Содержание:

Введение

Несмотря на то, что в настоящее время большинство документов составляется на компьютерах, задача создания полностью электронного документооборота ещё далека до полной реализации. Как правило, существующие системы охватывают деятельность отдельных организаций, а обмен данными между организациями осуществляется с помощью традиционных бумажных документов.

Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид.

Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл. Более предпочтительным, по сравнению с графическим, является текстовое представление информации. Этот вариант позволяет существенно сократить затраты на хранение и передачу информации, а также позволяет реализовать все возможные сценарии использования и анализа электронных документов. Поэтому наибольший интерес с практической точки зрения представляет именно перевод бумажных носителей в текстовый электронный документ.

Системы распознавания текстов (ocr-системы). Характеристика и функциональные возможности

С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический

файл – обычную картинку. Текст можно будет читать, распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

Основным методом перевода бумажных документов в электронную форму является сканирование. В результате сканирования получается графическое изображение, состоящее из точек, т.е. растровое изображение. Количество точек определяется как размером изображения, так и разрешением сканера.

Графический образ, получаемый после сканирования документа, иногда необходимо перевести в текст. Для этого используются специальные программные средства, называемые средствами распознавания образов. Из программ, способных распознавать текст на русском языке наиболее известной является ABBYY Fine Reader.

Преобразование документа в электронный вид происходит в три основных этапа. Каждый из этих этапов может выполняться программами как автоматически, так и под контролем пользователя.

- Сканирование. Запускается сканирующий модуль, настраиваются параметры сканирования (разрешение, размер, тип сканирования) и происходит собственно сканирование.
- Сегментация и распознавание текста. Прежде чем получить готовый текст, необходимо разбить фрагменты документа на блоки (текст, рисунок, таблица и т.д.), для того, чтобы правильно их распознать (преобразовать в текстовый документ).
- Проверка орфографии и передача текстового документа в нужное приложение для дальнейшей работы или сохранение в файл.

Методы распознавания символов

Если исходный документ имеет типографское качество, то задача распознавания решается методом сравнения с растровым шаблоном. При распознавании документов с низким качеством печати используется метод распознавания символов по наличиюв нихопределенных структурных элементов(отрезков, колец, дуг и др.).

Сканер (англ. scanner) – устройство, которое создаёт цифровое изображениесканируемого объекта. Полученное изображение может быть сохранено как графический файл, или, если оригинал содержал текст, распознано посредством программы распознавания текста и сохранено как текстовый файл.

В зависимости от способа сканирования объекта и самих объектов сканирования существуют следующие виды сканеров:

- Планшетные –наиболее распространённые, поскольку обеспечивают максимальное удобство для пользователя – высокое качество и приемлемую скорость сканирования. Представляет собой планшет, внутри которого под прозрачным стеклом расположен механизм сканирования.
- Барабанные– применяются в полиграфии, имеют большое разрешение (около 10 тысяч точек на дюйм). Оригинал располагается на внутренней или внешней стенке прозрачного цилиндра (барабана).
- Ручные – в них отсутствует двигатель, следовательно, объект приходится сканировать вручную, единственным его плюсом является дешевизна и мобильность, при этом он имеет массу недостатков – низкое разрешение, малую скорость работы, узкая полоса сканирования, возможны перекосы изображения, поскольку пользователю будет трудно перемещать сканер с постоянной скоростью.
- Сканеры штрих-кода– небольшие, компактные модели для сканирования штрих-кодов товара в магазинах.

Сканирование в сером является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если необходимо, чтобы содержащиеся в документе цветные элементы (картинки, цвет букв и фона) были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

FineReader – омнифонтовая (то есть система, распознающая символы практически любых размеров и начертаний) система оптического распознавания текстов. Это означает, что она позволяет распознавать тексты, набранные практически любыми шрифтами, без предварительного обучения. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати. FineReader имеет массы дополнительных функций и удобный интерфейс:

- распознавание текста;

- все найденные программой ошибки выделяются цветом. Затем программа производит проверку текста на наличие орфографических ошибок, и все некорректные слова подчеркивает красными линиями. Обнаруженные изображения программа выделяет красным цветом и в дальнейшем их не обрабатывает, а оставляет их такими, какие они есть, соответственно и передает их такими, как они получились при сканировании.
- Редактирование полученного документа.

Помимо редактирования формата отсканированной страницы пользователь может самостоятельно выделять области с текстом, картинки и таблицы, а затем распознавать обработанную страницу. В определенных условиях ручной режим определения типа блока может значительно повысить качество обрабатываемого документа. Выделяем необходимую часть отсканированной страницы и выбираем необходимый тип блока на этой панели. После ручной обработки необходимого объема материала запускаем распознавание. Программа допускает совместное использование автоматического и ручного определения типов блоков. Обработанный таким образом документ может быть сохранен в формате Word, Excel или Acrobat Reader.

Заключение

Говоря о системах распознания текста, главным образом выделяют лидера в данном направлении – компанию ABBYY.

Интеллектуальная система оптического распознавания ABBYY FineReader 9.0 позволяет быстро и точно переводить бумажные документы, цифровые фотографии документов и PDF-файлы в электронный вид. При распознавании ABBYY FineReader полностью сохраняет оформление документа: иллюстрации, картинки, списки и т. д. Полученные результаты можно исправлять в программах Microsoft Office, сохранять в разных форматах, отправлять по электронной почте и публиковать в интернете.

ABBYY FineReader представляет революционно новый подход к распознаванию документов. Теперь документ анализируется и обрабатывается целиком, а не постранично, что позволяет FineReader понять такие элементы его внутренней структуры, как верхние и нижние колонтитулы, сноски, подписи к картинкам и

диаграммам, стили, шрифты и т.д. Элементы исходного документа восстанавливаются в результирующем документе. Например, при сохранении в Word верхние и нижние колонтитулы, сноски воспроизводятся как соответствующие объекты в Word.

Система оптического распознавания ABBYY FineReader точно распознает и максимально полно сохраняет исходное оформление любого документа (в том числе с текстом на фоне картинок, с цветным текстом на цветном фоне, с обтеканием картинок текстом и т.д.)

Также ABBYY FineReader распознаёт документы на 179 языках, включая русский, английский, немецкий, французский, испанский, итальянский, шведский, финский, болгарский, венгерский, словацкий, чешский, башкирский, белорусский, казахский, украинский. Для 36 языков, предусмотрена проверка орфографии. Текст документа может быть составлен на двух и более языках. Пользователь может указать свой язык распознавания для каждого блока типа «текст» или для каждой ячейки таблицы.