УДК 004.021

АЛГОРИТМ ИЗВЛЕЧЕНИЯ ТАБЛИЧНОЙ ИНФОРМАЦИИ ИЗ ОТСКАНИРОВАННЫХ ДОКУМЕНТОВ НА ПРИМЕРЕ СПРАВКИ БЮРО ТЕХНИЧЕСКОЙ ИНВЕНТАРИЗАЦИИ О СОСТОЯНИИ ЗДАНИЯ

Качалин В.С., Панов Ю.Н., Калугин А.В.

ФГБОУ ВО «Московский авиационный институт (национальный исследовательский университет)», Москва, e-mail: vasiliy.kachalin@gmail.com

Автоматизация извлечения информации из таблиц отсканированного документа требует особого подхода, нежели обработка простого текста, и решения нескольких задач: исправление наклона отсканированного документа, обнаружение таблицы в теле документа, выделение ячеек таблицы, чтение информации из ячеек. В данной работе предлагается алгоритм по извлечению информации из таблиц, ячейки которых стоят в последовательности «характеристика – значение», на примере справки БТИ о состоянии здания. Устранение наклона отсканированного документа выполняется с помощью преобразования линий строк текста в сплошные линии, которые обладают своим углом наклона, на который поворачивается документ. Обнаружение таблицы происходит с помощью вычитания маски из изначального изображения документа. Выделение ячеек осуществляется с помощью поиска контуров. Чтение информации из ячеек выполняется средствами системы оптического распознавания символов, причем за один процесс чтения считываются две ячейки, первая содержит информацию о наименовании характеристики, вторая – ее значение. Вся извлеченная информация из таблицы по результатам работы алгоритма представлена в виде хеш-таблицы. Также было проведено тестирование алгоритма, которое подтвердило его работоспособность и позволило определить время выполнения, использование памяти и нагрузку на процессор на нескольких стендах.

Ключевые слова: извлечение, данные, таблица, форма 5, справка БТИ о состоянии здания, БТИ, наклон документа, отсканированный документ

ALGORITHM FOR EXTRACTING TABULAR INFORMATION FROM SCANNED DOCUMENTS ON THE EXAMPLE OF THE TECHNICAL INVENTORY BUREAU CERTIFICATE ON THE CONDITION OF THE BUILDING

Kachalin V.S., Panov Yu.N., Kalugin A.V.

Moscow Aviation Institute (National Research University), Moscow, e-mail: vasiliy.kachalin@gmail.com

Automating the extraction of information from the tables of a scanned document requires a special approach, rather than processing plain text, and solving several tasks: correcting the tilt of the scanned document, detecting a table in the body of the document, selecting table cells, reading information from cells. In this paper, an algorithm is proposed for extracting information from tables whose cells are in the sequence “characteristic – value”, using the example of the BTI certificate on the condition of the building. Elimination of the slope of the scanned document is performed by converting lines of text lines into solid lines, which have their own angle of inclination, by which the document is rotated. The table is detected by subtracting the mask from the original image of the document. The selection of cells is carried out by searching for contours. Reading information from cells is performed by means of an optical character recognition system, and two cells are read in one reading process, the first contains information about the name of the characteristic, the second contains its value. All extracted information from the table based on the results of the algorithm is presented in the form of a hash table. The algorithm was also tested, which confirmed its operability, and allowed to determine the execution time, memory usage and CPU load on several devices.

Keywords: extracting, data, table, form 5, BTI certificate on the condition of the building, BTI, document tilt, scanned document

При наличии большого объема бумажной документации недвижимого имущества возникает необходимость автоматизации процесса осмысленного извлечения различной информации. Под осмысленным извлечением информации в данной работе подразумевается отличающаяся друг от друга работа с разными ее типами. Если информация является сплошным текстом, то ее обработка не вызывает каких-либо затруднений, однако если информация представляет собой табличные данные, то возникают сложности при ее анализе: необходимо обнаружить таблицу на отсканированной странице документа, выделить ячейки таблицы для извлечения информации из каждой по отдельности, понять, что представляет собой извлеченная из ячейки информация.

