лучших решений для автоматизации промышленных процессов. С CODESYS можно создавать высокоэффективные автоматизированные системы для различных отраслей, от машиностроения до энергетики и логистики, что делает его востребованным инструментом на рынке.

Список использованных источников

1. CODESYS Group. CODESYS Development System V3.5 SP16 Documentation. CODESYS Group, 2022.

2. Heß, Manfred и Baumeister, Thomas. Introduction to CODESYS V3 for PLC Programming. CODESYS Group, 2021.

УДК 004.021

В.С. Качалин
Московский авиационный институт
Москва, Россия

УЛУЧШЕННЫЙ АЛГОРИТМ ИЗВЛЕЧЕНИЯ ТАБЛИЧНОЙ ИНФОРМАЦИИ ИЗ ОТСКАНИРОВАННЫХ ДОКУМЕНТОВ

Аннотация. *В статье предложено улучшение алгоритма извлечения табличной информации из отсканированных документов, которое позволяет применять алгоритм на таблицы, у которых каждая строчка описывает отдельный логический объект, а каждый столбец обозначает характеристики объектов.*

V.S. Kachalin
Moscow Aviation Institute
Moscow, Russia

IMPROVED ALGORITHM FOR EXTRACTING TABULAR INFORMATION FROM SCANNED DOCUMENTS

Abstract. *The article proposes an improvement in the algorithm for extracting tabular information from scanned documents, which allows to apply the algorithm to tables in which each row describes a separate logical object, and each column indicates the characteristics of objects.*

В современном мире, в котором отчетливо прослеживается тенденция к цифровизации большинства сфер жизни, большое количество неоцифрованных бумажных документов является наследием прошлого. Множество организаций переводят свой

документооборот в электронную среду, однако перенос уже имеющихся документов может оказаться проблемой из-за их большого количества. Самое простое решение – можно просто отсканировать документы и хранить их в таком виде на компьютере или сервере, но в этом случае невозможно выполнять поиск по содержимому документов или каким-либо образом взаимодействовать с ним. Необходимо извлекать информацию из документов и анализировать ее, чтобы обеспечить возможность более простой дальнейшей работы с оцифрованным документом.

С простым сплошным текстовым документом могут справиться системы оптического распознавания символов (OCR), однако если документ содержит табличные данные, то возникают сложности с логическим сопоставлением данных – считанная информация может быть неправильно интерпретирована, а это влечет за собой проблемы при последующей работе с документом сотрудников организации. Уже существует алгоритм, который позволяет извлекать табличную информацию из отсканированных документов [1], однако он имеет ограничение – данный алгоритм применим только для извлечения и анализа информации из таблиц, имеющих определенную структуру – у таких таблиц информация должна быть записана в двух соседних ячейках в виде «характеристика» – «значение». Цель данной работы – улучшить имеющийся алгоритм, чтобы его можно было применять и для таблиц, имеющих другую структуру, у которых каждая строчка описывает отдельный логический объект, а каждый столбец обозначает характеристики объектов. Пример рассматриваемой таблицы представлен на рис. 1.

В улучшенном алгоритме этапы, предшествующие чтению ячеек, идентичные оригинальным:

1. Устранение наклона отсканированного документа с помощью преобразования Хафа;
2. Формирование маски таблицы путем применения алгоритма RLSA и морфологических операций эрозия и дилатация;
3. Определение ячеек таблицы с помощью вычитания маски из изображения таблицы и применения метода поиска контуров Satoshi Suzuki [2].

Этаж	№	Помещение	Площадь			Высота помещения по обмеру
			жилая	вспом.	летние	
2	1	Аудитория		11,4		250
	2	Лаборатория		15,6		
	3	Коридор		50,0		
		Итого		77,0		
		Общая площадь	77,0			

Рис. 1 - Пример отсканированной офисным сканером таблицы, для которой применимо предлагаемое улучшение алгоритма

