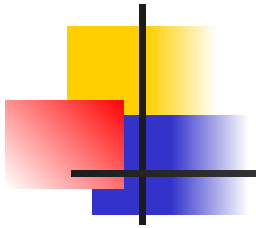
Системы оптического распознавания информации.

Борисов В.А.
КАСК – филиал ФГБОУ ВПО РАНХ и ГС
Красноармейск 2011 г.

Системы оптического распознавания текста (OCR)

- Optical Character Recognition — OCR-системы предназначены для автоматического ввода печатных документов в компьютер.

- Современные программы распознавания текста обеспечивают проверку орфографии, автоматическое форматирование текста и массу других дополнительных удобств.



ВОЗМОЖНОСТИ ПРОГРАММЫ FINEREADER

FineReader

- Омнифонтовая система оптического распознавания текстов.
- Позволяет распознавать тексты, набранные практически любыми шрифтами.

Особенности программы FineReader

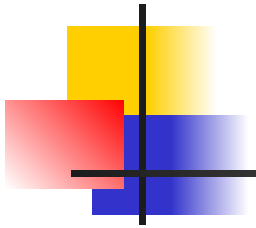
- Высокая точность распознавания и малая чувствительность к дефектам печати, что достигается благодаря применению технологии «целостного целенаправленного адаптивного распознавания».

Программа позволяет

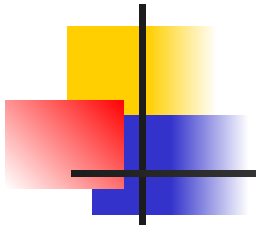
- распознавать с высокой точностью тексты более чем на 175 языках,
- выводить на печать исходное изображение и распознанный текст,
- сохранять отсканированное изображение в различных форматах,
- настраивать панели инструментов программы.

Программные продукты ABBYY FineReader

- FineReader Sprint,
- FineReader 6.0 Professional,
- FineReader 6.0 Corporate Edition,
- ABBYY FineReader 5.0 Pro for Mac.



ТЕХНОЛОГИЯ РАСПОЗНАВАНИЯ



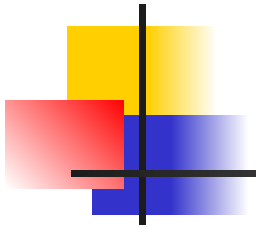
- Сложность машинного распознавания текстов заключается в том, что его невозможно построить по жесткому алгоритму хотя бы потому, что для написания одной и той же буквы существует множество вариантов написания.

Принцип целостности

- Распознаваемое изображение рассматривается как единый объект, состоящий из частей, связанных между собой пространственными соотношениями.

Принцип целенаправленности

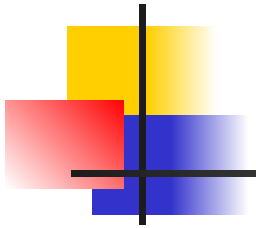
- Распознавание строится как процесс выдвижения и целенаправленной проверки гипотез об объекте, а принцип адаптивности подразумевает способность системы к самообучению.



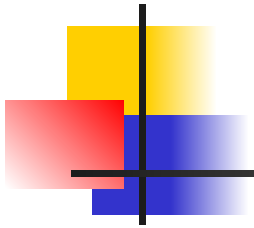
- Для выдвижения гипотез о том, что может представлять собой изображение, применяются так называемые *признаковые классификаторы*.

Признаковые классификаторы

- Используют ряд признаков, на основе которых программа вычисляет степень близости распознаваемого изображения и известных ей классов изображений, после чего выдает список подходящих классов, т. е. гипотезу о принадлежности объекта к тому или иному классу.



- Признаковые классификаторы применяются также и для повышения точности распознавания изображений с дефектами.



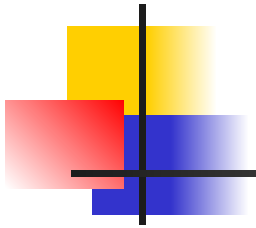
- Полученный набор классов последовательно проверяется структурным классификатором, анализирующим каждый символ.

