СИСТЕМА ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ДОКУМЕНТА

Модуль 2. Тема 5.
Козлов А.В.

Москва - 2014

Назначение систем оптического распознования 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).

Задача оптического распознования

- Перевод документов, научных публикаций, социальной информации, исторических изданий в электронный вид.
- Классификация документов.
- Накопление и хранение электронных документов.

ИСТОРИЯ

- 1929 году - Густав Таушек (*Gustav Tauschek*) получил патент на метод оптического распознавания текста в Германии;
- 1933 год - Гендель (*Paul W. Handel*) получил патент на свой метод в США ;
- 1935 год – Г. Таушек также получил патент США на свой метод;
- 1950 год - Дэвид Х. Шепард (*David H. Shepard*) - построил машину, решающую задачу преобразования печатных сообщений в машинный язык для обработки компьютером.
- 1955 год - Первая коммерческая система была установлена на «Ридерс Дайджест»
- 1965 год - «Ридерс Дайджест» и «Ар-Си-Эй» начали сотрудничество с целью создать машину для чтения документов, использующую оптическое распознавание текста, предназначенную для оцифровки серийных номеров купонов «Ридерс Дайджест», вернувшихся из рекламных объявлений.
- 1965 год - Почтовая служба Соединённых Штатов для сортировки почты использует машины, работающие по принципу оптического распознавания текста, созданные на основе технологий, разработанных исследователем Яковом Рабиновым.

ИСТОРИЯ

- 1971 год - Почта Канады использует системы оптического распознавания символов
- 1974 год - Рэй Курцвейл создал компанию «Курцвейл Компьютер Продактс», и начал работать над развитием первой системы оптического распознавания символов, способной распознать текст, напечатанный любым шрифтом.
- 1978 год - Компания «Курцвейл Компьютер Продактс» начала продажи коммерческой версии компьютерной программы оптического распознавания символов.
- 1992 год – Начало продажи первой коммерчески успешной программой, распознающей кириллицу, «AutoR» российской компании «ОКРУС» (ОС DOS).
- Конец 60-х годов – разработка и испытание шрифтонезависимого алгоритма распознования текста выпускниками МФТИ, биофизиками: Г. М. Зенкиным и А. П. Петровым

Системы оптического распознавания символов

При создании электронных библиотек и архивов путем перевода книг и документов в цифровой компьютерный формат, при переходе предприятий от бумажного к электронному документообороту, при необходимости отредактировать полученный по факсу документ используются системы оптического распознавания символов.

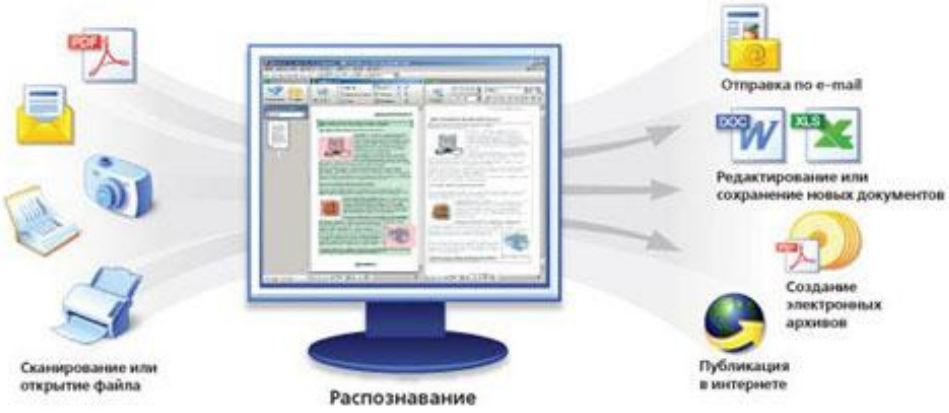
Оптическое распознавание символов *(англ. optical character recognition, OCR) – механический или электронный перевод изображений рукописного, машинописного или печатного текста в последовательность кодов, использующихся для представления в текстовом редакторе.* *(Википедия)*

С помощью сканера несложно получить изображение ницы текста в графическом файле.

Однако для получения документа в формате текстового файла необходимо провести распознование текста, т. е. преобразовать элементы графического изображения в последовательности текстовых символов.

