

Текст распознан!

При работе с документами часто возникает необходимость ввести в компьютер тексты из книг или газет. В некоторых случаях это занимает много времени. Поэтому удобнее воспользоваться софтом для распознавания печатного текста.

ПОЛЕЗНАЯ ИНФОРМАЦИЯ

Как работает OCR-система?

Распознавание текста документа, представленного в виде графического файла, происходит в несколько этапов. Сначала программа разбивает документ на блоки разных типов (изображения, текст, таблицы, штрихкоды и прочее). Затем текстовые и табличные блоки разбиваются на строки. Следующий этап — разделение строк на непрерывные области — символы. Программа сверяет их с эталонным и таким образом распознает.

Современные OCR-системы (Optical Character Recognition) позволяют распознавать тексты, набранные практически любыми шрифтами, в том числе и готическими, и при этом вам не понадобится предварительно «обучать» программу. Процедура достаточно проста: прежде всего бумажный документ необходимо отсканировать, а затем из полученного графического файла с помощью систем OCR-распознавания выделяется текст. Следовательно, для этого вам понадобится сканер и специальное программное обеспечение.

Выбор оборудования и сканирование

Выбирая сканер, помните, что для распознавания текста будет достаточно разрешения 600 dpi. Поэтому если вы планируете использовать сканер только для работы с документами, покупать дорогостоящее оборудование с разрешением более 1200 dpi, укомплектованное слайд-модулем, будет напрасной тратой средств. Лучше отдать предпочтение более простой и дешевой модели, например — планшетному сканеру **HP ScanJet 2400**.

Но помимо выбора наиболее подходящего для такой работы оборудования важно правильно отсканировать документ, ведь качество распознавания во многом зависит от качества исходного изображения. В этом случае основными параметрами сканирования будут **Тип изображения, Разрешение и Яркость**. В качестве первого выберите градации серого — в этом случае подбор оптимального значения яркости производится автоматически, что позволяет повысить качество полученного изображения. Можно сканировать и в черно-белом режиме, но при этом возможна потеря части графической информации, поэтому используйте его только для документов высочайшего качества печати. Цветной режим целесообразно выбирать, только если вам необходимо сохранить цветность элементов текста (буквы или фон). Разрешение более 400 dpi необходимо для сканирования текста, набранного мелким шрифтом (9 и менее пунктов). В других случаях будет достаточно 300 dpi. Наконец, если качество печати документа хорошее, выберите значение яркости 50%. Повышение или понижение этого параметра понадобится только для сканирования слишком темных или светлых документов соответственно.

Выбор программного обеспечения

Лидером среди программ для распознавания текста является российская разработка **ABBYY FineReader 7.0**. Несмотря на то что сегодня существуют и альтернативные программы, она продолжает оставаться наиболее популярной среди пользователей. К ее плюсам можно отнести русскоязычный интерфейс, многофункциональность и простоту в работе.

FineReader

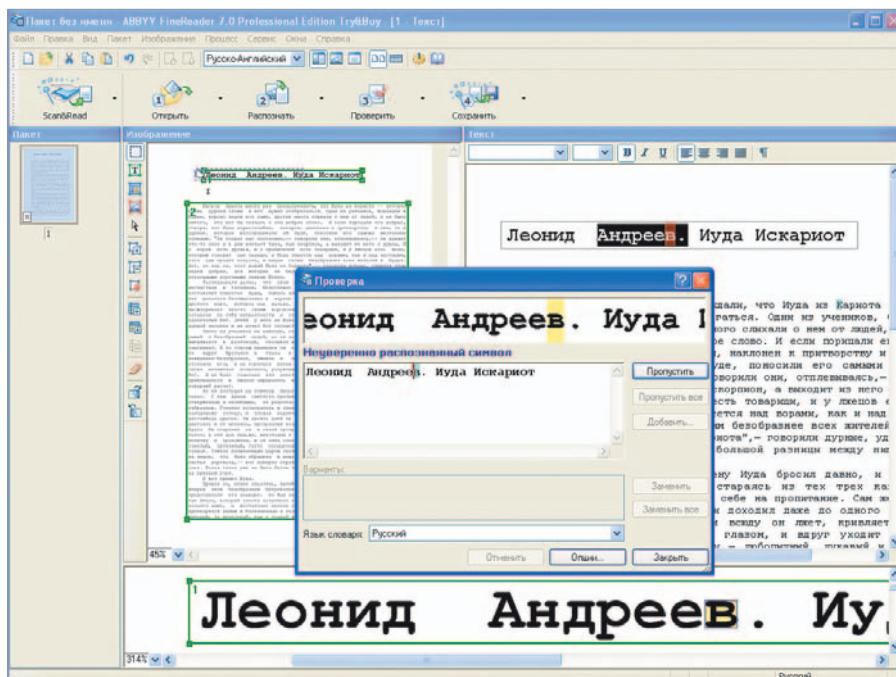
Версия ▶ 7.0 Professional Edition
Разработчик ▶ ABBYY Software House
Сайт ▶ www.abbyy.ru
ОС ▶ Windows 98SE/ME/NT/2000/XP
Объем дистрибутива ▶ 42,7 Мбайт
Интерфейс ▶ русскоязычный
Условия распространения ▶ commercial
Цена ▶ \$129

