

# **Системы оптического распознавания документов**





# **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

## 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.

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

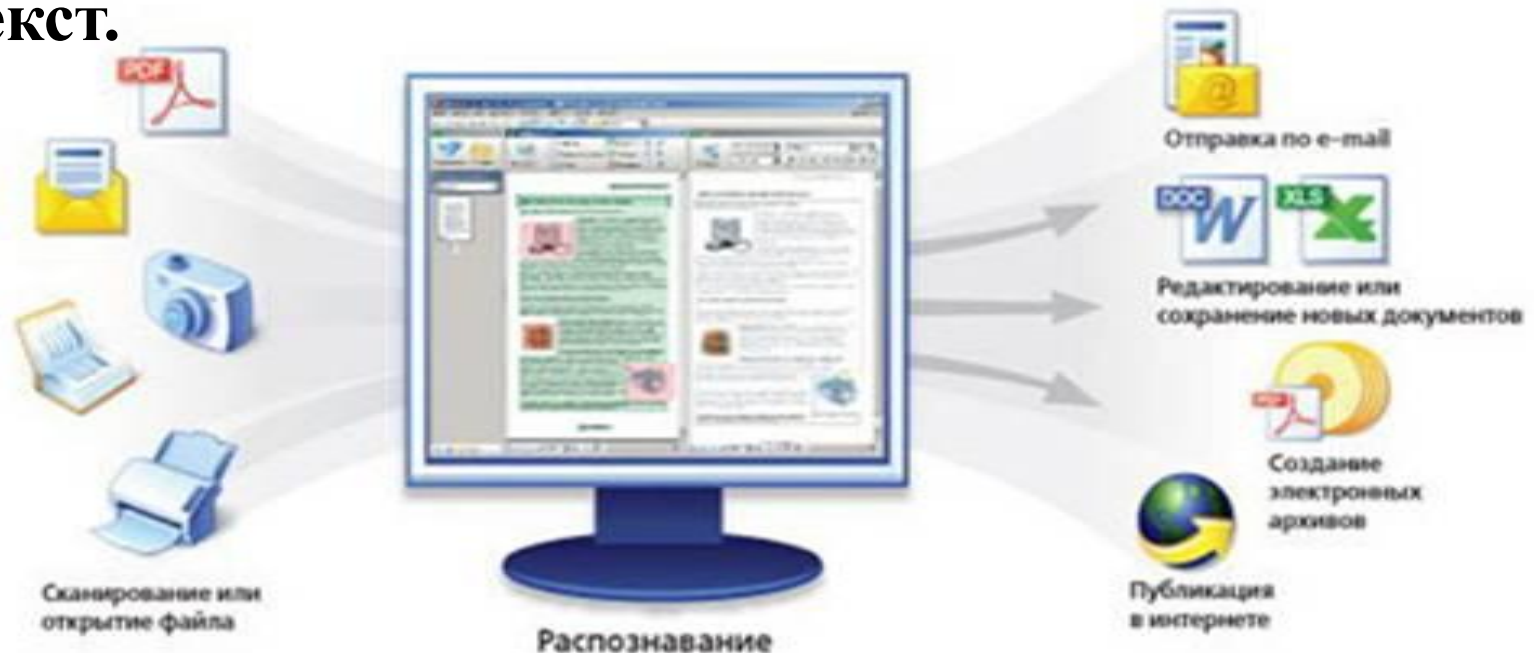


Однако для получения документа в формате текстового файла необходимо провести **распознавание текста**, т. е. преобразовать элементы графического изображения в последовательности текстовых символов.





- 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**

- 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.

**А В В Ф Я**



## *Хорошее качество текста*

### **Растровый метод распознавания текста**

- Растровое изображение каждого символа последовательно накладывается на растровые шаблоны символов, хранящиеся в памяти системы оптического распознавания. Результатом распознавания является символ, шаблон которого в наибольшей степени совпадает с изображением



Например, распознаваемый символ "Б" накладывается на растровые шаблоны символов (А, Б, В и т. д.)

