



## Как распознать... документ

*За последние 10-15 лет сменился по крайней мере десяток поколений персональных компьютеров. Их вычислительная мощность увеличилась в десятки тысяч раз. Цветные принтеры, плоские экраны, видеокамеры, лазерные диски... Эти обыденные вещи еще вчера относилось к разряду фантастики. Такой прогресс не может оставить равнодушным. Но есть одна область информационных технологий, которая за эти годы не претерпела столь стремительных перемен. Это – ввод данных, одна из ключевых проблем взаимодействия с компьютером. Сегодня, как и вчера, работа с клавиатурой остается основным, а иногда – и единственным способом формирования баз данных, документов и электронных таблиц. Увы... Но тот факт, что пальцы являются основным инструментом общения с ПК, еще долгое время будет оставаться актуальным. Решить эту проблему кардинально нам не под силу. Но кое-что сделать можно. Например, освоить альтернативные способы ввода информации в компьютер. В первую очередь – через сканирование изображений. Как правильно делать такую операцию, как задействовать для нее самые доступные средства, как получить приемлемый результат пойдет речь в этой статье.*

*Николай КАРПЕНКО, канд. техн. наук, заместитель главного редактора*

---

Вначале разберемся в двух моментах: для чего бухгалтеру сканировать и распознавать документы и какова бухгалтерская специфика этого процесса? Сканировать документы нужно для того, чтобы затем ввести их в компьютер. Вот несколько иллюстраций, когда это полезно. Заглянув с отчетом к вашему инспектору, вы увидели важное объявление. Что делать? Переписывать его от руки? Не мешало бы придумать что-нибудь более эффективное... Вам нужно срочно скопировать бланк, но ксерокс, как назло, недоступен. На глаза попалась животрепещущая заметка, инструктивное письмо, мало ли что еще? В любом из этих случаев материал проще всего сфотографировать, а потом разбираться, что с ним делать дальше.

А дальше есть два варианта. Первый – материал в виде фотокопии вы прочитали и больше он вам не нужен. В этом случае ваши дальнейшие действия – вытереть файл. Второй вариант: с материалом нужно работать. Например, откорректировать его в Word или в Excel, создать электронную форму и заполнить ее формулами. Вот для этого уже придется РАСПОЗНАТЬ графическое изображение, то есть превратить обезличенный набор точек в документ вполне конкретной структуры и содержания.

Так что же получит бухгалтер, уверенно овладев секретами сканирования и распознавания документов? СВОБОДУ. Свободу от необходимости куда-то бежать, что-то «ксерить», выстаивать нервную очередь у копировального агрегата... Все это останется в прошлой жизни. Одно нажатие кнопки – и вы забыли о проблеме, забрав копию документа с собой.

Относительно бухгалтерской специфики. Тексты приходится распознавать всем. А вот работа с табличным материалом – определенно бухгалтерский «конек». Поэтому ей нужно уделить повышенное внимание, работая над проблемой распознавания изображений. Есть и другие особенности. Потребность отсканировать листок-другой у бухгалтера обычно возникает не в тепличных условиях хорошо оснащенного офиса. Это может быть приемная, доска объявлений, кабинет инспектора – что угодно. А под рукой у бухгалтера будет не цветной сканер, а портативный фотоаппарат или камера в мобильном телефоне. Вот из этого и нужно исходить: скорее всего, отсканированное изображение будет нуждаться в дополнительной обработке. Поэтому одна из наших задач – научиться обрабатывать изображения до уровня, пригодного к распознаванию. Причем делать это с минимальными усилиями.

И последнее. Бухгалтер – не фотограф. И не дизайнер. Навыки владения фотокамерой ему не присущи, да и ни к чему. А вот прикладной аспект этого вопроса будет вовсе не лишним. Безусловно, премудрости фото процесса и обработки снимка ему не нужны. Но знать несколько конкретных рекомендаций, чтобы избежать массы ошибок при пересъемке документов, нужно. От этого зависит успешность всего процесса. От этого зависит время, которое затем придется потратить на обработку.

Поэтому одна из наших задач: научиться при помощи доступных средств получать максимальное качество отсканированных изображений.

Материал мы выстроим так: вначале разберемся, что такое оптическое распознавание, рассмотрим программы, которыми изображение можно распознать и как это сделать практически. И только затем проработаем рекомендации, как правильно перенести изображение и как быстро его обработать и привести к форме, пригодной для распознавания. Итак, обо всем по порядку.

## Что такое сканирование и распознавание

Сканирование – это процесс получения графической копии оригинального изображения. Отсканировать документ можно по-разному. Например, при помощи сканера или фотоаппарата. И в том и в другом случае результатом будет растровый графический файл. Этот файл может быть цветным или черно-белым, обладать различным разрешением, может быть представлен в разных форматах (чаще всего «\*.jpg» или «\*.tif»). Подробно о характеристиках графических файлов мы писали на страницах «Б&К», № 10,11 за 2007 г. Детали нам сейчас не важны. Важно понимать главное: растровый файл – это набор точек. То, что на экране или печати он выглядит как текст, еще ни о чем не говорит. Обработать этот «текст» программой Word или Excel не получится. Все, что вы сможете с ним сделать – вставить в виде готового объекта. Откорректировать растровый файл можно в специальной программе, – графическом редакторе. При этом работать с ним можно как с рисунком, а не текстовым документом.

**Важно!** Хотелось бы лишний раз напомнить разницу между текстом и отсканированным документом. Представьте, что вы набираете документ в Word, печатая его на клавиатуре. Каждый раз, нажимая клавишу, в компьютер «попадает» ее код. Собственно этот код и записан в «вордовском» документе. В данном случае для каждой буквы документа «известно» ее происхождение. Например, первый символ этого абзаца имеет код «0194», это буква «В», второй – код «0224», это символ «а» и т.д. При печати или выводе на экран выполняется ВИЗУАЛИЗАЦИЯ документа. То есть по содержимому, по кодам символов Word формирует их графическое изображение. Попросту говоря, «прорисовывает» буквы. Когда вы корректируете документ в Word, вы меняете коды символов в документе. Поэтому одно нажатие клавиши способно изменить один символ. В графическом изображении этого нет. Сканер ничего не печатает. Он вводит точки изображения. Расположившись в некотором порядке, эти точки выглядят как текст, и не более того. Тот факт, что какая-то часть этого изображения – это символ «В», какая-то часть – символ «а» компьютеру НЕИЗВЕСТНА. Для него эти символы – всего лишь выстроенный определенным образом набор пикселей. Поэтому и отредактировать эти точки программой Word нельзя.

