Распознавание образов

Поиск текстовых полей документа с помощью методов обработки изображений

Д.Г. Слугин, В.В. Арлазаров

Аннотация. В статье описан алгоритм поиска текстовых полей документа на примере национального паспорта РФ. Алгоритм основан на методах обработки изображений, в частности морфологической фильтрации, и проверке соответствия полученной структуры документа его шаблону. Приведены результаты работы на примере большого массива данных, представляющих собой изображения документов из видеопотока, сканы и фотографии, на которых достигнуты хорошие результаты. Алгоритм допускает обобщение на широкий класс документов, таких как ID карты, водительские удостоверения, визы и так далее.

Ключевые слова: *поиск полей документа, распознавание документа, видеопоток, обработка изображений, морфологические операции, преобразование Хафа, сопоставление шаблонов.*

Введение

В настоящее время распознавание документов на мобильных устройствах становится весьма актуальной задачей, в связи с ростом производительности данных устройств, улучшением разрешения и качества изображений, получаемых с камер. Это позволяет выполнять весь процесс распознавания, начиная от получения изображения и заканчивая выводом результата, полностью на устройстве, без передачи изображений на внешние сервера и облака, что повышает универсальность и безопасность приложений. Современным трендом в распознавании документов является использование не одного изображения документа (его фотографии или скана), а видеопотока в качестве исходных данных, производя распознавание на последовательном массиве кадров, содержащих изображения документа, в режиме реального времени [1]. Использование видеопотока имеет ряд преимуществ по отношению к одному изображению:

- наличие массива кадров повышает вероятность распознавания документа за счет выбора “наилучших” кадров из видеопотока
- комбинирование результатов, полученных на разных кадрах, для повышения качества распознавания
- возможность осуществления обратной связи в процессе съемки для её коррекции

Одновременно это накладывает и ряд ограничений на применяемые алгоритмы, особенно в плане скорости и устойчивости к возможным искажениям исходных данных [2].

Нахождение текстовых полей является важной частью процесса распознавания документа. От того, насколько точно было определено их положение, зависит финальный результат в целом. Сама задача поиска является достаточно сложной, существует несколько подходов к её решению [3], среди которых нет какого-либо универсального, выбор зависит от постановки задачи, исходных данных и возможных ограничений.

В статье рассматривается задача поиска текстовых полей на примере национального паспорта РФ, исходные данные - это в основном изображения из видеопотока, полученного с мобильных устройств, таких как телефоны, планшеты и так далее. Рассмотрим общую схему распознавания документа (рис. 1), содержащую необходимые этапы процесса.

Поиск текстовых полей располагается между этапом выделения зон документа, содержащих поля, и этапом распознавания полей. Он отвечает за поиск границ полей и определение, какому атри-

* Работа выполнена при финансовой поддержке РФФИ (проекты № 17-29-03263 и № 17-29-03170)

Рис. 1. Общая схема распознавания документа

буту документа соответствует каждое поле. Для успешной работы алгоритм поиска текстовых полей должен удовлетворять следующим критериям:

- Устойчивость к ошибкам нахождения зон документа. Определение границ документа является важным этапом распознавания документа, от которого зависит качество работы всех последующих этапов. Для поиска границ документа, особенно в случаях сложного окружения и фона, используются различные алгоритмы для достижения как можно более высокого уровня качества [4]. Несмотря на это, возможны ошибки, которые приводят к неправильному выделению границ документа и его зон. В задачах распознавания документа в видеопотоке документ обычно располагается в плоскости, почти перпендикулярной оптической оси устройства регистрации данных, следовательно, искажения зоны имеют в своей основе аффинные преобразования, такие как сдвиг и поворот.
- Устойчивость к изменению освещения. Большинство современных документов длительного пользования, таких как паспорта, ID карты, водительские удостоверения, представляют собой ламинированный бумажный или пластиковый носитель. В результате помимо теней от окружающих объектов даже при небольших углах отклонения освещенность документа может меняться в широких пределах во времени. Для документов с защитной светоотражающей пленкой возможно появление бликов, которые могут значительно ухудшать читаемость документа. В этом случае возможно применение алгоритмов детекции бликов и оценки качества как входного изображения в целом, так и отдельных полей документа по нахождению на них таких объектов [5].
- Устойчивость к фону документа. Многие документы обладают специальным защитным фоном – там называемый гильош – используемый для защиты от подделок. В этом случае текстовые поля печатаются поверх него и алгоритм должен уметь работать в данном случае.
- Скорость работы. Распознавание документов на мобильных платформах и устройствах-на-чипе, работающих в режиме реального времени, предъявляет высокие требования к оптимизации и скорости работы алгоритмов.
- Оценка результата. Помимо нахождения текстовых полей необходимо произвести их сопоставление с атрибутами документа и дать оценку такому сопоставлению. В случае невозможности корректного сопоставления алгоритм должен выдавать отказ.

