

Сканирование и распознавание текстовой и графической информации

В практической деятельности часто встречаются ситуации, когда необходимо перевести в электронный вид документ, напечатанный на бумаге. В этом случае можно просто набрать документ на компьютере, что довольно трудно, либо воспользоваться сканером — устройством, специально предназначенным для перевода документов в электронный вид.

Сканер – это аппарат для оцифровки бумажных (или других) носителей визуальной информации.

Виды сканеров:

- Сканер изображений — устройство для считывания двумерного изображения и представления его в растровой электронной форме
- 3D-сканер — устройство для считывания формы объёмного объекта
- Биометрические сканеры используются для целей идентификации личности

• Устройства автоматизированного считывания служебной информации: *Сканеры форм: для ввода информации с заполненных бланков (разновидность листовых). Очень высокое быстродействие (для анкет, опросных листов, избирательных бюллетеней).*

• Сканер штрихкода — устройство для считывания информации, представленной в виде штрих-кода

- Считыватель RFID-меток

RFID — это специальная технология идентификации, предоставляющей пользователям большие возможности. Самые распространенные RFID-метки, также как и штрих-коды, представляют самоклеящиеся rfid метки. Но если на штрих-кодах вся информация хранится в граф. виде, на метку данные заносят при помощи радиоволн. Radio Frequency IDentification, в переводе на русский - радиочастотная идентификация.

По существу сканер является устройством, которое воспринимает оптическую информацию, доступную для нашего зрения, и сначала преобразует ее в электрическую форму, а затем приводит к цифровому виду, пригодному для ввода в компьютер.

Технические характеристики сканеров

Разрешающая способность (разрешение). - чувствительность к мелким деталям воспроизводимого изображения. Чем выше разрешающая способность, тем точнее воспроизводится изображение оригинала.

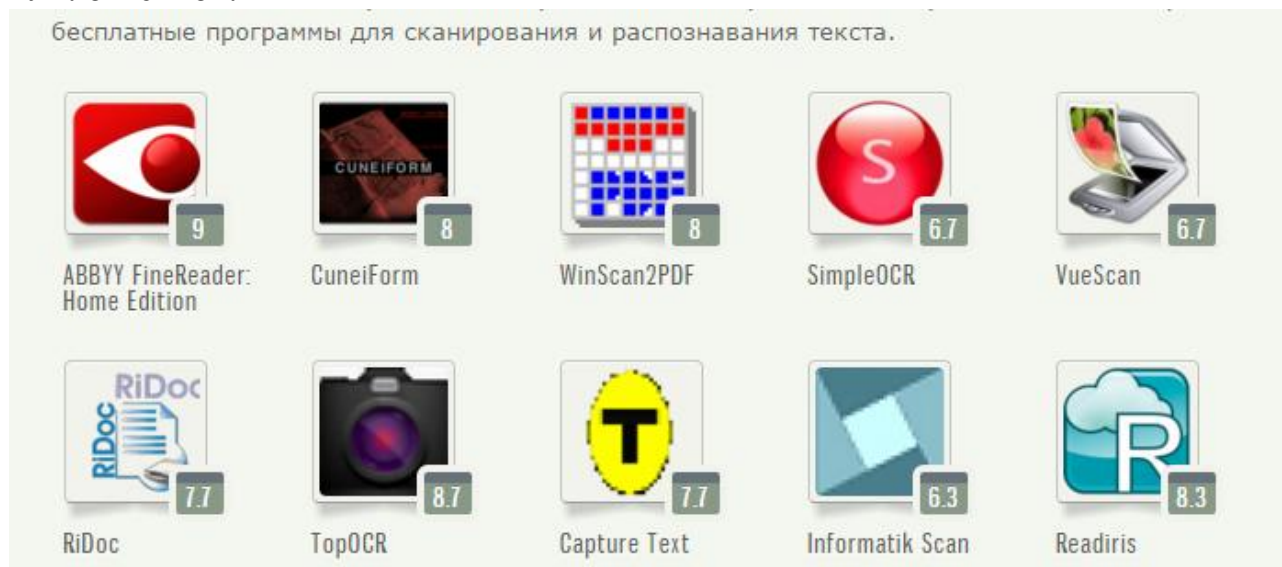
Производительность

Количество времени, затраченное на предварительный просмотр и собственно сканирование, может значительно влиять на производительность. Скорость сканирования зависит от таких факторов, как разрешение, размер и тип изображения (например, черно-белая или цветная фотография или же текст).