В рассматриваемых таблицах всегда присутствует шапка, содержимое которой характеризует значения в соответствующих столбцах. Как правило, шапка таблицы состоит из двух строк. Для последующего анализа необходимо прочитать содержимое шапки таблицы и сопоставить прочтенную информацию с координатами по оси X. Делается это следующим образом: массив, состоящий из ячеек таблицы, сортируется по возрастанию значения Y. Сортировка производится по возрастанию, потому что в цифровых изображениях координата Y увеличивается при перемещении сверху вниз. Такая сортировка массива позволяет распределить ячейки сверху вниз, причем ячейки, находящиеся на одной высоте, но с разными координатами X, будут следовать друг за другом. Это гарантирует что ячейки, относящиеся к одной строчке, будут расположены в массиве в виде непрерывной последовательности, что обеспечивает быстрый анализ таблицы. Затем необходимо прочитать с помощью системы OCR (например, Tesseract OCR) первую строчку таблицы, т.е. шапку, и записать каждую ячейку в виде ассоциативного массива, в котором бы хранились координаты ячейки шапки и прочитанный в них текст для дальнейшего определения метаинформации прочитанных значений ячеек таблицы. Конец строчки таблицы определяется изменением координаты Y анализируемой ячейки, т.е. если координата Y равна координате предыдущей ячейки, то эти две ячейки относятся к одной строчке. Если же координата Y текущей ячейки не равна координате предыдущей ячейки, то считается, что началась читаться новая строчка.

Стоит отметить, что текст в ячейке шапки может быть повернут на 90 градусов, из-за чего прочесть информацию не получится. Чтобы

учесть данную ситуацию необходимо перед чтением ячейки запустить Tesseract OCR в режиме OSD (Orientation and Script Detection – режим определения ориентации и алфавита текста). Однако Tesseract может не определить наклон из-за малого количества символов, чтобы избежать такого поведения можно создать пустое изображение и несколько раз по горизонтали и вертикали скопировать в него проверяемую ячейку. Это позволит увеличить количество символов и не изменит наклона текста. Затем при необходимости повернуть ячейку на нужный угол, определенный с помощью Tesseract OCR.

После прочтения первой строчки шапки необходимо прочитать ее вторую строчку. Для этого читаются ячейки с соответствующей координатой Y и проверяется их позиция по оси X относительно горизонтальных краев ячеек первой строчки. В случае расположения ячейки между началом и концом одной из ячеек первой строчки по оси X в соответствующий ассоциативный массив добавляется информация о наличии ячейки второй строчки, которая включает в себя текст и координаты только что прочитанной ячейки. Сохранять такую информацию можно в виде вложенного ассоциативного массива.

Далее последовательно читаются ячейки, принадлежащие третьей строчке (как было сказано ранее, принадлежность определяется одинаковыми значениями координаты Y). Координата по оси X прочтенной ячейки сверяется с такой же координатой у ячеек шапки. В случае совпадения координат и наличия у ячейки шапки вложенного ассоциативного массива, данная координата сверяется с соответствующими координатами ячеек второй строчки шапки. Таким образом определяется метаинформация для ячейки третьей строчки. После прочтения строчки данные записываются в ассоциативный массив, представляющий собой прочитанную строчку (которая представляет собой отдельный объект). В качестве ключа такого массива выступает метаинформация, а в качестве значения – непосредственно сама информация. Для всех последующих строчек данная процедура повторяется.

Предложенный алгоритм представлен в виде блок-схемы на рис. 2.

По итогу выполнения данного алгоритма получаются ассоциативные массивы, которые характеризуют строки проанализированной таблицы. Для удобства работы с такими ассоциативными массивами, их можно поместить в обычный массив типа «Список».

Данное улучшение алгоритма было протестировано на 100 таблиц аналогичных представленной ранее таблице на рисунке 1. В процессе тестирования сравнивалось содержимое отсканированных

таблиц и их распознанные алгоритмом представления, также уделялось внимание тому, чтобы в каждом ассоциативном массиве, присутствовала полная информация, находящаяся в соответствующей строчке таблицы и не было потерь данных. В результате тестирования были правильно сформированы все ассоциативные массивы, относящиеся ко всем 100 тестовым таблицам, и не было допущено ни одной ошибки, что свидетельствует о полной работоспособности данного улучшения.

В качестве заключения можно отметить практическую значимость данной работы, которая заключается в том, что предложенное улучшение существующего алгоритма извлечения табличной информации из отсканированных документов, расширяет область его применения, при этом не снижая его точность.

