Вычислительные технологии, 2024, том 29, № 6, с. 125–146. © ФИЦ ИВТ, 2024
Computational Technologies, 2024, vol. 29, no. 6, pp. 125–146. © FRC ICT, 2024
ISSN 1560-7534
eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2024.29.6.008

Распознавание таблиц неаннотированных PDF-документов на основе использования PDF-специфичных свойств

А. О. Шигаров
Институт динамики систем и теории управления им. В. М. Матросова СО РАН, 664033, Иркутск, Россия
Контактный автор: Шигаров Алексей Олегович, e-mail: shigarov@icc.ru
Поступила 20 сентября 2024 г., принята в печать 27 сентября 2024 г.

Сегодня PDF — это один из наиболее популярных форматов распространения печатно-ориентированных документов в электронной среде. PDF-документы часто являются неаннотированными: страницы представлены только низкоуровневыми инструкциями рендеринга текста и графики, они не сопровождаются аннотацией своих структурных компонентов (заголовков, абзацев, таблиц и пр.). Автоматическое восстановление такой аннотации может обеспечить доступность структурных компонентов. Последнее возможно при решении ряда задач, одной из которых является распознавание таблиц неаннотированных PDF-документов: обнаружение границ их строк, столбцов и ячеек.

В работе предложен метод распознавания таблиц неаннотированных PDF-документов. В отличие от имеющихся аналогов впервые означенная задача решается на базе использования PDF-специфичных свойств: порядка вывода текста, позиций перемещения пера и пр. Это позволило адаптировать к поставленной задаче некоторые известные подходы и методы, изначально ориентированные на растровые изображения и неформатированный текст, включая “кластеризацию слов”, обнаружение строк rows first, сегментацию пробельного пространства и анализ компонентов связности. Представленные результаты оценки производительности показывают эффективность решений, реализующих данный метод.

Ключевые слова: распознавание таблиц, извлечение таблиц, неструктурированные данные, документные таблицы, анализ компоновки страницы документа.

Цитирование: Шигаров А.О. Распознавание таблиц неаннотированных PDF-документов на основе использования PDF-специфичных свойств. Вычислительные технологии. 2024; 29(6):125–146. DOI:10.25743/ICT.2024.29.6.008.

Введение

Количество PDF-документов в мире исчисляется триллионами, по оценке Д. Джонсона, главы PDF-ассоциации (`https://pdfa.org/wp-content/uploads/2018/06/1330_Johnson.pdf`). Они предназначены для одинакового отображения информации на различных устройствах, но не для ее редактирования, поэтому содержимое их страниц представлено низкоуровневыми инструкциями рендеринга текста и графики. Несмотря на то, что современная версия формата PDF предусматривает возможность включения в PDF-документы аннотации (HTML-подобной разметки) структурных компонентов их

страниц (заголовков, абзацев, таблиц и пр.), многие PDF-документы остаются неаннотированными [1, 2]. Автоматическое восстановление такой аннотации может обеспечить доступность структурных компонентов в различных приложениях, требующих индексации и структурирования данных, представленных в PDF-документах. Последнее возможно при решении ряда задач анализа компоновки страниц PDF-документов, одной из которых является распознавание таблиц: обнаружение границ их строк, столбцов и ячеек.

В рассматриваемом случае задача распознавания таблиц формулируется следующим образом: имеется печатно-ориентированный PDF-документ с машиночитаемым текстовым содержимым, потенциально содержащий одну или несколько таблиц; требуется извлечь из него физическую структуру (т. е. ячейки с определенными позициями в пространстве строк и столбцов, характеристиками форматирования, текстовым и иным содержимым) каждой из них. Задача включает две взаимосвязанные подзадачи: обнаружение таблиц и обнаружение ячеек. Первая из них формулируется как выделение той части источника, которая представляет единственную таблицу и ничего другого. Вторая состоит в определении всех частей источника, которые представляют отдельные ячейки одной таблицы. Качество результатов первой подзадачи может быть улучшено за счет последующего распознавания ее ячеек, и, напротив, предварительное обнаружение таблиц позволяет улучшить результаты второй подзадачи. На выходе обеспечивается возможность представления физической структуры в редактируемом формате (HTML, CSV или др.).

Методы распознавания таблиц в печатно-ориентированных источниках, представленных в форматах растровых изображений, неформатированного текста и языков описания страниц, активно развиваются три последних десятилетия [3]. С 2013 г. периодически проводятся соревнования между реализациями таких методов. Следует отметить, что самые последние разработки показывают высокое качество результатов распознавания таблиц на устоявшихся соревновательных коллекциях данных ICDAR 2013–2021 [4]. Однако другие тесты производительности, включая SciTSR и IAIS, в целом выявляют недостаточную эффективность доступных решений [5–7]. Таким образом, рассматриваемая задача остается актуальной и в наши дни.

В данной работе предлагается новый метод распознавания таблиц в неаннотированных PDF-документах. В отличие от имеющихся аналогов впервые предлагается решить означенную задачу на базе использования PDF-специфичных свойств: порядка вывода текста, позиций перемещения пера и пр. Это позволило адаптировать к поставленной задаче некоторые известные подходы и методы анализа компоновки страниц печатно-ориентированных документов, изначально предназначенные для работы с растровыми изображениями и неформатированным текстом, включая “кластеризацию слов” T-Recs [8], обнаружение строк rows first, сегментацию пробельного пространства и анализ компонентов связности. Кроме того, оно позволило реализовать методику фильтрации кандидатных случаев, нацеленную на улучшение качества результатов предсказания ограничивающих рамок таблиц внутри страниц PDF-документов с помощью искусственных нейронных сетей. В ряде наших прошлых работ эти методики рассматривались по отдельности, а именно: сегментация страницы [9, 10], распознавание ограничивающих рамок таблиц на основе сопоставления строк [11] и обнаружения объектов [12, 13], распознавание структуры таблиц на основе сегментации пробельного пространства и анализа компонентов связности [14]. Данная работа представляет их суммарно как целостную конкретизацию шагов предлагаемого метода. Оставшаяся

часть настоящей работы организована следующим образом: представляется краткий обзор современного состояния исследований рассматриваемой области (разд. 1); излагается предлагаемый метод решения поставленной задачи (разд. 2); приводятся результаты оценки производительности решений, реализованных на основе предлагаемого метода, а также их количественное и качественное сравнение с имеющимися аналогами (разд. 3); в заключении делаются выводы.

1. Современное состояние исследований

Методы распознавания таблиц можно разделить на нисходящие и восходящие. Нисходящий подход состоит в том, что сперва выполняется обнаружение ограничивающей рамки некоторой таблицы внутри страницы документа, а затем разделение ее на строки, столбцы и ячейки. Восходящий подход, напротив, предполагает, что сперва выполняется обнаружение отдельных ячеек, строк и столбцов напрямую на странице документа, а затем уже они комбинируются в таблицу.

Известные методы основаны преимущественно на нисходящем подходе. Полагаясь на наличие разграфки, некоторые из них [15, 16] распознают линейки, отделяющие таблицу от остального содержимого страницы документа, а также разделяющие ее на ячейки. Однако, в общем случае, разграфка может быть неполной или вовсе отсутствовать. Поэтому предлагается анализировать размещение текста и/или пробельного пространства. Для этого одни методы [15, 16] используют правила анализа выравнивания блоков текста, а другие [17–19] предпочитают X-Y Cut-алгоритм [20], который строит проекционные профили блоков текста на осях X и Y. Восходящие методы можно разделить на две группы: одни [21–26] собирают таблицы из столбцов, другие [27–31] — из строк. В первом случае сперва собираются блоки текста на основе эвристик [21] или машинного обучения [26], а затем из них формируются столбцы на основе правил анализа выравнивания текста внутри таблицы. Во втором случае строки текста классифицируются на табличные и прочие на основе алгоритмов машинного обучения, в частности скрытой марковской модели [29], условных случайных полей [28] и метода опорных векторов [30]. Затем табличные строки группируются в таблицы на основе правил анализа выравнивания текста внутри таблицы [27].

Сегодня основное направление развития рассматриваемых методов связано с применением глубокого обучения [32]. Активно создаются искусственные нейронные сети (ИНС), способные обнаруживать таблицы [33–48] и распознавать их структуру [49–57]. Они реализуют различные современные архитектуры ИНС, в частности обнаружение объектов [33, 34, 36, 43–47, 55–57], семантическую сегментацию [35, 39, 40], деформируемые свертки [37, 48, 52], "полностью" сверточные [34, 37, 39], графовые [41, 42, 51], генеративно-состязательные [38], рекуррентные [50], расширенные свертки [53], Encoder-Decoder [49] и Encoder-Dual-Decoder модели [54]. Наиболее востребованными среди перечисленных являются архитектуры ИНС-обнаружения объектов, в том числе Faster R-CNN [58], Mask R-CNN [59], Cascade Mask R-CNN [60], YOLO [61], RetinaNet [62] и SSD [63]. Они разработаны для целей определения наличия объекта некоторого класса на изображении и нахождения его границ в виде ограничивающей рамки, ключевых точек или контура. При обнаружении объектов на изображениях для извлечения признаков используются сверточные нейронные сети, в том числе таких архитектур, как VGG [64], ResNet [65], DarkNet [61], ResNeXt [66] и EfficientNet [67]. Их базовые версии предобучены на больших массивах фотографических изображений, таких как

ImageNet [68], MS-COCO [69] и Pascal VOC [70]. Будучи предназначенными для обнаружения и классификации фотоизображений объектов реального мира, изначально они малоэффективны для работы с изображениями документов. Для того чтобы адаптировать их к рассматриваемой задаче, предлагается использовать технику трансферного обучения, называемую тонкой настройкой [33, 34]. Примеры адаптации различных архитектур можно найти в следующих работах: Faster R-CNN [33, 34, 36, 44, 45], YOLO [43, 46], Mask R-CNN [46], Cascade Mask R-CNN [47], SSD [46] и RetinaNet [46].