## *Плохое качество текста*



### **Структурный метод распознавания**

- При распознавании документов с **низким** качеством печати (машинописный текст, факс и т.д.) используется **метод распознавания структурных элементов** (отрезков, колец, дуг и др.) символов. В искаженном символьном изображении выделяются характерные детали и сравниваются со **структурными шаблонами** символов.

**И..Н**

- 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 углов, которые составляет третий отрезок с двумя другими.

## *Плохое качество текста*



### **Структурный метод распознавания**

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.



Например, распознаваемый символ "Б" накладывается на векторные шаблоны символов (А, Б, В и т. д.)



## 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 оптического распознавания форм

Единый государственный экзамен - 2005  
**Бланк регистрации**

Регистр. Код образовательного учреждения. Класс. Номер. Дата проведения ЕГЭ. Код пункта проведения ЕГЭ. Номер аудитории. Код предмета. Название предмета. Номер варианта. Служебная отметка.

Заполнить гелевой или капиллярной ручкой ЧЕРНЫМИ чернилами ЗАГЛАВНЫМИ ПЕЧАТНЫМИ БУКВАМИ по следующим образцам:  
А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я 1 2 3 4 5 6 7 8 9 0 X V I L

**ВНИМАНИЕ!** Данный бланк использовать только совместно с двумя другими бланками из данного пакета

Сведения об участнике единого государственного экзамена

Фамилия. Имя. Отчество.

Документ. Серия. Номер. Пол. ☐ Ж. ☐ М.

Резерв - 1. Резерв - 2. Резерв - 3. Факт выхода из аудитории во время экзамена. ☐

**ЗАМЕЧАНИЯ** участника ЕГЭ по процедуре проведения ЕГЭ

Заполнение НЕ ОБЯЗАТЕЛЬНО.

Отметьте ☐ замечания по проведению экзамена:

☐ Организация доставки участника в пункт проведения ЕГЭ при самостоятельном времени в пути более 1 часа. ☐ Присутствие в аудитории преподавателя образовательного учреждения, по которому проводится ЕГЭ

☐ Вскрытие доставочного пакета осуществлялось НЕ в присутствии участника ЕГЭ. ☐ Наличие нарушений дисциплины в аудитории

## FineReader Forms

Единый государственный экзамен - 2005  
**Бланк ответов № 1**

Заполнить гелевой или капиллярной ручкой ЧЕРНЫМИ чернилами ЗАГЛАВНЫМИ ПЕЧАТНЫМИ БУКВАМИ по следующим образцам:  
А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я 1 2 3 4 5 6 7 8 9 0

Регистр. Код образовательного учреждения. Класс. Номер. Дата проведения ЕГЭ. Код пункта проведения ЕГЭ. Номер аудитории. Код предмета. Название предмета. Номер варианта. Служебная отметка.

Получил участника ЕГЭ право выдачи бланка. 102

**ВНИМАНИЕ!** Данный бланк использовать только совместно с двумя другими бланками из данного пакета

Получив бланк ответов, участник должен проверить, чтобы в бланке не было повреждений, и убедиться, что бланк соответствует требованиям, указанным в инструкции к бланку.

Бланк ответов состоит из 10 страниц. В бланке ответов № 1 предусмотрено 10 страниц для ответов на вопросы экзамена. В бланке ответов № 2 предусмотрено 10 страниц для ответов на вопросы экзамена. В бланке ответов № 3 предусмотрено 10 страниц для ответов на вопросы экзамена. В бланке ответов № 4 предусмотрено 10 страниц для ответов на вопросы экзамена. В бланке ответов № 5 предусмотрено 10 страниц для ответов на вопросы экзамена.

Бланк ответов № 1. Бланк ответов № 2. Бланк ответов № 3. Бланк ответов № 4. Бланк ответов № 5.

Резерв - 1. Резерв - 2. Резерв - 3. Резерв - 4. Резерв - 5.

- **Бланком** называется стандартный лист бумаги, на котором размещается постоянная информация и отведено место для переменной.
- Сложность состоит в том, что необходимо распознать написанные от руки символы, довольно сильно различающиеся у разных людей.
- Кроме того система должна определить, к какому полю относится распознаваемый текст.



