###### **ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

###### **"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

###### **"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"**

###### **ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК**

САТАЕВ ДМИТРИЙ АЛЕКСАНДРОВИЧ

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

РАСПОЗНАВАНИЕ ТЕКСТА ИЗ ТАБЛИЦ В ДОКУМЕНТАХ

TEXT RECOGNITION FROM TABLES IN DOCUMENTS

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Анализ данных в девелопменте»

Научный руководитель

д-р …. наук/проф/должность

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Е. А. Анкудинова

Студент

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Д. А. Сатаев

Москва 2025

Содержание

Оглавление

**Элементы оглавления не найдены.**

# Введение

В условиях стремительной цифровизации бизнес-процессов и документооборота возрастает потребность в автоматизации обработки информации, представленной в неструктурированном виде. Особое место среди таких данных занимают таблицы, используемые для систематизации и представления важной информации в отчетах, актах, договорах и других типах документов. Несмотря на визуальную упорядоченность, автоматическая обработка таблиц по-прежнему представляет собой нетривиальную задачу. В отличие от обычного текста, таблицы могут иметь произвольное расположение в документе, содержать сложные вложенные структуры и сильно различаться по оформлению. Это затрудняет их интеграцию в информационные системы и требует значительных ресурсов при ручной обработке.

В связи с этим разработка универсального и надежного сервиса, способного находить таблицы в документе и распознавать их содержимое, является важной и актуальной задачей. Такой сервис может существенно упростить и ускорить процессы цифрового документооборота, снизить нагрузку на сотрудников, обеспечить более высокую точность и полноту извлекаемой информации.

Представляемая в рамках данной работы система направлена на решение этой задачи, фокусируясь на извлечении таблиц из изображений документов и преобразовании их в структурированный цифровой формат.

Для реализации поставленной цели необходимо провести анализ существующих подходов к извлечению и распознаванию таблиц, разработать архитектуру решения с разделением на отдельные модули детекции и распознавания, реализовать соответствующие компоненты системы, обеспечить их взаимодействие и провести оценку качества работы на реальных данных. Кроме того, в ходе работы важно определить области практического применения предложенного подхода и обозначить перспективы его дальнейшего развития.

1. Анализ предметной области.
   1. Обзор проблем автоматической обработки таблиц.

Несмотря на значительный прогресс в области компьютерного зрения и обработки документов, автоматическая обработка таблиц остаётся сложной и актуальной задачей. Таблицы, в отличие от сплошного текста, обладают как визуальной, так и логической структурой, которую необходимо корректно интерпретировать для дальнейшего анализа. Однако существуют ряд факторов, существенно затрудняющих этот процесс.

Во-первых, визуальное разнообразие таблиц представляет собой серьёзную проблему. Таблицы могут отличаться по количеству строк и столбцов, плотности содержимого, типу шрифтов, а также стилевому оформлению. Более того, документы, содержащие таблицы, часто представлены в виде изображений (например, отсканированных копий), где элементы таблицы могут быть искажены, сжаты или содержать артефакты, вызванные качеством сканирования.

Во-вторых, распознавание структуры таблицы требует понимания логических связей между её ячейками. Наличие объединённых ячеек, вложенных таблиц, различной ширины столбцов и высоты строк усложняет задачу восстановления исходной логики данных.

* 1. Существующие подходы и решения.

Amazon Textract

Amazon Textract — это облачный сервис, разработанный компанией Amazon в рамках платформы AWS (Amazon Web Services), предназначенный для интеллектуального извлечения данных из документов. Он способен автоматически определять текст, формы, таблицы и даже связи между полями. Особое внимание уделяется работе с таблицами: Textract может извлекать как структуру таблицы (строки и столбцы), так и её содержимое, независимо от положения на странице.

Преимущества:

1. Поддержка документов различных форматов (PDF, изображения).
2. Способность распознавать таблицы без явных границ.
3. Экспорт результатов в структурированном виде (JSON, CSV).
4. Масштабируемость и готовность к промышленному использованию.

Недостатки:

1. Ограниченная поддержка русскоязычных документов.
2. Затратная модель оплаты при большом объеме данных.
3. Невозможность локального развертывания.

Img2table

img2table — это библиотека с открытым исходным кодом на Python, предназначенная для извлечения таблиц из изображений и PDF-документов. В отличие от облачных сервисов, img2table работает локально и не требует подключения к внешним API, что делает её привлекательной для исследовательских проектов, систем с ограничениями по безопасности или разработчиков, стремящихся к полной контролируемости процесса.

