А. А. Захаренков, А. И. Саркисов, А. А. Ступников
Тюменский государственный университет, г. Тюмень
УДК 004.8, 004.91

СРАВНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ТЕКСТА В PDF-ДОКУМЕНТАХ

Аннотация. В работе представлено сравнение корректности распознавания текста различных моделей компьютерного зрения. Исследование проводилось на основе анализа библиотек PyTesseract, EasyOCR, OCRmyPDF и PyMuPDF для извлечения неразмеченного текста из PDF-документов. В ходе эксперимента оценивались точность и скорость работы инструментов с изображениями разного качества.

Ключевые слова: оптическое распознавание символов, OCR, PDF-документы, язык программирования Python.

Введение. Формат PDF является одним из наиболее распространенных стандартов для обмена документами благодаря своей универсальности и поддержке различных типов контента, включая текст, изображения и векторную графику. Однако значительная часть PDF-файлов содержит текст в виде сканированных изображений, что исключает возможность его прямого редактирования и поиска. Оптическое распознавание символов (OCR) играет ключевую роль в преобразовании таких документов в текстовый формат [5]. В связи с этим актуальной задачей является выбор оптимального решения для распознавания текста в PDF-документах, обеспечивающего высокую точность и скорость обработки.

Современные исследования в области OCR сосредоточены на повышении точности распознавания текста в условиях варьирующегося качества исходных изображений [2]. Tesseract, как один из наиболее популярных open-source движков OCR, демонстрирует высокую точность при работе с печатным текстом [1, 5], также существуют альтернативные решения, такие как EasyOCR и PyMuPDF. Их эффективность работы с PDF-документами требует дополнительной проверки [3, 4]. Целью данного исследования является сравнение различных методов оптического распознавания символов для последующего использования в веб-приложении, обеспечивающем загрузку, хранение и поиск по PDF-документам.

Проблема исследования. Поиск по тексту документа является важной возможностью инструментов для работы с PDF-документами, однако часть PDF-документов содержит неразмеченный текст, представленный изображением. Несмотря на наличие множества OCR-инструментов, их эффективность может значительно различаться в зависимости от качества исходного документа. Существующие платформы вроде Google Cloud Vision API и Azure Cognitive Services Vision обладают высокой точностью, но их использование ограничено. В связи с этим актуальной проблемой становится выбор оптимального решения для распознавания текста в PDF-документах.

Материалы и методы. Для сравнения были выбраны следующие инструменты оптического распознавания символов: PyTesseract, EasyOCR, OCRmyPDF, PyMuPDF. В качестве метрики качества распознавания текста использовалось расстояние Левенштейна, измеряющее по модулю разность между двумя последовательностями символов как минимальное количество односимвольных операций вставки, удаления, замены, необходимых для превращения одной последовательности символов в другую. Схожесть двух строк определялась по следующей формуле:

$$\left(1-\frac{\mathrm{lev}(S_1,S_2)}{\max(|S_1|,|S_2|)}\right)*100\%$$

где $\mathrm{lev}(S_1,S_2)$ — величина расстояния Левенштейна между двумя строками S_1 и S_2, $|S|$ — длина строки S в символах.

Результаты. В результате исследования была создана программа на языке программирования Python, измеряющая скорость и корректность оптического распознавания символов среди инструментов PyTesseract, EasyOCR, OCRmyPDF и PyMuPDF. Для измерения был выбран одностраничный PDF-документ с неразмеченным текстом в формате изображения, текст на русском и английском языках. Было проведено 10 измерений работы моделей на изображениях исходного разного качества, приведены результаты трех наиболее значимых измерений.

Для первого измерения был взят исходный PDF-документ с хорошо различимым текстом (рис. 1).


АБСТРАКТ
В работе сравнив
вырезанные слова с ма

Рис. 1. Образец текста первого измерения

По результатам первого эксперимента модели PyTesseract и OCRmyPDF показали лучшие значения точности, близкие к 100% (табл. 1, рис. 2). На работу PyTesseract ушло 4,3 секунды, OCRmyPDF — 14,8. Модели EasyOCR и PyMuPDF показали много худшие результаты.

Таблица 1

Результаты первого измерения

	Точность, %	*Время, с*
PyTesseract	98,79	4,3
EasyOCR	38,23	66,7
OCRmyPDF	99,30	14,8
PyMuPDF	17,44	3,5

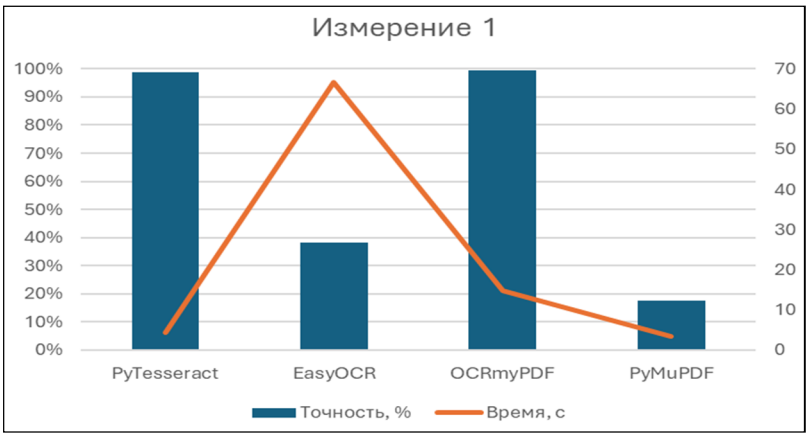

Рис. 2. Результаты первого измерения

Для проведения второго измерения у изображения исходного PDF-документа была увеличена контрастность с целью имитации визуальных недостатков (рис. 3).

