OCR-программы

Фолианты в мегабайты

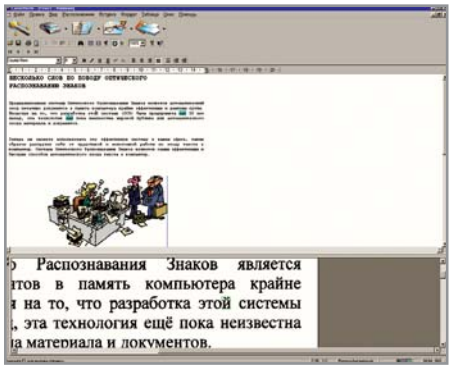
Сканер является одним из самых распространенных периферийных устройств. Это значит, что программы оптического распознавания текста (OCR) весьма популярны и востребованы среди пользователей. Мы вам расскажем обо всех современных разработках в этой области, поддерживающих русский язык.

Чип CHIP CD
Программы

В настоящее время существует пять программ оптического распознавания текста, работающих с кириллицей. Это отечественные программы CuneiForm и FineReader, продукт бельгийского происхождения Read I.R.I.S, а также два пакета от американской компании ScanSoft, являющейся одним из подразделений корпорации Xerox: OmniPage и TextBridge Classic. ScanSoft принадлежали также TextBridge Pro и Recognita OCR, и они тоже прекрасно распознавали кириллицу. Но дальнейшее их развитие прекращено, и, хотя эти программы в момент написания данной статьи продавались и поддерживались, их было решено не рассматривать. К тому же, скорее всего, к моменту выхода этого номера журнала их распространение будет прекращено. В общем-то, ScanSoft поступила логично: функциональность и TextBridge Pro, и Recognita была интегрирована в популярную OmniPage, а поддерживать один продукт намного проще, чем три.

Хотелось бы отметить, что все эти программы, за исключением OmniPage, часто являются модулями в продуктах других разработчиков. Например, Corel OCR&Trace является не чем иным, как CuneiForm. Фирма NewSoft включает в свои пакеты Presto! PageManager и SmartPanel for Scanner облегченную FineReader Sprint. Кроме того, NewSoft под своей маркой Presto! OCR Pro продает все ту же FineReader, хотя и довольно пожилую — 4-й версии. Read I.R.I.S, впрочем, тоже не самой новой версии и здорово урезанная, является составной частью фирменного программного обеспечения сканеров от HP. Но больше всего реинкарнаций у TextBridge Classic. Она известна также под псевдонимом MS Office Document Scanning и входит в состав по крайней мере трех систем управления архивами электронных документов: ScanSoft PaperPort, Avision PaperCom Document Manager и Kodak Imaging Professional.

Так что выбор конкретной программы уже не так прост, как три года назад, когда было всего лишь два продукта и отечественным пользователям работать с ними было весьма удобно и целесообразно (речь идет о FineReader и CuneiForm). Теперь таких приложений стало



▲ CuneiForm — простая в обращении и стабильная в работе OCR-система

уже пять, и у всех есть свои сильные и слабые стороны. Кроме того, каждой программе присуща своя уникальная особенность, отсутствующая у конкурентов. Надеюсь, данная статья поможет читателям сориентироваться в мире этих программ и выбрать оптимальную.

CuneiForm

+ удобный в освоении интерфейс
+ стабильная работа
– плохая автоматическая сегментация
– неудачное распознавание документов низкого качества

Программа поддерживает работу с 15 языками, распознает двуязычные русско-английские тексты. Есть версия с русским интерфейсом. Наряду с Windows поддерживается и платформа Mac. Однако продукт для компьютеров Apple называется по-другому — Tiger.

Главными достоинствами CuneiForm являются простота в освоении и надежность в работе. Любой неподготовленный пользователь может начинать работу с программой практически сразу. Очень хорошее качество сопроводительной документации, которую к тому же можно загрузить с сайта разработчиков в виде PDF-документа.

Зависания CuneiForm крайне редки. Среди пользователей, кстати, меньше всего жалоб на то, что эта программа не работает с тем или иным сканером. Точнее, их просто нет.

В поставку CuneiForm входит полнофункциональный текстовый процессор, с помощью которого более удобно править распознанный текст, чем со встроенными редакторами других программ распознавания. Имеется довольно мощная система контроля правописания. CuneiForm довольно бережно относится к иллюстрациям в документе, в то время как другие программы при их сохранении используют сильную компрессию с потерями, которая самым пагубным образом сказывается на их качестве.

