АНАЛИЗ РЫНКА СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ (OCR)

Нестеров А.С.

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Брянский государственный университет имени

академика И.Г. Петровского

город Брянск, Россия

Рассматриваются современные информационные системы оптического распознавания образов (OCR), производится сравнительно-сопоставительный анализ наиболее популярных OCR- сервисов на рынке, результаты анализа которого представлены в рамках показателя «качества распознавания». Были разработаны рекомендации в целях повышения качества распознавания.

Ключевые слова: OCR-системы, качество распознавания.

MARKET ANALYSIS OF MODERN INFORMATION SYSTEMS FOR OPTICAL CHARACTER RECOGNITION (OCR)

Nesterov A.S.

Federal Government budget educational institution of higher education "The

Bryansk State University named after academician I. G. Petrovsky"

Bryansk, Russia

A comparative analysis of the most popular optical recognition services on the market, the results of which are presented in the framework of the recognition quality indicator. Recommendations were adopted in order to improve the quality of recognition.

Keywords: OCR system, recognition quality.

OCR - системы — это системы для перевода изображений документов в редактируемый текст, который можно затем обрабатывать в текстовых и табличных редакторах, от англ. Optical Character Recognition — Оптическое распознавание символов. По сравнению с ручной перепечаткой текста, такие системы дают существенный выигрыш в скорости работы, к тому же делают меньше ошибок [1].

В настоящее время на рынке представлено большое количество OCR-систем. Нами были изучены и проанализированы основные программы оптического распознавания символов, затем были сформулированы следующие выводы:

- на рынке представлено большое количество систем оптического распознавания образов отечественного и зарубежного производителя;
- большинство представленных систем, распространяются на коммерческой основе;

- основная часть систем разработаны для работы с операционной системой Windows;

- проведенный анализ российского рынка информационных систем OCR показал, что наиболее популярными являются системы: Microsoft One Note 2010, SODA PDF OCR, Abbyy Fine Reader, Online OCR, SmartScore.

Сравнительно-сопоставительный анализ систем оптического распознавания символов (OCR)

Для проведения сравнительно-сопоставительного анализа мы подготовили пять файлов с растровым изображением текста различного разрешения (низкого: <100 dpi, среднего: 100 - 300 dpi, высокого:> 300 dpi) и выбрали следующие информационные системы оптического распознавания образов: Microsoft One Note, SODA PDF OCR, Abbyy Fine Reader, Online OCR, SmartScore.

Далее все подготовленные изображения были последовательно оцифрованы и распознаны в каждой из программ.

Для выполнения сравнительного анализа результатов оптического распознавания символов нами были выделены различные критерии. Критерии объединены в разделы, исходя из существенных признаков каждого из критериев.

Также для анализа была введена величина — качество распознавания, которая определяется по формуле:

$К = \frac{S - S_{ошиб}}{S} \cdot 100\%$, где $S_{ошиб}$ — это количество орфографических ошибок и S-количество слов в тексте (для Smart Score — количество ошибок в музыкальном тексте). Для подсчета количества слов и ошибок мы воспользовались сервисом text.ru.

Мы выяснили, что Microsoft One Note 2010, Abbyy Fine Reader, Smart Score предоставляют свои услуги небольшим и средним организациям [2], [6]. Стоит заметить, что программы Microsoft One Note 2010 и Abbyy Fine Reader также используют крупные организации. Так программа Abbyy Fine Reader имеет 3 программных пакета: Home Edition, Professional Edition, Corporate Edition, что делает это программу конкурентно способной во всех ценовых сегментах [2].

Все программы разработаны для работы с операционной системой Windows различных версий и Mac OS. Стоит отметить, что 3и информационные системы работают с операционной системой Android, это позволяет производить процесс распознавания документов на своем смартфоне в любом месте, не используя свой персональный компьютер. Это делает процесс распознавания текста более практичным, без использования дополнительных затрат.