Для того, чтобы превратить графический файл в «обычный» документ, его нужно РАСПОЗНАТЬ. Этим занимаются специальные программы оптического распознавания образов (OCR). Одной из самых распространенных программ этого класса является FineReader. Несмотря на отличия в деталях, цель этих программ одна. Они анализируют изображение, пытаются обнаружить фрагменты со «смысловым» содержанием. Это могут быть символы текста, таблицы, рисунки. Иными словами, программа распознавания пытается выяснить СТРУКТУРУ документа: найти буквы, слова, предложения, ориентируясь на специфику их внешнего вида. Затем она формирует текстовый документ, проверяет его по системе словарей, грамматических правил и предлагает пользователю для окончательного контроля. Когда проверка завершена, результат работы программы распознавания можно сохранить в формат текстового процессора Word или электронной таблицы Excel. И только после этого вы получите «нормальный» документ, с которым можете делать все, что угодно!

Конечно, на практике не все так гладко. Алгоритм распознавания – процесс сложный, некачественное изображение программа распознавания обработает с ошибками. Чтобы минимизировать их количество, изображение придется обработать в графическом редакторе. Но об этом чуть позже.

## Какие изображения можно распознать

Чтобы процесс распознавания был успешным, изображение должно соответствовать определенным требованиям, а именно.

- Графический файл должен иметь разрешение порядка 300 dpi. Кстати, это «стандартное» разрешение для цифрового фотоаппарата.

- Линейный размер изображения при этом разрешении должен быть сопоставим с оригиналом. Маловероятно, что вам удастся распознать страницу документа с разрешением 300 dpi размером с почтовую марку.

- Для распознавания текста изображение может быть черно-белым или в оттенках серого. Преимуществ работы с цветными изображениями в данном случае нет.

Это, так сказать, технические требования к рисункам для распознавания. Разумеется, есть и другие не менее важные условия для успешного выполнения этой работы. Например, изображение должно быть достаточно контрастным, буквы – с четко очерченными краями. Любой «мусор» на отсканированном изображении в виде пометок карандашом, пятен, дефектов бумаги программа распознавания попытается идентифицировать как символы. Ничего хорошего из этого не получится. По большому счету, устранение подобных дефектов – не наша задача. Мы должны постараться сделать так, чтобы их было как можно меньше.

## Распознаем документ

Вначале научимся распознавать хорошие изображения, которые НЕ ТРЕБУЮТ обработки. Такие изображения могут быть получены со сканера, или копированием фрагментов экрана. В качестве

примера для распознавания возьмем фрагмент текста с таблицей, который показан на рис. 1. Для распознавания применим программу «FineReader 8.0» компании АБВУУ (<http://www.abbyy.ua/>), скачав ознакомительную версию с <http://www.abbyy.ru/Download/>.

Окно программы показано на рис. 2. В нем представлено меню, панели инструментов. Ниже панели инструментов находятся пять крупных иконок для управления

### Какие изображения можно распознать

Чтобы процесс распознавания был успешным, графическое изображение должно соответствовать определенным требованиям, а именно.

Параметр	Значение	Комментарий
Разрешение	300 dpi	
Формат	«Jpg», «Tif», «Wmp»	Практически любой растровый формат
Тип	Цветное, оттенки серого, черно-белое	

Рис. 1

программой: «Scan&Read», «Открыть», «Распознать», «Проверить», «Microsoft Word». Основная часть окна разбита на четыре области. Перечислим их слева направо и сверху вниз. В левой части окна находится область страниц документа. Каждая страница – это отдельное изображение. Справа от области страниц расположено окошко «Изображение». В нем виден фрагмент отсканированного документа, а также его РАЗМЕТКА (о ней поговорим чуть позже). В правой части окна показан результат распознавания. Это – текстовый документ, его можно править прямо в программе FineReader. В нижней части окна расположена область с увеличенной копией изображения. Между областью страниц и окошком «Изображение» находится рабочее меню для разметки рисунка перед его распознаванием.