Цель исследования заключается в разработке алгоритма, который решает задачу по автоматическому извлечению табличных данных из отсканированных документов, таблицы которых имеют следующую структуру: отдельная осмысленная единица информации содержится в двух ячейках, которые идут друг за другом. В первой ячейке представлено наименование характеристики, а во второй ячейке – ее значение. Такой струк-

туре отвечает справка бюро технической инвентаризации (БТИ) о состоянии здания (также известная как форма 5), на примере которой продемонстрированы шаги алгоритма.

Материалы и методы исследования

В качестве технологий в данном алгоритме используются:

1. Метод Оцу – алгоритм определения порога бинаризации для изображения в оттенках серого. Метод весьма прост, но в то же время стабилен, из-за чего получил широкое применение в области обработки изображений [1].

2. Алгоритм RLSA (Run Length Smoothing Algorithm) – алгоритм, позволяющий преобразовывать пиксели на основе окрестности.

3. Преобразование Хафа – представляет собой алгоритм, который позволяет обнаруживать на изображении определенные фигуры.

4. Фильтр Кэнни – оператор, позволяющий выделять на изображении границы объектов. Обладает хорошей производительностью и на текущий момент является стандартом в обработке изображений [2].

5. Алгоритм Satoshi Suzuki – алгоритм, предназначенный для поиска контуров на изображениях.

6. Tesseract OCR – бесплатная компьютерная технология для распознавания написанного текста, которая показывает отличные результаты при работе с текстом, написанным кириллическими символами [3].

Прежде чем приступать к описанию работы алгоритма, необходимо декомпозировать задачу извлечения табличных данных на более мелкие задачи. Так, извлечение табличной информации должно содержать следующие этапы:

1. Устранение наклона документа – отсканированный документ может обладать некоторым углом наклона, который в свою очередь будет влиять на конечный результат алгоритма.

2. Обнаружение таблицы в отсканированном документе и выделение ячеек.

3. Чтение информации из ячеек таблицы.

Перед началом всех действий изображение отсканированного документа необходимо сделать черно-белым, так как это позволяет значительно уменьшить количество лишней информации. Для этого необходимо перевести отсканированный документ в оттенки серого, а затем с помощью порогового значения и метода Оцу преобразовать документ в черно-белый.

1. Устранение наклона отсканированного документа

Очевидно, что для того чтобы исправить наклон отсканированного документа, необходимо знать его угол. Существует множество методов, которые позволяют решить задачу вычисления угла наклона, например метод анализа профиля проекции, метод основанный на преобразовании Хафа, метод ближайших соседей [4].

Хорошим методом для определения угла наклона документа будет использование преобразования Хафа. Однако для начала надо подготовить отсканированный документ. Сначала с помощью RLSA алгоритма строки текста на черно-белом отсканированном документе преобразуются в широкие линии (RLSA заменяет фоновые пиксели в бинаризированном изображении пикселями переднего плана, если количество фоновых пикселей в окрестности не превышает порогового значения, т.е. удаляет маленькие окрестности пикселей [5]). Затем с помощью морфологической операции эрозия с документа удаляется лишняя информация (например, линии таблиц или пометки). В таком виде преобразование Хафа может неправильно определить линии текста, чтобы этого избежать, нужно с использованием оператора Кэнни сделать их «тоньше». Затем с помощью преобразования Хафа определяются линии (документ с нанесенными линиями строк текста представлен на рис. 1), а затем вычисляется их наклон. Так как линии могут быть распознаны некорректно, стоит принять за угол наклона документа медианное значение углов наклона всех линий. Исправление наклона документа заключается в повороте документа в обратную сторону на величину найденного угла.

2. Обнаружение таблицы на документе и выделение ячеек

Сам процесс обнаружения таблицы представляет собой формирование маски, которая впоследствии позволяет из исходного отсканированного документа получить скелет таблицы, то есть саму таблицу без ее содержимого. Далее скелет таблицы позволяет вычленить отдельные ячейки и работать с каждой самостоятельно.

