АНАЛИЗ СИСТЕМЫ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ НА ПРИМЕРЕ СИСТЕМЫ ABBYY FINEREADER

До Тхи Хань
Национальный исследовательский Томский политехнический университет
hanhdt21@gmail.com

В настоящее время, каждый день переводится огромное количество различных документов с бумаги в электронную форму, например: печатные тексты, платежные поручения, медицинская справка, официальные приглашения, различные опросные листы и т.д. При этом активно используются тысячи различных систем электронного документооборота практически во всех сферах деятельности. Поэтому при современных объемах потоков документов, без автоматизированной обработки, подобные операции немыслимы.

Одним из ключевых этапов во всех системах электронного документооборота и системах ввода печатных текстов является, распознавание текстовых символов - перевод информации из графической формы в текстовую форму. Распознавание текста является одним из направлений распознавания образов [3]. Распознавание текстовой информации играет важную роль в переводе печатного и рукописного текста в электронную. Целью этого является автоматизация документооборота или внедрение безбумажных технологий.

В настоящее время, проблема оптического распознавания символов (OCR) становится всё более актуальной при активном внедрении цифровой вычислительной техники и широким использованием текстовых процессоров.

Целью данной статьи является познакомление и исследование системы оптического распознавания символов (OCR), обеспечивающих высокое качество распознавания.

Сегодня, благодаря развитию технологий, именно с помощью сканера мы можем достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно, так как страница с текстом представляет собой обычную картинку. Можно текст читать и распечатывать, но невозможно его редактировать и форматировать. Поэтому, для того, чтобы получить документы в формате текстового файла необходимо провести распознавание текста.

«Оптическое распознавание символов (англ. optical character recognition, OCR) — механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные — последовательность кодов, использующихся для представления символов в компьютере (например, в текстовом редакторе)» [1]. Эта система предназначена для ввода печатного текста для печатных и электронных изданий.

Можно привести несколько пример системы оптического распознавания текста:

- ✓ Recognita Plus DTK фирмы Recognita Corporation (Венгрия),
- ✓ TextBridge фирмы Xerox Imaging Systems,
- ✓ TypeReader фирмы ExperVision (США),
- ✓ CharacterEyes фирмы Ligature (Израиль),
- ✓ IRIS OCR фирмы I.R.I.S. (Бельгия),
- ✓ Easy Reader фирмы Inovatic International (Франция),
- ✓ OmniPage Professional и WordScan Plus фирмы Caera (США) [4].

Наиболее известными программами класса «Системы оптического распознавания» в России являются OCR CuneiForm и ABBYY FineReader.

Система оптического распознавания текста имеет следующие преимущества:

- ✓ распознавать тексты
- ✓ корректно работать с текстами
- ✓ корректно распознавать не только четко набранные тексты, но и такие, качество которых не хорошее.

Системы OCR состоят из следующих основных блоков, предполагающих аппаратную или программную реализацию:

- ✓ блок сегментации;
- ✓ блок предобработки изображения;
- ✓ блок выделения признаков;
- ✓ блок распознавания символов;
- ✓ блок постобработки результатов распознавания.

Сначала необходимо выделять текстовые области, строки и разбиение связных текстовых строк на отдельные знакоместа, каждое из которых соответствует одному текстовому символу. Далее выделять текстовые фрагменты графического изображения страницы необходимо преобразовать в текст [5].

Для того, чтобы получить готовую электронную к редактированию копию любого печатного текста, программе OCR необходимо выполнить процедуры из множества отдельных операций:

- ✓ Сканирование и предварительная обработка изображения.
- ✓ Анализ структуры документа.
- ✓ Распознавание.

- ✓ Проверка результатов.
- ✓ Реконструкция документа (воссоздание его исходного вида).
- ✓ Экспорт.

Используя программы оптического распознавания текстов, можно редактировать текст, хранить документы в различных форматах, распечатывать материал, не теряя качества, анализировать информацию и применять к тексту электронный перевод, форматировать или преобразовать в речь. Следует отметить, что при работе с определенными шрифтами, система оптического распознавания требует калибровки.