Распознавать документы удобнее всего в режиме **Мастера Scan&Read**. Для его запуска выберите одноименный пункт меню **Процесс**. Сначала нужно выбрать исходное изображение: можно получить его непосредственно со сканера или открыть уже имеющийся графический файл. После добавления нужных документов в пакет выберите язык текста, на котором они написаны. Нажатие на кнопку **Далее** запускает процесс распознавания документа, после завершения которого вам предложат оценить качество работы. Для этого достаточно просмотреть полученный текст. Если многие буквы выделены бирюзовым цветом, распознавание прошло плохо, в этом случае нужно заново отсканировать исходный текст, но с измененными параметрами. Если же все в порядке, **FineReader** предложит проверить текст и сохранить его: программа поддерживает все наиболее распространенные файловые форматы: **DOC, RTF, HTML, PDF, TXT** и другие.

Распознавание журнальных статей

Очень часто возникает необходимость распознать текст из журнальных или газетных статей. Такие документы имеют ряд особенностей: при их создании используется полноцветная графика, часто они оформлены в несколько колонок.

Все это создает дополнительные трудности для программы OCR-распознавания текста. Чаще всего ПО неправильно выделяет различные блоки — области на странице, содержащие текст (зеленая рамка), рисунки (красная рамка), таблицы (синяя рамка) или штрихкоды (бледно-зеленая рамка). Иногда документ содержит отдельные изображения с текстом, который не нужно распознавать (например, весьма распространенные сегодня рекламные объявления). В этом случае удобнее всего вручную отметить такие области красным



↑ ABBYY FineReader 7.0 позволяет пополнить словарь неизвестными словами

цветом, или, если текст объявления вам необходим, а изображение — нет, вы можете самостоятельно разбить такой фрагмент материала на части.

Как это сделать? Перейдите к нужной странице документа (для этого выберите пункт меню **Файл** → **Открыть изображение** или **Файл** → **Сканировать изображение**), после чего выполните команду **Процесс** → **Анализ макета страницы** (также вы можете воспользоваться и комбинацией горячих клавиш **Ctrl + E**). Программа разделит всю область документа на различные блоки, а результат этих действий появится в окне **Изображение**.

Если вам необходимо удалить один из них, щелкните правой кнопкой мыши внутри рамки, выделяющей блок, и затем в появившемся контекстном меню выберите пункт **Удалить блок**.

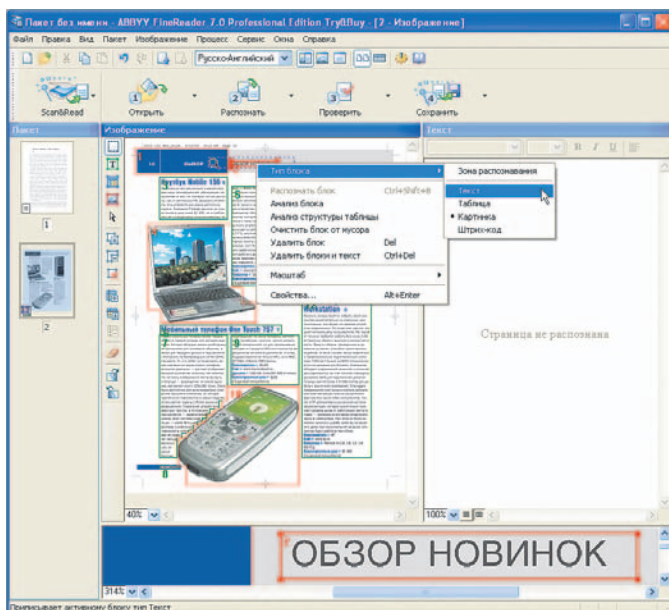
Обратите внимание, что это приведет к удалению только цветного прямоуголь-

ника, но не находящейся в нем информации. Чтобы удалить ее из документа, используется инструмент **Ластик (Изображение)** → **Выбрать инструмент** → **Ластик**: нужно активировать его и выделить лишний фрагмент.

