Жадаев А. Г.

Сканирование и распознавание текстов

Самоучитель по работе с ABBYY® FineReader 10

ДМК Пресс
издательство
Москва, 2010

УДК 32.973.26-018.2
ББК 004.4
Ж15

Ж15 **Жадаев А. Г.**

Сканирование и распознавание текстов. Самоучитель по работе с ABBYY® FineReader 10. – М.: ДМК Пресс, 2010. – 248 с.: ил.

ISBN 978-5-94074-595-2

Работать с электронными документами во многом удобнее и проще, чем с их бумажными аналогами. Электронный документ можно редактировать, использовать при создании собственных работ, его легко копировать и пересылать по электронной почте. Вместе с тем, многие материалы изначально доступны нам в нередактируемом виде (бумажные или отсканированные документы, цифровые фотографии). Программа ABBYY® FineReader – лучший инструмент для создания электронных копий любых печатных материалов: книг, справочников, журналов, договоров, бланков.

Книга включает описание приемов сканирования и распознавания разных оригиналов – от простых книжных страниц до сложно оформленных документов. А приведенные скриншоты программы позволят читателю быстро освоить интерфейс ABBYY® FineReader и получить практические навыки по работе с программой.

Изложение материала сопровождается практическими примерами. Читатели, которые еще не пробовали самостоятельно переводить печатные материалы в электронный вид, найдут в этой книге простое пошаговое руководство. Для тех же, кто хочет в совершенстве освоить работу с программой, книга откроет многочисленные тонкости настройки для эффективного использования ABBYY® FineReader.

УДК 32.973.26-018.2
ББК 004.4

ISBN 978-5-94074-595-2

Содержание

Глава 1 Текст и графика в компьютере

Любая информация в компьютере хранится и обрабатывается в цифровом виде. Хотя в этой книге повсюду упоминаются слова «текст» и «графика», на самом деле компьютер работает исключительно с цифровой информацией. Любая информация в компьютере хранится в виде файлов, а файл – всего лишь последовательность двоичных чисел (единиц и нулей). С помощью двоичных чисел можно закодировать все, что угодно: от команд, которые должен выполнять сам компьютер, до текста, изображений, музыки и фильмов.

Нужно лишь договориться, что в каждом случае будут означать группы двоичных чисел, из которых состоит файл, как их должны воспринимать и обрабатывать компьютерные программы. Алгоритм (правило), в соответствии с которым данные превращаются в цифры и помещаются в файл, называют *форматом файла*. Разумеется, для разного рода информации требуются разные способы кодирования. Более того, одну и ту же информацию можно представить в цифровом виде различными способами, поэтому форматов было изобретено много.

Например, Объединенная группа экспертов в области фотографии (Joint Photographic Experts Group) разработала набор спецификаций, позволяющих значительно сократить размер файла, в котором хранится изображение. Как вы уже догадались по первым буквам английского названия группы, в результате появился формат JPEG. С файлами этого формата работают и компьютерные программы, и цифровые фотоаппараты, и некоторые бытовые устройства.

В этой книге речь идет о файлах, содержащих данные двух видов. Это *текст* и *графика*.

Для компьютера **текст** – последовательность символов. Символами являются буквы разных алфавитов, цифры, знаки препинания. Пробел, разделяющий слова, – тоже символ. Каждому символу соответствует определенный числовой код.

С таким представлением текста компьютеру легко выполнять самые обычные математические операции. Например, он может найти в тексте все символы с определенным кодом, вставить в указанное место коды введенных с клавиатуры символов или заменить определенную последовательность на другую. На этом основана работа всех программ – текстовых редакторов. Простой пример – программа Блокнот (Notepad), поставляемая в составе ОС Windows. В текстовом редакторе можно выделить часть текста, скопировать ее в буфер обмена, а затем вставить в другое место того же документа или в другой документ; найти в тесте определенные последовательности символов и т. д.

Графика, или рисунок, в компьютерном представлении – набор отдельных точек, о каждой из которых известны ее положение и цвет.

ПРИМЕЧАНИЕ

В компьютерах представление цвета зависит от выбранной цветовой схемы.

Точки, из которых состоит изображение, называют пикселами (pixel, от англ. *PICture'S ELement или PICture Cell* – элемент или клетка изображения). Часто их зовут и просто точками (Dots). Таким образом, изображение является «картой точек», по-английски *Bitmap*. Графический файл содержит информацию обо всех точках изображения.

На рис. 1.1 показан пример рисунка «с точки зрения компьютера». Каждая клеточка изображает одну точку (пиксел). Мы видим, что ширина рисунка со-

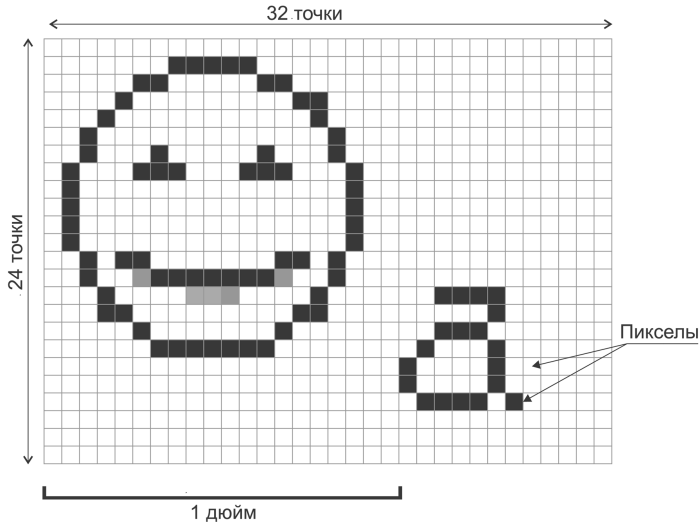

Рис. 1.1 ▾ Пример изображения

ставляет 32 точки, а высота – 24 точки. Белые, или «пустые», точки тоже считаются! В таком случае говорят, что размер этого рисунка – 32г24 точки (пиксела). Вместе с распространением цифровых камер размеры изображения чаще стали выражать общим числом точек. Например, размер этого рисунка – 768 пикселов, или примерно 0,0008 мегапиксела (Мпикс).