Недостатком CuneiForm по сравнению с FineReader и Recognita OCR является то, что она уверенно распознает лишь документы высокого и среднего качества. Использовать эту программу для работы с машинописным или газетным текстом практически невозможно. Единственное, в чем CuneiForm опережает другие пакеты, так это в обработке цветного текста на цветном фоне. И то, FineReader 5.0 уже догнала в этом CuneiForm, а FineReader 6.0, не говоря о новейшей 7-й версии, уже перегнала. Не предусмотрено распознавание с обучением. Кроме того, CuneiForm существенно отстает о FineReader по количеству поддерживаемых национальных алфавитов, их всего 15. То же самое относится и к количеству поддерживаемых форматов графических файлов. CuneiForm оказалась единственной программой, «не знающей» о существовании PNG и наряду с TextBrige не позволяющей распознавать содержимое PDF-документов. Еще одним недостатком данной программы является неработоспособность в среде эмулятора Wine.

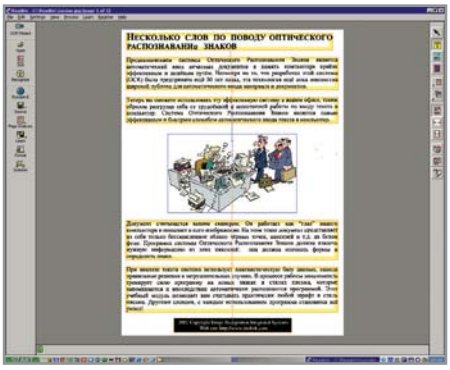
Основные замечания

Автоматическая сегментация не всегда срабатывала верно, приходилось определять блоки вручную (минус 1 балл). Сегментация и распознавание текста со сложным оформлением проходило существенно медленнее, чем с простым или несложным (минус 1 балл).

FineReader

+ великолепное качество распознавания
+ огромное количество поддерживаемых языков
– нестабильная работа в среде Windows 9x
– слишком сильное сжатие иллюстраций при экспорте

Программа очень известная и популярная, причем не только на постсовет- »



▲ I.R.I.S — самая быстрая программа распознавания текста

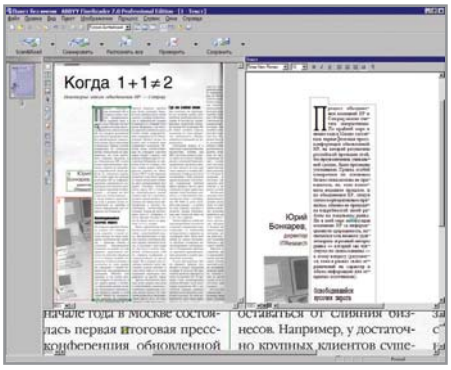
» ском пространстве. Прежде всего ее ценят за то, что она оптимальным образом сочетает простоту и удобство в обращении с великолепным качеством распознавания, особенно если документы низкого полиграфического качества. FineReader лучше всех задает настройки сканирования и сегментации распознаваемого документа, и вмешательство пользователя бывает необходимо довольно редко. Мне, например, приходилось сегментировать вручную лишь иллюстрации, которые представляли собой внешний вид диалоговых окон программ. А вот с определением блоков в тексте с многоколоночной верстkой FineReader почти всегда справлялась самостоятельно. Из плюсов отметим очень гибкую настройку экспорта в различные файловые форматы. Другим достоинством является правильная работа с подстрочными нижними индексами.

Программа позволяет работать с документами на 177 языках. Уникальной особенностью является поддержка «из коробки» практически всех алфавитов России и СНГ, за исключением грузинского. При этом FineReader позволяет распознавать документы с произвольным количеством и сочетанием языков. И только FineReader может правильно распознавать химические формулы. Пока способности в этой области у FineReader весьма ограниченны, но у других OCR-приложений нет и этого. Еще одним достоинством является правильная обработка документов с фоновыми рисунками. Это важно в большей степени для корпоративных пользователей (распознавание реквизитов денежных знаков и ценных бумаг), но может быть полезно и рядовым потребителям, которым приходится много работать с иллюстрированными изданиями. Кстати, если такого рода документов много, это важный аргумент для обновления FineReader до версии 7.0.