На этапе обнаружения таблицы также требуется применение RLSA. Данный алгоритм позволяет преобразовать строки текста внутри таблицы в сплошные широкие линии, которые используются в маске. Далее, с помощью морфологической операции эрозия удаляются мелкие дефекты сканирования, например пятно на стекле сканирующего устройства, и контур таблицы. Однако из-за этого линии строк текста уменьшаются в размере, чтобы вернуть их предыдущий размер, используется морфологическая операция дилатация. Таким образом формируется маска, которая представлена на рис. 2.

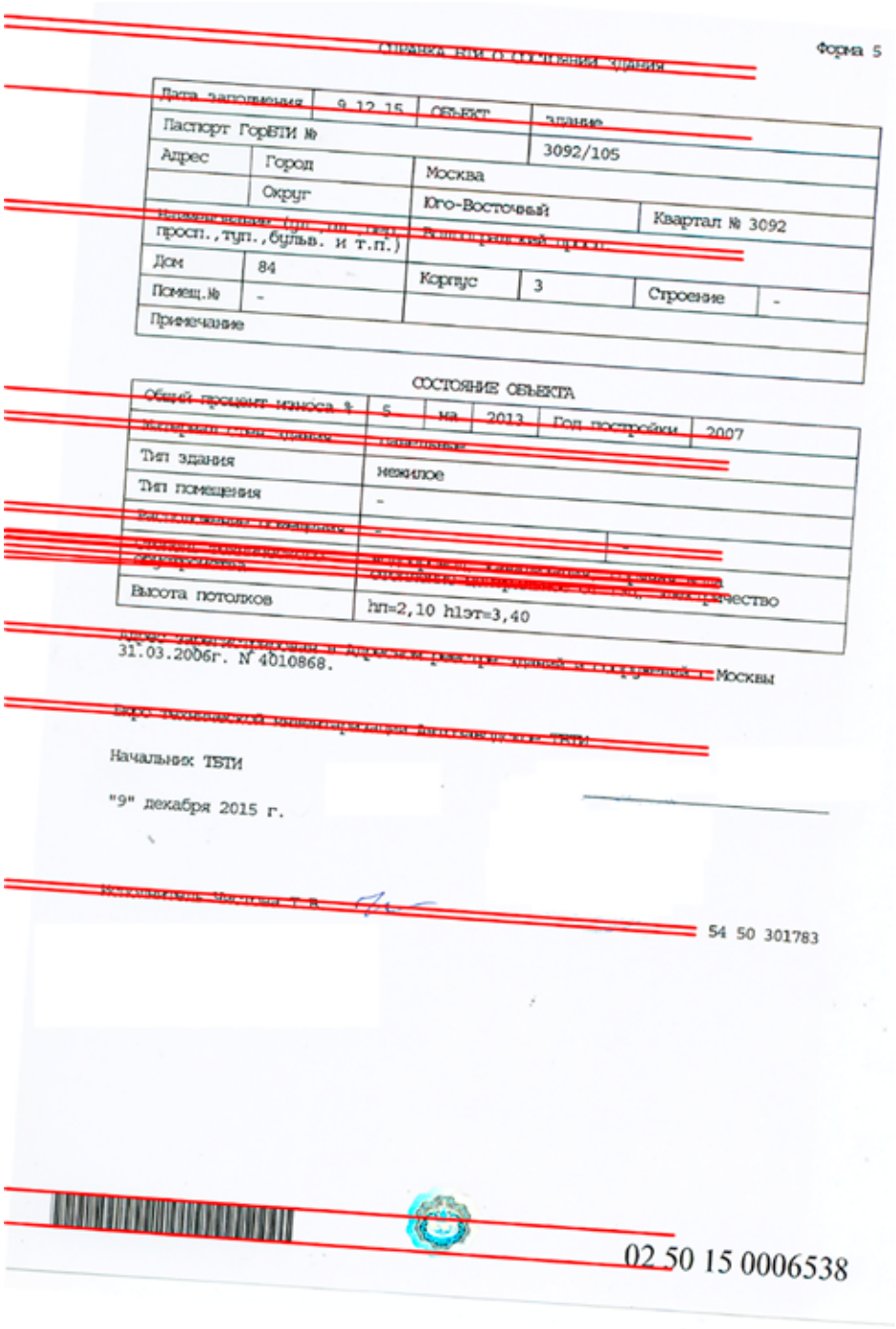

Рис. 1. Документ с нанесенными линиями строк текста

Затем из изначального изображения отсканированного документа вычитается маска, что позволяет оставить на изображении документа только скелет таблицы. На рис. 3 изображен документ, к которому применяется вышеописанный способ, результат которого представлен на рис. 4.