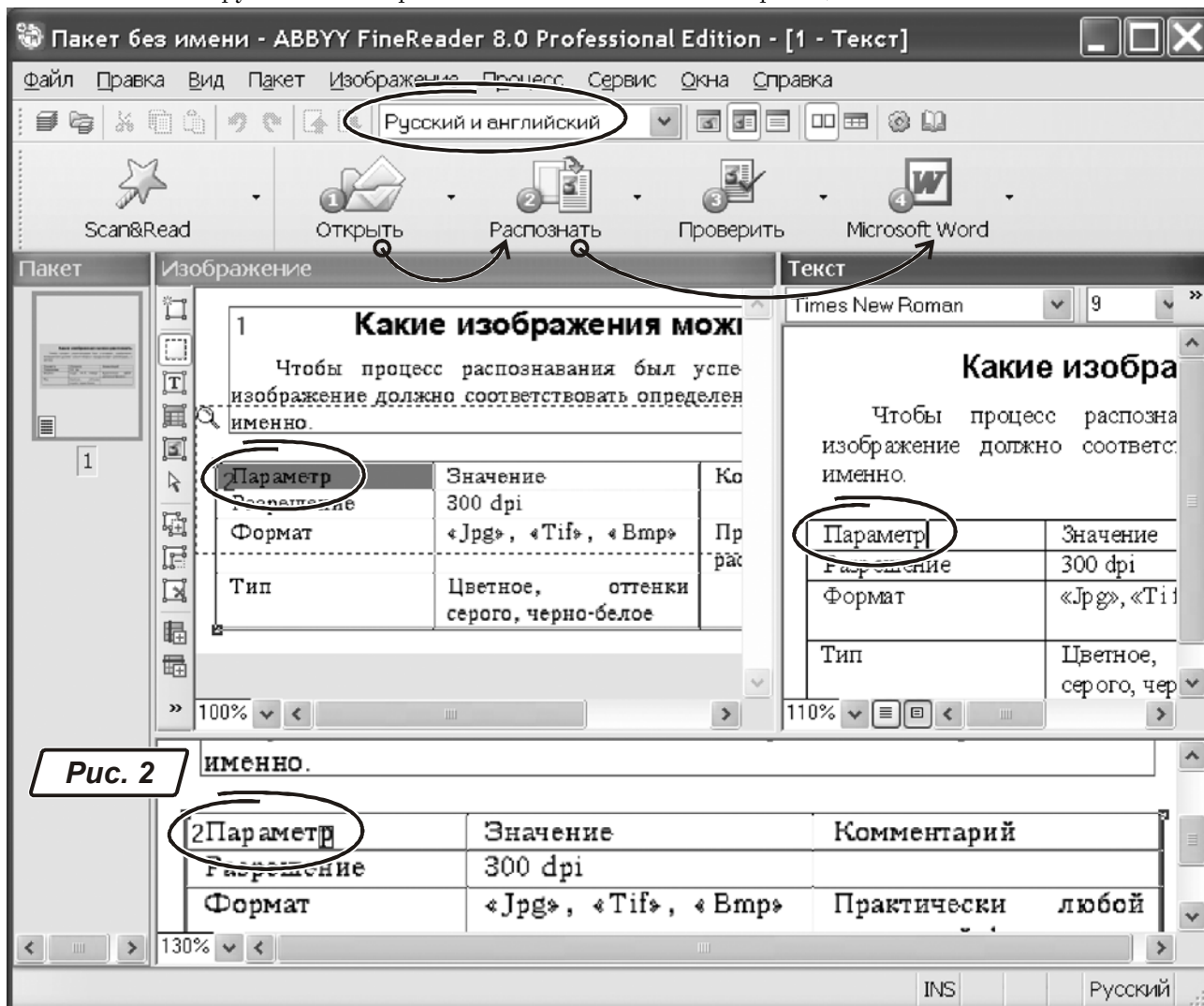
## ПОСЛЕДОВАТЕЛЬНОСТЬ ДЕЙСТВИЙ ПО РАСПОЗНАВАНИЮ ДОКУМЕНТА

Чтобы распознать отсканированный документ, нужно выполнить такие шаги.

1. Загрузить одно или несколько изображений в FineReader.
2. Подкорректировать изображение, чтобы очистить его от «мусора», устранить искажение строк и т.д. Этот этап нужен не всегда.
3. Разметить изображение, то есть указать, КАК FineReader должен интерпретировать отдельные участки рисунка. Обычно такая операция нужна при сложном форматировании документа. Для простых случаев операцию разметки FineReader прекрасно выполняет в автоматическом режиме.
4. Распознать документ.
5. Просмотреть результат работы. Если FineReader ошибся, можно подкорректировать документ или поправить разметку и повторить распознавание.
6. Импортировать результат распознавания во внешнее приложение, например в Word или Excel.

## ЗАГРУЗКА ИЗОБРАЖЕНИЯ

В нашем случае данные для распознавания находятся на жестком диске в виде графических файлов. Нажимаем кнопку «Открыть» (на рис. 2 она изображена с номером «1»). Появится окно Проводника. В нем находим папку, выделяем нужные файлы и щелкаем на кнопке «Открыть» или нажимаем «Enter». Загруженные изображения появятся в области страниц.



**Совет** Для быстрой загрузки изображений пользуйтесь комбинацией «Ctrl+O» (английская буква «O»!). Также можете воспользоваться меню «Файл → Открыть».

Кнопка «Scan&Read» служит для начала процесса сканирования документа. Если вы пользуетесь сканером, включите его, положите документ и нажмите «Scan&Read». Дальнейшая работа зависит от модели сканера и подробно описана в инструкции к нему. Через эту кнопку можно также и загрузить документ. Щелкните на маленьком треугольнике в левой части иконки «Scan&Read». Раскроется список возможных действий. Среди них будет и операция «Открыть и распознать». Точно так, раскрыв список действий кнопки «Открыть», вы увидите строку «Сканировать изображение... Ctrl+K». Таким образом, кнопки «Scan&Read» и «Открыть» во многом взаимозаменяют друг друга.

## АВТОКОРРЕКЦИЯ ИЗОБРАЖЕНИЯ

Операции корректировки изображения сосредоточены в меню «Изображение». Разберем те возможности, которые приходится использовать чаще всего.

**ПОВОРОТ ИЗОБРАЖЕНИЯ.** Текст при распознавании должен быть сориентирован «сверху вниз» и следовать «слева направо». Если страница документа отсканирована с поворотом, ее придется повернуть. Для этого войдите в «Изображение → Повернуть/Отразить изображение». Появится меню из пяти пунктов: «Повернуть по часовой стрелке», «Повернуть против часовой стрелки», «Повернуть на 180°», «Зеркально отразить относительно вертикали», «Зеркально отразить относительно горизонтали». На практике обычно хватает первых двух возможностей.

**КОРРЕКТИРОВКА РАЗРЕШЕНИЯ.** Если изображение отсканировано с «неправильным» разрешением, обратитесь к меню «Изображение → Исправить разрешение» или нажмите «Ctrl+Shift+T». Появится специальное окно с двумя параметрами. Каждый из них – это многопозиционный переключатель. Первый параметр называется «Выберите разрешение изображения». Он может принимать одно из четырех значений: «Отсканированное изображение (300dpi)», «Изображение, полученное по факсу (200dpi)», «Скриншот (120dpi)», «Другое разрешение». Выбираем первый вариант. Второй параметр называется «Исправить разрешение». Он определяет область действия инструмента и принимает одно из двух значений: «Выбранных изображений» или «Всех изображений пакета». Выбираем его значение по ситуации, нажимаем «Ok». Программа FineReader откорректирует разрешение выбранных изображений.

**ОЧИСТКА ОТ МУСОРА.** Теперь попробуем улучшить качество картинки. Сразу заметим: серьезные дефекты изображения FineReader не исправит. Он может слегка «почистить» изображение, удалив мелкие детали в виде соринок, пятнышек на бумаге и т.п. Причем эту операцию он сделает интеллектуально: текст останется в целостности и сохранности! Чтобы очистить изображение от мусора нужно:

- выделить одну или несколько страниц документа;

- обратиться к меню «Изображение → Обработать изображение → Очистить изображение от мусора».

**«ВЫРАВНИВАНИЕ» СТРОК.** Один из распространенных дефектов сканирования – искажение строк текста. Вот пример, когда он возникает. При сканировании книги края возле ее переплета были недостаточно ровными. Кстати, типичная ситуация. В этом случае строки текста на отсканированном изображении будут расположены не строго по прямой, а слегка «скруглено». Особенно заметен этот дефект будет по краям. Исправить его может операция устранения искажения строк. Правда, успешно справится с этой задачей она не всегда!

Чтобы выровнять строки в отсканированном изображении сделайте так.

- Выделите одну или несколько страниц документа.

- Обратитесь к меню «Изображение → Обработать изображение → Устранить искажение строк».

**Совет** Пользуйтесь инструментом выравнивания строк осторожно. В некоторых случаях он может отработать неправильно. Поэтому не применяйте выравнивание сразу ко всем страницам документа. Обработывайте их по одной, каждый раз просматривая полученный результат.