В данной статье для решения задачи применяются такие методы обработки изображений как морфологические операции. Этот подход ранее уже встречался в задачах поиска текстовых строк, например, в работах [6] и [7] применялись морфологические операции, однако нами предложена иная схема обработки, ориентированная на поиск именно полей документа с учетом особенности изображений, получаемых с мобильных устройств. Также предложено описание шаблона зоны документа и метод сопоставления найденных тестовых полей этому шаблону.

Постановка задачи

Рассмотрим постановку задачи поиска границ текстовых полей документа. Исходный документ состоит из одной или нескольких зон, каждая их которых содержит текстовые поля, составляющие содержимое документа. Зона состоит из строк, каждая из которых состоит из одного или несколько текстовых полей. Поля в строке отделены друг от друга расстоянием, характерным для данного документа, более близкие текстовые поля считаются одним общим полем. Количество строк и полей

на документе могут быть переменной величиной, сами текстовые поля выделены с помощью яркости и размеров. Требуется определить границы таких полей и сопоставить их соответствующим атрибутам документа. Рассмотрим решение задачи на примере национального паспорта РФ в качестве исходных данных. Для зоны данных страницы паспорта РФ (рис. 2) необходимо определить положение следующих полей: ФИО, пол, дата и место рождения. Поле отчество может отсутствовать, количество строк полей места рождения является переменной величиной, от одной до трех (такие поля выделены на изображении более светлыми рамками).

Рис. 2. Поля зоны данных национального паспорта РФ

Алгоритм решения поставленной задачи состоит из трех этапов:

- Предобработка изображения
- Выделение строк и текстовых полей
- Сопоставление найденных полей шаблону зоны документа

Предобработка изображения

Основная идея предобработки заключается в преобразовании исходного изображения таким образом, чтобы представить искомые текстовые поля в виде легко находимых и визуально определяемых объектов. Для решения этой задачи будем использовать морфологические операции. Основными морфологическими операциями являются дилатация и эрозия, которые выполняются для каждой точки изображения с вычислением max или min значений в её окрестности, заданной некоторым примитивом. Вычисление морфологии линейно зависит от размера примитива, поэтому является достаточно трудоёмкой операцией. Однако существует алгоритм Ван Херка [8], который позволяет вычислять морфологические операции с прямоугольным примитивом за время, не зависящее от его размеров. Для ускорения работы алгоритма возможно применение расширенных наборов команд, таких как SSE для процессоров семейства x86 и NEON для архитектуры ARM, что является особенно актуальным для мобильных устройств [9].

Пусть e прямоугольный примитив размером $[a,b]$

$$e=\{(x,y):|x|\le a,|y|\le b\}$$

Дилатация изображения f по примитиву e обозначается $f\oplus e$ и определяется как:

$$(f\oplus e)(x,y)=\max_{|s|\le a,|t|\le b}\{f(x+s,y+t)\}$$

Эрозия изображения f по примитиву e обозначается $f\otimes e$ и определяется как:

$$(f\otimes e)(x,y)=\min_{|s|\le a,|t|\le b}\{f(x+s,y+t)\}.$$

Рассмотрим еще две операции – размыкание и замыкание, которые являются комбинацией дилатации и эрозии. Размыкание множества f по примитиву e обозначается $f\circ e$ и определяется равенством:

$$f\circ e=(f\otimes e)\oplus e.$$

Замыкание множества f по примитиву e обозначается $f\bullet e$ и определяется как:

$$f\bullet e=(f\oplus e)\otimes e$$

Замыкание сглаживает контуры объекта, «обрывает» узкие перешейки и ликвидирует выступы, тогда как размыкание заполняет промежутки контуров и объединяет близкие объекты.