Чтобы изменить границы блоков, выберите **Изображение** → **Выбрать инструмент** → **Выделить зону распознавания**. Если программа неверно распознала тип элемента (например, выделила картинку зеленой рамкой), выберите нужный в пункте **Тип блока** контекстного меню. Для создания новых блоков имеются подходящие инструменты (**Изображение** → **Выбрать инструмент**), с помощью которых вы сможете выделить нужную область. Таким образом, потратив несколько минут на корректировку разбивки документа на различные блоки, можно добиться гораздо более качественного распознавания текста.

Возможные проблемы

Полученный результат	Описание	Устранение недостатка
	Буквы четкие, хорошо различимые. Текст легко читается	Изображение хорошего качества и подходит для распознавания
	Буквы тонкие, «разорванные», некоторые элементы отсутствуют	Отсканируйте документ еще раз, уменьшив значение яркости, либо в градациях серого
	Буквы расплывчатые, нечеткие, склеенные. Текст трудно читается	Отсканируйте документ еще раз с большим значением яркости или в градациях серого



← В контекстном меню блока вы найдете все необходимые для работы с ним инструменты

Резка документа на части

Бывают ситуации, когда за один проход сканируется несколько документов: разворот книги, несколько визитных карточек или мелких бумаг. В таком случае для повышения качества распознавания текста рекомендуется разрезать полученное изображение на части так, чтобы каждая из них содержала только один документ. Конечно, вы можете сделать это и в любом графическом редакторе, например, встроенном **Paint** или популярном **Adobe Photoshop**. Но гораздо удобнее воспользоваться все той же программой **FineReader**, которая предоставляет пользователям такую возможность. Для начала выполните команду **Изображение → Разбить изображение**. В появившемся окне предусмотрена автоматическая разбивка разворота книги (исходное изображение в пакете заменяется двумя новыми, в каждом — по одной странице) и нескольких визитных карточек (в результате для каждой будет создана новая страница). Для этого предназначены кнопки, расположенные в правой части окна: **Разбить на две страницы** и **Разбить на визитки**.

Если программе не удается правильно выполнить автоматическую разбивку, вы можете сделать это вручную. Для добавления горизонтальных и вертикальных разделителей, а также для их удаления предусмотрены кнопки в правой части окна. Нажав на нужную, наведите курсор мыши на изображение документа — на нем появится полоска разделителя. Если около курсора находится крестик — в этом месте можно поставить разделитель, если же вы видите перечеркнутый круг — значит, здесь это сделать нельзя. Например, чтобы разделить книжный разворот на 2 страницы, выберите инструмент **Добавить вертикальный разделитель** и щелкните левой кнопкой мыши в пространстве между страницами. После нажатия на кнопку **ОК** документ будет разбит на два отдельных листа. Можно перемещать установленные разделители с помощью инструмента **Выбор объекта** или убрать их все инструментом **Удалить разделители**.

[illegible]

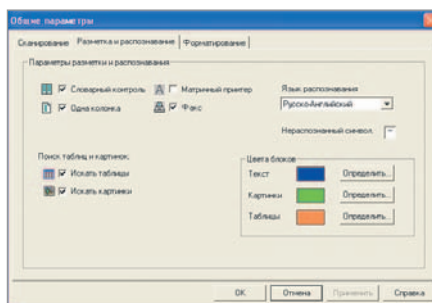
CuneiForm

Версия ► 2000 Professional
Разработчик ► Cognitive Technologies, Ltd.
Сайт ► www.cuneiform.ru
ОС ► Windows 95/98SE/ME/NT/2000/XP
Объем дистрибутива ► 16 Мбайт
Интерфейс ► русскоязычный
Условия распространения ► commercial
Цена ► \$129

Cognitive CuneiForm 2000 Professional — еще одна российская OCR-разработка. Эта программа является одним из основных конкурентов описанной выше **ABBYY FineReader 7.0**. Часто она поставляется в комплекте со сканерами.

Для распознавания в **CuneiForm** также предусмотрен специальный мастер, для начала работы с ним нажмите на кнопку **Мастер распознавания**, расположенную на **Панели инструментов** главного окна. Но в автоматическом режиме программа часто работает некорректно, с большим количеством ошибок, поэтому лучше выполнить все действия самостоятельно, уделив большое внимание настройке работы, — подходящие значения параметров будут зависеть от качества отсканированного документа.