Этими возможностями мы пока и ограничимся.

## РАЗМЕТКА ДОКУМЕНТА

Информацию в электронном документе можно условно разделить на три категории: текст, таблица и рисунок. При распознавании программе FineReader желательно «знать», с чем она имеет дело. В принципе, анализируя страницу, выявить ее структуру FineReader может и самостоятельно. Но при сложном форматировании возможны ошибки. Например, – если в документ вставлено изображение окна Windows, которое содержит ТЕКСТОВОЕ меню. FineReader наверняка попытается распознать строки этого меню как обычный текст. Хотя по смыслу ничего такого делать не нужно.

**Совет** В простых случаях опирайтесь на автоматическую разметку текста. Это сэкономит время и силы. Если форматирование идет в несколько колонок, перемежается рисунками, таблицами, делайте разметку сами. Исправлять ошибки FineReader-а получится дольше. Поэтому есть смысл освоить «ручную» разметку документа. Тем более, что сделать это совсем не сложно.

Посмотрите на панель инструментов для разметки документа, она находится слева от области страниц. Начнем с иконки с буквой «Т»: это инструмент для определения ТЕКСТОВОГО блока. Щелкните на иконке левой кнопкой мыши. Поставьте указатель в том месте на ИЗОБРАЖЕНИИ

документа, где должен начинаться фрагмент текста. Удерживая нажатой левую кнопку мыши, обведите текстовый блок. Отпустите левую кнопку мыши. На документе появилась прямоугольная область зеленого цвета. Она должна ОХВАТЫВАТЬ фрагмент изображения, который FineReader теперь будет рассматривать как текст. На рис. 2 текстовый блок охватывает всю часть документа, за исключением таблицы. Он помечен номером 1. Чтобы изменить размер блока выделите его и растяните за угол или границу так, как вы меняете размер окна Windows.

Под иконкой создания текстового блока расположены инструменты для блоков «Таблица» и «Рисунок». Работа с ними такая же, как и создание текстового блока. Отличие состоит в алгоритме обработки соответствующих участков изображения программой FineReader.

На изображении можно создать несколько блоков разного типа. Чтобы выделить тот или иной блок на панели разметки есть специальный ИНСТРУМЕНТ ВЫБОРА, его иконка выглядит в виде «стрелки». Щелкаем на иконке левой кнопкой мыши, щелкаем стрелкой на блоке и он становится активным. Активный блок можно переместить, можно изменить его размеры или удалить. Через контекстное меню правой кнопки мыши ТИП активного блока можно изменить. Под «стрелкой» расположены две иконки, на них видны знаки «+» и «-». Их используют соответственно для добавления к выделенному блоку нового участка или для удаления фрагмента из блока. Пользоваться этими средствами очень просто. Выбираем инструмент, выделяем блок так, чтобы он пересекался с уже имеющимся. В зависимости от нашего выбора новое выделение будет добавлено к имеющемуся блоку или удалено из него. Две последние (самые нижние) иконки на панели разметки нужны для работы с таблицами. С их помощью можно создавать и удалять вертикальные и горизонтальные линии разметки в табличном блоке. Но об этом чуть позже.

Чтобы УДАЛИТЬ блок разметки нужно выделить его и нажать клавишу «Del».

Относительно областей типа «Рисунок», здесь все понятно: эти участки изображения FineReader распознавать не будет. Он перенесет их в результирующий документ «как есть». А вот с таблицами, основным бухгалтерским объектом, имеет смысл познакомиться подробнее. Рекомендуем придерживаться такой последовательности действий при распознавании таблиц.

- Выбираем инструмент определения блока типа «Таблица».

- Аккуратно обводим фрагмент изображения, где располагается таблица. Лишнее пространство при этом лучше не захватывать, но внешняя граница таблицы должна полностью находиться внутри блока.

- Щелкаем внутри созданного блока правой кнопкой мыши. Из контекстного меню (рис. 3) выбираем «Анализ структуры таблицы».

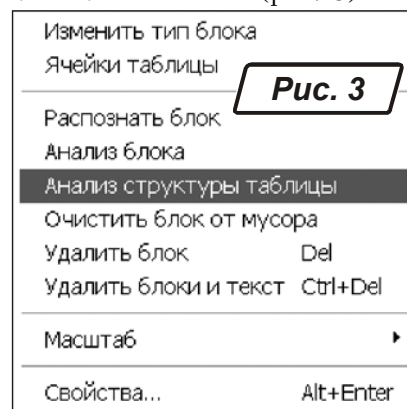
- Оцениваем результат разметки. Внимательно смотрим, как FineReader определил отдельные ячейки. Для этого есть смысл последовательно щелкать левой кнопкой мыши по ячейкам таблицы и анализировать правильность их расположения.

- Если FineReader все сделал правильно, переходим к следующему этапу работы.

А что делать в том случае, если при разметке таблицы обнаружились ошибки? Например, FineReader не распознал бледную линию границы и объединил несколько ячеек в одну? Для этого на панели разметки блоков есть инструменты: «Добавить вертикальную линию», «Добавить горизонтальную линию» и «Удалить линии». Пользоваться ими очень просто. Щелкните на иконке «Добавить вертикальную линию». Переведите указатель мыши в окошко «Изображение» на таблицы. Вместе с указателем по изображению будет перемещаться предполагаемая линия разметки. Поставьте ее на нужное место и щелкните левой кнопкой мыши. Таким же способом добавьте все отсутствующие вертикальные линии в разметке таблицы. Чтобы закончить работу в режиме добавления линий выберите другой инструмент разметки.

По такому же алгоритму разберитесь с горизонтальными линиями. Инструмент «Удалить линии» поможет стереть лишние линии разметки таблицы.

При работе с таблицами часто приходится объединять или разделять ячейки таблицы. Сделать это проще всего через правую кнопку мыши. Поставьте указатель на ячейку таблицы. Щелкните правой кнопкой мыши. Из контекстного меню (рис. 3) выберите пункт «Ячейки таблицы». В нем есть три раздела: «Разбить ячейки», «Объединить ячейки», «Объединить строки». Если FineReader ошибочно слил несколько ячеек в одну, воспользуйтесь возможностью «Разбить ячейки». Если потребу-



ется объединение ячеек в одно целое, выделите их и обратитесь к пункту «Объединить ячейки» контекстного меню.

