УДК 004.9

СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДОКУМЕНТОВ

Кизилова Э.А., студент гр. 12002541, I курс
Нестерова Т.А., студент гр. 12002540, I курс,
Быканов И.Г., студент гр. 12002373, III курс
Научный руководитель: Федоров В.И., к.т.н., доцент
Белгородский государственный национальный исследовательский университет, г. Белгород, Россия

Автоматизация обработки документов является одной из ключевых задач в области компьютерного зрения и обработки естественного языка. За последние семьдесят лет технологии распознавания текста прошли путь от простых систем шаблонного сопоставления до современных нейросетевых архитектур, способных не только извлекать текст с изображения, но и понимать его семантику. Эволюция подходов к обработке документов отражает общий тренд развития искусственного интеллекта: от жёстко запрограммированных правил к обучаемым моделям, от изолированных компонентов к end-to-end системам, от анализа текста к совместной обработке визуальной и текстовой информации. В контексте задачи автоматизации документооборота приёмной комиссии критически важно понимание возможностей и ограничений различных подходов к обработке документов, что позволяет обосновать выбор технологического решения для практической реализации системы.

Появление в 2020-2023 годах крупных мультимодальных моделей (Large Multimodal Models, LMM), способных совместно обрабатывать изображения и текст, открыло принципиально новые возможности для задач документального анализа. В отличие от традиционных OCR-систем, которые "видят" лишь набор символов, Vision-Language модели (VLM) обладают семантическим пониманием содержимого документа.

Современные VLM строятся на базе трансформерной архитектуры (Vaswani et al., 2017 [1]) с двумя основными компонентами.

Vision Encoder — преобразует изображение в последовательность векторных представлений (embeddings):

- ViT (Vision Transformer) — разбивает изображение на патчи 16×16 пикселей, каждый патч обрабатывается как токен;
- CLIP-подобные энкодеры — обучаются на задаче сопоставления изображений и текстовых описаний.
- Language Model Decoder — генерирует текстовый ответ на основе визуальных представлений:

- Авторегрессивная генерация (GPT-стиль): каждый следующий токен предсказывается с учётом предыдущих

- Механизм cross-attention для связывания визуальных и текстовых модальностей

В ходе сравнительного анализа инструментов для распознавания текста с изображений выявлена принципиальная проблема традиционных оптических распознавателей — низкая надёжность обработки кириллического текста в условиях вариативности качества сканов, шрифтов и оформления официальных документов. Как показали эксперименты на выборке паспортов гражданина РФ, Tesseract и EasyOCR демонстрируют снижение точности на документах со сложной структурой и низким качеством изображений и могут допускать ошибки при распознавании кириллических символов, что делает извлечённый текст непригодным для автоматической дальнейшей обработки без ручной коррекции. Визуально-лингвистическая модель olmoOCR-2, обученная, в том числе, на документах с кириллическим наполнением и использующая контекстное понимание структуры паспорта, обеспечивает корректное распознавание кириллических реквизитов и возвращает данные в структурированном формате, готовом к импорту в информационную систему без этапа постобработки.

Метод автоматического определения ориентации документа на основе каскадной детекции лиц и локализации ключевых точек, решает следующую проблему. При загрузке сканов документов в систему пользователи часто снимают документы на телефон в произвольной ориентации. Паспорт может быть перевёрнут на 90°, 180°, 270° и т.д. Традиционные OCR-системы эффективнее работают с правильно ориентированными изображениями - при неправильном повороте точность распознавания падает или система вообще не может распознать текст. Предлагаемый метод решает эту проблему за счёт использования фотографии владельца как опорного элемента — лицо человека всегда находится в определённой части паспорта (применимо к документам конкретных стран), и его детекция [1] не требует текстового распознавания. Это позволяет быстро определить правильную ориентацию документа перед основной OCR-обработкой.

Метод работает в два этапа.

Этап 1: Грубая ориентация (проверка 4 поворотов). Система последовательно поворачивает изображение на 0°, 90°, 180° и 270°, и для каждого варианта пытается обнаружить лицо с помощью каскадного классификатора Хаара [1,2] (встроенный в OpenCV алгоритм). Для улучшения качества детекции изображение масштабируется [34] в 2 раза — это помогает найти даже небольшие фотографии.

Этап 1 метода автоматической ориентация документов на основе каскадной детекции лиц: грубая ориентация изображения документа на рисунке 1.