Подобные архитектуры обеспечивают возможность тонкой настройки весов отдельных слоев предобученных моделей, не изменяя бóльшую часть их слоев. Например, Faster R-CNN позволяет дообучить два последних полносвязных слоя, один из которых отвечает за уточнение координат ограничивающей рамки, а другой — за классификацию объекта внутри рамки. Такая тонкая настройка может быть выполнена на небольших наборах данных, включающих от сотен до десятков тысяч примеров изображений страниц документов с эталонной разметкой представленных в них таблиц: UW3/UNLV [71], Marmot [72], ICDAR-2013 [73], ICDAR-2017 POD [74], ICDAR-2019 [75], SciTSR [5] и др. Обучение целевых нейросетевых моделей также можно выполнить с помощью более крупных наборов данных, включающих от сотен тысяч до миллионов примеров: TableBank [45], PubLayNet [76], PubTabNet [54] и др.

Текущее состояние исследований все еще не позволяет сегодня реализовать готовые к использованию в общем случае решения распознавания таблиц в неаннотированных PDF-документах, качество результатов которых не требовало бы ручной корректировки. Полная автоматизация может быть достигнута только в отдельных случаях. Некоторые из предлагаемых решений предоставляют возможность настройки своих параметров для улучшения качества получаемых результатов при работе с конкретными источниками. Опубликованные эксперименты показывают, что иногда они могут дать результаты, сопоставимые по качеству с работой оператора.

К наименее проработанным вопросам следует отнести изучение возможности применения PDF-специфичной информации. Сегодня основное направление исследований связано с растровыми изображениями. В действительности PDF-документ можно растеризовать, сведя таким образом задачу к более общей. Однако данный процесс неизбежно сопровождается потерей информации. По сравнению с растром PDF предоставляет дополнительную информацию (текст, шрифты, линейки и пр.), которая может улучшить качество результатов в некоторых случаях. В частности, она может использоваться на этапе предобработки для сегментации страницы PDF-документа (обнаружения заголовков, колонок и абзацев текста), а также на этапе постобработки для фильтрации кандидатных случаев. Таким образом, изучение возможности применения PDF-специфичной информации для распознавания таблиц в неаннотированных PDF-документах является актуальным направлением исследований.

2. Предлагаемый метод

Для входных данных (неаннотированных PDF-документов) должны выполняться следующие условия: страница документа имеет так называемую манхэттенскую компоновку, любая непустая ячейка таблицы наполнена исключительно машиночитаемым текстом, наличие разграфки не обязательно. Метод опирается на нисходящий подход: обнаружение таблиц предваряет распознавание их структуры. Он реализован рядом мето-

дик сегментации страницы, обнаружения и распознавания структуры таблиц, в каждой из которых применяется PDF-специфичная информация. Рассмотрим их подробнее.

Сегментация страницы документа. Доступные изначально напечатанные символы объединяются в однокомпонентные блоки текста — слова (рис. 1, *а, б*), те, в свою очередь, группируются в многокомпонентные блоки — строки и абзацы (рис. 1, *в, г*). Предлагается адаптация известного метода кластеризации слов в неформатированном тексте T-Recs [8] к специфике PDF-документов. Он включает два этапа: первый — начальное формирование многокомпонентных и многострочных блоков; второй — коррекция типовых ошибочных случаев, которые могут возникнуть в результате выполнения первого этапа. В адаптированной версии, на первом этапе, в качестве ограничений на восстанавливаемый блок используется PDF-специфичная информация: порядок вывода текста, позиции перемещения пера и пр. (рис. 2, *а, б*). По сравнению с оригинальным методом T-Recs [8] настоящая адаптация включает дополнительную проверку кандидатных блоков. Предполагается, что подлинный блок удовлетворяет ряду условий:

- порядок вывода символьных позиций не должен разрываться;
- при их выводе должен использоваться шрифтовой комплект одной гарнитуры (размеры и начертания шрифтов могут отличаться);

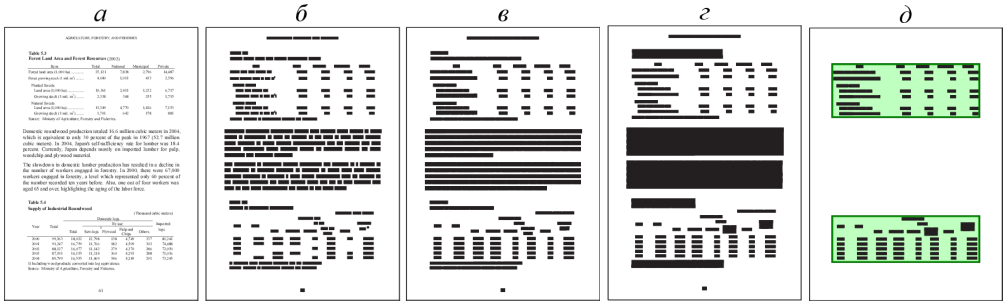

Рис. 1. Синтез блоков текста: символы (*а*), слова (*б*), строки (*в*), абзацы (*г*). Обнаружение ограничивающих рамок таблиц (*д*)
Fig. 1. Text block synthesis: characters (*а*), words (*б*), lines (*в*), paragraphs (*г*). Detection of table bounding boxes (*д*)

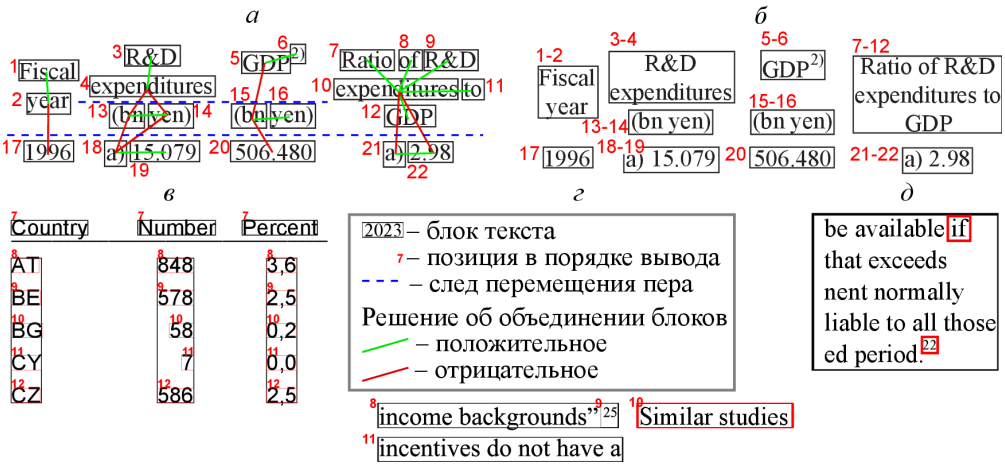

Рис. 2. Начальное формирование блоков: исходные (*а*), целевые (*б*). Типы ошибочных случаев: дупликация порядка вывода текста (*в*), изоляция (*г*) и наложение (*д*) блоков текста
Fig. 2. Initial formation of blocks: source (*а*), target (*б*). Types of error cases: duplication in the text rendering order (*в*), isolation (*г*) and overlapping (*д*) of text blocks

- внутри ограничивающей рамки нет линеек разграфки, а также нет вертикальных следов перемещения пера;
- символьные позиции могут располагаться в соседних строках, при этом значение междустрочного интервала оценивается приближенно по шрифтовым метрикам исходных символьных позиций.

Кроме того, в отличие от оригинального метода T-Recs, предлагаемая адаптация рассчитана на другое представление входных данных. Поэтому ее применение может приводить к другим типам ошибок. Второй этап дополнен новыми алгоритмами коррекции соответствующих ошибочных случаев (рис. 2, *в, д*).

По восстановленным блокам текста сегментируется пробельное пространство. Среди полученных сегментов выбираются промежутки между колонками текста. Некоторые блоки распознаются как разделители, а именно заголовки по ключевым словам (“таблица”, “рисунок” и др.) и крупные абзацы текста, размеры которых превышают заданный порог. В результате область поиска таблиц может быть сужена до некоторой части внутри страницы. Одна страница может быть разделена на несколько изолированных областей поиска. Из области поиска исключаются односимвольные блоки, соответствующие однотипным знакам маркированных списков. (Более подробно данная методика излагается в работе [10].)

Обнаружение таблиц на основе правил. В заданной области поиска блоки группируются в строки по пересечению проекций на ось Y (рис. 3). Среди них выбираются только многокомпонентные (т. е. включающие по два и более блока). Предполагается, что любые соседние строки одной таблицы имеют схожее размещение промежутков пробельного пространства. Предлагается двухэтапное сопоставление строк: сперва группируются соседние многокомпонентные строки, затем аналогичным образом — сами группы. Каждая из полученных групп принимается за одну целую таблицу. (Более подробно данная методика излагается в работах [9, 11].)