## РАСПОЗНАЕМ ИЗОБРАЖЕНИЕ

Когда разметка завершена, нажмите «Ctrl+Shift+R» – распознать все страницы. Это действие можно выполнить через меню «Процесс → Распознать → Распознать все» или через список доступных действий кнопки «Распознать».

**Важно!** Проследите, какой ЯЗЫК установлен для распознавания. Выбрать язык можно прямо на панели инструментов FineReader. Например, на рис. 2 установлены «Русский и английский». Щелчком мыши раскройте список. Укажите вариант, подходящий для вашего случая.

Результат распознавания можно посмотреть в окошке «Текст» (рис. 2). Внимательно изучите его. Вначале убедитесь, что FineReader правильно обработал структуру документа. Если это не так, откорректируйте разметку и заново распознайте страницу.

**Совет** При повторном распознавании не обязательно заново обрабатывать весь документ. Вы можете ограничиться текущей или несколькими выделенными страницами. Выделите страницу и щелкните на кнопке «Распознать» или нажмите «Ctrl+R».

## СОХРАНЯЕМ РЕЗУЛЬТАТ

Работа над большими документами может потребовать много времени. В этом случае ее выполнят в несколько этапов. Чтобы сохранить результат работы на диск в программе FineReader используется понятие «пакет». Пакет объединяет исходные изображения, их разметку, результаты обработки в том состоянии, которое было на момент его создания. Сохранить пакет можно через меню «Файл → Сохранить пакет как...». Появится окно, идентичное окну Проводника. Выбираем папку, указываем имя пакета и результат записан.

**Важно!** Пакет объединяет много файлов. Чтобы не вносить сумятицу, не сохраняйте пакет в папку, где уже есть какие-то данные. Лучше сохранить пакет в отдельном каталоге.

## ИМПОРТИРУЕМ РЕЗУЛЬТАТ

Импорт нужен для сохранения результатов во внешнем приложении или буфере обмена. Чтобы сохранить текущую страницу в Word щелкните на последней иконке с номером «4» и надписью «Microsoft Word» (рис. 2). Откроется окно Word с распознанным документом.

Это – не единственный способ импортирования данных. Щелкните на значке раскрытия списка кнопки «Microsoft Word». Появятся такие варианты: «Мастер сохранения результатов», «Сохранить страницы...», «Передать страницы в...», «Передать все страницы в...», «Опции».

Через «Сохранить страницы...» можно записать результат распознавания на диск. При записи есть возможность выбрать все или текущую страницу, а также определить формат файла для сохранения (в нашей редакции FineReader мы насчитали 17 вариантов...). Как правило, достаточно пунктов «Передать страницы в...» и «Передать все страницы в...». Эти две возможности отличаются только тем, что в первом случае импортируется текущая или несколько выделенных страниц. Во втором случае FineReader передаст во внешнее приложение все страницы документа. Оба меню содержат несколько разделов. Их количество зависит от состава установленных программ на вашем компьютере. Можно смело ограничить свой выбор тремя: «Microsoft Word», «Microsoft Excel» и «Буфер обмена». Выберите любой из указанных пунктов. FineReader вызовет приложение и передаст в него результат распознавания.

## ЕЩЕ РАЗ О СЛОЖНЫХ ТАБЛИЦАХ

Таблицы неоднородной структуры в практике бухгалтера – вещь обыденная. Поэтому, завершая тему распознавания документов, хотелось бы дать совет, как с ними работать.

**Совет** Прежде чем распознавать сложную таблицу постарайтесь увидеть ее «базовую структуру». Найдите тот «оптимальный» набор строк и колонок, который впоследствии минимизирует

затраты на разделение и объединение ячеек. Когда структура прояснится, выполните ручную разметку таблицы.

Выбор структуры очень важен. Не спешите начинать обработку документа, пока не продумаете этот вопрос. Взгляните для примера на бланк налоговой накладной, изображенный на рис.4. Попробуйте распознать документ «в лоб», в автоматическом режиме. Скорее всего, вы получите кашу из текстовых полей, фрагментов таблиц и т.п. Привести в порядок такую абракадабру будет сложнее, чем набрать бланк «с нуля». В то же время проблема решается просто. На наш взгляд, большинство фрагментов документа укладывается в 11 колонок. Они пронумерованы и показаны в верхней части рисунка. Мы поступили так. При разметке создали блок типа «Таблица». Инструментами «Добавить вертикальную линию», «Добавить горизонтальную линию» разделили его на ячейки. В предположении, что в таблице будет 11 колонок. Причем эту операцию сделали вручную! После этого просмотрели таблицу, объединяя отдельные ячейки заголовка в группы. Результат распознавания такого сложного документа был великолепен. Напомним, что объединять и разделять ячейки удобно через правую кнопку мыши (пункты «Разбить ячейки», «Объединить ячейки»). Учтите, если вы удачно выбрали структуру для ручной разметки, самые сложные таблицы вам не страшны.

1	2	3	4	5	6	7	8	9	10	11	
Оригінал (видається покупцю)				ЗАТВЕРДЖЕНО							
Копія (залишається у продавця)				наказом ДПА України							
(Непотрібне виділити поміткою "X")				від 30.05.97 р. N 163(у редакції наказу ДПА України							
				від 08.10.98 р. N 469)							
<b>ПОДАТКОВА НАКЛАДНА</b>											
Дата виписки податкової накладної				05.05.07			Порядковий номер		505		
				Продавець					Покупець		

Рис. 4

Розділ	Дата відвантаження (виконання, поставки, послуги) товарів (послуг) продавця	Номенклатура поставки товарів (послуг) продавця	Одиниця виміру товару	Кількість (об'єм, обсяг)	Ціна поставки одиниці продукції без урахування ПДВ	Обсяги поставки (база оподаткування) без урахування ПДВ, що підлягають оподаткуванню за ставками				Загальна сума коштів, що підлягає оплаті
						20 %	0 % (реалізація на митній території України)	0 % (експорт)	Звільнення від ПДВ	
1	2	3	4	5	6	7	8	9	10	11
I	05.05.07	Маркер	шт.	10	1,5	15,00				
		Ручка	шт.	9	1,8	16,20				
		Ластик	шт.	4	0,65	2,60				
		Усього по розділу I					33,80			33,80
II	Товаротранспортні									1,80

## Как получить изображение

### НАКЛОН И ОСВЕЩЕНИЕ

Если вы пользуетесь сканером, смело пропустите этот раздел или прочитайте его «по диагонали». Материал этот для тех, кто собирается «сканировать» документы цифровым фотоаппаратом или камерой мобильного телефона. Вот несколько простых советов по организации процесса.