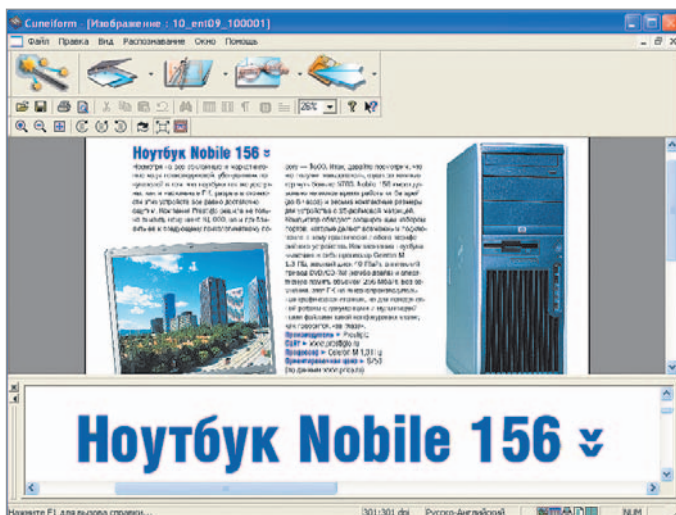
Итак, прежде всего необходимо открыть уже имеющийся графический файл (для этого выберите пункт главного меню **Файл → Открыть**) или отсканировать бумажный документ (**Файл → Сканировать**). После этого с помощью инструмента **Область распознавания** (для работы с ним выберите одноименный пункт меню **Распознавание**) выделяем ту часть документа, которую вам необходимо обрабо-



↑ **Указав подходящие параметры, вы сможете значительно улучшить качество распознавания**

тать. Теперь необходимо выполнить разметку, то есть, аналогично созданию различных блоков в программе **ABBYY FineReader 7.0**, вы должны разделить всю область документа на отдельные элементы: текст, изображения и другие. Для этого предназначен специальный инструмент **Авторазметка** (вы можете воспользоваться одноименным пунктом меню **Распознавание** или кнопкой **Разметка** на **Панели инструментов**). Программа сама выделит элементы. При необходимости корректируем размещение блоков вручную — все необходимые для этого инструменты собраны в меню **Правка**. После этого нажмите на кнопку **Распознавание**.

Проверка полученного текста в программе реализована неважно — удобнее выполнить ее в **MS Word** (**Сервис → Правка описания**) или любом текстовом редакторе. Для этого можно либо сохранить файл (**Файл → Сохранить**), либо экспортировать распознанные данные в указанные приложения **MS Office** (для этого предназначены соответствующие пункты меню **Распознавание**).



← **Интерфейс и инструментальная база Cognitive CuneiForm 2000 Professional значительно проще, чем ABBYY FineReader 7.0**

Программы для распознавания текста позволяют не только значительно упростить работу с бумажными документами, но и сэкономить драгоценное время и силы. Выбор того или иного программного продукта зависит от ваших личных потребностей и предпочтений. Если с необходимостью распознавать текст вы сталкиваетесь часто и собираетесь покупать специальную программу, то лучше отдать предпочтение **ABBYY FineReader 7.0**. Если же такая задача появляется редко, а в комплекте со сканером вы бесплатно получили программу **Cognitive CuneiForm 2000 Professional**, то ее возможностей будет вполне достаточно.

Игорь Буряк

✓ **оптимальный выбор**

HP ScanJet 2400



Как и весь модельный ряд сканеров Hewlett-Packard, HP ScanJet 2400 отличается от конкурентов высокой надежностью и отличным ПО, входящим в комплект. Использование высококачественной электроники вместе с поддержкой интерфейса USB 2.0 гарантирует высокую скорость работы без ущерба качеству сканирования. Если вы не хотите тратить средства на приобретение дополнительных программ для работы со сканером, если вам важно качество отсканированного материала, а каждая секунда на счету, эта модель именно для вас.

Интерфейс ► USB 2.0
Производитель ► Hewlett-Packard
Размер области сканирования ► 210x297 мм (A4)
Оптическое разрешение ► 1200 dpi
Цветопередача ► 48 бит
Поддерживаемые ОС ► Windows 98SE/ME/2000/XP
Слайд-модуль ► нет
Размеры ► 615x458x275 мм
Вес ► 2,2 кг
Слайд-модуль ► commercial
Средняя цена ► \$80