Рисунок 1 — Этап 1 метода автоматической ориентация документов на основе каскадной детекции лиц: грубая ориентация изображения документа.

Этап 2: Тонкая коррекция наклона (рис. 2). После определения грубой ориентации документ может иметь небольшой наклон (до 10°) из-за неровного сканирования. Для финальной коррекции используется библиотека MediaPipe [3], которая определяет 468 ключевых точек лица

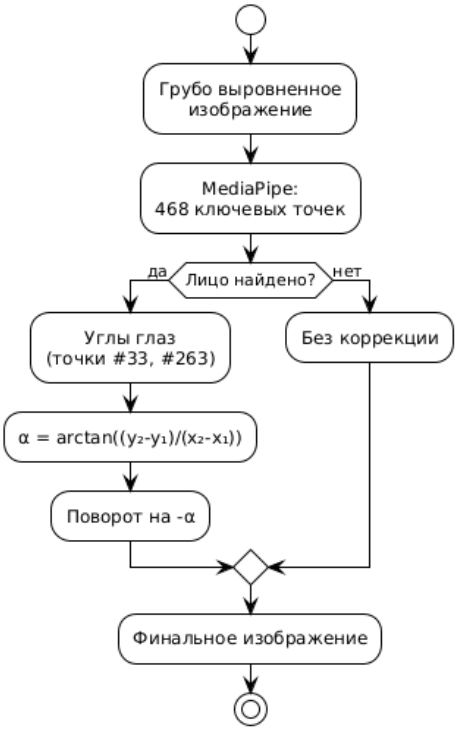

Рисунок 2 — Этап 2 метода автоматической ориентация документов на основе каскадной детекции лиц: грубая ориентация изображения документа

Метод контекстно-ориентированного извлечения структурированных данных с использованием Vision-Language моделей. Традиционные OCR-системы возвращают неструктурированный текст, распознанный с изображения, без понимания семантики документа. Для автоматической работы системы приёмной комиссии требуются структурированные данные в формате, пригодном для непосредственной записи в базу данных. Предлагаемый метод заменяет парсинг использованием Vision-Language модели (VLM), которая одновременно анализирует изображение и текстовую инструкцию. Вместо извлечения «сырого» текста модели задаётся задача на естественном языке: извлечь конкретные поля и вернуть их строго в заданном формате. Модель самостоятельно находит нужные фрагменты документа и формирует структурированный ответ. Основными этапами метода являются:

- предобработка изображения;
- формирование промпта;
- передача данных в модель;
- генерация ответа;
- постобработка;
- нормализация данных.

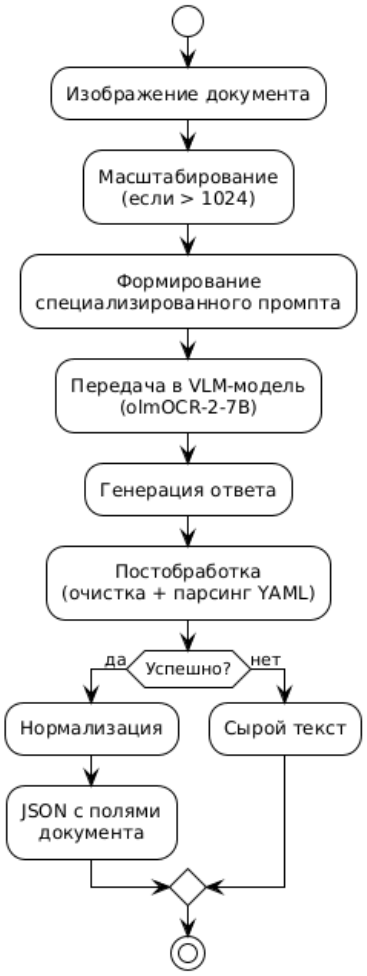

Рисунок 3 — Блок-схема метода контекстно-ориентированного извлечения структурированных данных с использованием Vision-Language моделей.

Метод сегментированной обработки композитных документов с разнородными зонами. Российский паспорт содержит зоны с различной ориентацией текста: основные поля (ФИО, дата рождения, место выдачи) расположены горизонтально, а серия и номер паспорта напечатаны вертикально красным шрифтом справа на развороте. Обработка всего документа единым запросом к VLM-модели приводит к существенному снижению точности распознавания вертикального текста.