Глубина цвета

Глубина цвета – это характеристика, показывающая, насколько точна информация о цвете каждой точки отсканированного изображения, и отражающая разрядность аналого-цифрового преобразователя.

С помощью сканера можно считывать как графические изображения, так и текстовую информацию. Для этих целей используются специальные программы, например Fine Reader, позволяющие преобразовывать текст в файлы с высокой производительностью и точностью, с распознаванием до 16 млн. оттенков.



Сканирование – технологический процесс, в результате которого создается графический образ бумажного документа. Минимальный элемент изображения интерпретируется сканером как цветная (или серая) точка и в файле хранится растровое изображение исходного документа, количество точек определяется размером изображения и разрешением сканера.

Отсканированное изображение документа не является текстовым, а лишь набором разноцветных точек. Распознавание текста производится с помощью специальных программных средств – средств распознавания образов.

Наиболее широко известными и распространенными являются программы Fine Reader (ABBYU) и Cunei Form (Cognitive). Fine Reader обеспечивает высокое качество распознавания и удобство применения, позволяет объединять сканирование и распознавание в одну операцию.

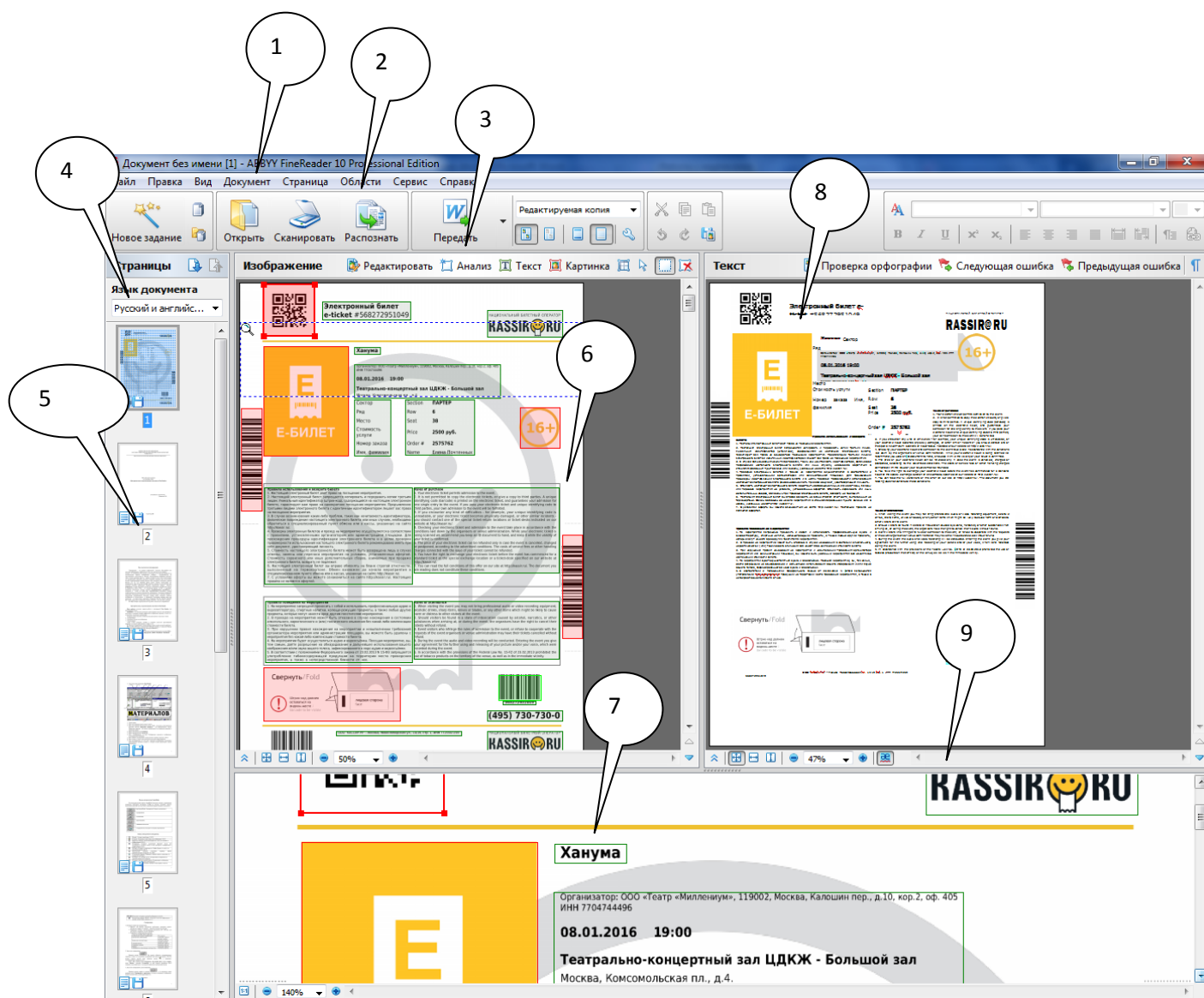
ABBYY FineReader — система оптического распознавания текстов (Optical Character Recognition – OCR). Предназначена как для автоматического ввода печатных документов в компьютер, так и для конвертирования PDF–документов и фотографий в редактируемые форматы.