В найденной таблице определяются ячейки, для этого используется метод поиска контуров, предложенный Satoshi Suzuki [6]. Также определить ячейки можно с помощью другого метода – оконного преобразования Хафа [7].

3. Чтение информации из ячеек таблицы

Сама информация должна сохраняться в хеш-таблице, где ключом является наименование характеристики, а значением – ее значение.

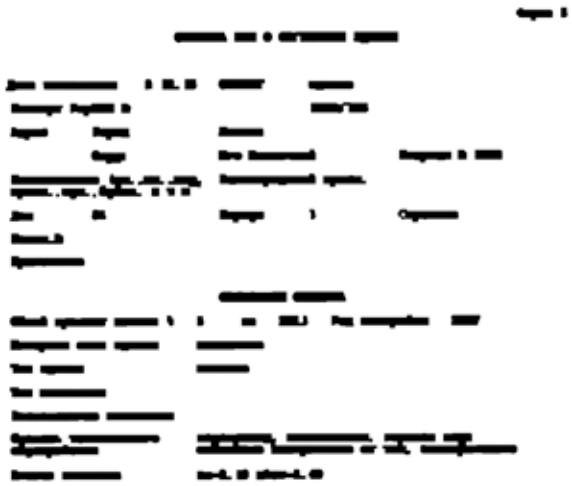

Рис. 2. Маска для выделения таблицы

Дата заполнения	9.12.15	ОБЪЕКТ	здание		
Паспорт ГорБТИ №			3092/105		
Адрес	Город	Москва			
	Округ	Юго-Восточный		Квартал № 3092	
Наименование (ул.,пл.,пер, просп.,туп.,бульв. и т.п.)		Волгоградский просп.			
Дом	84	Корпус	3	Строение	-
Помещ.№	-				
Примечание					

СОСТОЯНИЕ ОБЪЕКТА

Общий процент износа %	5	на	2013	Год постройки	2007
Материал стен здания	панельные				
Тип здания	нежилое				
Тип помещения	-				
Расположение помещения	-			-	
Степень технического обустройства	водопровод, канализация, горячая вода отопление центральное от тэц, электричество				
Высота потолков	hп=2,10 h1эт=3,40				

Рис. 3. Оригинальная таблица

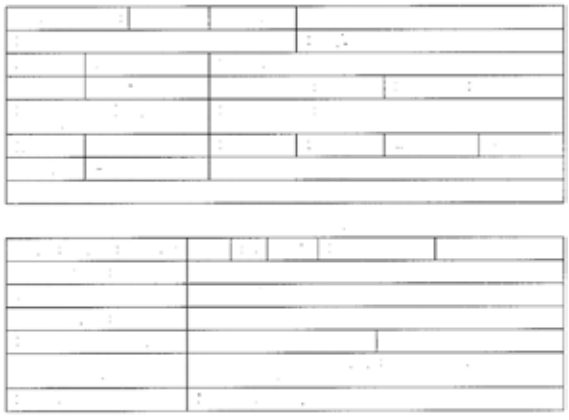

Рис. 4. Полученный «скелет» таблицы

Предполагается, что обнаруженные контуры, которые содержат вписанные в себя другие контуры, не являются ячейками таблицы, поэтому при анализе игнорируются. Далее идет проход по всем ячейкам в направлении сверху вниз и слева направо, в ходе которого с помощью системы оптического распознавания символов Tesseract OCR определяются интересующие данные и их метаинформация. Делается это следующим образом: берется ячейка, которая должна быть обработана, из нее считывается информация, которая определяет метаинформацию, и сразу же читается следующая ячейка, которая определяет саму информацию, затем обе прочитанных ячейки записываются в хеш-таблицу, как пара «ключ – значение».