Обнаружение таблиц на основе машинного обучения. Другая методика состоит в использовании ИНС-обнаружения объектов на изображениях с последующей фильтрацией кандидатных случаев. Первая часть может базироваться на настройке доступных ИНС известной архитектуры Faster R-CNN (этот подход впервые предложен S. Schreiber и др. [34]). Настройка сводится к обучению двух полносвязных слоев (классификации объектов и регрессии координат регионов) на небольших наборах предметных данных. Обучающая выборка комбинируется из доступных коллекций изображений документов с размеченными регионами таблиц (UNLV, Marmot, ICDAR-2017 POD и др.). Для увеличения обучающей выборки предлагается применять аугментацию примеров на основе аффинных преобразований. (Более подробно процесс настройки данной ИНС рассмотрен в работе [12].) Во второй части предлагается прибегнуть к бинарной классификации графовых представлений кандидатных случаев на ложноположительные и истинно-положительные. (Вершины такого графа соответствуют блокам текста, реб-

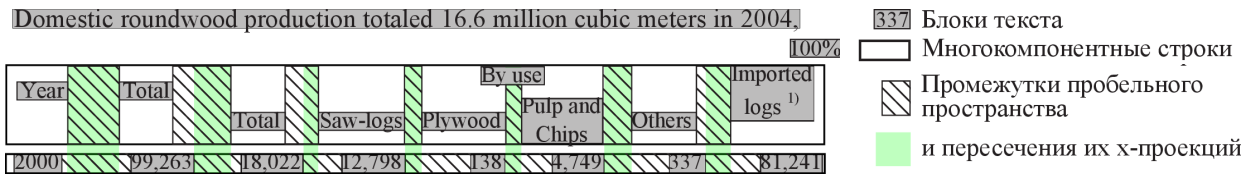

Рис. 3. Обнаружение таблиц на основе сопоставления строк
Fig. 3. Table detection based on row matching

ра — пересечениям их проекций на ось X. Два примера графов кандидатных случаев представлены на рис. 4.) Бинарная классификация может быть реализована с помощью ансамбля деревьев решений. При классификации используется ряд признаков графового представления кандидатных случаев, таких как: количество вершин, компонентов связности и ребер, а также агрегации типа сумма, среднее, максимум, минимум, медиана, стандартное отклонение, несмещенная дисперсия над x- и y-позициями середин ограничивающих рамок блоков текста, степенями вершин и весами ребер. Следует отметить, что предлагаемая методика требует обучения с учителем. Как и в случае с настройкой ИНС, классификатор кандидатных случаев может быть обучен на небольших выборках. Более подробно процесс фильтрации кандидатных случаев рассмотрен в работе [12].

Сегментация таблицы. Предлагается две методики на основе правил: сегментация пробельного пространства и анализ компонентов связности. В первом случае сперва восстанавливаются линейки разграфки таблицы по промежуткам пробельного пространства, затем формируются ячейки по пересечениям полученных линеек (рис. 5). Во втором случае выбираются блоки текста, односвязные по проекции на ось X, каждый компонент связности соответствует одному столбцу (рис. 6, *а*). Аналогично выбираются блоки, односвязные по проекции на ось Y, каждый компонент связности соответствует одной строке (рис. 6, *б*). Когда несколько ячеек пересекают один блок текста, они объединяются (рис. 6, *в*). Более подробно данные методики излагаются в работе [14].

Программная реализация. Изложенные методики распознавания таблиц реализованы в виде программного обеспечения (TabbyPDF [83]), исходный код которого опубликован в открытом доступе под свободными лицензиями (`https://github.com/tabbydoc/tabbypdf`, `https://github.com/tabbydoc/tabbypdf2`). Это программное обеспечение

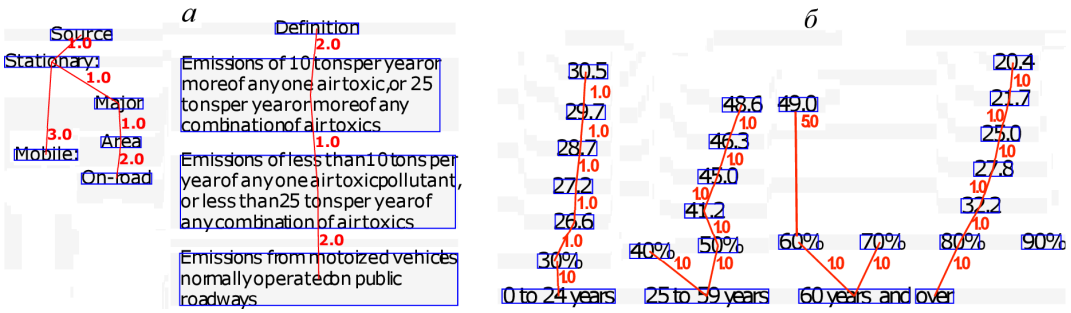

Рис. 4. Фильтрация кандидатных случаев на основе классификации их графовых представлений: истинно-положительный случай (*а*); ложноположительный случай (*б*)

Fig. 4. Filtering candidate cases based on the classification of their graph representations: true positive case (*а*); false positive case (*б*)

Year	Total	Domestic logs					Imported logs [1]
		Total	By use				
			Saw-logs	Plywood	Pulp and chips	Others	
2000	99,263	18,022	12,798	138	4,749	337	81,241
2002	88,127	16,077	11,142	279	4,370	286	72,050
2003	87,191	16,155	11,214	360	4,293	288	71,036
2004	89,799	16,555	11,469	546	4,249	291	73,245
2005	85,857	17,176	11,571	863	4,426	316	68,681

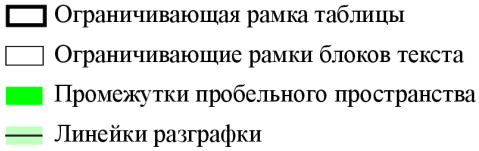

Рис. 5. Обнаружение ячеек на основе сегментации пробельного пространства

Fig. 5. Cell detection based on white space segmentation

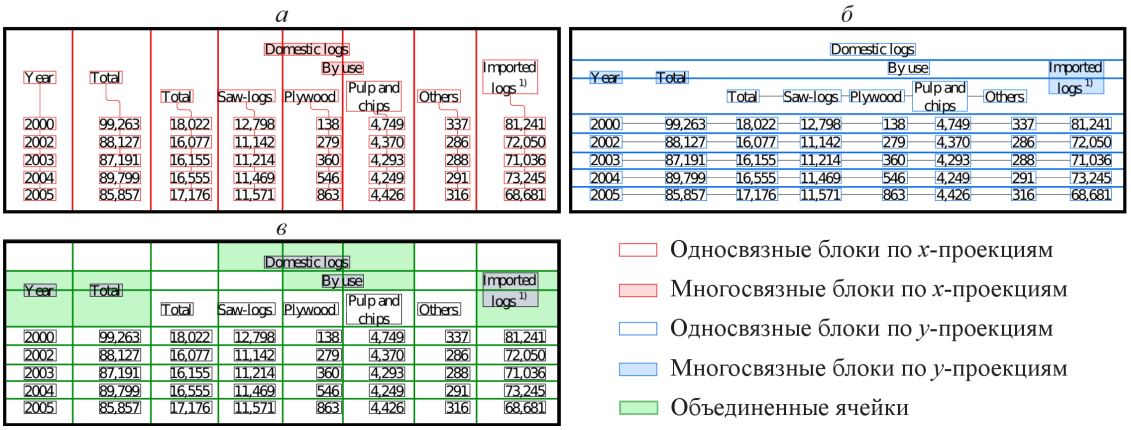

Рис. 6. Обнаружение ячеек на основе анализа компонентов связности: разделение на столбцы по x-проекциям блоков текста ($а$); разделение на строки по y-проекциям блоков текста ($б$); объединение ячеек, содержащих части одного блока текста ($в$)

Fig. 6. Cell detection based on connected component analysis: partitioning into columns by x-projections of text blocks ($а$); partitioning into rows by y-projections of text blocks ($б$); merging cells containing parts of one text block ($в$)

позволило выполнить оценку производительности предлагаемых решений и сравнить их с аналогами.

3. Оценка производительности и сравнение с аналогами

Используются следующие метрики качества результатов распознавания таблиц: точность (P), полнота (R) и F-мера (F_1):

$$P = \frac{tp}{tp + fp}, \qquad R = \frac{tp}{tp + fn}, \qquad F_1 = 2\frac{P \cdot R}{P + R},$$

где количество tp — истинно-положительных, fp — ложноположительных и fn — ложноотрицательных исходов, значения которых вычисляются в результате сопоставления результатов (набора однотипных объектов R), произведенных тестируемым решением, с эталонными данными теста производительности (набором однотипных объектов GT). Исход сравнения некоторого объекта из R считается истинно-положительным, если есть

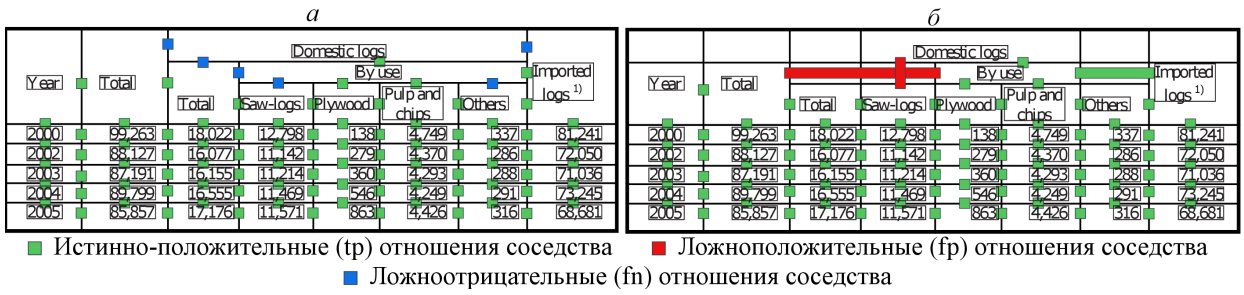

Рис. 7. Подсчет отношений соседства между ячейками в соответствии с методикой M. Göbel и др. [77]: эталонная ($а$) и восстановленная ($б$) таблицы

Fig. 7. Calculation of neighborhood relations between cells according to the method M. Göbel et al. [77]: a reference table ($а$) and recovered one ($б$)

Т а б л и ц а 1. Оценка производительности на тесте ICDAR-2013
Table 1. Performance evaluation using the ICDAR-2013 test

Метрика	Обнаружение таблиц			Сегментация таблиц*	
	СС	ОО	ОО+Ф	СПП	АКС
P_{doc}	0.7605	0.8651	0.9703	0.9180	0.9499
R_{doc}	0.8172	0.9795	0.9795	0.9121	0.9233
F_1	0.7878	0.9187	**0.9748**	0.9150	**0.9364**

* Используются эталонные ограничивающие рамки таблиц.
Оцениваются следующие варианты: СС — сопоставление строк; ОО — обнаружение объектов с помощью ИНС; ОО+Ф — ОО с фильтрацией кандидатных случаев; СПП — сегментация пробельного пространства; АКС — анализ компонентов связности.