Программа FineReader выпускается отечественной компанией ABBYY Software (www.bitsoft.ru). Эта программа предназначена для распознавания текстов на русском, английском, немецком, украинском, французском и многих других языках, а также для распознавания смешанных текстов.

Программа имеет ряд удобных возможностей. Она позволяет объединять сканирование и распознавание в одну операцию, работать с пакетами документов (или с многостраничными документами) и с бланками. Программу

можно обучать для повышения качества распознавания неудачно напечатанных текстов или сложных шрифтов. Она позволяет редактировать распознанный текст и проверять его орфографию.

1. Заголовок окна программы
2. Горизонтальное меню
3. Панель инструментов Главная
4. Область для указания языка входного текста.
5. Окно Страницы
6. Окно «Изображение». В окне Изображение выводится изображение, подлежащее распознаванию.
7. Окно Крупный план - в этом окне можно рассмотреть детали изображения
8. Окно «Текст». В окно Текст выводятся результаты распознавания.
9. Свойства области



Последовательность сканирования

Процесс ввода документов с помощью программы ABBYY FineReader состоит из трех этапов: получение изображения, распознавание и сохранение результатов распознавания.

1. команда **Сканировать** - запускается специальный режим сканирования во время которого система контролирует действия пользователя и подсказывает ему, что надо делать, чтобы получить тот или иной результат. В результате сканирования появится окно Изображение, содержащее "фотографию" страницы. Затем документ разбивается на блоки, содержащее цельные фрагменты. Эту операцию программа может выполнить автоматически, хотя разбиение не всегда проходит удачно. Границы и типы блоков можно устанавливать вручную.

2. команда **Распознать** - выделенные при сканировании области анализируются. Результат распознавания Вы увидите в окне Текст. Если программа выделила некоторые блоки неправильно, проще и быстрее редактировать неправильно размеченные блоки, используя инструмент для редактирования блоков, чем удалять блоки и выделять их заново вручную.

Неправильно отображаемые символы

Если в окне Текст программы FineReader символы отображаются голубым цветом это означает, что приложение сомневается в правильности распознавания, необходимо визуально проверить работу приложения или запустить проверку орфографии.

3. команда **Передать в ...** - результат сканирование направляется в то или иное программное средство для последующего редактирования или сохранения.

Последовательность выполнения сканирования документов

- включить сканер;
- открыть крышку сканера и положить документ (лицом вниз);
- выполнить команду Пуск - устройства и принтеры, контекстное меню по устройству - начать сканирование;

- установить параметры сканирования (цветовой формат, разрешение);

Используйте разрешение 300 dpi для стандартных текстов (размер шрифта 10pts. и больше) и разрешение 400-600 dpi для текстов с меньшим шрифтом (9pts. и меньше). Сканирование в сером режиме рекомендуется для повышения качества распознавания. При сканировании в сером режиме яркость регулируется автоматически.

- импортировать полученное изображение (скан сохраняется в указанную папку);

- полученное изображение открыть с помощью FineReader;

- выполнить команду распознать (если распознавание не началось автоматически);

- передать полученный документ в приложение для последующего форматирования.

PDF документ

Одним из наиболее популярных форматов представления электронных документов в Internet, архивах и т.д. является формат **PDF** (*Portable Document Format*). PDF является распространенным форматом для пересылки документов по электронной почте или публикации документов на web-сайтах. Естественно, что при публикации на web-сайтах очень важна высокая скорость

открытия документов. Документ, сохраненный из программы FineReader в формате PDF, отвечает подобным требованиям.

Для чтения и редактирования PDF документов используются приложения Adobe Acrobat Reader или Foxit Reader.