УДК 004.932.75'1

РАЗРАБОТКА СИСТЕМЫ РАСПОЗНАВАНИЯ РУКОПИСНЫХ И ПЕЧАТНЫХ ДАННЫХ С КОРРЕКЦИЕЙ ИСКАЖЕНИЙ

Сагайдак А.А. (Университет ИТМО)
Научный руководитель - доцент, Фёдоров Д.А. (Университет ИТМО)

Введение. OCR (Optical Character Recognition) представляет собой технологию для преобразования рукописного и печатного текста с изображения в машиночитаемый формат. Текст, полученный в результате обработки изображения, можно редактировать, искать, преобразовывать или использовать в других цифровых приложениях. Несмотря на высокую точность современных систем OCR, при обработке искаженных сканов, бликов и сложных документов остаются определенные проблемы. В частности, для титульных листов Всероссийских проверочных работ (ВПР) важно не только распознать текст, но и корректно извлечь данные из таблиц и других структурированных элементов. Оптическое распознавание символов уже реализовано в таких системах как: ABBYY FineReader, Tesseract, Kofax OmniPage. Но отсутствие возможности предобработки изображения для улучшения его качества - существенный недостаток, из-за которого встает вопрос о разработке методики для устранения дефектов таких как: изображение под углом, блики, пометки ручкой (в случае, когда текст написан на полях, либо выходит за пределы клетки), некорректное распознавание рукописного текста и структуры таблицы, искажения от сканов книг и низкое качество изображений, что значительно позволит повысить точность распознавания текста.

Основная часть. Одной из ключевых задач при разработке системы автоматического распознавания титульных листов ВПР является обеспечение высокой точности извлечения данных в сложных условиях. Для корректного распознавания печатных символов предлагаемая система в соответствии с методикой использует нейросетевые алгоритмы для предобработки изображений, включая коррекцию перспективы, устранение бликов и восстановление геометрии документа. Для эффективного распознавания рукописных символов необходимо предварительно выделить область каждой цифры на изображении, после чего произвести преобразование, при котором каждая цифра будет представлена в виде черно-белого изображения размером 28x28 пикселей с расположением цифры по центру и размером 20x20 пикселей. Важно отметить, что система не будет учитывать дефекты, возникшие из-за ручного заполнения документа, такие как перекрытия, исправления, пометки, зачеркивания и подтёки чернил. Однако цифры, выступающие за границы таблицы распознаваться будут. Это позволяет минимизировать ошибки на этапе распознавания. Дополнительно применяются адаптивные алгоритмы выделения областей интереса, позволяющие точно определять позиции таблиц с баллами и техническими данными. После распознавания информация преобразуется в структурированный формат (JSON) и передается через API для дальнейшей обработки. Важной частью системы является пользовательский интерфейс (UI) для загрузки документов на бэкэнд и просмотра результатов распознавания. Пользователи смогут отслеживать процесс обработки и при необходимости вносить коррективы (например, в случае ошибочного распознавания, пользователь сможет редактировать данные вручную), что повысит точность и удобство работы с системой.

Выводы. В ходе исследования особенностей работы OCR в программных продуктах были выбраны подходящие технологии и алгоритмы для создания системы распознавания рукописных и печатных данных с коррекцией искажений. В результате проведенной разработки была сформулирована методика, устраняющая ключевые проблемы, характерные для существующих OCR-систем, с помощью применения нейросетевых алгоритмов. Система эффективно предобрабатывает изображение перед распознаванием текста на нем, тем самым

повышается точность и надежность извлечения данных, обеспечивая корректное распознавание как печатных, так и рукописных символов даже в условиях искаженных изображений. Это значительно снижает вероятность ошибок при внесении оценок и автоматизирует обработку ВПР.

Список использованных источников:

1. Как извлечь текст из сканов: OCR, нейросети и их возможности [Электронный ресурс]. – 2024. – URL: https://habr.com/ru/companies/documenterra/articles/869938/ (дата обращения: 14.02.2025)
2. Предобработка изображений для OCR [Электронный ресурс]. – 2023. – URL: https://vc.ru/u/1627433-nazhmutdin-gumuev/696910-predobrabotka-izobrazhenii-dlya-ocr (дата обращения: 14.02.2025)
3. 9 Best OCR Software of 2025 (Free and Paid Tools) [Электронный ресурс]. – 2024. – URL https://theecmconsultant.com/best-ocr-software/ (дата обращения: 14.02.2025)
4. MNIST-MIX: A Multi-language Handwritten Digit Recognition Dataset – 2020. – URL: https://arxiv.org/abs/2004.03848 (дата обращения: 14.02.2025)
5. Understanding MNIST: A Comprehensive Overview of the Popular AI Data Set [Электронный ресурс]. – 2024. – URL: https://www.perplexuty.com/data-sets-and-libraries-mnist (дата обращения: 14.02.2025)
6. Каковы проблемы при обнаружении и извлечении текста из рукописных изображений? [Электронный ресурс]. – 2023. – URL: https://ru.eitca.org/artificial-intelligence/eitc-ai-gvapi-google-vision-api/understanding-text-in-visual-data/detecting-and-extracting-text-from-handwriting/examination-review-detecting-and-extracting-text-from-handwriting/what-are-the-challenges-in-detecting-and-extracting-text-from-handwritten-images/ (дата обращения: 14.02.2025)