- Сначала необходимо распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и т. д.

- Далее выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст.



ХОРОШЕЕ КАЧЕСТВО ТЕКСТА
РАСТРОВЫЙ МЕТОД РАСПОЗНАВАНИЯ ТЕКСТА

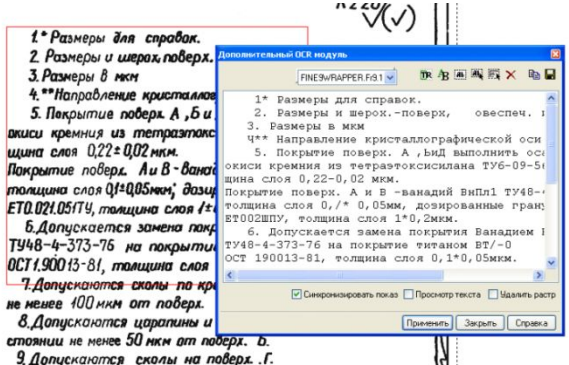
Если исходный документ имеет типографское качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений), то задача распознавания решается методом сравнения с растровым шаблоном.

- Сначала растровое изображение страницы разделяется на изображения отдельных символов.
- Затем каждый из них последовательно накладывается на шаблоны символов, имеющихся в памяти системы, и выбирается шаблон с наименьшим количеством точек, отличных от входного изображения.

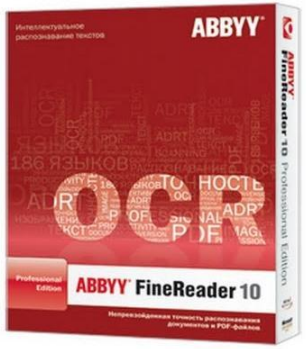
ПЛОХОЕ КАЧЕСТВО ТЕКСТА
СТРУКТУРНЫЙ МЕТОД РАСПОЗНАВАНИЯ

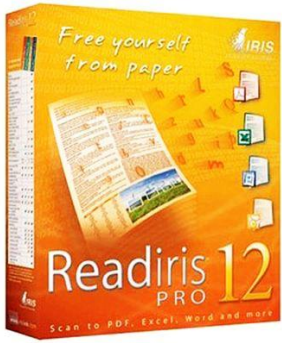
При распознавании документов с низким качеством печати (машинописный текст, факс и т. д.) используется метод распознавания символов по наличию в них определенных структурных элементов (отрезков, колец, дуг и др.).

Любой символ можно описать через набор параметров, определяющих взаимное расположение его элементов. Например, буква «Н» и буква «И» состоят из трех отрезков, два из которых расположены параллельно друг другу, а третий соединяет эти отрезки. Различие между буквами в величине улов, которые составляет третий отрезок с двумя другими. При распознавании структурным методом в искаженном символьном изображении выделяются характерные детали и сравниваются со структурными шаблонами символов. В результате выбирается тот символ, для которого совокупность всех структурных элементов и их расположение больше всего соответствуют распознаваемому символу.



ПРОГРАММЫ РАСПОЗНАВАНИЯ ТЕКСТА





Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (Optical Character Recognition - OCR).

Современная OCR должна уметь многое: распознавать тексты, набранные не только определенными шрифтами, но и самыми экзотическими, вплоть до рукописных. Уметь корректно работать с текстами, содержащими слова на нескольких языках, корректно распознавать таблицы. И самое главное — корректно распознавать не только четко набранные тексты, но и такие, качество которых, мягко говоря, далеко от идеала. Например, текст с пожелтевшей газетной вырезки или третьей машинописной копии. Само собой, распознать текст — это еще полдела. Не менее важно обеспечить возможность сохранения результата в файле популярного текстового (или табличного) формата — скажем, формата Microsoft Word.

Наиболее распространенные системы оптического распознавания символов, например, **ABBYY FineReader** и **CuneiForm** от Cognitive, используют как растровый, так и структурный методы распознавания. Кроме того, эти системы являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов) и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

Существует также системы On-line распознавания текста: **Online OCR** и **ABBYY FineReader Online** (http://www.onlineocr.ru , http://finereader.abbyyonline.com)

СИСТЕМЫ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ФОРМ

При проведении Единого государственного экзамена, при заполнении налоговых деклараций и т. д. используются различного вида бланки с полями. Рукописные тексты (данные вводятся в поля печатными буквами от руки) распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные базы данных.