Рассмотрим предлагаемый алгоритм предобработки изображения (рис. 3), каждый этап согласно нумерации и получаемые результаты (рис. 4).

(1) Текстовые поля документов в большинстве случаев представляют собой последовательные наборы букв, яркость которых отличается от яркости фона. Таким образом, первым шагом осуществляем переход от цветного изображения к полутоновому. Если фон имеет характерный цвет, то вместо усреднения каналов предпочтительно взять данные конкретного цветового канала для лучшего контрастирования фона от текста.

Рис 3. Алгоритм предобработки изображения

(2) Производится масштабирование исходной зоны к заданному размеру шаблона зоны (рис. 4а). Это делается по следующим соображениям:
- Исходные изображения имеют различные размеры, в зависимости от источников данных, поэтому требуется их унификация.
- Размер шаблона выбирается небольшим, но одновременно текстовые поля должны быть достаточно отличимыми от фона. Это позволяет значительно ускорить

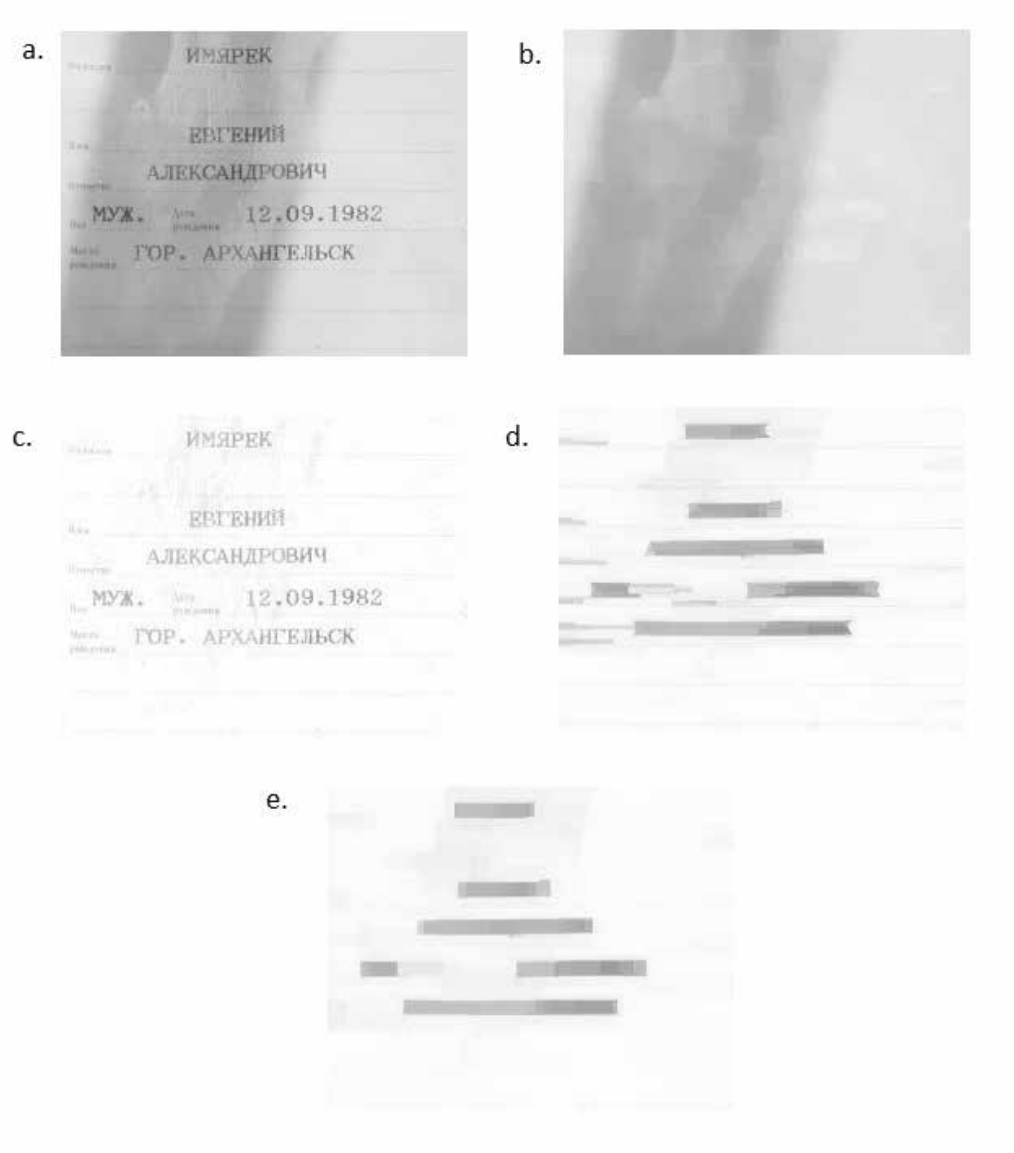

Рис. 4. Результаты предобработки изображения по этапам

операции с такой зоной по сравнению с исходным изображением. Для примера обычный размер шаблона зоны данных паспорта РФ составляет 250x250 точек.

- Масштабирование позволяет сгладить особенности гильоша документа и преобразовать его в близкий к однородному фон.

(3) Производится отделение фона от текста путём замыкания изображения шаблона зоны с примитивом небольшого размера $[4,4]$, в результате получаем почти однородное изображение (рис. 4b), состоящее из усредненного цвета фона и сохраняющее особенности неравномерности освещения различных частей документа.

(4) Инвертируем изображение фона

(5) Складываем полученное изображение фона с изображением шага (2) и получаем результат, где фон стал близок к белому, а текстовые поля остались контрастными (рис. 4c).

