УДК 004.93

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВОЙ ИНФОРМАЦИИ ИЗ СТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Ахмад М.[1] (Университет «ИТМО»**)**
Научный руководитель — д-р техн. наук, профессор Алиев Т. И.[1]
[1]Университет «ИТМО»
Научный консультант— к.т.н., доцент Салех Хади Мухаммед[2]
[2]НИУ ВШЭ

Mahmodahmad003@mail.ru

Введение

Автоматизированное извлечение ключевой информации из структурированных документов (чеки, удостоверения личности и др.) важно для современных систем документооборота. Задача ограничена не только точностью, но и условиями развертывания: из-за конфиденциальности данные часто нельзя передавать сторонним сервисам, поэтому требуется надежное on-premise решение с полным контролем хранения и обработки [1]. Цель — по изображению документа получить нормализованный JSON с парами «ключ–значение» (дата, сумма, идентификатор, продавец, налог/VAT) для учета и аналитики. Сложность обусловлена разнообразием макетов, многоязычностью и деградациями качества изображений. Подходы делятся на OCR-конвейеры [2] и OCR-free/vision-language методы, извлекающие поля напрямую и уменьшающие ошибки OCR [3].

Основная часть

Данная работа включает реализации двух практических стратегий извлечения. Первая — классический конвейер (detection + OCR). Этот конвейер включает предобработку изображения (включая коррекцию ориентации), детекцию полей, OCR на обнаруженных регионах [4] и постобработку. Постобработка выполняет нормализацию полей (даты, числа, валюты, идентификаторы) и формирует JSON-вывод, чтобы результаты были стабильными и напрямую пригодными для использования системами.

Второй конвейер — подход на основе мультимодальной модели (Vision-Language). Модель vision-language получает изображение документа и извлекает значимые атрибуты напрямую, что может улучшить производительность в сложных случаях и снизить зависимость от качества OCR [3].

Подходы сравниваются с использованием метрик точности извлечения полей и устойчивости к шуму. Классические конвейеры подвержены распространению ошибок: когда OCR не справляется с изображениями, итоговое качество результатов резко снижается, особенно для критических полей (даты, суммы). В отличие от этого, OCR-free/VLM подходы могут вводить ошибки согласованности (например, несогласованные ответы в зависимости от простоты используемых моделей).

На основе практического сравнения предлагается гибридное решение: детекция обеспечивает надежную, контролируемую локализацию; распознавание на основе VLM улучшает восстановление текста в сложных условиях съемки; а постобработка/валидация обеспечивает согласованные результаты в виде пригодного для машинной обработки JSON.

Выводы

Для извлечения ключевой информации из структурированных документов были проанализированы и оценены два подхода в условиях ограничений по приватности, on-premise развертыванию и стоимости: классический OCR-конвейер и vision-language конвейер. Сильные и слабые стороны обоих из них были показаны с точки зрения точности извлечения, устойчивости к шуму и вариативности макетов, а также стабильности вывода. На основе проведенного анализа разработано гибридное решение, объединяющее детекцию полей и макета с распознаванием на основе VLM, с добавлением модуля валидации, что улучшает качество извлечения, предлагаемое двумя подходами, и формирует JSON-вывод, подходящий для производственного развертывания в системах управления документами, бухгалтерского учета и аналитики. Результаты могут быть использованы для реализации и тестирования on-premise решения в организациях, которые не могут использовать облачную обработку из-за требований безопасности и конфиденциальности.

Литература

1. Регламент (ЕС) 2016/679 Европейского парламента и Совета от 27 апреля 2016 г. о защите физических лиц при обработке персональных данных и о свободном обращении таких данных (GDPR) [Электронный ресурс]. – Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679 (дата обращения: 20.02.2026).
2. Huang Z., Chen K., He J., Bai X., Karatzas D., Lu S., Jawahar C. V. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction [Электронный ресурс]. – 2021. – arXiv:2103.10213. – DOI: 10.48550/arXiv.2103.10213 (дата обращения: 20.02.2026).
3. Kim G., Hong T., Yim M., Nam J. Y., Park J., Yim J., Hwang W., Yun S., Han D., Park S. OCR-Free Document Understanding Transformer // Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022: Proceedings. – Lecture Notes in Computer Science. – 2022. – Vol. 13688. – P. 498–517. – DOI: 10.1007/978-3-031-19815-1_29.
4. Smith R. An Overview of the Tesseract OCR Engine // Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007). – 2007. – Vol. 2. – P. 629–633. – DOI: 10.1109/ICDAR.2007.4376991.

Примечание. Для перевода использовались инструменты ИИ/машинного перевода. Автор несёт полную ответственность за содержание и итоговую редакцию текста.