Структурный эталон

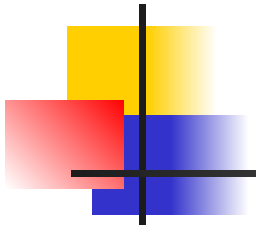
- Описывает символ как комбинацию структурных элементов (отрезок, дуга, кольцо, точка), находящихся в определенных отношениях между собой.

Процесс распознавания

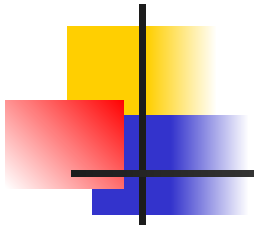
- Делится на этапы выделения структурных элементов в изображении и сопоставлении их с эталоном.



- Если в окончательный список попало более одной гипотезы, они попарно сравниваются с помощью дифференциальных классификаторов.



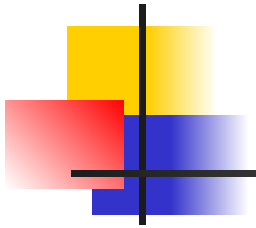
- Если структурный классификатор при распознавании символов не может однозначно выбрать одну из двух букв с похожим написанием, то между этими конкурирующими гипотезами делается дифференциальный выбор.



- С завершением работы дифференциального классификатора заканчивается распознавание и начинается этап проверки итогового списка гипотез.

Окончательная стадия распознавания

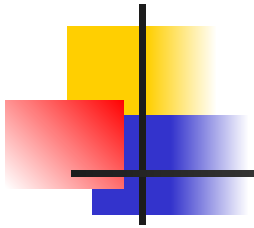
- Осуществляется системой контекста — при наличии некоторого количества распознанных букв из слова программа, используя словарь, может «догадаться», что это за слово.



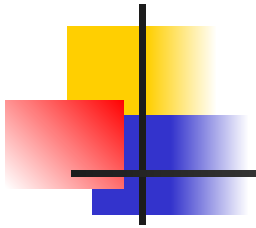
ОРГАНИЗАЦИЯ РАБОТЫ В FINEREADER

Пакет

- Является основой работы FineReader.
- Содержит всю информацию о распознаваемом документе.
- Представляет собой набор страниц документа и может содержать около тысячи страниц.

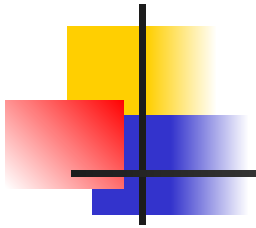


- В один пакет для удобства работы рекомендуется объединять изображения, логически связанные между собой, например страницы одной книги.

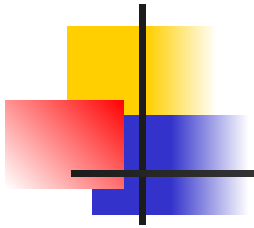


- В окне Пакет виден список страниц, входящих в открытый пакет.
- Для просмотра страницы нужно щелкнуть мышью по ее изображению или номеру, при этом откроются файлы, которыми данная страница представлена в пакете.

- Страницы в окне Пакет могут быть представлены пиктограммами или уменьшенным изображением страницы.



- Если исходное изображение представляет собой негатив, оно может быть инвертировано, далее производится очистка от «мусора» — мелких дефектов изображения.



- Если не нужна цветность, то цветные изображения сводятся к черно-белым, что экономит место на диске и ускоряет процесс распознавания.

Анализ макета страниц пакета

- FineReader анализирует ориентацию страницы и переворачивает изображение, если это необходимо, а также выделяет блоки - области, которые при дальнейшем анализе будут интерпретироваться как текст, таблицы или рисунки.

Распознавание текста и таблиц

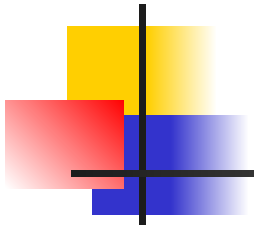
- Является «сердцем» FineReader и обеспечивает ее уникальность, однако этот процесс совершенно незаметен пользователю.