(6) Производим размыкание полученного изображения с примитивом размера [s,*0*], где *s* – приблизительная ширина символа. В результате такого преобразования получаем изображение, представляющее собой склеенные в компоненты текстовые поля на однородном фоне (рис. 4d).

(7) Произведем замыкание изображения с примитивом размера [0,*h*/2], где *h* – приблизительная высота символа. Это преобразование уберёт выступы компонент, а также небольшие статические тексты и связи с ними. В результате получается выходное изображение, на котором текстовые поля представлены в виде визуально отличимых компонент на почти однородном фоне (рис. 4e).

(8) В ряде случаев, связанных с ошибками нахождения границ документа или проблемами при печати данных, текстовые поля документа могут иметь значительные углы наклона, что затрудняет их поиск и последующее распознавание. Для определения угла наклона и коррекции полей используется быстрое преобразование Хафа [10], после чего осуществляем поворот изображения.

Выделение строк и текстовых полей

Рассмотрим алгоритм выделения текстовых строк и полей, в качестве исходных данных получающий результат предобработки изображения с предыдущей стадии (рис. 5).

(1) Исходное изображение содержит характерные компоненты текстовых полей. Для отделения их от фона вычисляем пороговое значение одним из методов бинаризации, например, методом Отсу [11]. Данное значение используется для подсчета гистограмм на следующих шагах алгоритма.

(2) Вычисляем вертикальную гистограмму, суммируя лишь те значения, которые меньше порога, так как компоненты текста темнее фона (рис. 6).

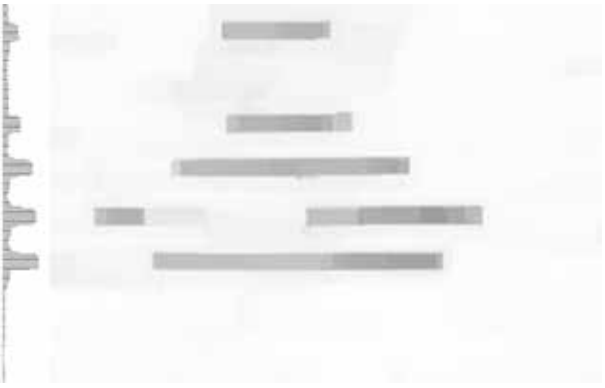

Рис. 6. Вычисление вертикальной гистограммы

(3) Вычисляем верхние и нижние границы строк, содержащих текст, находя выбросы на гистограмме. Для отсечения случайных выбросов используются такие характеристики как минимальные и максимальные размеры высоты символов. Выбросы, размеры которых значительно меньше минимальной вы-