Т а б л и ц а 2. Сравнение с аналогами на тесте ICDAR-2013
Table 2. Comparison with analogues using the ICDAR-2013 test

Решение	Обнаружение таблиц			Решение	Распознавание таблиц[1]		
	P_{doc}	R_{doc}	F_1		P_{doc}	R_{doc}	F_1
FineReader*	0.973	0.997	0.985	FineReader	0.871	0.883	0.877
Kavasidis et al. [40]	0.981	0.975	0.978	OmniPage	0.846	0.838	0.842
ОО+Ф[2]	**0.970**	**0.979**	**0.975**	**ОО+Ф/АКС**[2]	0.849	0.824	**0.839**
DeepDeSRT [34]	0.974	0.961	0.968	Tabula [82]	0.869	0.808	0.837
TableNet [39]	0.970	0.963	0.966	Tab.IAIS [81]	0.918	0.762	0.832
OmniPage*	0.957	0.964	0.966	**СС/СПП**[2]	0.834	0.830	**0.832**
Tran et al. [78]	0.964	0.952	0.958	Acrobat*	0.816	0.726	0.768
Silva et al. [79]	0.929	0.983	0.955	Nitro	0.846	0.679	0.753
Hao et al. [80]	0.922	0.972	0.946	Silva et al. [79]	0.687	0.705	0.696
Nitro*	0.940	0.932	0.936	pdf2table [22]	0.575	0.595	0.585

[1] Используются автоматически обнаруженные ограничивающие рамки таблиц.
[2] Предлагаемые решения на основе ИНС-обнаружения объектов с фильтрацией кандидатных случаев (ОО+Ф), сопоставления строк (СС).
* Индустриальные программные продукты.

Т а б л и ц а 3. Сравнение с аналогами по качественным характеристикам
Table 3. Comparison with analogues by quality characteristics

Метод	Подход	Формат	Не требуется		PDF-специфика				
			OCR	Разграфка	Т	Ш	Л	ПВ	ПП
DeepDeSRT [34]	ГО	РИ	−	+	−	−	−	−	−
GraphTSR* [5]	П/ГО	ИР	+	+	+	+	+	+	+
pdf2table [22]	П	ИР	+	+	+	−	−	−	−
Tab.IAIS [81]	П	РИ	−	−	−	−	+	−	−
TabbyPDF	**П/ГО**	**ИР**	+	+	+	+	+	+	+
TableNet [39]	П/ГО	РИ	−	+	−	−	−	−	−
Tabula [82]	П	ИР	+	+	+	−	+	−	−
Tran et al. [78]	П	РИ	−	−	−	−	+	−	−

П — правила, ГО — глубокое обучение; РИ — растровые изображения, ИР — инструкции рендеринга PDF-документа. Поддерживаемая PDF-специфика: Т — машиночитаемый текст, Ш — шрифтовые свойства, Л — линейки разграфки (в том числе восстанавливаемые из растра), ПВ — порядок вывода текста, ПП — перемещение пера.
* Базируется на извлечении блоков текста из PDF-документа, выполняемого программным обеспечением, разработанным на основе предлагаемого метода (TabbyPDF [83]).

идентичный ему объект в наборе GT, и ложноположительным — в противном случае. Исход сравнения некоторого объекта из GT считается ложноотрицательным, если среди объектов набора R нет идентичного ему.

Оценка производительности решений распознавания таблиц, реализованных на основе предлагаемого метода, выполнена с помощью известной методики M. Göbel и др. [77] (рис. 7). Вычисляются средние значения точности (P_{doc}) и полноты (R_{doc}) среди документов:

$$P_{doc} = \frac{P_1 + \cdots + P_n}{n}, \quad R_{doc} = \frac{R_1 + \cdots + R_n}{n},$$

где P_i — точность, а R_i — полнота, измеренная на i-м документе. Оценивается качество обнаружения и сегментации/распознавания таблиц. В первом случае сопоставляются печатаемые символы внутри предсказанных и эталонных ограничивающих рамок таблиц, а во втором — отношения соседства текстового содержимого обнаруженных ячеек.

Результаты оценки предлагаемых решений распознавания таблиц получены на тесте ICDAR-2013 [73] (табл. 1). Обнаружение таблиц с помощью ИНС дает лучшие результаты по сравнению с методикой на основе сопоставления строк. Однако при высокой полноте (около 98 %) точность остается низкой (менее 87 %) из-за большого количества ложноположительных предсказаний ИНС. Фильтрация кандидатных случаев позволила значительно поднять точность (до 97 %). Методики сегментации пробельного пространства и анализа компонентов связности дают близкое качество обнаружения ячеек.

Качественное сравнение с аналогами приводится в табл. 2 и 3. Оно охватывает только часть конкурентных методов, которые в целом дают общую картину отличий предлагаемого метода от аналогов по качественным характеристикам. PDF-специфичная информация, такая как порядок вывода текста, позиции перемещения пера и пр., используется только решениями, реализованными на основе предлагаемого в данной работе метода, а также сторонним аналогом GraphTSR [5]. Однако сам этот аналог в свою очередь базируется на извлечении блоков текста из PDF-документа, выполняемого программным обеспечением, разработанным на основе предлагаемого метода (TabbyPDF [83]). Остальные аналоги применяют исключительно отдельные особенности PDF-документов (текст и линейки) или вовсе не применяют, прибегая к их растеризации.

Заключение

Впервые изучены вопросы применимости PDF-специфичных свойств (порядка вывода текста, позиций перемещения пера и пр.) к комплексной задаче распознавания таблиц в неаннотированных PDF-документах. Насколько известно автору, никто до сих пор не предлагал использовать именно такие свойства в означенном ключе. Показано, что они могут быть полезны на всех этапах процесса распознавания таблиц, а именно при сегментации страниц документов, обнаружении ограничивающих рамок таблиц и распознавании структуры их ячеек. Вовлечение PDF-специфичных свойств позволило адаптировать к поставленной задаче некоторые известные подходы и методы, изначально ориентированные на растровые изображения и неформатированный текст, включая кластеризацию слов, обнаружение строк rows first, сегментацию пробельного пространства и анализ компонентов связности. В совокупности они обеспечивают распознавание таблиц в неаннотированных PDF-документах.

Представленные результаты оценки производительности показывают эффективность решений, реализованных на основе предлагаемого метода. Количественное сравнение с аналогами свидетельствует об их соответствии современному уровню технологического развития в рассматриваемой области. В то же время качественное сравнение выявляет следующие преимущества перед аналогами. Реализация предлагаемого метода распознавания таблиц не требует предварительной настройки параметров и обучения с учителем. Однако при наличии готовых нейросетевых моделей они могут заменить алгоритмы обнаружения таблиц на основе правил. При этом качество окончательных результатов может быть улучшено за счет применения фильтрации кандидатных случаев.

Дальнейшая исследовательская работа может проводиться в направлении развития ИНС-моделей, базирующихся на применении PDF-специфичных свойств. Модель GraphTSR [5], реализованная исследователями из Пекинского технологического института, может послужить тем, кто возьмется за эту работу, и некоторым примером для вдохновения.

Список литературы

[1] **Turró M.R.** Are PDF documents accessible? Information Technology and Libraries. 2008; 27(3):25–43. DOI:10.6017/ital.v27i3.3246.

[2] **Nganji J.T.** The Portable Document Format (PDF) accessibility practice of four journal publishers. Library & Information Science Research. 2015; 37(3):254–262. DOI:10.1016/j.lisr.2015.02.002.

[3] **Shigarov A.** Table understanding: problem overview. WIREs Data Mining and Knowledge Discovery. 2023; 13(1):e1482. DOI:10.1002/widm.1482.

[4] **Yepes A., Zhong P., Burdick D.** ICDAR 2021 competition on scientific literature parsing. Processing of the 16th International Conference on Document Analysis and Recognition. Part IV. Switzerland: Lausanne; 2021: 605–617. DOI:10.1007/978-3-030-86337-1_40.

[5] **Chi Z., Huang H., Xu H.-D., Yu H., Yin W., Mao X.-L.** Complicated table structure recognition. arXiv preprint, arXiv:1908.04729. 2019. DOI:10.48550/arXiv.1908.04729. Available at: `https://arxiv.org/abs/1908.04729`.

[6] **Adams T., Namysl M., Kodamullil A.T., Behnke S., Jacobs M.** Benchmarking table recognition performance on biomedical literature on neurological disorders. Bioinformatics. 2021; 38(6):1624–1630. DOI:10.1093/bioinformatics/btab843.

[7] **Zhang M., Perelman D., Le V., Gulwani S.** An integrated approach of deep learning and symbolic analysis for digital PDF table extraction. Processing of the 25th International Conference on Pattern Recognition. Italy: Milan; 2021: 4062–4069. DOI:10.1109/ICPR48806.2021.9413069.

[8] **Kieninger T.** Table structure recognition based on robust block segmentation. Document Recognition V. USA: San Jose; 1998: 22–32. DOI:10.1117/12.304642.

[9] **Shigarov A., Fedorov R.** Simple algorithm page layout analysis. Pattern Recognition and Image Analysis. 2011; 21(2):324–327. DOI:10.1134/S1054661811021008.

[10] **Шигаров А.О., Парамонов В.В.** Сегментация текста неразмеченных PDF-документов. Вычислительные технологии. 2022; 27(5):69–78. DOI:10.25743/ICT.2022.27.5.007.

[11] **Бычков И.В., Ружников Г.М., Хмельнов А.Е., Шигаров А.О.** Эвристический метод обнаружения таблиц в разноформатных документах. Вычислительные технологии. 2009; 14(2):58–73.