На том же примере продемонстрируем еще одну характеристику компьютерного изображения – *разрешение*. Для сравнения под рисунком помещена линейка длиной 1 дюйм (2,54 см). На рисунке ей соответствуют 20 точек. Таким образом, разрешение этого изображения составляет 20 точек на дюйм, или 20 DPI (Dots Per Inch – точек на дюйм).

Возникает вопрос: как же связаны размер изображения и его разрешение? У цифрового изображения есть лишь «истинный размер в пикселах». О разрешении можно говорить, только если мы одновременно уточним, с оригинала какого размера, в сантиметрах или дюймах, было получено (отсканировано, сфотографировано) это изображение, или каков должен быть размер изображения при его выводе на экран. На практике любой графический файл содержит служебную информацию (заголовок), где именно это и указано: размер в пикселах и разрешение в DPI одновременно.

Для работы в программе FineReader изображение обычно получают со сканера или цифровой камеры. При этом в настройках сканера задают разрешение получаемого изображения, например 150 или 300 DPI. Размер изображения будет зависеть от размера сканируемого оригинала. При сканировании с разрешением 300 DPI стандартной книжной страницы получится изображение размером примерно 2400×1600 точек, или, по «фотографической» терминологии, около 4 Мпикс.

Цифровой фотоаппарат выдает снимки определенного размера. Часто цифровые камеры позволяют регулировать размер снимка. Максимально возможный размер является одной из самых важных ее характеристик. В обиходе такой параметр называют «разрешением», хотя на самом деле это именно *размер* выходного изображения в точках.

Нужно различать *размер изображения* и *размер файла*. Первый выражается в количестве точек (пикселов) и характеризует размер самого изображения. Размер (объем) файла показывает, сколько места занимает файл на диске, и измеряется в байтах (килобайтах, мегабайтах). Размер файла зависит не только от размера изображения, но и от формата файла, а также использованных алгоритмов сжатия: если при сохранении изображения применяется сжатие информации, то файл получается меньше.

Форматы файлов

Сложилось так, что разработчики программ регулярно создавали и предлагали новые форматы файлов. В настоящее время существуют сотни различных форматов, каждый из которых обладает своими особенностями, достоинствами и, возможно, недостатками по сравнению с другими. Тип файла (документа) – бо-

лее общая характеристика содержимого файла, например «текст», «изображение», «звук», «архивы» и т. д. Так, к графическим форматам, в которых компьютер работает с изображениями, относятся форматы **BMP**, **TIFF**, **GIF**, **JPEG** и многие другие.

Любая программа способна открывать, обрабатывать и сохранять файлы только определенных типов и форматов. Некоторые программы поддерживают лишь один-два формата, но чаще приложения работают с целым рядом форматов. Рассмотрим некоторые типы и форматы файлов (документов), которые могут нам встретиться при работе с программой FineReader. Заметим, что от версии к версии набор поддерживаемых форматов, как графических, так и текстовых, может изменяться.

Текстовые форматы

Самым старым и простым типом файлов является текстовый. По сути, текстовый файл представляет собой только последовательность символов (букв, цифр, знаков препинания и математических символов), отображаемых на экране или выводимых на печать. Помимо них, в текстовом файле могут содержаться «символ перевода строки», «символ разрыва строки», «символ табуляции» и «символ конца страницы», которые не отображаются на экране и не распечатываются на бумаге. Это так называемые «непечатаемые символы».

Текстовый файл не содержит информации ни о размере или начертании шрифта, ни о выравнивании строк или отступе первой строки абзаца. То, как отображается текст на экране, зависит только от настроек программы, в которой открыт этот файл.

Текстовые файлы имеют расширение **TXT**. Работу с такими файлами должны поддерживать все программы, которые способны так или иначе обрабатывать текстовую информацию. На основе текстового файла было создано еще несколько форматов для более узких применений.

Файлы формата **CSV** (*Comma Separated Values* – значения, разделенные запятыми) являются такими же текстовыми файлами, как файлы формата txt, но специально предназначены для хранения данных в виде списков или таблиц. Каждая строка считается строкой таблицы, а запятые разделяют текст на части, которые должны быть помещены в отдельные ячейки. Когда вы открываете такой файл обычным текстовым редактором, например Блокнотом, то видите обычный текст. Если же открыть файл **CSV** программой для работы с электронными таблицами, например Microsoft Excel, то такая программа воспримет все запятые как разделители ячеек и автоматически разместит данные в ячейки таблицы. Иногда для разделения значений используется точка с запятой. При открытии или сохранении файла программы предлагают уточнить, какой символ следует считать разделителем.

В формате **HTML** (*HyperText Markup Language* – язык разметки гипертекста) создаются файлы, которые служат основой веб-страниц и предназначены для просмотра с помощью программ-браузеров. Эти файлы имеют расширение **HTML** или **HTM**. По существу, это тоже текстовые файлы, но в тексте есть спе-