Способы получения изображения во всех системах одинаковые. Изображения можно получать со сканера или многофункционального устройства, с помощью цифрового фотоаппарата, фотокамеры мобильного телефона (с матрицей от 2 МП и функцией автофокуса), PDF-файлы. Стоит отметить, что информационная система SODA PDF OCR предлагает производить распознавания документов из форматов doc, jpeg, PDF, но фактически распознавание происходит из формата PDF. Форматы doc и jpeg не поддерживаются в данной программе.

Пользовательский интерфейс во всех системах разный. Заметно, что разработчики в зависимости от финансовой состоятельности компании делают соответственно такой же интерфейс. В SODA PDF OCR интерфейс полностью основан на интерфейсе пакета программ Microsoft Office 2016 года [4].

В системах Microsoft One Note 2010 и Abbyy Fine Reader, Smart Score полностью свой интерфейс. Пользователям доступны возможности перетаскивания страниц, изменение выбранных изображений, изменение области распознавания и т.д.

Интерфейс в Online OCR не доработан. Не структурированное главное меню, отсутствие возможности загружать для распознавания несколько изображений и возможности автоматического выбора языка распознавания.

В программе Smart Score нет в наличии русского языка для интерфейса, что делает работу в программе для людей, не знающих английский язык очень трудной.

Тип лицензии во всех системах либо бесплатный, либо коммерческий. Это связано с тем набором функций, которые предлагает производитель.

Проведем дополнительный сравнительный анализ информационных систем оптического распознавания образов по критерию «Качество распознавания» (Таблица 1). Для этого мы будем находить значение качества распознавания, и составим гистограмму на которой будут показаны значения данного параметра в зависимости от номера файла и использованной системы распознавания. Качество распознавания для информационной системы Smart Score также будет определяться аналогично другим системам, но ключевую роль будет играть количество фальшивых нот, иначе говоря, ошибки в музыкальном тексте.

Таблица 1

Показатели качества распознавания

Информационная система	1	2	3	4	5
SODA PDF OCR	1%	0,5%	0,4%	0,5%	0,01%
Online OCR	83%	89%	88%	87%	61%

Abbyy Fine Reader	98%	98%	99%	85%	0%
Microsoft One Note 2010	63%	64%	39%	41%	1%
Smart Score	99%	98%	70%	60%	0%

Результаты 0% в таблице говорят о том, что программы не справились с распознаванием предложенного файла. На рисунке 1 представлена гистограмма с показателями качества распознавания всех программ.

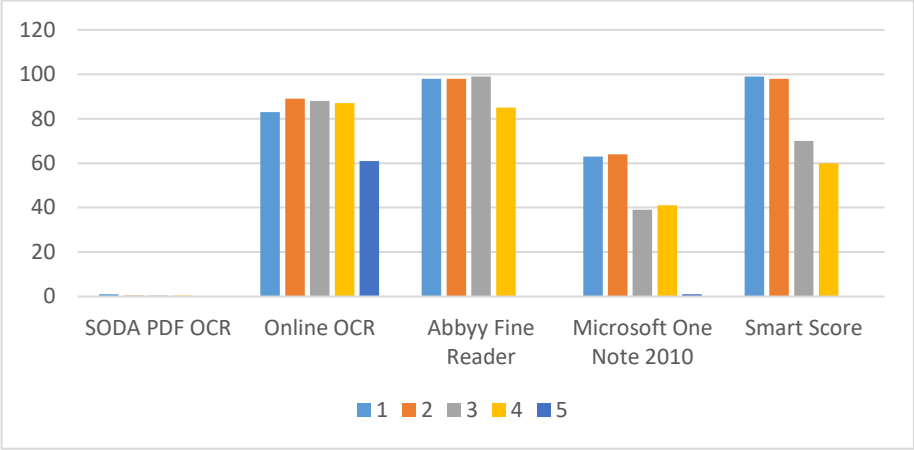

Рисунок 1 – Гистограмма с показателями качества распознавания анализируемых программ

На основе рисунка 1 и таблицы 1 можно сформулировать следующие выводы:

1. Качество распознавания не зависит от типа лицензии. Действительно, коммерческие программы предоставляют широкий спектр инструментов работы с OCR-технологией и гарантируют хорошее распознавание. Так системы: Online OCR, Abbyy Fine Reader, Smart Score на высоком уровне справились с распознаванием текста, но Abbyy Fine Reader и Smart Score не справились с распознаванием последнего файла, когда Online OCR также на высоком уровне распознал изображения, причем Online OCR обладает бесплатной лицензией.

