среднее, сигма	99.7%	25
6	Минимум, среднее, сигма, коэффициентом асимметрии, коэффициентом эксцесса	99.82%	32

Табл. 2

Более подробная статистика использованных методов

Номер	Метод	Среднее расстояние	Точность и полнота		F-мера
1	Минимум	1.19	0.83	0.72	0.77
2	Среднее	2.1	0.86	0.83	0.84
3	Минимум по верхней половине и нижней половине	2.57	0.92	0.87	0.89
4	Минимум, среднее, максимум	0.91	0.96	0.94	0.95
5	Минимум, среднее, сигма	0.56	0.997	0.996	0.996
6	Минимум, среднее, сигма, коэффициентом асимметрии, коэффициентом эксцесса	0.41	0.998	0.997	0.997

ним из основных этапов процесса OCR документа, его ускорение и уточнение являются важной задачей для выполнения распознавания на мобильных устройствах. Наиболее приемлемое соотношение качества скорости было достигнуто для 3-х функций: минимального, среднего и сигма. В верхней части этих функций использовались простые сверточные нейронные сети. Наш метод обладает хорошей обобщающей способностью и может использоваться без дополнительной подготовки по другим типам документов, взятых с камеры мобильного телефона.

Литература

1. *Хиромичи Фудзисава*. Сорок лет исследований в области распознавания символов и документов – промышленная перспектива, распознавание паттернов.
2. *Доерманн Дэвид, Томбр Карл.* «Справочник по обработке и распознаванию документов». Springer-Verlag, Лондон, 2014.
3. *Лян Цзянь, Дэвид Доерманн и Хупинг Ли*. Анализ текста и документов на основе камер: обзор // Международный журнал анализа и распознавания документов (IJDAR), 2005.
4. *Лу Тонг и Паляйнакоте, Шивакумара, Тан Чу Лим и Лю Вэньин*. Обнаружение видеотекста, Springer-Verlag, Лондон, 2014.
5. *Скорюкина Н., Николаев Д.П., Шешкус А., Полевой Д.* Прямое обнаружение документов в реальном времени на мобильных устройствах, Proc. SPIE 9445, Седьмая международная конференция по машиностроению (ICMV 2014), 12 февраля 2015 г.
6. *Лимонова Е., Ильин Д. и Николаев Д.* Улучшение производительности нейронной сети на SIMD-архитектурах. Восьмая международная конференция по машинному зрению. Барселона, Испания. 2015.
7. *Лимонова Е., Шешкус А. и Николаев Д.* Вычислительная оптимизация сверточных нейронных сетей с использованием архитектуры разделенных фильтров // Международный журнал прикладных инженерных исследований. 2016.
8. *Yi Lu*. Сегментация печатных машин. Обзор, Распознавание образов, том. 28, вып. 1, с. 67 –80, 1995.
9. *Ричард Г. Кейси и Эрик Леколинет*. Обзор методов и стратегий сегментации символов, IEEE-транзакции по анализу шаблонов и машинной разведке, том. 18, с. 690–706, 1996.
10. *Grafmller M., Beyerer J.* Сегментация печатных серо-масштабных матричных символов // Материалы 14-й мировой многоконференции по системной, кибернетике и информатике WMSCI 2010 (т. II, с. 8791).
11. *LeBourgeois F.* Robust Multifont OCR System от изображений уровня серого, Международная конференция по анализу и распознаванию документов, vol. 0, p. 1, 1997.
12. *Ye Q. и Doermann D.* Обнаружение и распознавание текста в образах: обзор, IEEE-транзакции по анализу шаблонов и машинной разведке, том. 37, вып. 7, с. 1480-1500, 1 июля 2015 г.
13. *Yin X.C., Zuo Z.Y., Tian S. и Liu C.L.* Обнаружение текста, отслеживание и распознавание в видео: Всестороннее обследование, транзакции IEEE по обработке изображений, том. 25, вып. 6, стр. 2752-2773, июнь 2016 г.

Ильин Дмитрий Алексеевич. Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия. Аспирант. Количество печатных работ: 7. Область научных интересов: машинное обучение, обработка данных, распознавание изображений. E-mail: dmitry.ilin@phystech.edu

Fast words boundaries localization in text fields for low quality document images

D.A. Ilin[1]

[1] Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. The paper examines the problem of word boundaries precise localization in document text zones. Document processing on a mobile device consists of document localization, perspective correction, localization of individual fields, finding words in separate zones, segmentation and recognition. While capturing an image with a mobile digital camera under uncontrolled capturing conditions, digital noise, perspective distortions or glares may occur. However, the problem of word boundaries localization has to be solved at run-time on mobile CPU with limited computing capabilities under specified restrictions. The method presented in this paper solves a more specialized problem than the task of finding text on natural images. It uses local features, a sliding window and a lightweight neural network in order to achieve an optimal algorithm speed-precision ratio. The duration of the algorithm is 12 ms per field running on an ARM processor of a mobile device. The error rate for boundaries localization on a test sample of 8000 fields is 0.3%.

Keywords: *localization, image, document processing, computer vision.*
DOI: 10.14357/20790279180522

References

1. *Hiromichi Fujisawa*. Forty Years of Research in Character and Document Recognition – an Industrial Perspective, Pattern Recognition.
2. *David Doermann, Karl Tombre*. Handbook of Document Image Processing and Recognition, Springer-Verlag, London, 2014.
3. *Liang, Jian, David Doermann, and Huiping Li*. Camera-based analysis of text and documents: a survey, International Journal of Document Analysis and Recognition (IJDAR), 2005.
4. *Lu, Tong and Palaiahnakote, Shivakumara and Tan, Chew Lim and Liu, Wenyin*. Video Text Detection, Springer-Verlag, London, 2014.
5. *Natalya Skoryukina, Dmitry P. Nikolaev, Alexander Sheshkus,Dmitry Polevoy*. Real time rectangular document detection on mobile devices, Proc. SPIE 9445, Seventh International Conference on Machine Vision (ICMV 2014), 94452A (February 12, 2015).
6. *Limonova, E., Ilin, D. and Nikolaev, D.* 2015. Improving neural network performance on SIMD architectures. Eighth International Conference on Machine Vision. Barcelona, Spain.
7. *Limonova, E., Sheshkus, A. and Nikolaev, D.* 2016. Computational optimization of convolutional neural networks using separated filters architecture. International Journal of Applied Engineering Research.
8. *Yi Lu*. Machine printed character segmentation An overview, Pattern Recognition, vol. 28, no. 1, pp. 6780, 1995.
9. *Richard G. Casey and Eric Lecolinet*. A Survey of Methods and Strategies in Character Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, pp. 690-706, 1996.
10. *Grafmller, M., Beyerer, J.* Segmentation of printed gray scale dot matrix characters, Proceedings of 14th world multi-conference on systemics, cybernetics and informatics WMSCI 2010 (Vol. II, pp. 87-91).
11. *F. LeBourgeois*. Robust Multifont OCR System from Gray Level Images, International Conference on Document Analysis and Recognition, vol. 0, p. 1, 1997.
12. *Q. Ye and D.* Doermann Text Detection and Recognition in Imagery: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 7, pp. 1480-1500, July 1 2015.
13. *X. C. Yin, Z. Y. Zuo, S. Tian and C. L. Liu*. Text Detection, Tracking and Recognition in Video: A Comprehensive Survey, IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2752-2773, June 2016.

D.A. Ilin. Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, PhD candidate Moscow, Russia, number of publications: 7, areas of interest: machine learning, data processing, image recognition. E-mail: dmitry.ilin@phystech.edu