Сложность состоит в том, что необходимо распознавать символы, написанные от руки, а они довольно сильно различаются у разных людей. Кроме того, система должна определить, к какому полю относится распознаваемый текст.

Системы распознавания рукописного текста. С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста. Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.

OCR-ПРИЛОЖЕНИЯ

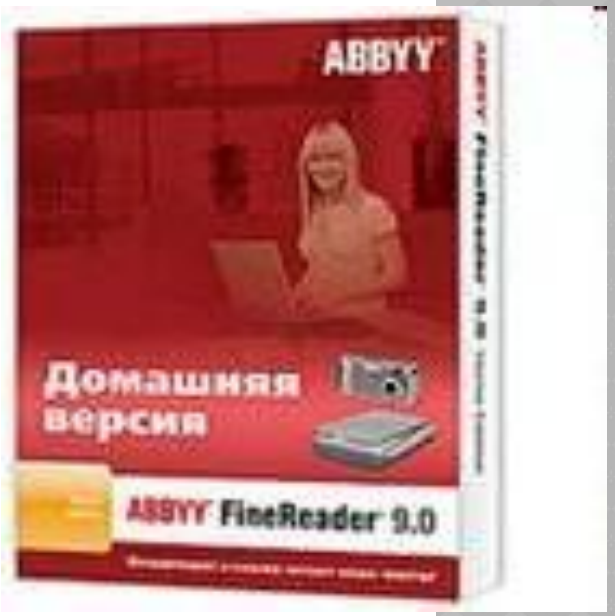
- **Это приложения, которые производят сканирование и распознавание текста, от англ. Optical Character Recognition - Оптическое распознавание символов**
- Это программы для перевода **изображений** документов в редактируемый текст, который можно затем обрабатывать в текстовых и табличных редакторах. По сравнению с ручной перепечаткой текста, такие программы дают существенный выигрыш в **скорости** работы, к тому же делают меньше ошибок. **Еще одно достоинство** - возможность сохранить иллюстрации, а они иногда не менее важны, чем текст документа.

OCR CUNEIFORM

- Это **бесплатная** программа сканирования и распознавания текста российского разработчика Cognitive Technologies.
- **OCR CuneiForm** обеспечивает быстрое, удобное и качественное распознавание текста с сохранением исходного вида документа. Поддерживается распознавание с более 20 языков, среди них русский, украинский, английский, немецкий, французский, испанский, итальянский, португальский, шведский, финский, сербский, хорватский, польский, а также распознавание смешанного русско-английского текста.

Скачать бесплатно программу сканирования и распознавания текста OCR Cuneiform 12 (freeware) с DepositFiles
http://depositfiles.com/files/sj9pt7q6x

ABBYY FINEREADER



- **Популярная программа распознавания текста российской компании ABBYY**
- Программа производит распознавание текста с более **180 языков**, для **38** из них предусмотрена встроенная проверка орфографии. Начиная с версии **Professional**, распознаются иврит, японский, тайский, китайский языки. Finereader открывает файлы графических форматов (TIFF, JPG, PFD, PNG и др.) в том числе **DjVu** – компактный формат для хранения отсканированных документов, книг.