[12] **Cherepanov I., Mikhailov A., Shigarov A., Paramonov V.** On automated workflow for finetuning deep neural network models for table detection in document images. Processing of the 43rd International Convention on Information, Communication and Electronic Technology. Croatia: Opatija; 2020: 1130–1133. DOI:10.23919/MIPRO48935.2020.9245241.

[13] **Mikhailov A., Shigarov A., Rozhkov E., Cherepanov I.** On graph-based verification for PDF table detection. Processing of the 2020 Ivannikov ISPRAS Open Conference. Russia: Moscow; 2020: 91–95. DOI:10.1109/ISPRAS51486.2020.00020.

[14] **Shigarov A., Mikhailov A., Altaev A.** Configurable table structure recognition in untagged PDF documents. Processing of the 2016 ACM Symposium on Document Engineering. Austria: Vienna; 2016: 119–122. DOI:10.1145/2960811.2967152.

[15] **Fang J., Gao L., Bai K., Qiu R., Tao X., Tang Z.** A table detection method for multipage PDF documents via visual separators and tabular structures. Processing of the 11th International Conference on Document Analysis and Recognition. China: Beijing; 2011: 779–783. DOI:10.1109/ICDAR.2011.304.

[16] **Bart E.** Parsing tables by probabilistic modeling of perceptual cues. Processing of the 10th IAPR International Workshop on Document Analysis Systems. Australia: Gold Coast; 2012: 409–414. DOI:10.1109/DAS.2012.67.

[17] **Cesarini F., Marinai S., Sarti L., Soda G.** Trainable table location in document images. Processing of the International Conference on Pattern Recognition. Vol. 3. Canada: Quebec; 2002: 236–240. DOI:10.1109/ICPR.2002.1047838.

[18] **Wang Y., Phillips I., Haralick R.** Table structure understanding and its performance evaluation. Pattern Recognition. 2004; 37(7):1479–1497.

[19] **Klampfl S., Jack K., Kern R.** A comparison of two unsupervised table recognition methods from digital scientific articles. D-Lib Magazine. 2014; 20(11/12). DOI:10.1045/november14-klampfl.

[20] **Jaekyu H., Haralick R.M., Phillips I.T.** Recursive X-Y cut using bounding boxes of connected components. Processing of the 3rd International Conference on Document Analysis and Recognition. Vol. 2. Canada: Montreal; 1995: 952–955.

[21] **Kieninger T., Dengel A.** The T-Recs table recognition and analysis system. Document Analysis Systems: Theory and Practice. Japan: Nagano; 1999: 255–270. DOI:10.1007/3-540-48172-9_21.

[22] **Yildiz B., Kaiser K., Miksch S.** pdf2table: a method to extract table information from PDF files. Processing of the 2nd Indian International Conference on Artificial Intelligence. India: Pune; 2005: 1773–1785.

[23] **Hassan T., Baumgartner R.** Table recognition and understanding from PDF files. Processing of the 9th International Conference on Document Analysis and Recognition. Vol. 2. Brazil: Parana; 2007: 1143–1147. DOI:10.1109/ICDAR.2007.4377094.

[24] **Shafait F., Smith R.** Table detection in heterogeneous documents. Processing of the 9th IAPR International Workshop on Document Analysis Systems. USA: Boston; 2010: 65–72. DOI:10.1145/1815330.1815339.

[25] **Deckert F., Seidler B., Ebbecke M., Gillmann M.** Table content understanding in SmartFIX. Processing of the 11th International Conference on Document Analysis and Recognition. China: Beijing; 2011: 488–492. DOI:10.1109/ICDAR.2011.104.

[26] **Bansal A., Harit G., Roy S.D.** Table extraction from document images using fixed point model. Processing of the Indian Conference on Computer Vision Graphics and Image Processing. India: Bangalore; 2014: 67:1–67:8. DOI:10.1145/2683483.2683550.

[27] **Handley J.** Table analysis for multiline cell identification. Document Recognition and Retrieval VIII. USA: San Jose; 2000: 34–43. DOI:10.1117/12.410853.

[28] **Pinto D., McCallum A., Wei X., Croft W.B.** Table extraction using conditional random fields. Processing of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. Canada: Toronto; 2003: 235–242. DOI:10.1145/860435.860479.

[29] **e Silva A.C., Jorge A.M., Torgo L.** Design of an end-to-end method to extract information from tables. International Journal of Document Analysis and Recognition. 2006; 8(2–3):144–171. DOI:10.1007/s10032-005-0001-x.

[30] **Liu Y., Mitra P., Giles C.L.** Identifying table boundaries in digital documents via sparse line detection. Processing of the 17th ACM Conference on Information and Knowledge Management. USA: Napa Valley; 2008: 1311–1320. DOI:10.1145/1458082.1458255.

[31] **Oro E., Ruffolo M.** PDF-TREX: an approach for recognizing and extracting tables from PDF documents. Processing of the 10th International Conference on Document Analysis and Recognition. Spain: Barcelona; 2009: 906–910. DOI:10.1109/ICDAR.2009.12.

[32] **Hashmi K.A., Liwicki M., Stricker D., Afzal M.A., Afzal M.A., Afzal M.Z.** Current status and performance analysis of table recognition in document images with deep neural networks. IEEE Access. 2021; (9):87663–87685. DOI:10.1109/ACCESS.2021.3087865.

[33] **Gilani A., Qasim S.R., Malik I., Shafait F.** Table detection using deep learning. Processing of the 14th International Conference on Document Analysis and Recognition, Vol. 1. Japan: Kyoto; 2017: 771–776. DOI:10.1109/ICDAR.2017.131.

[34] **Schreiber S., Agne S., Wolf I., Dengel A., Ahmed S.** DeepDeSRT: deep learning for detection and structure recognition of tables in document images. Processing of the 14th International Conference on Document Analysis and Recognition. Vol. 1. Japan: Kyoto; 2017: 1162–1167. DOI:10.1109/ICDAR.2017.192.

[35] **He D., Cohen S., Price B., Kifer D., Giles C.L.** Multi-scale multi-task FCN for semantic page segmentation and table detection. Processing of the 14th International Conference on Document Analysis and Recognition. Vol. 1. Japan: Kyoto; 2017: 254–261. DOI:10.1109/ICDAR.2017.50.

[36] **Arif S., Shafait F.** Table detection in document images using foreground and background features. Processing of the Digital Image Computing: Techniques and Applications. Australia: Canberra; 2018: 1–8. DOI:10.1109/DICTA.2018.8615795.

[37] **Siddiqui S.A., Malik M.I., Agne S., Dengel A., Ahmed S.** DeCNT: deep deformable CNN for table detection. IEEE Access. 2018; (6):74151–74161. DOI:10.1109/ACCESS.2017.

[38] **Li Y., Gao L., Tang Z., Yan Q., Huanget Y.** A GAN-based feature generator for table detection. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 763–768. DOI:10.1109/ICDAR.2019.00127.

[39] **Paliwal S.S., Vishwanath D., Rahul R., Sharma M., Vig L.** TableNet: deep learning model for end-to-end table detection and tabular data extraction from scanned document images. 2019 International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 128–133. DOI:10.1109/ICDAR.2019.00029.

[40] **Kavasidis I., Pino C., Palazzo S., Rundo F., Giordano D., Spampinato C.** A saliency-based convolutional neural network for table and chart detection in digitized documents. Image Analysis and Processing. 2019: 292–302. DOI:10.1007/978-3-030-30645-8_27.

[41] **Holeček M., Hoskovec A., Baudiš P., Klinger P.** Table understanding in structured documents. 2019 International Conference on Document Analysis and Recognition Workshops. Vol. 5. Australia: Sydney; 2019: 158–164. DOI:10.1109/ICDARW.2019.40098.

[42] **Riba P., Dutta A., Goldmann L., Fornés A., Ramos O., Lladós J.** Table detection in invoice documents by graph neural networks. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 122–127. DOI:10.1109/ICDAR.2019.00028.

[43] **Huang Y., Yan Q., Li Y., Chen Y., Wang X., Gao L., Tang Z.** A YOLO-based table detection method. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 813–818. DOI:10.1109/ICDAR.2019.00135.

[44] **Sun N., Zhu Y., Hu X.** Faster R-CNN based table detection combining corner locating. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 1314–1319. DOI:10.1109/ICDAR.2019.00212.

[45] **Li M., Cui L., Huang S., Wei F., Zhou M., Li Z.** TableBank: table benchmark for image-based table detection and recognition. Processing of the 12th Language Resources and Evaluation Conference. France: Marseille; 2020: 1918–1925.

[46] **Casado-García Á., Domínguez C., Heras J., Mata E., Pascual V.** The benefits of close-domain fine-tuning for table detection in document images. Document Analysis Systems. Springer Nature; 2020: 199–215. DOI:10.1007/978-3-030-57058-3_15.

[47] **Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure K.** CascadeTabNet: an approach for end-to-end table detection and structure recognition from image-based documents. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. USA: Seattle; 2020: 2439–2447. DOI:10.1109/CVPRW50498.2020.00294.

[48] **Agarwal M., Mondal A., Jawahar C.V.** CDeC-Net: composite deformable cascade network for table detection in document images. Processing of the 25th International Conference on Pattern Recognition. Italy: Milan; 2021: 9491–9498. DOI:10.1109/ICPR48806.2021.9411922.

[49] **Deng Y., Rosenberg D., Mann G.** Challenges in end-to-end neural scientific table recognition. 2019 International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 894–901. DOI:10.1109/ICDAR.2019.00148.

[50] **Khan S.A., Khalid S.M.D., Shahzad M.A., Shafait F.** Table structure extraction with bidirectional gated recurrent unit networks. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1366–1371. DOI:10.1109/ICDAR.2019.00220.