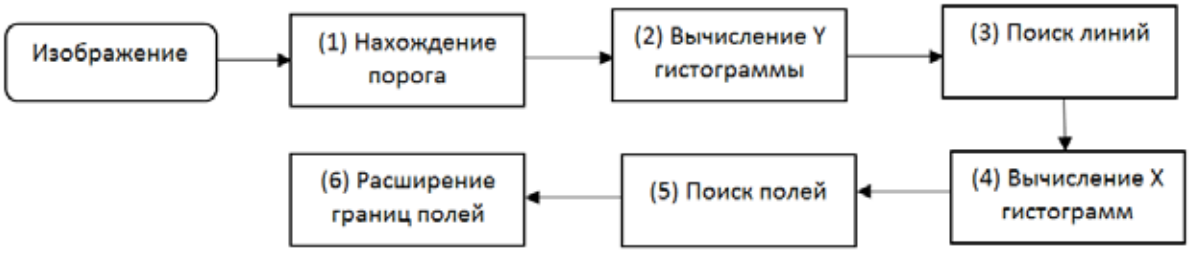

Рис 5. Алгоритм поиска текстовых полей

соты, интерпретируем как случайные. Если же они больше максимального размера, то это случай слипания строк и потребуется дополнительный анализ для нахождения точек разрезания путём определения перепадов уже внутри самого выброса.

(4) Для каждой найденной строки вычисляем горизонтальную гистограмму с использованием порога отсечения, аналогично (2).

Рис. 7. Вычисление горизонтальной гистограммы

(5) Находим границы полей внутри строк с помощью анализа горизонтальной гистограммы для каждой строки (рис. 7). Для отсечения выбросов используется минимальный допустимый размер текстового поля (как минимум ширина одного символа). Для отделения одного текстового поля от другого внутри одной строки установим такую характеристику как минимальное расстояние между полями. Это возможно, ввиду того что в большинстве существующих документов для визуального отличия текстовых полей, содержащих различные данные, используется отделение их друг от друга на расстояние, значительно превышающее размер символа. Таким образом, близкие компоненты сливаются в одно поле, далекие друг от друга поля останутся отдельными полями и в результате получаем прямоугольники найденных полей документа.

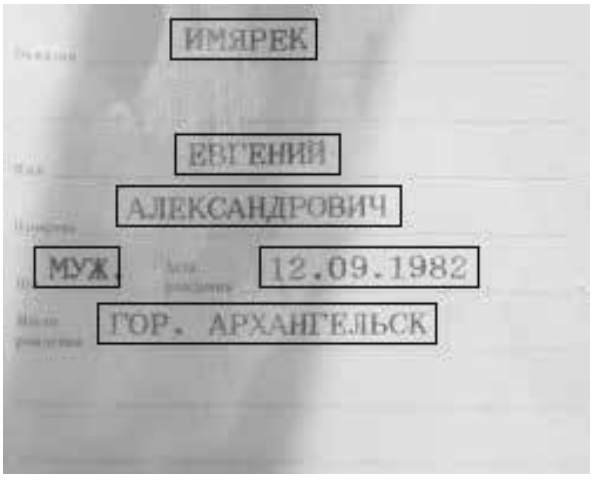

Рис. 8. Результат работы алгоритма поиска текстовых полей

(6) Морфологические операции могут “съедать” часть элементов первых и последних символов полей, а также терять над и под строчные символы (например, умляуты) ввиду их малости. Чтобы избежать этого, границы найденных прямоугольников немного расширяются по горизонтали и вертикали (рис. 8).

Сопоставление найденных полей шаблону зоны документа

Приведенный выше метод поиска текстовых полей выдает как результат некоторый набор координат строк и найденных в них полей, но также необходимо сопоставить найденные текстовые поля атрибутам документа, с учетом возможного частичного их отсутствия. Например, для рассматриваемой зоны паспорта РФ “отчество” может отсутствовать, а количество строк, содержащих данные о “месте рождения”, варьируется от одной до трёх. Для решения этой задачи предлагается производить оценку полученного набора строк и полей на соответствие шаблону, описывающему зону документа. Для создания описания шаблона используется геометрическая модель соответствия строк и полей друг другу. Рассматриваем зону документа как набор строк, а каждую строку как набор полей. Для строк вводится отношение выше/ниже, для полей – левее/правее. Строки и поля могут быть необязательными, также для них могут вводиться дополнительные характеристики, такие как характерные размеры, положение внутри зоны (например, поле прижато к левому краю) и другие, все они используются для оценки полученных данных шаблону зоны. После этого производится рекурсивный перебор всех возможных вариантов сопоставления строк и полей, найденных на изображении, со строками и полями на шаблоне, с оценкой каждого такого сопоставления (рис. 9).