Однако имеется много жалоб на неустойчивую работу программы, особенно в среде Windows 9x. Много шума наделала также неработоспособность версий 4 и 5 в Windows XP (хотя они и вышли до нее) без использования специальной программной заплатки или ручного редактирования системного реестра. FineReader производит «недопустимую операцию» чаще, чем любой из конкурирующих пакетов. Очень часто программа отказывается принимать данные со сканера. Подобные проблемы отмечены со многими моделями от Hewlett-Packard, некоторыми сканерами Mustek с USB-подключением и оборудованием от ряда других производителей. В частности, серьезные проблемы испытывали владельцы ручных сканеров. Для решения этих задач надо было загружать и устанавливать программные патчи, причем довольно увесистые. Иногда после инсталляции FineReader требуется переустановка программного обеспечения сканера.

Однако разработчики не стоят на месте. Например, в новейшей версии 7.0 наконец решили проблему снижения качества распознавания в том случае, если образы документов загружаются не со сканера, а в виде файла на диске. Плюс ко всему существенно быстрее стал производиться импорт файлов в формате PDF.

Испытательная версия есть на сайте разработчика (www.abbyy.ru), но по истечении срока бесплатного использования необходимо будет приобрести коробочную. Кроме того, FineReader 7 является единственной программой из обзора, которая требует активации через Интернет. Как мы уже знаем, эта процедура часто создает существенные неудобства для законопослушных пользователей, тем более что пираты ее легко обходят. У более ранних версий проблем с установкой было еще больше: необходимо было запускать инсталлятор со специальной дискеты. В результате FineReader было невозможно поставить на компьютеры без флоппи-дисковода, портативные ноутбуки, где нельзя одновременно использовать флоппи и оптический накопители, а также в среде Wine Unix-подобных операционных систем, например Linux.



▲ FineReader — безусловный лидер в области качества распознавания текста

Но все же именно FineReader является лучшей программой распознавания текста, которую можно применять в наших условиях. Отказываться от ее услуг целесообразно лишь тогда, когда из-за проблем с оборудованием она нестабильно работает и это не удается исправить никоим образом.

Основные замечания

Оценки были снижены лишь за падение качества иллюстраций при экспорте.

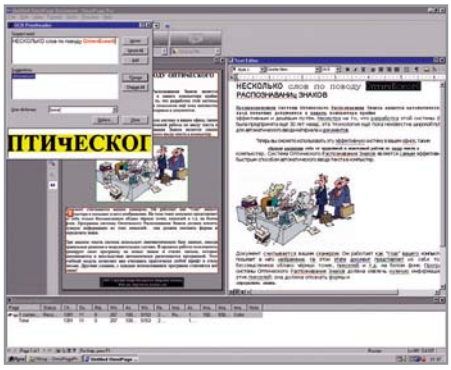
Read I.R.I.S

+ поддержка восточных языков
+ быстрая работа
– отсутствие средств для контроля распознавания
– плохая автоматическая сегментация

Read I.R.I.S является единственной массовой программой распознавания текста, которая позволяет работать с документами на восточных языках с ориентацией текста справа налево или использующих слоговое письмо. Для этого достаточно загрузить специальный модуль с официального сайта.

Программа уверенно распознает документы со сложной верстkой, содержащие таблицы и иллюстрации. Сегментацию документа можно доверить самой программе, а можно и провести вручную. Кроме того, поддерживается распознавание с обучением.

Read I.R.I.S обладает самым высоким быстродействием из всех рассмотренных нами программ. В среднем на обра- »



▲ OmniPage — наиболее популярная OCR-программа на североамериканском рынке

» ботку одной страницы уходит около 40 секунд (у других программ — до двух минут). Read I.R.I.S также очень быстро импортирует PDF-документы.

Имеются средства редактирования изображений: фильтр удаления «мусора», средства увеличения резкости, балансировки яркости, контрастности, гаммы. Это существенно больше, чем у других программ аналогичного назначения.

Есть бесплатная демонстрационная версия с ограниченным сроком использования (www.irisusa.com). Ее можно зарегистрировать и пользоваться как полностью функциональным приложением. К сожалению, модули для работы с кириллицей в поставку испытательной версии не входят, и их нельзя загрузить с официального сайта. Однако эта проблема решаема с помощью перемещения в каталог, куда установлена Read I.R.I.S, файла rus.ytr. Его можно позаимствовать у ближайшего владельца относительно нового сканера HP, в поставку которого входит данная программа.