[51] **Qasim S., Mahmood H., Shafait F.** Rethinking table recognition using graph neural networks. International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 142–147. DOI:10.1109/ICDAR.2019.00031.

[52] **Siddiqui S.A., Fateh I.A., Rizvi S.T.R., Dengel A., Ahmed S.** DeepTabStR: deep learning based table structure recognition. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 1403–1409. DOI:10.1109/ICDAR.2019.00226.

[53] **Tensmeyer C., Morariu V., Price B., Cohen S., Martinez T.** Deep splitting and merging for table structure decomposition. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 114–121. DOI:10.1109/ICDAR.2019.00027.

[54] **Zhong X., Bavani E.S., Yepes A.J.** Image-based table recognition: data, model, and evaluation. Processing of the 16th European Conference on Computer Vision. UK: Glasgow; 2020: 564–580. DOI:10.1007/978-3-030-58589-1_34.

[55] **Zheng X., Burdick D., Popa L., Zhong X., Wang N.X.R.** Global Table Extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. Processing of the 2021 IEEE Winter Conference on Applications of Computer Vision. 2021: 697–706. DOI:10.1109/WACV48630.2021.00074.

[56] **Hashmi K.A., Stricker D., Liwicki M., Afzal M.N., Afzal M.Z.** Guided table structure recognition through anchor optimization. IEEE Access. 2021; (9):113521–113534. DOI:10.1109/ACCESS.2021.3103413.

[57] **Raja S., Mondal A., Jawahar C.V.** Table structure recognition using top-down and bottom-up cues. Processing of the 16th European Conference on Computer Vision. Pt XXVIII. UK: Glasgow; 2020: 70–86. DOI:10.1007/978-3-030-58604-1_5.

[58] **Ren S., He K., Girshick R., Sun J.** Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015; 39(6):1137–1149. DOI:10.1109/TPAMI.2016.2577031.

[59] **He K., Gkioxari G., Dollar P., Girshick R.** Mask R-CNN. 2017 IEEE International Conference on Computer Vision. Italy: Venice; 2017: 2980–2988. DOI:10.1109/ICCV.2017.322.

[60] **Cai Z., Vasconcelos N.** Cascade R-CNN: high quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021; 43(5):1483–1498. DOI:10.1109/TPAMI.2019.2956516.

[61] **Redmon J., Farhadi A.** YOLOv3: an incremental improvement. arXiv Preprint. arXiv:1804.02767. DOI:10.48550/arXiv.1804.02767. Available at: `https://arxiv.org/abs/1804.02767`.

[62] **Lin T.-Y., Goyal P., Girshick R., He K., Dollar P.** Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020; 42(2):318–327. DOI:10.1109/TPAMI.2018.2858826.

[63] **Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A.C.** SSD: single shot multibox detector. Processing of the 14th European Conference on Computer Vision. Pt I. Netherlands: Amsterdam; 2016: 21–37. DOI:10.1007/978-3-319-46448-0_2.

[64] **Simonyan K., Zisserman A.** Very deep convolutional networks for large-scale image recognition. Processing of the 3rd International Conference on Learning Representations. USA: San Diego; 2015: 1–14.

[65] **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. Processing of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. USA: Las Vegas; 2016: 770–778. DOI:10.1109/CVPR.2016.90.

[66] **Xie S., Girshick R., Dollar P., Tu Z., He K.** Aggregated residual transformations for deep neural networks. Processing of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. USA: Honolulu; 2017: 5987–5995. DOI:10.1109/CVPR.2017.634.

[67] **Tan M., Le Q.V.** EfficientNet: rethinking model scaling for convolutional neural networks. Processing of the 36th International Conference on Machine Learning. Long Beach, USA; 2019: 6105–6114.

[68] **Deng J., Dong W., Socher R., Li L.-J., Li K., Li F.-F.** ImageNet: a large-scale hierarchical image database. Processing of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. USA: Miami; 2009: 248–255. DOI:10.1109/CVPR.2009.5206848.

[69] **Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollar P., Zitnick C.L.** Microsoft COCO: common objects in context. Processing of the 13th European Conference on Computer Vision. Switzerland: Zurich; 2014: 740–755. DOI:10.1007/978-3-319-10602-1_48.

[70] **Everingham M., van Gool L., Williams C.K.I., Winn J., Zisserman A.** The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision. 2010; 88(2):303–338. DOI:10.1007/s11263-009-0275-4.

[71] **Shahab A., Shafait F., Kieninger T., Dengel A.** An open approach towards the benchmarking of table structure recognition systems. Processing of the 9th IAPR International Workshop on Document Analysis Systems. USA: Boston; 2010: 113–120. DOI:10.1145/1815330.1815345.

[72] **Fang J., Tao X., Tang Z., Qiu R., Liu Y.** Dataset, ground-truth and performance metrics for table detection evaluation. Processing of the 10th International Workshop on Document Analysis Systems. Australia: Gold Coast; 2012: 445–449. DOI:10.1109/DAS.2012.29.

[73] **Göbel M., Hassan T., Oro E., Orsi G.** ICDAR 2013 table competition. Processing of the 12th International Conference on Document Analysis and Recognition. USA: Washington; 2013: 1449–1453. DOI:10.1109/ICDAR.2013.292.

[74] **Gao L., Yi X., Jiang Z., Hao L., Tang Z.** ICDAR2017 competition on page object detection. Processing of the 14th International Conference on Document Analysis and Recognition. Japan: Kyoto; 2017: 1417–1422. DOI:10.1109/ICDAR.2017.231.

[75] **Gao L., Huang Y., Dejean H., Meunier J.-L., Yan Q., Fang Y., Kleber F., Lang E.** ICDAR 2019 competition on table detection and recognition (cTDaR). Processing of the International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1510–1515. DOI:10.1109/ICDAR.2019.00243.

[76] **Zhong X., Tang J., Yepes A.J.** PubLayNet: largest dataset ever for document layout analysis. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1015–1022. DOI:10.1109/ICDAR.2019.00166.

[77] **Göbel M., Hassan T., Oro E., Orsi G.** A methodology for evaluating algorithms for table understanding in PDF documents. Processing of the ACM Symposium on the Document Engineering. France: Paris; 2012: 45–48. DOI:10.1145/2361354.2361365.

[78] **Tran D.N., Tran T.A., Oh A., Kim S.H., Na I.S.** Table detection from document image using vertical arrangement of text blocks. International Journal of Contents. 2015; 11(4):77–85. DOI:10.5392/IJoC.2015.11.4.077.

[79] **e Silva A.C.** Parts that add up to a whole: a framework for the analysis of tables: Ph.D. thesis. Edinburgh, UK: University of Edinburgh; 2010: 266.

[80] **Hao L., Gao L., Yi X., Tang Z.** A table detection method for PDF documents based on convolutional neural networks. Processing of the 12th IAPR Workshop on Document Analysis Systems. 2016: 287–292. DOI:10.1109/DAS.2016.23.

[81] **Namysl M., Esser A., Behnke S., Kohler J.** Flexible table recognition and semantic interpretation system. Processing of the 17th International Conference on Computer Vision Theory and Applications. 2021. DOI:10.5220/0010767600003124.

[82] **Nurminen A.** Algorithmic extraction of data in tables in PDF documents: Ms. thesis. Tampere, Finland: Tampere University of Technology; 2013: 64.

[83] **Shigarov A., Altaev A., Mikhailov A., Paramonov V., Cherkashin E.** TabbyPDF: web-based system for PDF table extraction. Processing of the 24th International Conference on Information and Software Technologies. Lithuania: Vilnius; 2018: 257–269. DOI:10.1007/978-3-319-99972-2_20.

Вычислительные технологии, 2024, том 29, № 6, с. 125–146. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 6, pp. 125–146. © FRC ICT, 2024 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2024.29.6.008

Table recognition in untagged PDF documents using PDF-specific features

A. O. Shigarov

Matrosov Institute for System Dynamics and Control Theory SB RAS, 664033, Irkutsk, Russia
Corresponding author: Alexey O. Shigarov, e-mail: shigarov@icc.ru
Received September 20, 2024, accepted September 27, 2024.

Abstract

Nowadays, PDF is one of the most popular formats for distributing print-oriented documents in the electronic environment. PDF documents are often untagged, i. e. pages are represented only by low-level instructions for rendering text and graphics and are not accompanied by annotations of their structural components (headings, paragraphs, tables, etc.). Automatic recovering for such annotations can ensure the accessibility of structural components. The latter is possible as a result of solving a number of tasks, one of which is recognizing tables in untagged PDF documents: detecting the boundaries of their rows, columns, and cells. This paper proposes a method for recognizing tables in untagged PDF documents. Unlike existing analogues, it is originally proposed to solve the stated task based on the use of PDF-specific features such as text output order, pen movement positions, etc. This proposal allowed adapting some known approaches and methods to the declared task, initially oriented towards raster images and unformatted text, including "word clustering", "rows first" detection, whitespace segmentation, and connected component analysis. The presented performance evaluation results demonstrate the effectiveness of solutions implementing this method. The presented results of the performance evaluation demonstrate the efficiency of the solutions implemented based on the proposed method. Quantitative comparison with analogues indicates their compliance with the current level of technology development in the area under consideration. At the same time, qualitative comparison reveals the following advantages over analogues. The implementation of the proposed table recognition method does not require preliminary parameter adjustment and supervised learning. However, if ready-to-use neural network models are available, they can replace rule-based table detection algorithms. At the same time, the quality of the final results can be improved by applying filtering of candidate cases.

Keywords: table recognition, table extraction, unstructured data, document tables, document page layout analysis.

Citation: Shigarov A.O. Table recognition in untagged PDF documents using PDF-specific features. Computational Technologies. 2024; 29(6):125–146. DOI:10.25743/ICT.2024.29.6.008. (In Russ.)