- Стоимость программы **3990** рублей

OMNIPAGE

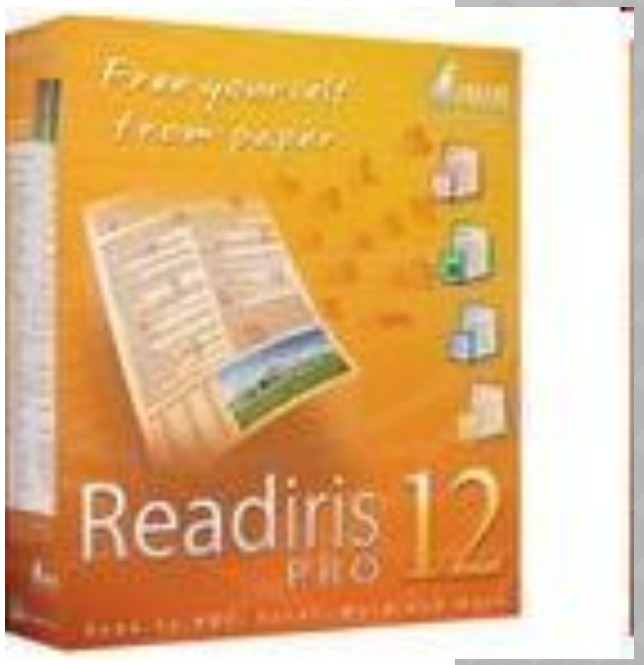
- Популярная программа распознавания текста **российской компании ABBYY**
- Программа отличается высокой скоростью и точностью распознавания. Распознаются более **120** языков с различными алфавитами: **латинский, греческий алфавиты, кириллица, китайский, японский и корейский** языки. Как и FineReader, OmniPage уверенно распознает документы, полученные с помощью цифровых камер с помощью технологии коррекции изображения "3D Correction".

- Стоимость программы **6090** рублей (150 евро)

READIRIS

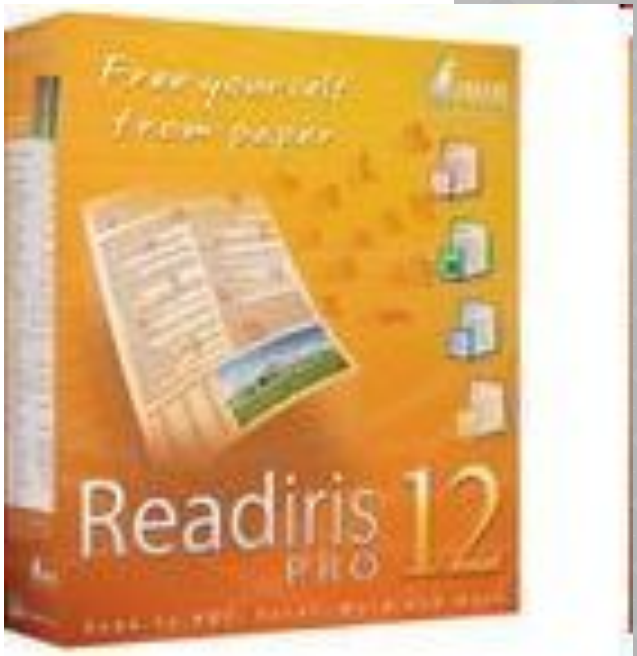
- Программа сканирования и распознавания текста **компании I.R.I.S.**
- Поддерживается распознавание текста с более **120 языков** распознавания, включая русский, а также ближневосточные языки - **арабский, иврит, фарси** (в версии Middle-East) и **японский, китайский, корейский** (в версии Asian). Есть версия Readiris для **Macintosh**.
- Вместе с поддержкой распознавания популярных форматов картинок, распознаются файлы **PDF** и **DjVu**.
- Стоимость программы

3845-14875 рублей (**129** $-**499** $)



MICROSOFT OFFICE DOCUMENT IMAGING

- Программа распознавания текста компании **Microsoft**
- Программа Document Imaging способна работать только с **двумя** языками: английским и языком локализации самого MS Office. Для поддержки других языков необходимо дополнительно устанавливать пакет **Multilingual User Interface** (MUI). **OCR** настроек в программе практически нет, программа в автоматическом режиме поддерживает распознавание типа и размера шрифтов, картинок и простых таблиц.
- Стоимость программы входит в стоимость пакета MS Office.



Источники литературы:

- 1. Богданов В., Ахметов К. Системы распознавания текстов в офисе. // Компьютер-пресс — 1999 №3, с.40-42.
- 2. Павлидис Т. Алгоритмы машинной графики и обработки изображений. М:, Радио и связь, 1986
- 3. Shani U. Filling Regions in Binary Raster Images — a Graph-theoretic Approach. // SIGGRAPH'80, pp 321-327.
- 4. Merrill R.D. Representation of Contours and Regions for Efficient Computer Search. // CACM, 16 (1973), pp. 69-82.
- 5. Pavlidis T. Filling Algorithms for Raster Graphics. // CGIP, 10 (1979), pp. 126141.
- 6. http://expscan.narod.ru/
- 7 . http://ru.wikipedia.org/wiki/OCR