2. Качество распознавания зависит от разрешения изображения.

Если обратить внимание на рисунок 1, то можно заметить зависимость: чем ниже разрешение, тем хуже распознавание. Системы Online OCR, Abbyy Fine Reader на хорошем уровне справились с распознаванием изображений высокого качества (>83%) и среднего качества (>85%), в то время как у других систем начались заметные затруднения в распознавании файлов среднего и низкого разрешений. Стоит заметить, что Abbyy Fine Reader и Smart Score не способны распознавать файлы низкого разрешения. Это

свидетельствует наличие ошибок в программах при работе с распознаванием файлов. Так, Abbyy Fine Reader уведомлял пользователя о низком разрешении файла, а Smart Score и вовсе отказывался производить распознавание, тем самым экстренно отключая программу.

Для повышения качества распознавания можно применять следующие методы:

- устранение перекосов изображения, полученного со сканера;
- удаление пустых страниц многостраничного документа;
- разделение двойных страниц, если отсканирован сразу весь разворот книги;
- очистка изображения удалением случайного «мусора» — излишних точек на поле документа, особенно вблизи границ символов;
- цифровое увеличение изображения;
- автоматический поворот страниц, вошедших в сканер не той стороной;
- обрезка «черноты» по краям документа, размер которого оказался меньше поля сканера, и т. д.

3. Системы OCR используются не только в распознавании текста, но и в системах распознавания рукописного ввода, сканерах отпечатках пальцев, системах распознавания лиц (Face Recognition), системах распознавания речи (Speech Recognition).

4. Если изображение получено с сканера или многофункционального устройства рекомендуются системы Abbyy Fine Reader, Microsoft One Note, которые показали высокое качество распознавания документов с разрешением 100 – 300 dpi.

5. Если изображение получено с камеры телефона или фотоаппарата рекомендуется система Online OCR, в связи с высоким качеством распознавания документов с разрешением <100 dpi.

Список литературы:

1. В. В. Трофимов. Информационные технологии: учебник для академического бакалавриата / под ред. В. В. Трофимова. – М.: Издательство Юрайт, 2009. – 624 с.
2. Официальный сайт Abbyy Fine Reader [Электронный ресурс]. – режим доступа: https://www.abbyy.com/ru-ru/finereader/. – (Дата обращения: 10.11.2019).
3. Официальный сайт Online OCR [Электронный ресурс]. – режим доступа: https://www.onlineocr.net/ru/. – (Дата обращения: 11.11.2019).
4. Официальный Soda PDF OCR [Электронный ресурс]. – режим доступа https://www.sodapdf.com/ocr-pdf/ . – (Дата обращения: 03.11.2019).
5. Спецификация Abbyy Fine Reader 12 [Электронный ресурс]. – режим доступа: https://support.abbyy.com/hc/ru/articles/360004047940. – (Дата обращения: 05.11.2019).

6. Основные задачи в One Note 2010 [Электронный ресурс]. – режим доступа:https://support.office.com/ru-ru/article/Основные-задачи-в-onenote-2010-29a50122-eb92-4eaf-8a39-ae5f01094ddc. – (Дата обращения: 10.11.2019).
7. Копирование текста из вставленных изображений в One Note для Mac [Электронный ресурс]. – режим доступа: https://support.office.com/ru-ru/article/Копирование-текста-из-вставленных-изображений-в-onenote-для-mac-b840c9a0-6f25-423c-bbb5-f240cc07d4db. – (Дата обращения: 18.11.2019).
8. Мультимедиа [Электронный ресурс]. – режим доступа: http://soft-lenta.ru/index.php?newsid=1146389194. – (Дата обращения:10.11.2019).