При отсутствии такого сопоставления или очень низкой оценки выдаётся отказ, при наличии несколько вариантов – выбирается вариант с наибольшей оценкой или рассматривается дополнительная задача выбора. Отказ в большинстве случаев означает, что границы документа и его зоны были найдены неверно или исходное изображение слишком низкого качества, а данные – не читаемые.

Рассмотрим оценку качества работы алгоритма на примере набора изображений паспортов РФ, полученных из различных источников: скан, фотография, изображение с веб-камеры или мобильного устройства. Часть изображений представляет собой черно-белые ксерокопии плохого качества.

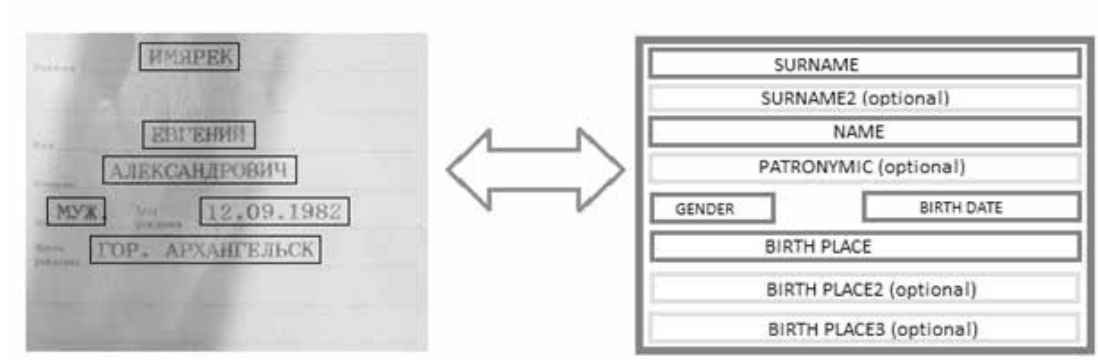

Рис. 9. Найденные поля (слева) и шаблон зоны документа (справа)