Вышеописанный алгоритм представлен на рис. 5.

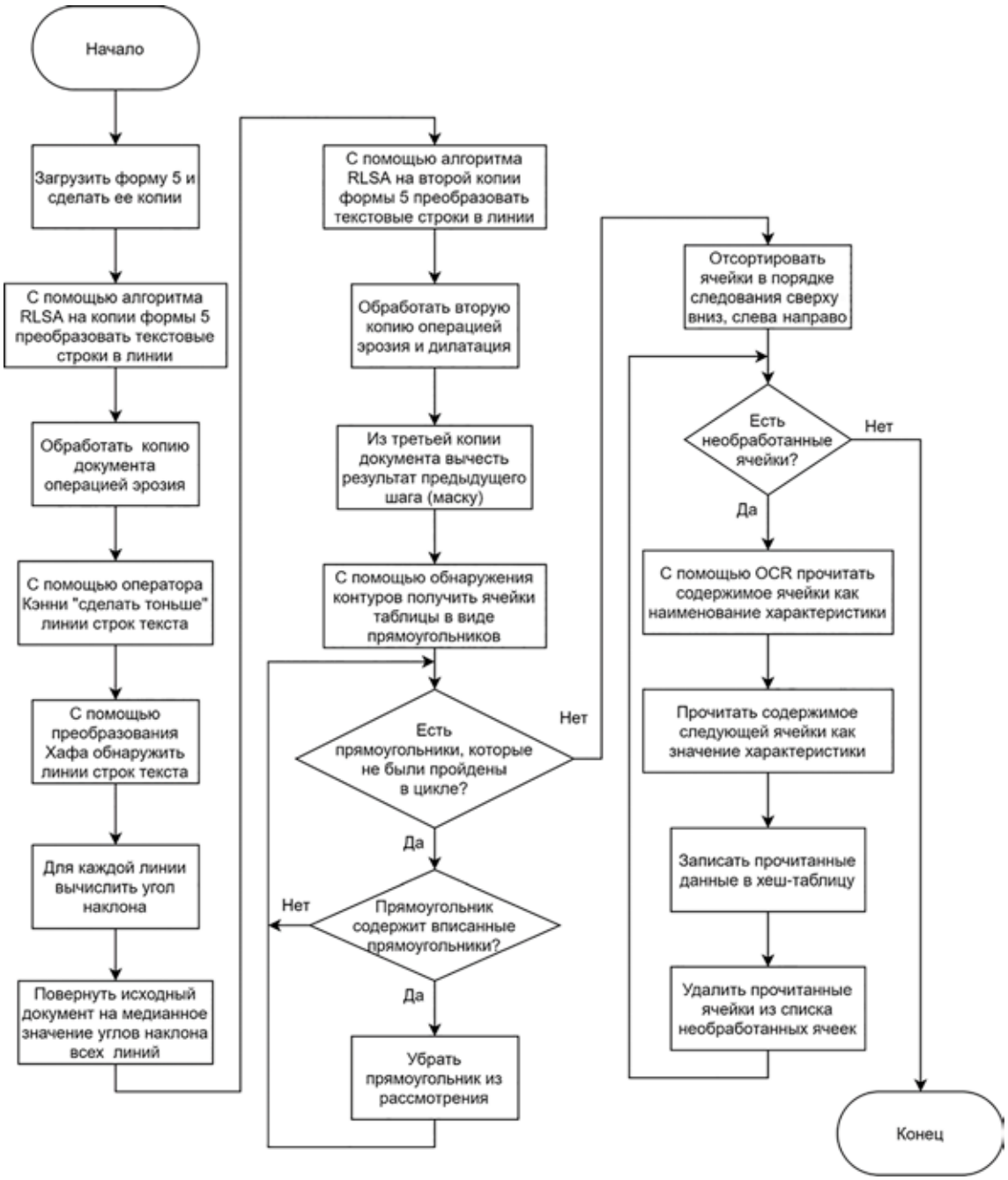

Рис. 5. Алгоритм анализа формы 5

Результаты тестирования предложенного алгоритма

Время выполнения, с		Использование памяти, Мб		Нагрузка на процессор, %	
Стенд № 1	Стенд № 2	Стенд № 1	Стенд № 2	Стенд № 1	Стенд № 2
24,6	13,8	14,5	14,6	9,9	9,7