Недостатков у этого продукта три. Прежде всего, отсутствуют средства контроля распознавания. Для вычитки документа надо загружать текстовый процессор. Автоматическая сегментация работает не очень хорошо, так что лучше ей не доверять документы со сложной версткой. Особенно плохо обстоит дело с текстом в несколько колонок. Единственным выходом в этом случае будет отказаться от полного сохранения документа при его экспорте.

Основные замечания

Полное отсутствие штатных средств контроля распознавания (минус 1 балл). Автоматическая сегментация документов со сложным оформлением, особенно содержащих таблицы, сопровождалась множеством ошибок, приходилось проводить эту процедуру только вручную (минус 2 балла). Запуск процедуры обучения для каждого неуверенно распознанного символа (минус 1 балл). В то же время были подняты оценки за чрезвычайно высокое быстродействие (плюс 1 балл).

OmniPage

+ грамотная работа с иллюстрациями
+ поддержка большого количества языков
– слабая проверка орфографии
– неудобная ручная сегментация

Эта программа является лидером на североамериканском рынке. Во многом продукт близок к отечественному пакету FineReader, в том числе по интерфейсу. Однако OmniPage 11, или версия для Mac, может работать не только в среде классической Mac OS, но и в среде новейшей Mac OS X. Последняя версия OmniPage 12 позволяет распознавать текст на 115 языках с алфавитами на основе всех видов латиницы, кириллицы и греческого. Правильно распознаются многоязычные документы, в том числе и с произвольным сочетанием фрагментов на разных языках.

Важное достоинство OmniPage — бережное отношение к иллюстрациям. Во всяком случае, картинки не подвергаются сильному сжатию, а при преобразовании изображения в монохромное используется псевдосмешение.

Недостатков у этой программы довольно много. Наиболее значимы два: высокая цена и невозможность приоб- »

Важный момент

Как мы тестировали

Тестирование состояло в распознавании документов различного качества и оформления. Полиграфическое качество делилось на следующие градации:

- **Высокое.** Документ напечатан типографским способом или на лазерном принтере, на лощеной или мелованной бумаге. Нет «грязи», границы символов четкие, отсутствуют лигатуры.
- **Среднее.** Документ отпечатан типографским способом на офсетной бумаге или на струйном принтере. Фон неоднороден, контур символов может быть искажен, имеются лигатуры.
- **Низкое.** Документ отпечатан на матричном принтере, пишущей машинке, лазерном принтере или копировальном аппарате с «севшим» картриджем, а также типографским способом на газетной бумаге. Фон неоднороден, имеют место заливки контура символов или, наоборот, контур нечеткий.

Оформление документов также было разделено на три типа:

- **Простое.** Текст размещен в одну колонку, отсутствуют какие-либо текстовые и графические врезки, а также таблицы. Типичный пример — страница в художественной книге небольшого формата.
- **Несложное.** Текст набран в одну колонку, но документ содержит иллюстрации, таблицы, сноски и колонтитулы. Типичный пример — официальная и бизнес-документация на фирменных бланках, научная и учебная литература небольшого формата.
- **Сложное.** Текст сверстан в несколько колонок, содержит таблицы, текстовые и графические врезки, в том числе и непрямоугольной формы. Типичный пример — статья в иллюстрированном журнале или газете, текст научного или справочного издания.

Всего каждая программа тестировалась на девяти типах документов разного уровня качества и оформления. При этом считывание образа документа проводилось как со сканера, так и в виде файла на диске.

Результат распознавания оценивался по 10-балльной системе, от 0 (абсолютно неприемлемый результат) до 9 (программа не допустила ни одной ошибки, и результат распознавания не требовал доводки). Какая-либо обработка образов документов в сторонних пакетах не проводилась, использовались только штатные средства программы приема изображений со сканера и самой OCR-системы.

» ретения полной версии в России, во всяком случае, для Windows. Редакция для Mac OS появляется в продаже довольно оперативно. Хотелось бы также напомнить, что сокращенные версии OmniPage не позволяют распознавать кириллицу. Плюс ко всему в OmniPage используется довольно слабый движок проверки орфографии. Если в слове больше двух ошибок, варианта замены предложено не будет. Неудобна и громоздка процедура ручной сегментации и переопределения блоков. Программа путает переносы и дефисы. Кроме того, OmniPage невозможно установить в среде Wine или с помощью CrossoverOffice.