Преимущества:

1. Поддержка документов в формате изображения или PDF.
2. Детекция таблиц с границами и извлечение их структуры.
3. Интеграция с OCR-системой Tesseract для распознавания текста внутри ячеек.
4. Сохранение результатов в виде Pandas DataFrame или Excel-файлов.
5. Простота в установке и использовании.

Недостатки:

1. Чувствительная к качеству изображений и сканов.
2. Менее точна по сравнению с современными нейросетевыми решениями.
   1. Выводы и результаты по главе.

Автоматическая обработка таблиц в документах остаётся сложной и актуальной задачей, несмотря на наличие большого количества инструментов и технологий. Основные сложности связаны с визуальным разнообразием таблиц, отсутствием единых стандартов разметки, сложной структурой содержимого и вариативностью форматов документов. Особенно это проявляется при работе с отсканированными изображениями, где таблицы могут быть частично повреждены, искажены или вовсе не иметь чётких визуальных границ.

1. Архитектура и проектирование системы.
   1. Общая архитектура системы.

При проектировании системы для распознавания текста из таблиц в документах основной целью является создание модульного, масштабируемого и легко расширяемого решения, которое можно адаптировать под различные форматы документов и сценарии использования. Предполагается реализация архитектуры, основанной на разделении логики на независимые сервисы, каждый из которых отвечает за свою часть обработки данных.

В качестве архитектурной основы выбирается микросервисный подход. Это позволяет разделить систему на два ключевых компонента — сервис детекции таблиц и сервис распознавания структуры и содержимого таблиц. Такой подход обеспечивает гибкость при развитии системы, возможность параллельной разработки и масштабирования отдельных модулей, а также повышает устойчивость к сбоям.

Сервис детекции таблиц проектируется как отдельный компонент, который принимает на вход изображение документа (в формате PNG, JPEG или страницы PDF) и возвращает координаты областей, содержащих таблицы. Для реализации этого компонента планируется обучить модель глубокого обучения на соответствующем датасете. Детекция выполняется с учётом требований к скорости обработки и точности при работе с различными типами документов.

Следующим этапом обработки является интерпретация и структурирование содержимого в обнаруженных таблицах. Проектируется сервис, который будет получать изображения таблиц (обрезанные по координатам из предыдущего этапа), выполнять OCR и восстанавливать структуру строк и столбцов. Для OCR предполагается использовать Tesseract или его аналоги, а для анализа структуры — методы постобработки на основе визуальных признаков и координат текста.

Сервисы взаимодействуют друг с другом через REST API. Такой подход обеспечивает простоту интеграции, возможность распределённого запуска и независимого масштабирования. Также предусматривается возможность последовательной обработки через очередь задач или REST-запросы от клиента к каждому сервису поэтапно.

Проектируемая архитектура ориентирована на максимальную адаптивность и возможность замены отдельных компонентов без изменения всей системы. Такой подход позволяет использовать как готовые модели, так и собственные, дообученные под конкретные наборы документов. В дальнейшем система может быть дополнена пользовательским интерфейсом для ручной аннотации, а также механизмами активного обучения и самообучающейся разметки.

* 1. Выбор инструментов и технологий.

Для реализации системы автоматического извлечения и распознавания таблиц из изображений документов были выбраны технологии, обеспечивающие баланс между качеством обработки, гибкостью настройки, простотой внедрения и возможностью доработки. В рамках проектируемой архитектуры инструменты и библиотеки подбирались отдельно для каждого из компонентов.

* + 1. Сервис детекции таблиц (DetectionService).

Цель: локализовать таблицы на изображении документа с высокой точностью и устойчивостью к различным видам искажения (качество скана, наклон, отсутствие границ и др.).

Для решения задачи автоматического обнаружения таблиц на изображениях документов в проектируемой системе выбрана модель YOLO (You Only Look Once), как одна из наиболее эффективных и широко применяемых архитектур в области детекции объектов.

YOLO относится к одноэтапным алгоритмам обнаружения объектов, в которых детекция и классификация выполняются в рамках одного прохода нейронной сети. Это обеспечивает высокую скорость обработки изображений без существенной потери точности, что делает YOLO подходящим выбором как для исследовательских, так и для прикладных систем.

* + 1. Сервис распознавания данных в таблице (RecognitionService).
    2. Общие инструменты.