Лидером в области распознавания текста является программа FineReader от компании ABBYY. «ABBYY FineReader — система оптического распознавания символов, разработанная российской компанией ABBYY» [2].

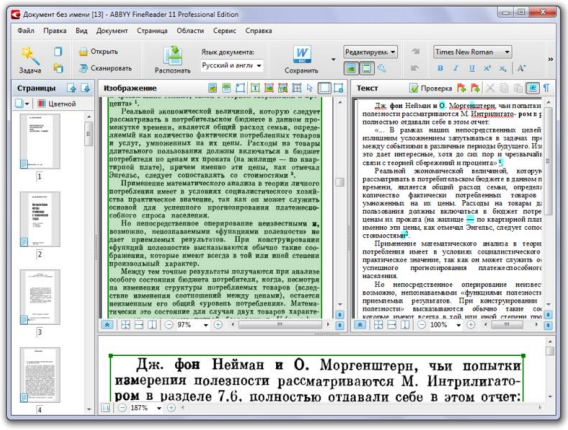

Рис.1. Оптическое распознавание символов ABBYY FineReader

Программа позволяет извлекать текстовые данные из цифровых изображений (фотографий, результатов сканирования, PDF-файлов). Она быстро переведет бумажные документы или PDF-файлы в удобный для редактирования формат. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати. Это достигается благодаря применению технологии «целостного целенаправленного адаптивного распознавания» (3 приципа: целостность, целенаправленность, адаптивность).

Программа ABBYY FineReader обладает такими преимуществами, такие как: скорость и высокая точность распознавания, поддержкой многих мировых языков, распознаванием сфотографированных документов, автоматической обработкой документов, сохранением документов в различных форматах, а также отправкой в интернет-хранилища, проверкой результатов распознавания и т.д.

С помощью программы ABBYY FineReader процесс обработки документов состоит из четырех этапов: получение изображения; распознавание документа; проверка и редактирование полученного текста; сохранение результатов распознавания.

Краткое руководство пользователя ABBYY FineReader 12 описывается на 2 рисунке

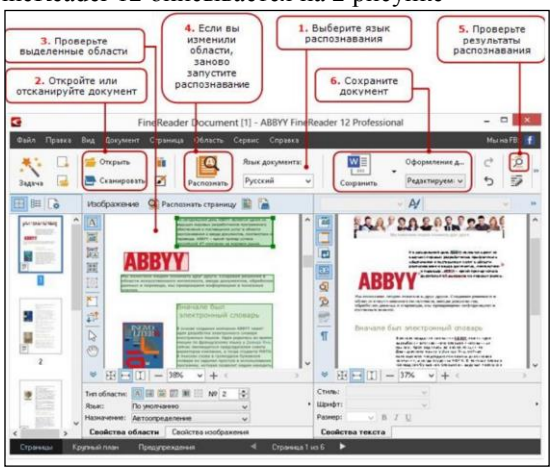

Рис.1. Краткое руководство пользователя ABBYY FineReader

Таким образом, в настоящее время широко используются программы класса «системы оптического распознавания текстов» для ввода печатных текстов. Эта программа является одной из наиболее перспективных областей применения искусственного интеллекта, распознавания образов и компьютерного зрения.

Литература

1. Оптическое распознавание символов [Электронный ресурс] // Википедия. Свободная энциклопедия URL: http://vi.wikipedia.org/wiki/Оптическое_распознавание_символов
2. ABBYY FineReader [Электронный ресурс] // Википедия. Свободная энциклопедия URL: http://vi.wikipedia.org/wiki/ABBYY_FineReader
3. Колесников С. Распознавание образов. Общие сведения /Газета «Компьютер-Информ». Программное обеспечение. http://www.ci.ru/
4. Распознавание текста [Электронный ресурс]. – Режим доступа: http://it-claim.ru/Education/Course/Lingvistika/Lecture/Lecture11.pdf, свободный.
5. Оптическое распознавание символов (OCR) [Электронный ресурс]. – Режим доступа: wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_(OCR), свободный.