- Ахиллесова пята любого фотоаппарата – чувствительность к освещению. Чем больше света, тем лучше результат. Найдите возможность хорошо осветить документ, прежде чем скопировать его.
- Добейтесь РАВНОМЕРНОГО освещения страницы документа со всех сторон. Иногда зеркальце с противоположной стороны от источника света помогает решить проблему.
- Следите, чтобы тень от вашего тела (или его части) не падала на документ. Убрать темное пятно на листе будет непросто. Научитесь критически оценивать предмет, который фотографируете.
- Следите за наклоном камеры (телефона). Типичная ошибка при перефотографировании документов – съемка под углом. В этом случае геометрия изображения будет искажена. Вместо прямо-



угольного листа вы получите подобие трапеции. Устранить этот дефект можно. Но лучше держите камеру строго перпендикулярно плоскости документа.

### ЕСЛИ У ВАС ЕСТЬ ВСПЫШКА

- Используйте ее. Это даст достаточно света, изображение получится четким, контрастным. Но главное – все изображения будут примерно одинаковые. Поэтому корректировать их придется по одной схеме. А это упрощает задачу.

- При работе со вспышкой НЕ снимайте с близкого расстояния. Скорее всего, вспышка «пересветит» кадр, документ получится слишком светлым. Мы советуем «сканировать» с расстояния 70-90 см. Точное значение вам подскажет опыт общения с вашей моделью фотокамеры. Чтоб оригинал не получился очень мелким, выберите степень увеличения (зуммирования) в 1,5, максимум в 2 раза. Тогда и вспышка кадр не «пересветит» и геометрические искажения, характерные для оптики недорогой фотокамеры, будут минимальны.

- Максимально используйте площадь кадра. Следите, чтобы не было больших полей между краями листа и границами изображения. Кстати, этот критерий вам подскажет степень зуммирования.

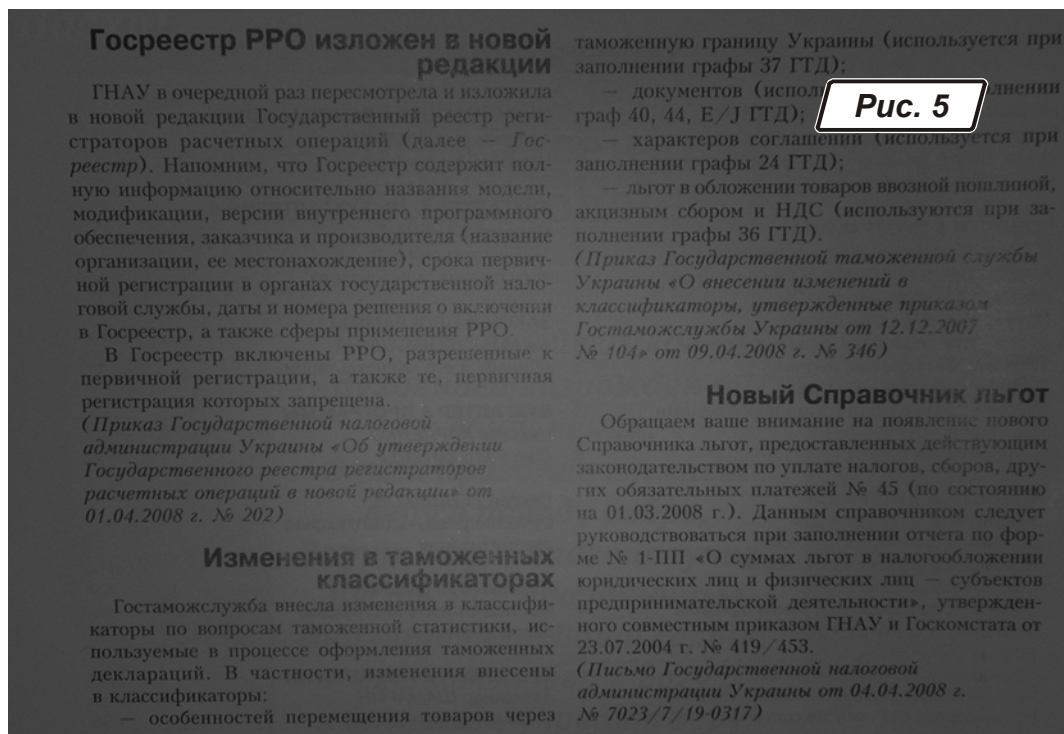
- Не устанавливайте предельное увеличение: скорее всего, возникнут проблемы с резкостью изображения.

И последнее. Запомните, от качества съемки зависит объем работы по корректировке изображения. Чем аккуратнее вы пересняли оригинал, тем меньше придется его корректировать. Потратив считанные секунды на улучшение условий освещения, вы сэкономите часы работы над улучшением фотокопии. Возможно, в вашей модели фотоаппарата есть встроенные средства для улучшения качества изображений, в т.ч. – при перефотографировании документов. Об этом читайте в инструкции к фотоаппарату. Мы же остановимся на тех проблемах, которые возникают НЕАВИСИМО от технических деталей. Это в первую очередь, нарушение геометрии, света и тона.

### Как отредактировать изображение

Итак, документы отсняты, файлы переписаны на жесткий диск. Прежде чем начать распознавание посмотрим, чем мы располагаем. А получили мы, скорее всего, изображение как на рис. 5.

Этот файл получен при помощи камеры мобильного телефона с разрешением 2 Мегапиксела в

