

УДК 004.932.75'1

РАЗРАБОТКА СЕРВИСА РАСПОЗНАВАНИЯ СТРУКТУРИРОВАННЫХ ДАННЫХ ТИТУЛЬНЫХ ЛИСТОВ ВСЕРОССИЙСКИХ ПРОВЕРОЧНЫХ РАБОТ С ПРИМЕНЕНИЕМ НЕЙРОСЕТЕВЫХ МЕТОДОВ

Смирнов Д.А. (ИТМО) Сагайдак А.А. (ИТМО)

Научный руководитель - кандидат технических наук, доцент Федоров Д.А. (ИТМО)

Введение. Для преобразования изображений, содержащих текст, в том числе рукописный, в редактируемый машиночитаемый формат активно используются технологии Optical Character Recognition (OCR). Однако в случаях, когда изображение включает структурированные данные, такие как таблицы, функционала распознавания текста недостаточно. В таких ситуациях применяют методы для разметки таблиц с последующим распознаванием их содержимого [1-3]. Существуют коммерческие закрытые решения, справляющиеся с данной проблемой, но недоступность их кода для адаптации под конкретную задачу и платная основа делает их менее предпочтительными. В то же время есть не столь приспособленные открытые решения, но в совокупности предоставляющие необходимые инструменты для данной задачи [4-8].

Основная часть. Разрабатываемый сервис автоматизированного распознавания титульных листов ВПР основан на использовании нейросетевых методов для обработки изображений и извлечения структурированной информации. Основная задача заключается в корректном выделении и валидации областей интереса и распознавании их содержимого - основных заголовков и таблиц, что требует надежных и регулируемых алгоритмов предобработки и постобработки данных [9-10]. Расположение областей интереса в титульных листах предсказуемо из-за постоянной структуры и не требует специальных средств для распознавания – достаточно заранее определить координаты, в которые эта область попадает. Для распознавания печатного текста без сложной структуры достаточно предобученной модели открытого решения Tesseract. В случае с областями, содержащими информацию в табличном представлении, необходимо использование средств распознавания их структуры и содержания, которое в данном случае частично является рукописным, что усложняет задачу и требует применения соответствующих обученных моделей. Из-за отсутствия подготовленных данных, которые могли бы служить для определения точности разработанного решения, необходимо внедрение средств валидации и коррекции с возможностью дообучения используемых моделей. Обработанные данные преобразуются в унифицированный формат (JSON) и передаются через API для дальнейшего использования внешними средствами. Сервис обеспечивает управление процессом распознавания, хранение результатов и возможность верификации данных пользователем, что позволит адаптировать систему к новым типам документов и повысить точность извлечения информации в будущем.

Выводы. Проведен анализ проблемы автоматизированного распознавания титульных листов ВПР и методов ее решения с применением гибридного OCR-подхода. Рассмотрены этапы предобработки изображений, сегментации таблиц и распознавания рукописного текста, выявлены ограничения стандартных OCR-методов. Определены возможности повышения точности с использованием машинного обучения и языковых моделей, что формирует основу для дальнейшей разработки адаптивной системы распознавания.

Список использованных источников:

1. Kazdar T., Jmal M., Souidene W., Attia R. Table Recognition in Scanned Documents // Computational Collective Intelligence. ICCCI 2022. Lecture Notes in Computer Science, vol. 13501. Springer, Cham, 2022.
2. Kasem M., Abdallah A., Berendeyev A., Elkady E., Abdalla M., Mahmoud M., Hamada M., Nurseitov D., Taj-Eddin I. Deep learning for table detection and structure recognition: A survey // Preprint submitted to Elsevier. 2022.
3. Kim G., Hong T., Yim M., Nam J., Park J., Yim J., Hwang W., Yun S., Han D., Park S. OCR-free Document Understanding Transformer // NAVER CLOVA. 2022.
4. opencv // github URL: <https://github.com/opencv/opencv>
5. tesseract // github URL: <https://github.com/tesseract-ocr/tesseract>
6. donut // github URL: <https://github.com/clovaai/donut>
7. TableNet // github URL: <https://github.com/AmanSavaria1402/TableNet>
8. PaddleOCR // github URL: <https://github.com/PaddlePaddle/PaddleOCR>
9. Давлетов А. Р. Современные методы машинного обучения и технология OCR для автоматизации обработки документов // Вестник науки. 2023. № 10 (67).
10. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. 2020.