Проблема возникает по двум причинам. Во-первых, Vision-Language модели обучаются преимущественно на изображениях с горизонтальным текстом, что делает распознавание вертикальных надписей менее надёжным. Во-вторых, во время практической реализации, предварительная смена ориентации изображения, которая приводит текст в горизонтальное положение, многократно повышала точность распознавания этого текста. На рисунке 4 представлена блок-схема метода сегментированной обработки композитных документов с разнородными зонами.

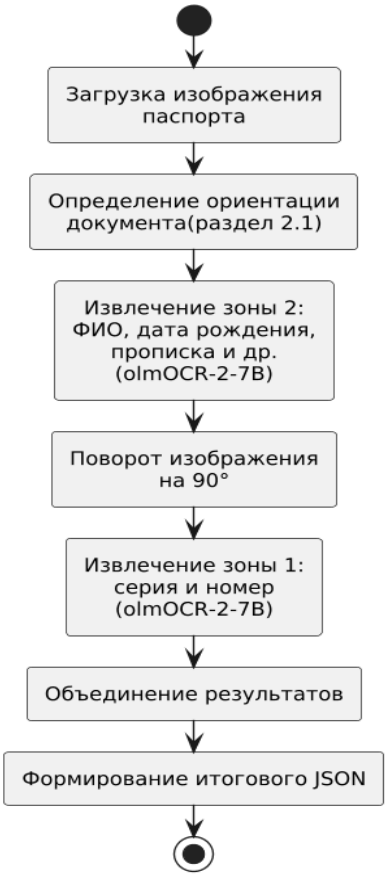

Рисунок 4 — Блок-схема метода сегментированной обработки композитных документов с разнородными зонами

Интерфейс содержит три поля загрузки файлов для паспорта, СНИЛС и документа об образовании, каждое из которых принимает изображения в форматах JPEG, PNG или PDF. Под полями загрузки расположены две кнопки: основная кнопка «Обработать документы» инициирует отправку файлов на обработку, дополнительная кнопка «Проверить состояние сервиса» позволяет

убедиться в доступности микросервиса перед загрузкой документов. На рисунке 5 представлен интерфейс страницы формы ИИ ассистента.

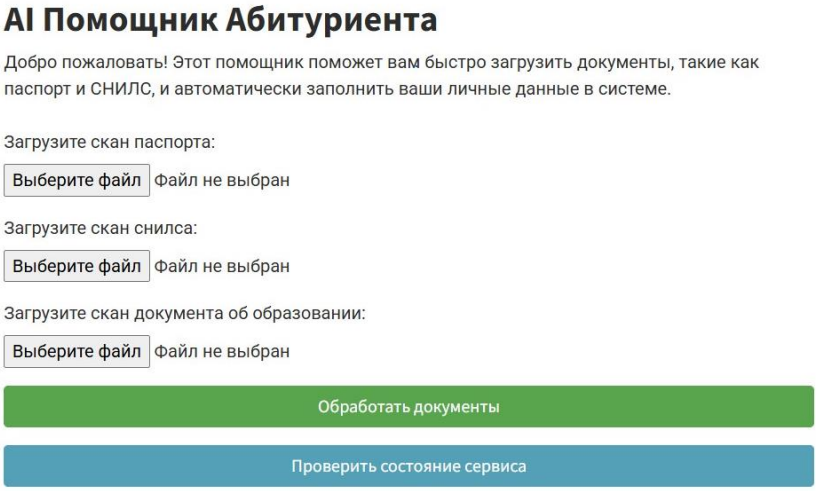

Рисунок 5 — Интерфейс страницы формы ИИ ассистента.

Клиентская логика реализована через виджет AIAssistantWidget. После успешной отправки документов интерфейс отображает полноэкранное модальное окно с индикаторами прогресса обработки. Модальное окно содержит три блока состояния, по одному для каждого типа документа. Каждый блок показывает название документа, анимированный индикатор загрузки в процессе обработки, иконку успешного завершения при получении результата или сообщение об ошибке при сбое обработки. Виджет каждые три секунды отправляет AJAX-запрос к контроллеру get_job_status, передавая массив идентификаторов активных задач и получая актуальную информацию о статусе каждой из них. На рисунках 4 и 5 представлены интерфейсы страницы формы ИИ ассистента после успешной отправки и успешной обработки приложенных сканов документов.