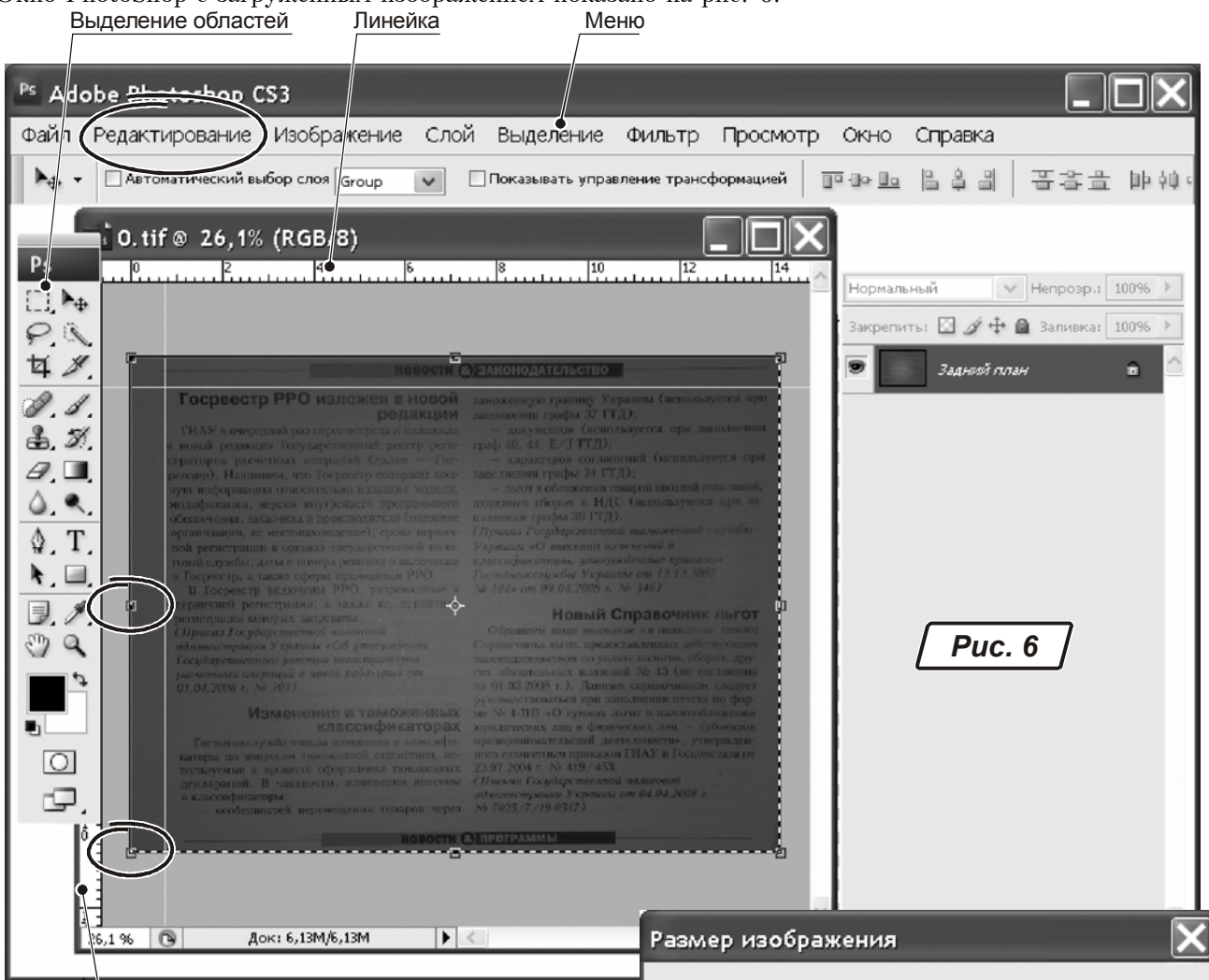


коридоре ИКК «Фактор». Освещение искусственное, света было мало. Конечно, высокого качества

ожидать не приходится. На изображении есть все типичные проблемы для этого случая. Центр и края засвечены неравномерно. Бумага выглядит серой, в то время как реально она белого цвета. Есть геометрические искажения, текст на странице имеет трапециевидную форму. Посмотрим, что можно сделать. Для работы нам понадобится хороший редактор растровой графики. Мы остановили свой выбор на PhotoShop CS3. Пожалуй, это лучший профессиональный редактор. И, хотя нам понадобятся всего несколько его функций, лучше доверять испытанным программным продуктам. PhotoShop – разнообразная и сложная программа. О ее возможностях ходят легенды. Мы не собираемся на страницах нашей газеты подробно их описывать. Нам нужен только результат. Поэтому основные приемы корректировки мы будем давать без подробных комментариев. Принимайте их, как есть. Действуйте по принципу «делай как я» и все будет в порядке.

**ЗАГРУЖАЕМ ФАЙЛ**

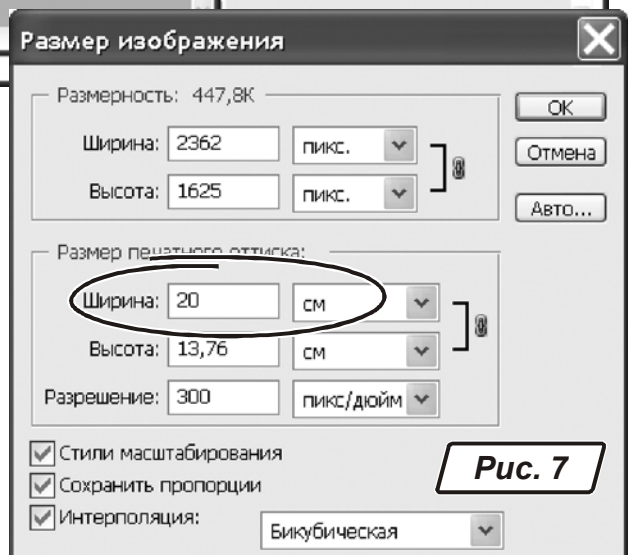
Вызываем PhotoShop через «Пуск → Программы». Открываем в нем изображение. Для этого ждем «Ctrl+O», появится окно Проводника. В нем находим файл, двойным щелчком открываем его. Окно PhotoShop с загруженным изображением показано на рис. 6.



**Рис. 6**

**КОРРЕКТИРУЕМ РАЗМЕР ИЗОБРАЖЕНИЯ**

Это первое, с чего мы начнем. И вот почему. Матрица нашей камеры была 2 Мегапикселя. Это означает, что ЛЮБОЙ оригинал, занимая весь



**Рис. 7**

кадр, будет размером примерно 9x12 см в разрешении 300 dpi. Сейчас имеет смысл вернуть его размеры к первоначальным: крупные буквы FineReader распознает лучше.

Нажимаем «Alt+Ctrl+I» («Изображение → Размер изображения»). Появится окно, показанное на рис. 7. Измеряем (приблизительно) размер БУМАЖНОГО оригинала. Достаточно определить одно измерение, например – ширину. Вводим полученное значение в поле «Ширина».

**Важно!** Флажки «Интерполяция» и «Сохранить пропорции» должны быть включены. Значение параметра «Разрешение» должно быть не менее «300 dpi».

В нашем случае ширина оригинала была около 21 см. Поэтому в поле «Ширина» вводим значение «21» и нажимаем «Ок».

**Важно!** Если вы все сделали правильно, размер изображения (в мегабайтах!) должен измениться.

## УДАЛЯЕМ ЛИШНЕЕ

- Нажимаем «М», букву «М» жмем на английской клавиатуре! Мы выбрали инструмент «Прямоугольное выделение».
  - Удерживая левую кнопку мыши, обводим с небольшим запасом полезную площадь изображения.
  - Отпускаем левую кнопку мыши. Вокруг выделенной области появится пунктирная рамка.
  - Вызываем меню «Изображение → Кадрировать».
- Photoshop удалит часть документа, расположенную вне выделенной области.