## Системы оптического распознавания форм

- Для обработки бланков предназначено специальное приложение **FineReader Forms**.
- Для распознавания содержимого бланка необходимо предварительно создать шаблон формы.

### Сервис/ Шаблоны

- Шаблон используют на этапе сегментации. Сегментация в данном случае состоит в наложении шаблона.
- Положение шаблона корректируется в соответствии с тем, насколько ровно был размещён бланк при сканировании.
- Заключительный этап состоит в распознавании содержимого бланка.



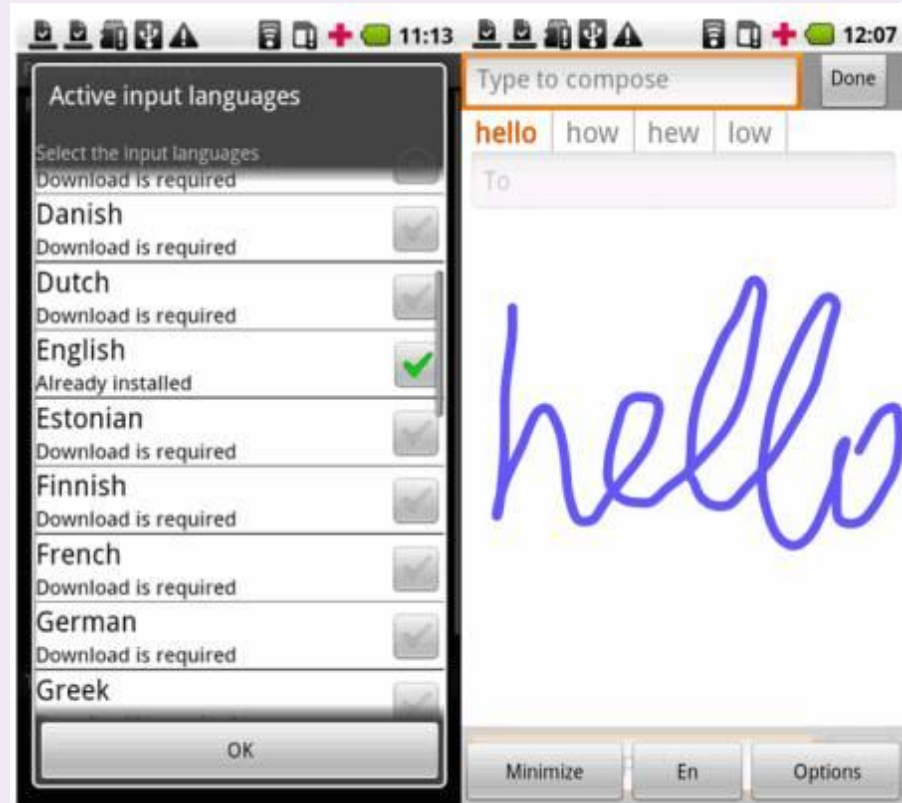
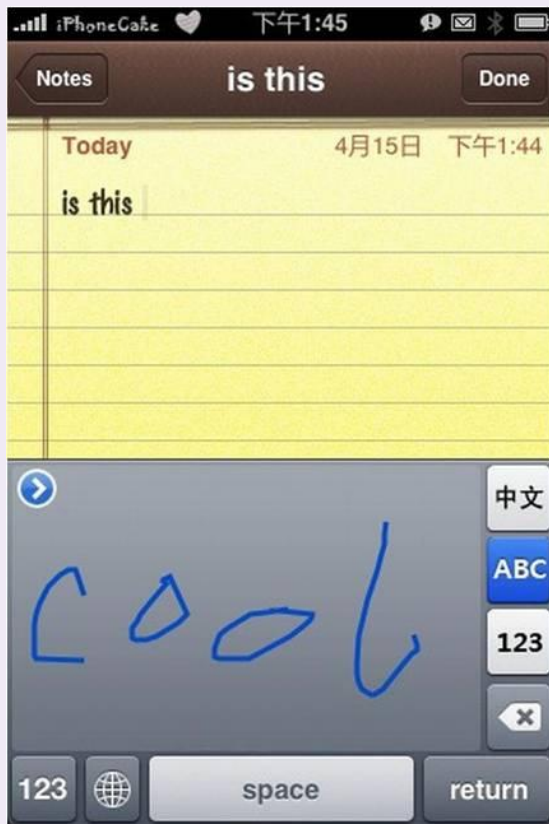
## Системы распознавания рукописного текста

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.