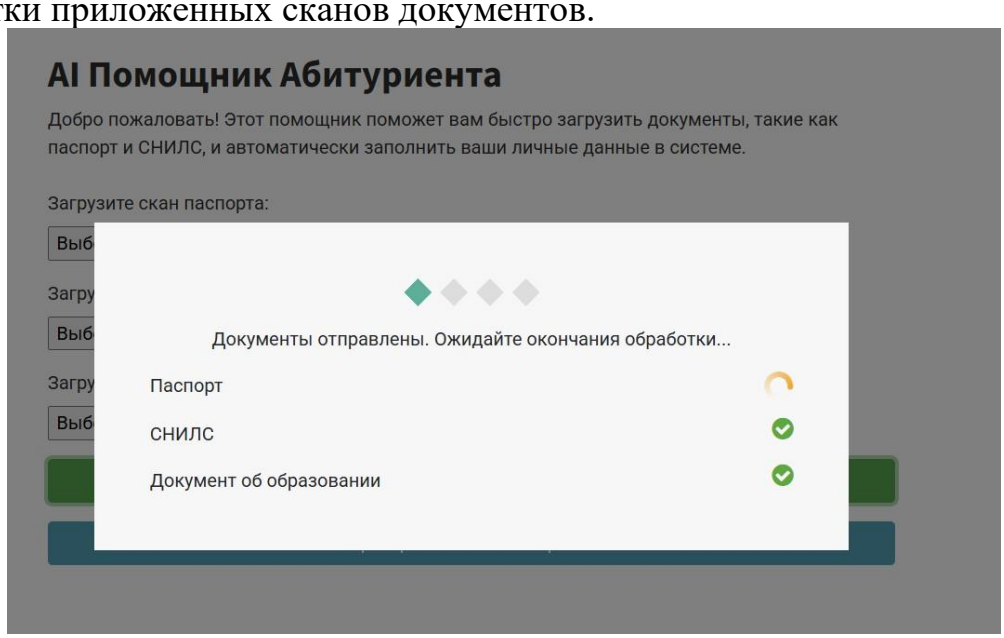

Рисунок 6 — Интерфейс страницы формы ИИ ассистента после успешной отправки

Опрос статуса задач продолжается до тех пор, пока все задачи не перейдут в финальное состояние «завершена» или «ошибка». Виджет отслеживает количество завершённых задач и обновляет соответствующие индикаторы в интерфейсе: заменяет анимацию загрузки на зелёную иконку галочки при успешной обработке или красный текст ошибки при сбое. Механизм опроса реализован с защитой от сетевых сбоев: при возникновении ошибки соединения виджет повторяет запрос до пяти раз, прежде чем показать пользователю сообщение о проблемах связи с сервером.

Разработанный комплекс из трёх взаимосвязанных методов интеллектуальной обработки документов использует фотографию владельца как опорный элемент и применяет двухэтапную коррекцию через каскадный классификатор Хаара и MediaPipe для обработки изображений, загружаемых в произвольной ориентации. Метод контекстно-ориентированного извлечения данных на базе Vision-Language модели olmOCR-2-7B обеспечивает прямое получение структурированных данных в формате YAML без разработки специализированных парсеров, при этом качество извлечения критически зависит от полноты промпта. Метод сегментированной обработки композитных документов решает проблему низкой точности распознавания вертикального текста через раздельную обработку зон с различной ориентацией и применением специализированных промптов для каждой зоны.

Предложенные методы образуют последовательный конвейер обработки документов и могут быть интегрированы в архитектуру портала приёмной комиссии на базе Odoo 15 через микросервисную модель взаимодействия, где задачи обработки передаются на специализированный GPU-сервер с установленной моделью olmOCR-2-7B, а результаты возвращаются в ERP-систему для автоматического заполнения полей анкеты абитуриента.

Список литературы:

1. Vaswani, A. Attention Is All You Need [Электронный ресурс] / ArXiv. – 2017. URL: https://arxiv.org/abs/1706.03762 (дата обращения: 10.02.2026).

2. Распознавание лиц с использованием каскадов Хаара [Электронный ресурс] / CyberLeninka. – 2026. – URL: https://cyberleninka.ru/article/n/raspoznavanie-lits-s-ispolzovaniem-kaskadov-haara (дата обращения: 12.01.2026).

3. Google. MediaPipe Face Mesh [Электронный ресурс] / Google LLC. – 2025. – URL: https://developers.google.com/mediapipe/solutions/vision/face_landmarker (дата обращения: 12.12.2025).