Проверка правописания

- «На суд» пользователя выносятся слова, которых нет в словаре системы, а также символы, в точности распознавания которых программа не уверена.

Сохранение и экспорт результатов распознавания

- Вся информация, включая распознанный текст и его форматирование, автоматически сохраняются в пакете вместе с исходным изображением и сведениями о макете страниц.

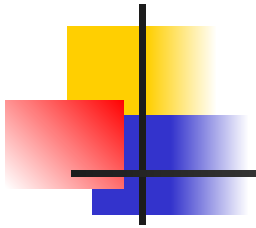


СКАНИРОВАНИЕ ИЗОБРАЖЕНИЙ

- Для сканирования изображения документа кладем на стекло сканера страницу с текстом или книгу и нажимаем кнопку Сканировать (Scan) или в меню Файл выберем пункт Сканировать.

Качество распознавания

- Зависит от того, насколько хорошее изображение получено при сканировании, что достигается установкой основных параметров сканирования — типа изображения, разрешения и яркости.



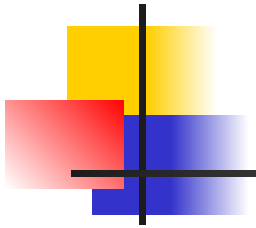
- Черно-белый тип изображения обеспечивает более высокую скорость сканирования, но при этом теряется часть информации о буквах, что может привести к ухудшению качества распознавания на документах среднего и низкого качества печати.

Настройки

- инвертирование изображения,
- очистку от «мусора»,
- автоматическое определение ориентации текста на изображении.

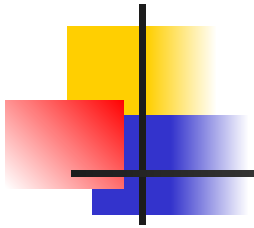
- При распознавании изображение должно иметь стандартную ориентацию, т. е. текст должен читаться сверху вниз и строки должны быть горизонтальными.

- После завершения сканирования изображение окажется включенным в конец пакета, если не активна опция Запрашивать номер страницы перед добавлением в пакет, а его пиктограмма отобразится на панели пакета.

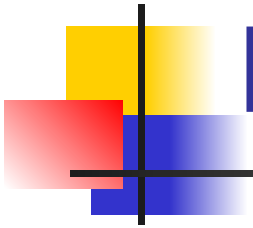


АНАЛИЗ МАКЕТА СТРАНИЦ

- Определение ориентации текста при установке соответствующей опции производится автоматически, хотя можно сделать это и вручную путем поворота исходного изображения.



- отдельными блоками выделяются таблицы и рисунки, которые не подлежат распознаванию;
- четкое выделение блоков позволяет максимально корректно сохранить макет исходной страницы при передаче распознанного документа во внешние приложения.



Блоки

- Заключенные в рамки участки изображения.
- Блоки выделяют для того, чтобы указать программе, какие участки отсканированной страницы надо распознавать и в каком порядке.
- Также по ним воспроизводится исходное оформление страницы.

Типы блоков

- зона распознавания,
- текст,
- таблица,
- картинка,
- штрих-код.

Графики с подписями осей

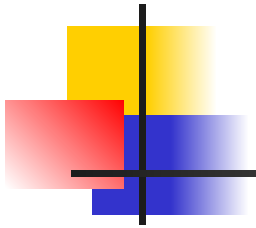
- FineReader отдает предпочтение тексту и выделяет подписи как текстовый блок, оставляя сам график без внимания или же выделяя как рисунок какую-либо его часть.

Сложные математические или химические формулы

- При работе с документами, содержащими формулы, их приходится выделять как рисунки.