References

1. **Turró M.R.** Are PDF documents accessible? Information Technology and Libraries. 2008; 27(3):25–43. DOI:10.6017/ital.v27i3.3246.
2. **Nganji J.T.** The Portable Document Format (PDF) accessibility practice of four journal publishers. Library & Information Science Research. 2015; 37(3):254–262. DOI:10.1016/j.lisr.2015.02.002.
3. **Shigarov A.** Table understanding: problem overview. WIREs Data Mining and Knowledge Discovery. 2023; 13(1):e1482. DOI:10.1002/widm.1482.
4. **Yepes A., Zhong P., Burdick D.** ICDAR 2021 competition on scientific literature parsing. Processing of the 16th International Conference on Document Analysis and Recognition. Pt IV. Switzerland: Lausanne; 2021: 605–617. DOI:10.1007/978-3-030-86337-1_40.

5. **Chi Z., Huang H., Xu H.-D., Yu H., Yin W., Mao X.-L.** Complicated table structure recognition. arXiv preprint, arXiv:1908.04729. 2019. DOI:10.48550/arXiv.1908.04729. Available at: https://arxiv.org/abs/1908.04729.
6. **Adams T., Namysl M., Kodamullil A.T., Behnke S., Jacobs M.** Benchmarking table recognition performance on biomedical literature on neurological disorders. Bioinformatics. 2021; 38(6):1624–1630. DOI:10.1093/bioinformatics/btab843.
7. **Zhang M., Perelman D., Le V., Gulwani S.** An integrated approach of deep learning and symbolic analysis for digital PDF table extraction. Processing of the 25th International Conference on Pattern Recognition. Italy: Milan; 2021: 4062–4069. DOI:10.1109/ICPR48806.2021.9413069.
8. **Kieninger T.** Table structure recognition based on robust block segmentation. Document Recognition V. USA: San Jose; 1998: 22–32. DOI:10.1117/12.304642.
9. **Shigarov A., Fedorov R.** Simple algorithm page layout analysis. Pattern Recognition and Image Analysis. 2011; 21(2):324–327. DOI:10.1134/S1054661811021008.
10. **Shigarov A.O., Paramonov V.V.** Page text segmentation in untagged PDF documents. Computational Technologies. 2022; 27(5):69–78. DOI:10.25743/ICT.2022.27.5.007. (In Russ.)
11. **Bychkov I.V., Rugnikov G.M., Hmelnov A.E., Shigarov A.O.** A heuristic method of table detection in documents of various formats. Computational Technologies. 2009; 14(2):58–73. (In Russ.)
12. **Cherepanov I., Mikhailov A., Shigarov A., Paramonov V.** On automated workflow for finetuning deep neural network models for table detection in document images. Processing of the 43rd International Convention on Information, Communication and Electronic Technology. Croatia: Opatija; 2020: 1130–1133. DOI:10.23919/MIPRO48935.2020.9245241.
13. **Mikhailov A., Shigarov A., Rozhkov E., Cherepanov I.** On graph-based verification for PDF table detection. Processing of the 2020 Ivannikov ISPRAS Open Conference. Russia: Moscow; 2020: 91–95. DOI:10.1109/ISPRAS51486.2020.00020.
14. **Shigarov A., Mikhailov A., Altaev A.** Configurable table structure recognition in untagged PDF documents. Processing of the 2016 ACM Symposium on Document Engineering. Austria: Vienna; 2016: 119–122. DOI:10.1145/2960811.2967152.
15. **Fang J., Gao L., Bai K., Qiu R., Tao X., Tang Z.** A table detection method for multipage PDF documents via visual separators and tabular structures. Processing of the 11th International Conference on Document Analysis and Recognition. China: Beijing; 2011: 779–783. DOI:10.1109/ICDAR.2011.304.
16. **Bart E.** Parsing tables by probabilistic modeling of perceptual cues. Processing of the 10th IAPR International Workshop on Document Analysis Systems. Australia: Gold Coast; 2012: 409–414. DOI:10.1109/DAS.2012.67.
17. **Cesarini F., Marinai S., Sarti L., Soda G.** Trainable table location in document images. Processing of the International Conference on Pattern Recognition. Vol. 3. Canada: Quebec; 2002: 236–240. DOI:10.1109/ICPR.2002.1047838.
18. **Wang Y., Phillips I., Haralick R.** Table structure understanding and its performance evaluation. Pattern Recognition. 2004; 37(7):1479–1497.
19. **Klampfl S., Jack K., Kern R.** A comparison of two unsupervised table recognition methods from digital scientific articles. D-Lib Magazine. 2014; 20(11/12). DOI:10.1045/november14-klampfl.
20. **Jaekyu H., Haralick R.M., Phillips I.T.** Recursive X-Y cut using bounding boxes of connected components. Processing of the 3rd International Conference on Document Analysis and Recognition. Vol. 2. Canada: Montreal; 1995: 952–955.
21. **Kieninger T., Dengel A.** The T-Recs table recognition and analysis system. Document Analysis Systems: Theory and Practice. Japan: Nagano; 1999: 255–270. DOI:10.1007/3-540-48172-9_21.
22. **Yildiz B., Kaiser K., Miksch S.** pdf2table: a method to extract table information from PDF files. Processing of the 2nd Indian International Conference on Artificial Intelligence. India: Pune; 2005: 1773–1785.
23. **Hassan T., Baumgartner R.** Table recognition and understanding from PDF files. Processing of the 9th International Conference on Document Analysis and Recognition. Vol. 2. Brazil: Parana; 2007: 1143–1147. DOI:10.1109/ICDAR.2007.4377094.
24. **Shafait F., Smith R.** Table detection in heterogeneous documents. Processing of the 9th IAPR International Workshop on Document Analysis Systems. USA: Boston; 2010: 65–72. DOI:10.1145/1815330.1815339.

25. **Deckert F., Seidler B., Ebbecke M., Gillmann M.** Table content understanding in SmartFIX. Processing of the 11th International Conference on Document Analysis and Recognition. China: Beijing; 2011: 488–492. DOI:10.1109/ICDAR.2011.104.
26. **Bansal A., Harit G., Roy S.D.** Table extraction from document images using fixed point model. Processing of the Indian Conference on Computer Vision Graphics and Image Processing. India: Bangalore; 2014: 67:1–67:8. DOI:10.1145/2683483.2683550.
27. **Handley J.** Table analysis for multiline cell identification. Document Recognition and Retrieval VIII. USA: San Jose; 2000: 34–43. DOI:10.1117/12.410853.
28. **Pinto D., McCallum A., Wei X., Croft W.B.** Table extraction using conditional random fields. Processing of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. Canada: Toronto; 2003: 235–242. DOI:10.1145/860435.860479.
29. **e Silva A.C., Jorge A.M., Torgo L.** Design of an end-to-end method to extract information from tables. International Journal of Document Analysis and Recognition. 2006; 8(2–3):144–171. DOI:10.1007/s10032-005-0001-x.
30. **Liu Y., Mitra P., Giles C.L.** Identifying table boundaries in digital documents via sparse line detection. Processing of the 17th ACM Conference on Information and Knowledge Management. USA: Napa Valley; 2008: 1311–1320. DOI:10.1145/1458082.1458255.
31. **Oro E., Ruffolo M.** PDF-TREX: an approach for recognizing and extracting tables from PDF documents. Processing of the 10th International Conference on Document Analysis and Recognition. Spain: Barcelona; 2009: 906–910. DOI:10.1109/ICDAR.2009.12.
32. **Hashmi K.A., Liwicki M., Stricker D., Afzal M.A., Afzal M.A., Afzal M.Z.** Current status and performance analysis of table recognition in document images with deep neural networks. IEEE Access. 2021; (9):87663–87685. DOI:10.1109/ACCESS.2021.3087865.
33. **Gilani A., Qasim S.R., Malik I., Shafait F.** Table detection using deep learning. Processing of the 14th International Conference on Document Analysis and Recognition, Vol. 1. Japan: Kyoto; 2017: 771–776. DOI:10.1109/ICDAR.2017.131.
34. **Schreiber S., Agne S., Wolf I., Dengel A., Ahmed S.** DeepDeSRT: deep learning for detection and structure recognition of tables in document images. Processing of the 14th International Conference on Document Analysis and Recognition. Vol. 1. Japan: Kyoto; 2017: 1162–1167. DOI:10.1109/ICDAR.2017.192.
35. **He D., Cohen S., Price B., Kifer D., Giles C.L.** Multi-scale multi-task FCN for semantic page segmentation and table detection. Processing of the 14th International Conference on Document Analysis and Recognition. Vol. 1. Japan: Kyoto; 2017: 254–261. DOI:10.1109/ICDAR.2017.50.
36. **Arif S., Shafait F.** Table detection in document images using foreground and background features. Processing of the Digital Image Computing: Techniques and Applications. Australia: Canberra; 2018: 1–8. DOI:10.1109/DICTA.2018.8615795.
37. **Siddiqui S.A., Malik M.I., Agne S., Dengel A., Ahmed S.** DeCNT: deep deformable CNN for table detection. IEEE Access. 2018; (6):74151–74161. DOI:10.1109/ACCESS.2017.
38. **Li Y., Gao L., Tang Z., Yan Q., Huanget Y.** A GAN-based feature generator for table detection. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 763–768. DOI:10.1109/ICDAR.2019.00127.
39. **Paliwal S.S., Vishwanath D., Rahul R., Sharma M., Vig L.** TableNet: deep learning model for end-to-end table detection and tabular data extraction from scanned document images. 2019 International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 128–133. DOI:10.1109/ICDAR.2019.00029.
40. **Kavasidis I., Pino C., Palazzo S., Rundo F., Giordano D., Spampinato C.** A saliency-based convolutional neural network for table and chart detection in digitized documents. Image Analysis and Processing. 2019: 292–302. DOI:10.1007/978-3-030-30645-8_27.
41. **Holeček M., Hoskovec A., Baudiš P., Klinger P.** Table understanding in structured documents. 2019 International Conference on Document Analysis and Recognition Workshops. Vol. 5. Australia: Sydney; 2019: 158–164. DOI:10.1109/ICDARW.2019.40098.
42. **Riba P., Dutta A., Goldmann L., Fornés A., Ramos O., Lladós J.** Table detection in invoice documents by graph neural networks. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 122–127. DOI:10.1109/ICDAR.2019.00028.