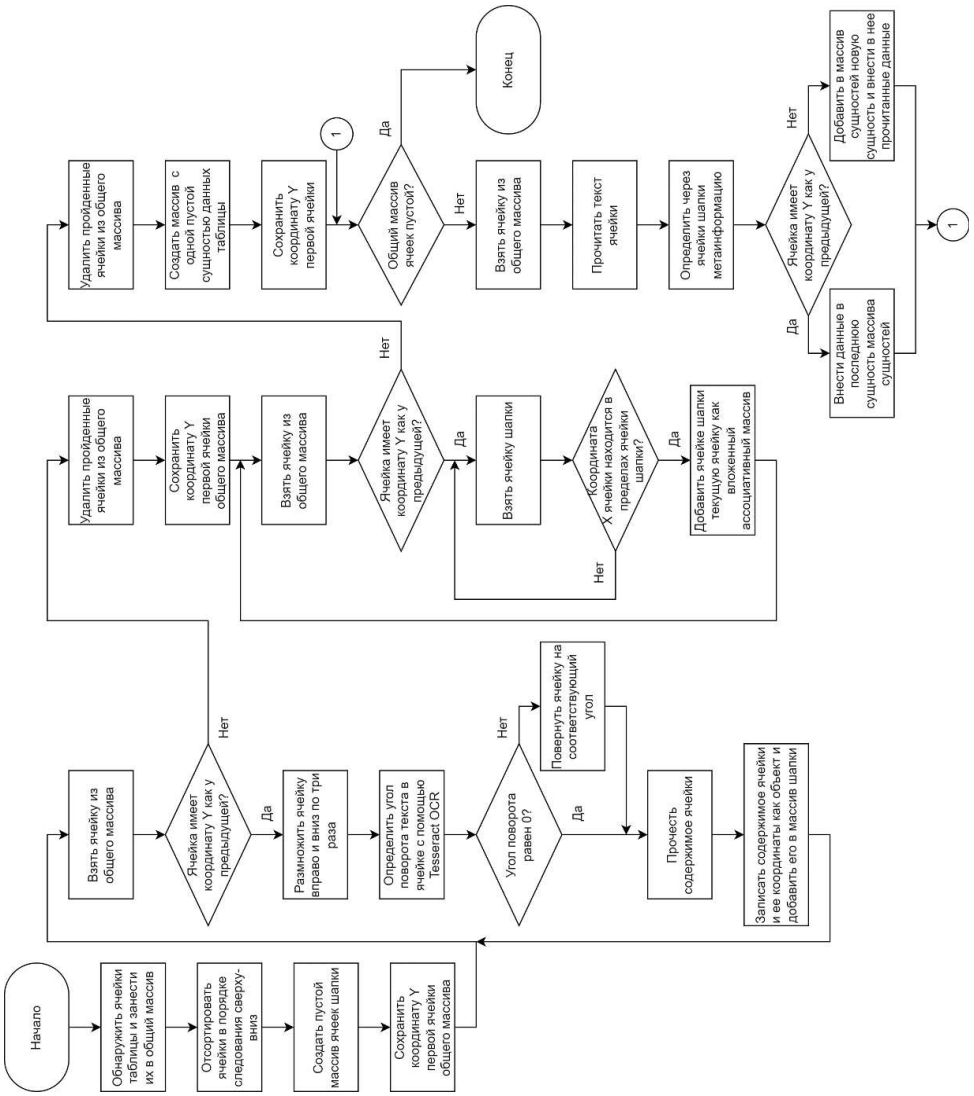

Рис. 2 - Улучшенный алгоритм извлечения табличной информации из отсканированных документов

Список использованных источников

1. Качалин В.С., Панов Ю.Н., Калугин А.В. Алгоритм извлечения табличной информации из отсканированных документов на примере справки бюро технической инвентаризации о состоянии здания // Современные наукоемкие технологии. 2022. № 11. С. 46-51.
2. 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.

УДК 330.336: 330.332

Т.В. Каштелян
Белорусский государственный технологический университет
Минск, Беларусь

ОНЛАЙН-ТОРГОВЫЕ ПЛОЩАДКИ ЭКОСИСТЕМНЫХ УСЛУГ В ЭКОНОМИЧЕСКОМ РАЗВИТИИ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Аннотация. *В статье на основе анализа проблем лесного сектора Беларуси обосновывается необходимость создания онлайн-торговых площадок экосистемных услуг. В связи с институциональными изменениями и технологическими возможностями демонстрируется важность вовлечения различных отраслей в экологическую экономику.*

T.V. Kashtelyan
Belarusian State Technological University
Minsk, Belarus

ONLINE TRADING MARKETPLACES OF ECOSYSTEM SERVICES IN ECONOMIC DEVELOPMENT: PROBLEMS AND PROSPECTS

Abstract. *The article based on the analysis of the problems of the forestry sector in Belarus. In connection with substantiates the need to create online trading platforms for ecosystem services and institutional changes technological capabilities, the importance of involving various industries in the ecological economy is demonstrated.*

На постсоветском пространстве возникают различные современные элементы структуризации территорий, включая экосистемную составляющую. Однако одни государства быстро адаптируются под системы поиска средств для сбалансированного эколого-экономического развития, а другие – медленно. В