# Системы распознавания рукописного текста



# **Программы оптического распознавания текста**





# Программы оптического распознавания документов





# Принцип работы сканера

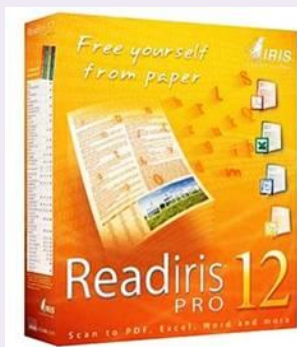
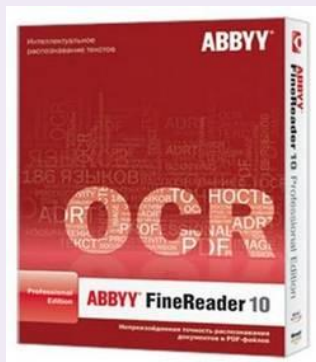
Принцип работы сканера состоит в следующем: в результате преобразования света получается электрический сигнал, содержащий информацию об активности цвета в исходной точке сканируемого изображения. После оцифровки аналогового сигнала в АЦП цифровой сигнал через аппаратный интерфейс сканера идет в компьютер, где его получает и анализирует программа для работы со сканером. После окончания одного такого цикла (освещение оригинала — получение сигнала — преобразование сигнала — получение его программой) источник света и приемник светового отражения перемещается относительно оригинала.



# Программы распознавания текста

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

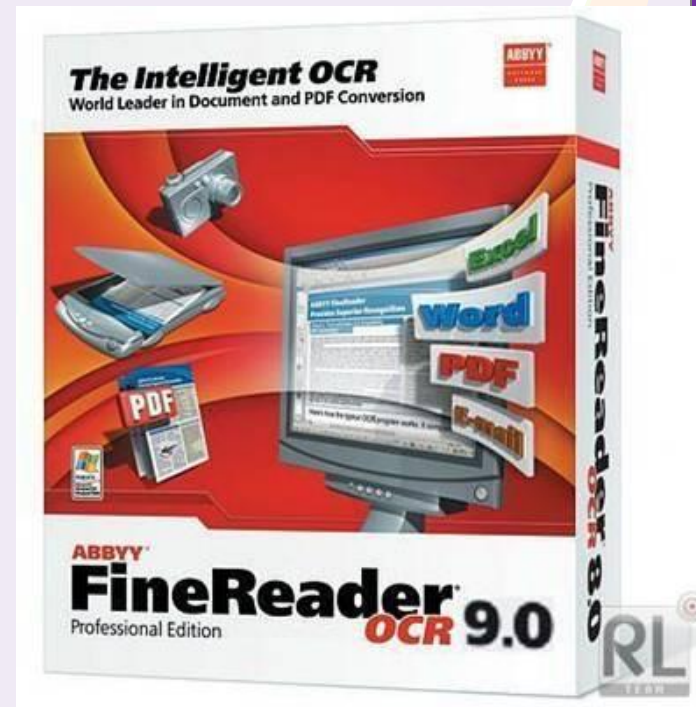
Современная OCR должна уметь многое: распознавать тексты, набранные не только определенными шрифтами, но и самыми экзотическими, вплоть до рукописных. Уметь корректно работать с текстами, содержащими слова на нескольких языках, корректно распознавать таблицы. И самое главное — корректно распознавать 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.





# ABBY FineReader

- Популярная проприетарная программа распознавания текста 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.





# Процесс обработки FineReader

- **Сканирование** (сканер, цифровой фотоаппарат, цифровая видеокамера).
- **Сегментация** - выделение блоков на изображении.
- **Распознавание** – неоднозначно опознанные символы выделяются цветом.
- **Проверка ошибок**- можно провести проверку грамматики.
- **Сохранение** результатов в виде отформатированного или неотформатированного документа, или прямой передачи в другое приложение - WORD, Excel в буфер обмена Windows.