Рис. 3. Образец текста второго измерения

В результате качество оптического распознавания текста у моделей снизилось (табл. 2, рис. 4), модели PyTesseract и OCRmyPDF по-прежнему показали лучшие результаты точности и скорости исполнения: 93% за 4,5 сек и 96% за 12,7 сек соответственно.

Таблица 2

Результаты второго измерения

	Точность, %	*Время, с*
PyTesseract	92,76%	4,5
EasyOCR	33,62%	51,9
OCRmyPDF	95,95%	12,7
PyMuPDF	17,53%	3,6

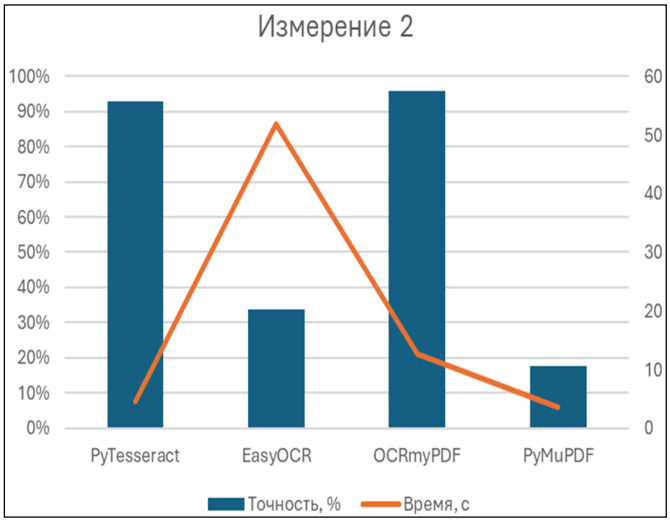

Рис. 4. Результаты второго измерения

Третий эксперимент проводился над изображением с эффектом шума (рис. 5), затрудняющего распознавание текста.

Рис. 5. Образец текста третьего измерения

В результате измерения наибольший показатель точности оказался у модели PyTesseract — 71% за 6,6 сек, точность модели OCRmyPDF оказалась меньше — 62% за 17,8 сек (табл. 3, рис. 6).

Таблица 3

Результаты третьего измерения

	Точность, %	*Время, с*
PyTesseract	71,21%	6,6
EasyOCR	18,18%	70,7
OCRmyPDF	62,11%	17,8
PyMuPDF	17,77%	4,4

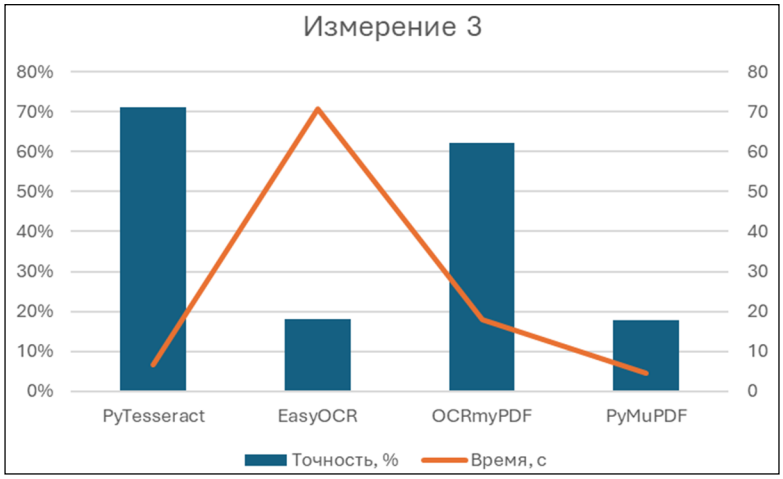

Рис. 6. Результаты третьего измерения

Заключение. В ходе исследования наилучшие показатели точности и скорости продемонстрировала модель PyTesseract, что позволяет сделать выбор в ее пользу для интеграции в инструмент для работы с PDF-документами.

СПИСОК ЛИТЕРАТУРЫ

1. Smith, R. An Overview of the Tesseract OCR Engine // Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR). — 2007. — Vol. 2. — P. 629–633. — DOI: 10.1109/ICDAR.2007.4376991.
2. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H. What Is Wrong With Scene Text Recognition Model Comparisons? // IEEE International Conference on Computer Vision (ICCV). — 2019. — P. 4715–4723. — DOI: 10.1109/ICCV.2019.00481.
3. Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., Pletschacher, S. A Survey of OCR Evaluation Tools and Metrics // International Journal of Document Analysis and Recognition (IJDAR). — 2021. — Vol. 24. — No. 1. — P. 1–20.
4. Rice, S.V., Jenkins, F.R., Nartker, T.A. The Fourth Annual Test of OCR Accuracy. — Las Vegas : Information Science Research Institute, 1996. — 35 p. — (Report No. TR-96-01).
5. Dome, S., Sathe, A.P. Optical Character Recognition using Tesseract and Classification // 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) : proceedings : March 5–7, 2021, Pune, India. — Piscataway : IEEE, 2021. — P. 153–158. — ISBN 978-1-7281-8519-4. — DOI: 10.1109/ESCI50559.2021.9397008.