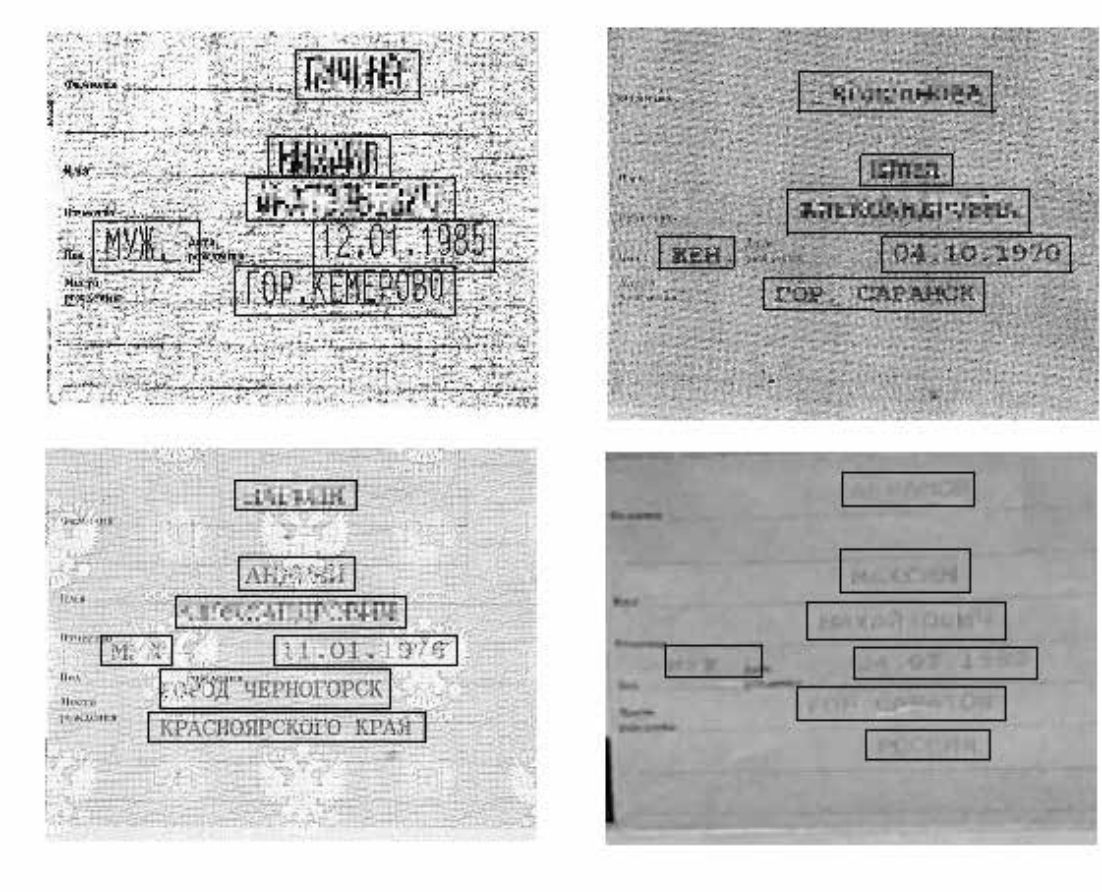

Рис. 10. Поиск полей в сложных случаях

Табл. 1.

Результаты работы на массиве изображений паспортов РФ

	Всего изображений	Определены границы документа	Найдены поля документа
Число	2796	2711	2675
Процент	100%	96,96%	95,67%

Как следует из табл. 1, границы документа были определены на 96,96%, а поля найдены на 95,67% от общего числа изображений. Качество работы самого алгоритма, посчитанное от числа найденных документов, составляет 2675/2711 = **98,67%**. Стоит отметить, что алгоритм достаточно успешно справляется со сложными случаями на зашумленных или плохо читаемых документах (рис. 10).

Рис. 11. Результаты поиска полей на различных документах

Заключение

Предложенный алгоритм поиска текстовых полей на документах показал свою работоспособность и высокое качество. Важным его плюсом является достаточная универсальность и высокая скорость работы, а также устойчивость к фону документов, освещению, искажениям зон полей. На данный момент алгоритм используется для поиска полей на более чем 70 различных документах со всего мира, на рис. 11 приведены примеры таких документов.

Литература

1. *Полевой Д., Булатов К., Скорюкина Н., Чернов Т., Арлазаров В.В., Шешкус А.* «Ключевые аспекты распознавания документов с использованием малоразмерных цифровых камер», Вестник РФФИ, 2016, № 4 (92), стр. 97-108.
2. *Арлазаров В.В., Жуковский А., Кривцов В., Николаев Д., Полевой Д.* «Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видео камер для распознавания документов», Информационные технологии и вычислительные системы, 2014, №3, стр. 71-81.
3. *Junga K., Kimb K.I., Jain A.K.,* “Text information extraction in images and video: a survey”, Pattern Recognition, 2004, pp. 977-997.
4. *Zhukovsky A., Arlazarov V., Postnikov V., Krivtsov V.* «Segments Graph-Based Approach for Smartphone Document Capture» Proceedings SPIE. Eighth International Conference on Machine Vision (ICMV 2015), 2015, V. 9875, 98750P, pp. 1-7.
5. *Чернов Т.С.* «Детектирование и фильтрация бликов в задачах распознавания документов с мобильных устройств», Труды Института системного анализа РАН, 2017, Т. 67, № 1, стр. 67-74.
6. *Wu J.C., Hsieh J.W., Chen Y.S.,* “Morphology-based text line extraction”, Machine Vision and Applications, 2008, pp. 195-207.
7. *Rodolfo P.* dos Santos, Gabriela S. Clemente, Tsang Ing Ren, George D.C. Calvalcanti, “Text Line Segmentation Based on Morphology and Histogram Projection”, Proceedings of ICDAR ‘09, 2009, pp. 651-655.
8. *M. van Herk.* “A Fast Algorithm for Local Minimum and Maximum Filters on Rectangular and Octagonal Kernels”, Pattern Recognition Letters, 1992, pp. 517-521.
9. *Limonova E., Terekhin A., Nikolaev D., Arlazarov V.V.* «Fast Implementation of Morphological Filtering Using ARM NEON Extension», International Journal of Applied Engineering Research, 2016, V. 11, № 24, pp. 11675-11680.
10. *Nikolaev D.P., Karpenko S.M., Nikolaev I.P. and Nikolayev P.P.,* “Hough Transform: Underestimated

Tool in the Computer Vision Field", Proceedings of ECMS '08, 2008, pp. 238-246.
11. *Otsu N.*, «A threshold selection method from gray-level histograms», IEEE Trans. Sys., Man., Cyber. 9, 1979, pp. 62-66.