Основные замечания

Необходима ручная корректировка результатов автоматической сегментации (минус 1 балл). Ошибки в системе контроля результатов распознавания — знаки переносов воспринимаются как дефисы, есть ошибки в словарях, система контроля не предлагает замену в случае, если в слове допущено более двух ошибок (минус 2 балла).

TextBridge Classic

+ невысокая цена
– полная потеря оформления распознаваемого документа
– отсутствие средств редактирования
– нестабильная работа

Недостатки TextBridge Classic общие для всех облегченных систем распознавания текста. Более того, при распознавании полностью теряется оформление документа. Например, совсем игнорируются внедренные иллюстрации, многоколоночная верстка, информация о шрифтах.

▲ Программа TextBridge Classic — самое слабое звено в нашем обзоре

Таблицы эта программа воспринимает как текстовые абзацы с табуляцией. Не предусмотрено средств редактирования распознанного текста. Для этого надо вызывать текстовый процессор. TextBridge Classic не позволяет проводить и ручную сегментацию документа. Плюс ко всему распознавать цветные образы документов просто нельзя, а при обработке документов, отсканированных в полутоновом сером режиме, количество ошибок превышает все разумные пределы. Очень высока чувствительность к неправильно установленной яркости сканирования. Программа способна работать с документами только высокого и среднего качества. Машинописные копии, распечатки, сделанные на матричном принтере, газетный текст, ксерокопии TextBridge Classic распознает очень плохо, во всяком случае, по сравнению с FineReader Sprint. К большому сожалению, программа не может распознавать текст в двуязычных документах. При использовании TextBridge Classic в Windows XP возможен целый ряд проблем. Отмечают, например, аварийное завершение работы при использовании фирменного программного обеспечения сканеров Umax Astra 4500.

Так что этот продукт для очень невзыскательного пользователя. Применять Xerox TextBridge Classic как основную программу распознавания нельзя.

Основные замечания

Невозможность ручной сегментации документа (минус 1 балл). Ограниченные возможности по сохранению оформления документа (минус 1 балл). Полное отсутствие средств контроля результатов распознавания (минус 1 балл).

Итоги

Как видно из нашего обзора, абсолютного лидера среди OCR-программ не наблюдается. У каждой из них имеются как достоинства, так и недостатки. CuneiForm, например, можно смело рекомендовать пользователям, которым не требуется распознавание документов слишком низкого качества, но для которых на первом месте стоит простота в освоении и стабильность в работе. FineReader лучше всего распознает тексты, причем даже самого отвратительного качества, однако иногда огорчает неприятными сюрпризами при установке и работе.

Read I.R.I.S — самая быстрая и невзыскательная к системным ресурсам программа, для нее вполне достаточно 486-го процессора, к тому же это почти единственный выбор для тех, кому приходится работать с текстами на восточных языках, с письменностью справа налево, но у нее, к сожалению, отсутствуют средства контроля распознанного текста.

Две последние программы вряд ли найдут своего пользователя. OmniPage способна отпугнуть слишком высокой ценой, а TextBridge Classic, несмотря на низкую цену, просто не выдерживает критики. ■ ■ ■ **Федосей Сапожников**

Общие сведения	CUNEIFORM_	FineReader OCR	Readiris Pro 8	OmniPage Pro 12 Office	TextBridge
Название	**CuneiForm 2000**	**FineReader 7.0**	**Read I.R.I.S Pro 8.0**	**OmniPage Pro 12.0 Office**	**TextBridge Classic**
Разработчик	Cognitive Technologies	ABBYY Software House	I.R.I.S	ScanSoft	ScanSoft
Сайт разработчика	www.cuneiform.ru	www.abbyy.ru	www.irisusa.com	www.scansoft.com	www.scansoft.com
Цена, $	120 (Professional), 249 (Master)	129 (Professional), 259 (Corporate)	129	599	26
ОС	Windows, Mac OS	Windows, Mac OS	Windows	Windows, Mac OS	Windows
Оценки при тестировании (• — простое; •• — несложное; ••• — сложное)					
Полиграфическое качество (по вертикали) и оформление документов (по горизонтали)					
	• •• •••	• •• •••	• •• •••	• •• •••	• •• •••
Высокое	6 5 4	8 8 7	6 6 2	6 6 4	5 4 0
Среднее	8 8 6	8 8 8	8 7 4	7 7 6	6 5 0
Низкое	9 9 7	9 8 8	8 7 5	7 7 6	7 6 0