Результаты исследования и их обсуждение

Предложенный алгоритм был представлен в виде программы, написанной на языке Python версии 3.7.3, и протестирован на 50 отсканированных справках БТИ о состоянии здания, которые были созданы специально для проверки алгоритма и не являются реальными документами. Тестирование проводилось на двух стендах со следующими характеристиками:

1. Стенд № 1:

− центральный процессор: Intel Core i5 8500 с тактовой частотой 3 ГГц;

− оперативная память: DDR4 24 Гб с частотой 1 ГГц.

2. Стенд № 2:

− центральный процессор: AMD Ryzen 7 3700X с тактовой частотой 3,59 ГГц;

− оперативная память: DDR4 32 Гб с частотой 1,6 ГГц.

В таблице представлены усредненные результаты проведенного тестирования.

На всех 50 отсканированных справках БТИ о состоянии здания были правильно определены ячейки таблицы, однако из-за выборки маленького размера нельзя однозначно утверждать, что данный алгоритм с вероятностью 100 % не потеряет табличную информацию, но можно отметить, что все же данная вероятность стремится к этому числу.

Результат данной работы в виде предложенного алгоритма имеет практическую ценность – данный алгоритм можно использовать при работе с документами, содержащими таблицы, ячейки которых стоят в последовательности «характеристика – значение», что позволяет быстро извлекать информацию, при этом не теряя метаданные для каждой прочитанной характеристики. Также при некоторой модификации алгоритма его можно использовать как универсальный метод по извлечению информации из таблицы любого вида.

Научная новизна работы заключается в том, что предложен алгоритм, который решает задачу извлечения информации из таблиц, ячейки которых стоят в последовательности «характеристика – значение».

В будущем планируется разработка аналогичных алгоритмов для анализа других документов, относящихся к области недвижимого имущества.

Заключение

Предложенный алгоритм справляется с поставленной перед ним задачей – извлечение информации из таблиц, ячейки которых стоят в последовательности «характеристика – значение», и на выходе предоставляет хеш-таблицу, содержащую извлеченную информацию.

Список литературы

1. Huang M., Yu W., Zhu D. An Improved Image Segmentation Algorithm Based on the Otsu Method. 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. 2012. P. 135–139. DOI: 10.1109/SNPD.2012.26.

2. Xu Q., Varadarajan S., Chakrabarti C., Karam L.J. A Distributed Canny Edge Detector: Algorithm and FPGA Implementation. IEEE Transactions on Image Processing. 2014. Vol. 23. No. 7. P. 2944–2960. DOI: 10.1109/TIP.2014.2311656.

3. Качалин В.С., Панов Ю.Н., Попов Н.-Л.Э. Сравнительный анализ различных систем оптического распознавания символов при работе с текстом, написанным с помощью кириллического алфавита // Современные наукоемкие технологии. 2022. № 8. С. 65–70. DOI: 10.17513/snt.39268.

4. Al-Khatatneh A.M., Pitchay S.A., Al-qudah M. A Review of Skew Detection Techniques for Document. 17th UKSIM-AMSS International Conference on Modelling and Simulation. 2015. P. 316–321. DOI: 10.1109/UKSim.2015.73.

5. Tian Y., Gao C., Huang X. Table Frame Line Detection in Low Quality Document Images Based on Hough Transform. The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014). 2014. P. 818–822. DOI: 10.1109/ICSAI.2014.7009397.

6. Suzuki S., Keiichi A. Topological structural analysis of digitized binary images by border following . Computer Vision, Graphics, and Image Processing. 1985. Vol. 30. P. 32–46. DOI: 10.1016/0734-189X(85)90016-7.

7. Jung C.R., Schramm R. Rectangle Detection based on a Windowed Hough Transform. Proceedings. 17th Brazilian Symposium on Computer Graphics and Image Processing. 2004. P. 113–120. DOI: 10.1109/SIBGRA.2004.1352951.