Слугин Дмитрий Геннадьевич. Научный сотрудник ИСА ФИЦ ИУ РАН. Окончил в 2000 г. МГУ. Количество печатных работ: 8. Область научных интересов: системный анализ, алгоритмы обработки изображений, распознавание образов. E-mail: slugindm@gmail.com

Арлазаров Владимир Викторович. Заведующий лабораториуй ИСА ФИЦ ИУ РАН. К.т.н. Окончил в 1999 г. НИТУ «МИСиС». Количество печатных работ: 26. Область научных интересов: распознавание образов, обработка изображений, системы массового обслуживания. E-mail: vva777@gmail.com

Text fields extraction based on image processing

D.G. Slugin, V.V. Arlazarov

Abstract. This paper presents text fields extraction algorithm for Russian citizen passport. The algorithm is based on image processing, such as morphology, and template matching. The experimental results on the dataset, contains a large number of images from video streams, scanners and photo cameras, show effectiveness and very good score rate of the proposed algorithm. The method may be generalized on the large set of documents such as ID cards, driving licenses, visas and etc.

Keywords: text fields extraction, document recognition, video stream, image processing, morphology, Hough transform, template matching

References

1. *Polevoy D., Bulatov K., Skorykina N., Chernov T., Arlazarov V., Sheshkus A.* «Key Aspects of Document Recognition Using Small Digital Cameras», Herald of the RFBR, 2016, №4 (92), pp. 97-108.
2. *Arlasarov V.V., Zhukovsky A.E., Krivtsov V.E., Nikolaev D.P., Polevoy D.V.* « Analysis of features of the use of fixed and mobile small-sized digital video camera for OCR », Information Technologies and Computing Systems, 2014, №3, pp. 71-81.
3. *Junga K., Kimb K.I., Jain A.K.,* "Text information extraction in images and video: a survey", Pattern Recognition, 2004, pp. 977-997.
4. *Zhukovsky A., Arlazarov V., Postnikov V., Krivtsov V.* «Segments Graph-Based Approach for Smartphone Document Capture» Proceedings SPIE. Eighth International Conference on Machine Vision (ICMV 2015), 2015, V. 9875, 98750P, pp. 1-7.
5. *Chernov T. S.* «Glare detection and filtering in document recognition tasks on mobile devices», ISA RAS proceedings, 2017, Vol. 67, № 1, pp. 67-74.
6. *Wu J.C., Hsieh J.W., Chen Y.S.,* "Morphology-based text line extraction", Machine Vision and Applications, 2008, pp. 195-207.
7. *Rodolfo P.* dos Santos, Gabriela S. Clemente, Tsang Ing Ren, George D.C. Calvalcanti, "Text Line Segmentation Based on Morphology and Histogram Projection", Proceedings of ICDAR '09, 2009, pp. 651-655.
8. *M. van Herk.* "A Fast Algorithm for Local Minimum and Maximum Filters on Rectangular and Octagonal Kernels", Pattern Recognition Letters, 1992, pp. 517-521.
9. *Limonova E., Terekhin A., Nikolaev D., Arlazarov V.V.* «Fast Implementation of Morphological Filtering Using ARM NEON Extension», International Journal of Applied Engineering Research, 2016, V. 11, № 24, pp. 11675-11680.
10. *Nikolaev D.P., Karpenko S.M., Nikolaev I.P. and Nikolayev P.P.,* "Hough Transform: Underestimated Tool in the Computer Vision Field", Proceedings of ECMS '08, 2008, pp. 238-246.
11. *Otsu N.*, «A threshold selection method from gray-level histograms», IEEE Trans. Sys., Man., Cyber. 9, 1979, pp. 62-66.

Slugin D. G. Federal Research Centre "Computer Science and Control Systems" The Institute for Systems Analysis of Russian Academy of Sciences, Moscow, Russia. Researcher. slugindm@gmail.com

Arlazarov V.V. Federal Research Centre "Computer Science and Control Systems" The Institute for Systems Analysis of Russian Academy of Sciences, Moscow, Russia. Head of laboratory. va777@gmail.com