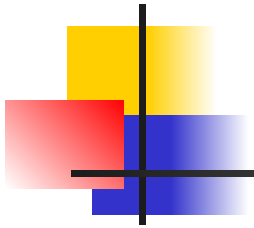
Плохой оригинал

- Подобные ошибки могут быть исправлены на этапе работы с макетом, поскольку сделать это проще, чем впоследствии редактировать готовый текст.



- Изменять размеры или форму существующих блоков можно, потянув мышью за их границы.

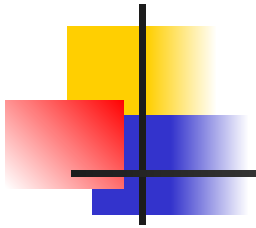
- Изменить тип блока позволяет «всплывающее» меню, появляющееся после щелчка мышью по пиктограмме в углу блока, обозначающего его тип.



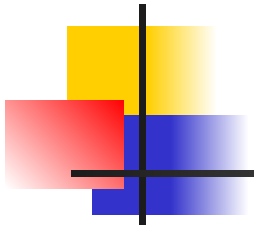
РАСПОЗНАВАНИЕ ТЕКСТА

Задача распознавания

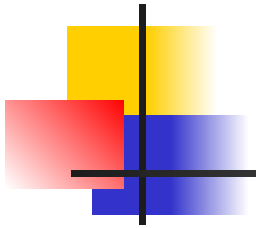
- Преобразовать отсканированное изображение в текст, сохранив при этом оформление страницы.



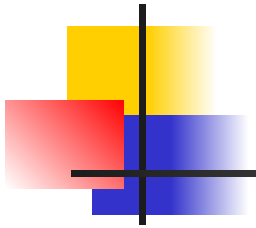
- Язык, на котором будет проводиться распознавание, выбирается на основной панели инструментов.



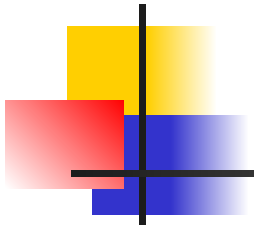
- Помимо языка оригинала, модуль распознавания учитывает и тип печати, который по умолчанию определяется автоматически, но при необходимости может быть установлен и вручную.



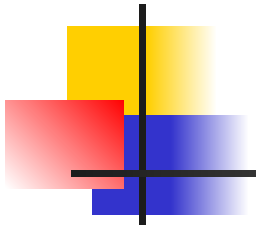
ПРОВЕРКА ПРАВОПИСАНИЯ И СОХРАНЕНИЕ РЕЗУЛЬТАТОВ РАБОТЫ



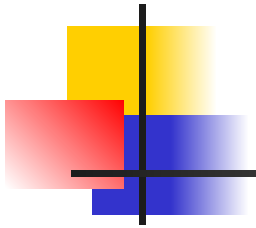
- Модуль распознавания анализирует не только отдельные символы, но и целые слова, используя при этом встроенный словарь.



- Работа со словами, неизвестными системе, и с неуверенно распознанными символами осуществляется в модуле проверки правописания.



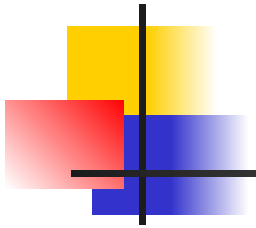
- После окончания проверки правописания следует определить, в каком формате сохранять полученные результаты.



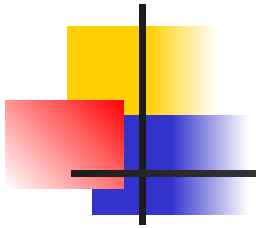
ДРУГИЕ OCR-СИСТЕМЫ

Предварительное сканирование позволяет

- выделить мышью область сканирования;
- выбрать режим сканирования;
- выставить параметры яркости, контраста или выбрать автоматическое определение этих параметров;
- запустить основное сканирование.



- Подбор настроек сканера уменьшает количество неверно распознанных букв до вполне приемлемого качества сканирования и распознавания.



- Особенно важен подбор оптимальной яркости при сканировании достаточно большого объема текста низкого качества.