Системы распознавания текста

Технология обработки текстовой информации

Необходимость в системах распознавания символов

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 и 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.

Программы распознавания текста

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

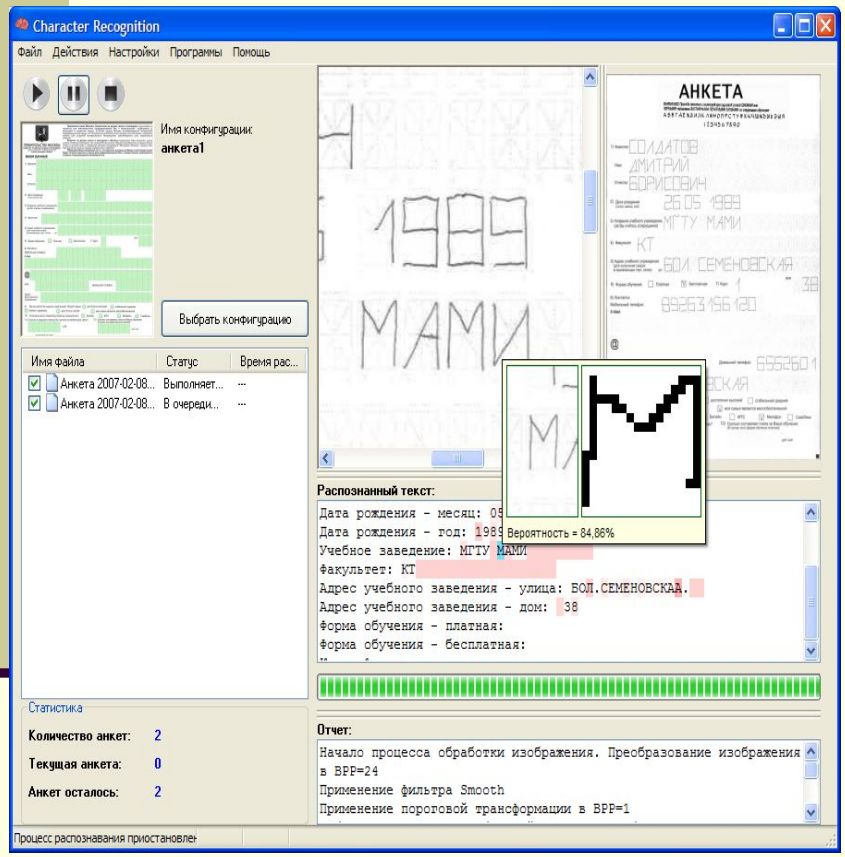
в электронный вид происходит в три основных этапа. Каждый из этих этапов может выполняться программами как автоматически, так и под контролем пользователя.

1. Сканирование. Запускается сканирующий модуль, настраиваются параметры сканирования (разрешение, размер, тип сканирования) и происходит собственно сканирование.

2. 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).

3. 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 наличию** в них **определенных структурных элементов** (отрезков, колец, дуг и др.).

Сканер

Ска́нер (англ. scanner) — устройство, которое создаёт цифровое *изображение* сканируемого 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).

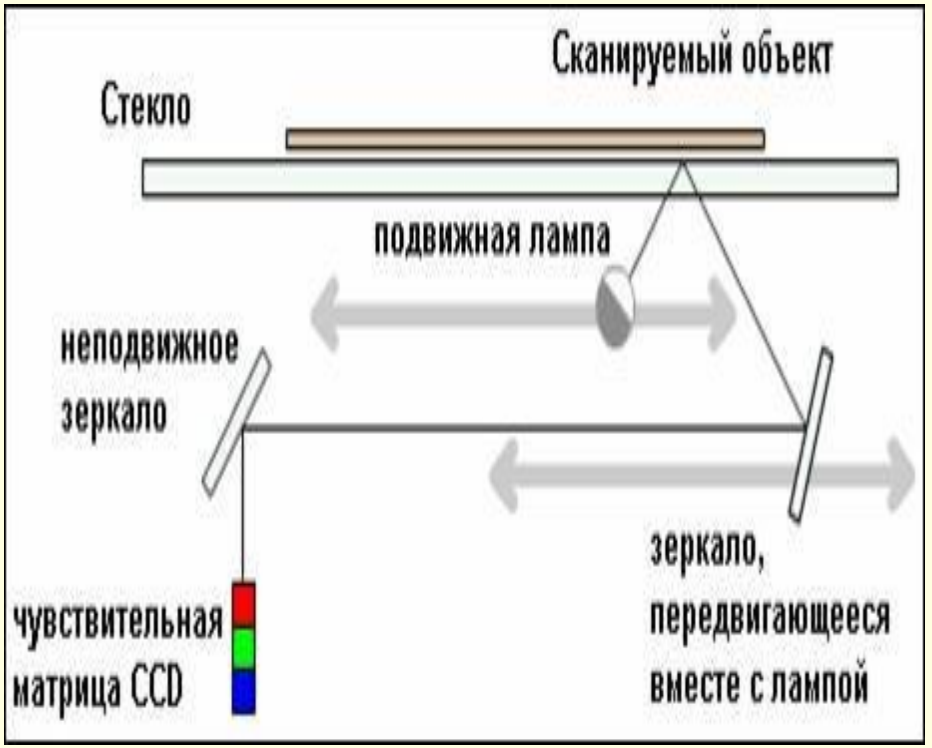
Ручные — в них отсутствует двигатель, следовательно, объект приходится сканировать вручную, единственным его плюсом является дешевизна и мобильность, при этом он имеет массу недостатков — низкое разрешение, малую скорость работы, узкая полоса сканирования, возможны перекосы изображения, поскольку пользователю будет трудно перемещать сканер с постоянной скоростью.