## ПРАВИМ ГЕОМЕТРИЮ

- Ставим указатель мыши на левую ЛИНЕЙКУ окна с изображением (рис. 6).
  - Удерживая левую кнопку мыши, перемещаем ее указатель на область документа. Появится вертикальная направляющая.
  - Повторяем операцию, перетащив направляющую с верхней линейки на документ.
- Таким образом, у нас появились ориентиры для контроля перпендикулярности.
- Нажимаем «Ctrl+A» (символ «А» жмем на английской клавиатуре!). Мы выделили весь документ. Вокруг него появится пунктирная рамка.
  - Нажимаем «Ctrl+T» («Редактирование → Свободная трансформация»). Вокруг изображения появится прямоугольная область с прямоугольными метками на краях и на каждой стороне. Растягивая область, мы увеличиваем размер изображения в том или ином направлении. Но нам нужно другое: ИСКАЗИТЬ ЕГО. Для этого нужно «растянуть» нижнюю часть изображения вправо.
  - Ставим указатель мыши на прямоугольную метку в правом нижнем углу. Нажимаем «Ctrl» и тащим метку вправо, пока текст не выровняется относительно направляющих. Очень наглядно.

**Совет** Познакомьтесь с возможностями трансформации в PhotoShop самостоятельно, открыв с меню «Редактирование → Трансформирование...». Уверен, вы обнаружите много полезного.

## Выравниваем яркость

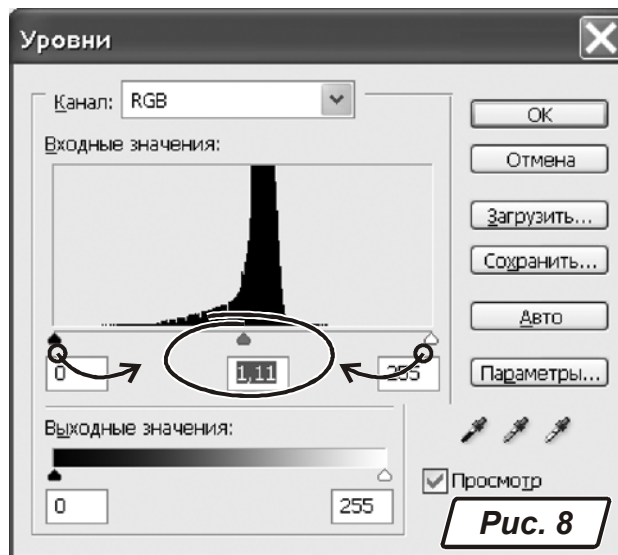
- Нажимаем «Shift+D», снять выделение (на всякий случай...).
- Смотрим на значок выделения областей (рис. 6). Если он прямоугольный, то жмем «М», затем «Shift+M». Значок выделения станет овальным. Если значок был овальный, просто нажимаем «М».
- Удерживая нажатой левую кнопку мыши, обводим более яркий фрагмент изображения (в центре). Отпускаем левую кнопку мыши.
- Нажимаем «Q». Изображение покрывается красной «поволокой». Останется только «отверстие» в центре с ЧЕТКИМИ краями. Как раз эту четкость мы сейчас и уберем.
- Нажимаем «Фильтр → Размытие → Размытие по Гауссу». Появится окно с единственным параметром. «Радиус». Увеличивайте этот параметр, чтобы добиться очень плавного размытия краев «отверстия». Обычно значение радиуса составляет 150-200 пикселей.
- Нажмите «Ок», чтобы закрыть окно.
- Нажмите «Q». Красная «маска» пропадет. Мы перешли в режим работы с изображением.

- Нажмите «Ctrl+I» (инвертировать выделенную область).
- Нажмите «Ctrl+L». Появится окно, изображенное на рис. 8. Ставим указатель мыши на центральный ползунок и, удерживая нажатой левую кнопку, плавно перемещаем его вправо. Край изображения должны стать светлее.

**Важно!** Описанный прием требует определенного опыта. От того, насколько удачно построена выделенная область, как плавно она размыта, зависит результат выравнивания. Вы можете повторить все действия с другим размером выделенной области. Но в любом случае описанный способ позволяет получить приемлемый результат.

- Нажимаем «Ok».
- Снимаем выделенную область «Ctrl+D».
- Снова выбираем «Ctrl+L».
- Осветляем изображение, перемещая правый ползунок влево. Чтобы повысить плотность текста, перемещаем левый ползунок навстречу – вправо.

**Важно!** Не стремитесь получить текст на идеально белом фоне. Текст на сероватом фоне FineReader прекрасно распознает. Главное, чтобы фон был более-менее равномерным.



- Чтобы закрыть окно, нажимаем «Ok».
- Сохраняем документ «Shift+Ctrl+S» («Сохранить как...») или «Ctrl+S» (просто «Сохранить»).

Фрагмент того, что получилось после использования описанных приемов, показан на рис. 9. Обратите внимание: этот результат получен при помощи **МОБИЛЬНОГО ТЕЛЕФОНА** с более чем скромной фотокамерой... Остается только распознать файл.

На этом мы завершим первое знакомство с системами оптического распознавания текстов. Надеемся, они не раз сослужат хорошую службу в нелегкой работе бухгалтера.

*Ждем ваших вопросов, писем, предложений и замечаний на [bk@id.factor.ua](mailto:bk@id.factor.ua), или на форуме редакции [www.bk.factor.ua/forum](http://www.bk.factor.ua/forum).  
Успешной работы!*

## Госреестр РРО изложен в новой редакции

ГНАУ в очередной раз пересмотрела и изложила в новой редакции Государственный реестр регистраторов расчетных операций (далее – *Госреестр*). Напомним, что Госреестр содержит полную информацию относительно названия модели, модификации, версии внутреннего программного обеспечения, заказчика и производителя (название организации, ее местонахождение), срока первичной регистрации в органах государственной налоговой службы, даты и номера решения о включении в Госреестр, а также сферы применения РРО.

В Госреестр включены РРО, разрешенные к первичной регистрации, а также те, первичная регистрация которых запрещена. (Приказ Государственной налоговой администрации Украины «Об утверждении Государственного реестра регистраторов расчетных операций в новой редакции» от 01.04.2008 г. № 202)

**Рис. 9**