43. **Huang Y., Yan Q., Li Y., Chen Y., Wang X., Gao L., Tang Z.** A YOLO-based table detection method. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 813–818. DOI:10.1109/ICDAR.2019.00135.
44. **Sun N., Zhu Y., Hu X.** Faster R-CNN based table detection combining corner locating. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 1314–1319. DOI:10.1109/ICDAR.2019.00212.
45. **Li M., Cui L., Huang S., Wei F., Zhou M., Li Z.** TableBank: table benchmark for image-based table detection and recognition. Processing of the 12th Language Resources and Evaluation Conference. France: Marseille; 2020: 1918–1925.
46. **Casado-García Á., Domínguez C., Heras J., Mata E., Pascual V.** The benefits of close-domain fine-tuning for table detection in document images. Document Analysis Systems. Springer Nature; 2020: 199–215. DOI:10.1007/978-3-030-57058-3_15.
47. **Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure K.** CascadeTabNet: an approach for end-to-end table detection and structure recognition from image-based documents. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. USA: Seattle; 2020: 2439–2447. DOI:10.1109/CVPRW50498.2020.00294.
48. **Agarwal M., Mondal A., Jawahar C.V.** CDeC-Net: composite deformable cascade network for table detection in document images. Processing of the 25th International Conference on Pattern Recognition. Italy: Milan; 2021: 9491–9498. DOI:10.1109/ICPR48806.2021.9411922.
49. **Deng Y., Rosenberg D., Mann G.** Challenges in end-to-end neural scientific table recognition. 2019 International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 894–901. DOI:10.1109/ICDAR.2019.00148.
50. **Khan S.A., Khalid S.M.D., Shahzad M.A., Shafait F.** Table structure extraction with bidirectional gated recurrent unit networks. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1366–1371. DOI:10.1109/ICDAR.2019.00220.
51. **Qasim S., Mahmood H., Shafait F.** Rethinking table recognition using graph neural networks. International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 142–147. DOI:10.1109/ICDAR.2019.00031.
52. **Siddiqui S.A., Fateh I.A., Rizvi S.T.R., Dengel A., Ahmed S.** DeepTabStR: deep learning based table structure recognition. Processing of the 15th International Conference Document Analysis and Recognition. Australia: Sydney; 2019: 1403–1409. DOI:10.1109/ICDAR.2019.00226.
53. **Tensmeyer C., Morariu V., Price B., Cohen S., Martinez T.** Deep splitting and merging for table structure decomposition. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 114–121. DOI:10.1109/ICDAR.2019.00027.
54. **Zhong X., Bavani E.S., Yepes A.J.** Image-based table recognition: data, model, and evaluation. Processing of the 16th European Conference on Computer Vision. UK: Glasgow; 2020: 564–580. DOI:10.1007/978-3-030-58589-1_34.
55. **Zheng X., Burdick D., Popa L., Zhong X., Wang N.X.R.** Global Table Extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. Processing of the 2021 IEEE Winter Conference on Applications of Computer Vision. 2021: 697–706. DOI:10.1109/WACV48630.2021.00074.
56. **Hashmi K.A., Stricker D., Liwicki M., Afzal M.N., Afzal M.Z.** Guided table structure recognition through anchor optimization. IEEE Access. 2021; (9):113521–113534. DOI:10.1109/ACCESS.2021.3103413.
57. **Raja S., Mondal A., Jawahar C.V.** Table structure recognition using top-down and bottom-up cues. Processing of the 16th European Conference on Computer Vision. Pt XXVIII. UK: Glasgow; 2020: 70–86. DOI:10.1007/978-3-030-58604-1_5.
58. **Ren S., He K., Girshick R., Sun J.** Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015; 39(6):1137–1149. DOI:10.1109/TPAMI.2016.2577031.
59. **He K., Gkioxari G., Dollar P., Girshick R.** Mask R-CNN. 2017 IEEE International Conference on Computer Vision. Italy: Venice; 2017: 2980–2988. DOI:10.1109/ICCV.2017.322.
60. **Cai Z., Vasconcelos N.** Cascade R-CNN: high quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021; 43(5):1483–1498. DOI:10.1109/TPAMI.2019.2956516.

61. **Redmon J., Farhadi A.** YOLOv3: an incremental improvement. arXiv Preprint. arXiv:1804.02767. DOI:10.48550/arXiv.1804.02767. Available at: `https://arxiv.org/abs/1804.02767`.
62. **Lin T.-Y., Goyal P., Girshick R., He K., Dollar P.** Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020; 42(2):318–327. DOI:10.1109/TPAMI.2018.2858826.
63. **Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A.C.** SSD: single shot multibox detector. Processing of the 14th European Conference on Computer Vision. Pt I. Netherlands: Amsterdam; 2016: 21–37. DOI:10.1007/978-3-319-46448-0_2.
64. **Simonyan K., Zisserman A.** Very deep convolutional networks for large-scale image recognition. Processing of the 3rd International Conference on Learning Representations. USA: San Diego; 2015: 1–14.
65. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. Processing of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. USA: Las Vegas; 2016: 770–778. DOI:10.1109/CVPR.2016.90.
66. **Xie S., Girshick R., Dollar P., Tu Z., He K.** Aggregated residual transformations for deep neural networks. Processing of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. USA: Honolulu; 2017: 5987–5995. DOI:10.1109/CVPR.2017.634.
67. **Tan M., Le Q.V.** EfficientNet: rethinking model scaling for convolutional neural networks. Processing of the 36th International Conference on Machine Learning. Long Beach, USA; 2019: 6105–6114.
68. **Deng J., Dong W., Socher R., Li L.-J., Li K., Li F.-F.** ImageNet: a large-scale hierarchical image database. Processing of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. USA: Miami; 2009: 248–255. DOI:10.1109/CVPR.2009.5206848.
69. **Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollar P., Zitnick C.L.** Microsoft COCO: common objects in context. Processing of the 13th European Conference on Computer Vision. Switzerland: Zurich; 2014: 740–755. DOI:10.1007/978-3-319-10602-1_48.
70. **Everingham M., van Gool L., Williams C.K.I., Winn J., Zisserman A.** The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision. 2010; 88(2):303–338. DOI:10.1007/s11263-009-0275-4.
71. **Shahab A., Shafait F., Kieninger T., Dengel A.** An open approach towards the benchmarking of table structure recognition systems. Processing of the 9th IAPR International Workshop on Document Analysis Systems. USA: Boston; 2010: 113–120. DOI:10.1145/1815330.1815345.
72. **Fang J., Tao X., Tang Z., Qiu R., Liu Y.** Dataset, ground-truth and performance metrics for table detection evaluation. Processing of the 10th International Workshop on Document Analysis Systems. Australia: Gold Coast; 2012: 445–449. DOI:10.1109/DAS.2012.29.
73. **Göbel M., Hassan T., Oro E., Orsi G.** ICDAR 2013 table competition. Processing of the 12th International Conference on Document Analysis and Recognition. USA: Washington; 2013: 1449–1453. DOI:10.1109/ICDAR.2013.292.
74. **Gao L., Yi X., Jiang Z., Hao L., Tang Z.** ICDAR2017 competition on page object detection. Processing of the 14th International Conference on Document Analysis and Recognition. Japan: Kyoto; 2017: 1417–1422. DOI:10.1109/ICDAR.2017.231.
75. **Gao L., Huang Y., Dejean H., Meunier J.-L., Yan Q., Fang Y., Kleber F., Lang E.** ICDAR 2019 competition on table detection and recognition (cTDaR). Processing of the International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1510–1515. DOI:10.1109/ICDAR.2019.00243.
76. **Zhong X., Tang J., Yepes A.J.** PubLayNet: largest dataset ever for document layout analysis. Processing of the 15th International Conference on Document Analysis and Recognition. Australia: Sydney; 2019: 1015–1022. DOI:10.1109/ICDAR.2019.00166.
77. **Göbel M., Hassan T., Oro E., Orsi G.** A methodology for evaluating algorithms for table understanding in PDF documents. Processing of the ACM Symposium on the Document Engineering. France: Paris; 2012: 45–48. DOI:10.1145/2361354.2361365.
78. **Tran D.N., Tran T.A., Oh A., Kim S.H., Na I.S.** Table detection from document image using vertical arrangement of text blocks. International Journal of Contents. 2015; 11(4):77–85. DOI:10.5392/IJoC.2015.11.4.077.
79. **e Silva A.C.** Parts that add up to a whole: a framework for the analysis of tables: Ph.D. thesis. Edinburgh, UK: University of Edinburgh; 2010: 266.

80. **Hao L., Gao L., Yi X., Tang Z.** A table detection method for PDF documents based on convolutional neural networks. Processing of the 12th IAPR Workshop on Document Analysis Systems. 2016: 287–292. DOI:10.1109/DAS.2016.23.
81. **Namysl M., Esser A., Behnke S., Kohler J.** Flexible table recognition and semantic interpretation system. Processing of the 17th International Conference on Computer Vision Theory and Applications. 2021. DOI:10.5220/0010767600003124.
82. **Nurminen A.** Algorithmic extraction of data in tables in PDF documents: Ms. thesis. Tampere, Finland: Tampere University of Technology; 2013: 64.
83. **Shigarov A., Altaev A., Mikhailov A., Paramonov V., Cherkashin E.** TabbyPDF: web-based system for PDF table extraction. Processing of the 24th International Conference on Information and Software Technologies. Lithuania: Vilnius; 2018: 257–269. DOI:10.1007/978-3-319-99972-2_20.