Сканеры штрих-кода — небольшие, компактные модели для сканирования штрих-кодов товара в магазинах.

Принцип действия планшетных сканеров

Сканируемый объект кладётся на стекло планшета сканируемой поверхностью вниз. Под стеклом располагается подвижная лампа, движение которой регулируется шаговым двигателем.

Свет, отражённый от объекта, через систему зеркал попадает на чувствительную матрицу (CCD — Couple-Charged Device), далее на АЦП и передаётся в компьютер. За каждый шаг двигателя сканируется полоска объекта, потом все полоски объединяются программным обеспечением в общее изображение.



Характеристики сканеров

- **Формата сканируемой поверхности:** A4 (стандартный печатный лист), A3, слайд-сканеры под формат пленки 13x18 и 18x24…
- **Оптическое разрешение.** Разрешение измеряется в точках на дюйм (dots per inch — dpi). Указывается два значения, например 600x1200 dpi, горизонтальное — определяется матрицей CCD, вертикальное — определяется количеством шагов двигателя на дюйм.
- **Интерполированное разрешение**. Искусственное разрешение сканера достигается при помощи программного обеспечения. Его практически не применяют, потому что лучшие результаты можно получить, увеличив разрешение с помощью графических программ после сканирования. Используется производителями в рекламных целях.
- **Скорость работы**. Измеряется в страницах в минуту, при этом имеются в виду страницы определенного формата и определенное разрешение сканнера, из числа возможных.
- **Глубина цвета**. Определяется качеством матрицы CCD и разрядностью АЦП. Измеряется количеством оттенков, которые устройство способно распознать. 24 бита соответствует 16777216 оттенков. Современные сканеры выпускают с глубиной цвета 24, 30, 36 бит. Несмотря на то, что графические адаптеры пока не могут работать с глубиной цвета больше 24 бит, такая избыточность позволяет сохранить больше оттенков при преобразованиях картинки в графических редакторах.

Оптимальное разрешение при сканировании

Оптимальным разрешением для обычных текстов является - 300 dpi и 400-600 dpi для текстов, набранных мелким шрифтом (9 и менее пунктов).

Сканирование в сером является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если Вы хотите, 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.

ABBYY FineReader

FineReader - омнифонтовая 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.

Автоматический перевод текста

Идея автоматического перевода текстов с одного языка на другой зародилась с появлением первых компьютеров. Если бы полноценный перевод был возможен, то значительно упростилось бы общение между народами. Но это очень сложная задача, о полном решении которой пока говорить рано.

Программы автоматического перевода позволяют переводить отдельные слова и строить смысловые связи в предложениях, не всегда учитывая те или иные особенности языка. Поэтому они предназначены лишь для общего ознакомления с содержанием документа.

Программные средства автоматического перевода можно условно разделить на две основные категории:

1. Компьютерные словари. Назначение их - предоставить значения неизвестных слов быстро и удобно для пользователя.

2. Системы автоматического перевода - позволяют выполнять автоматический перевод связного текста. В ходе работы программа использует словари и наборы грамматических правил, обеспечивающих наилучшее качество перевода.

Сердюкова Татьяна Александровна

1 квалификационная категория.

Ставропольский край г. Ставрополь
МОУ лицей 8.

srtanja71@mail.ru

www.lic8.stavedu.ru