ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ИРКУТСКОЙ ОБЛАСТИ «БРАТСКИЙ ПРОМЫШЛЕННЫЙ ТЕХНИКУМ»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по выполнению практических работ
ПО ПРОФЕССИОНАЛЬНМУ МОДУЛЮ
ВЫПОЛНЕНИЕ РАБОТ ПО ОДНОЙ ИЛИ НЕСКОЛЬКИМ ПРОФЕССИЯМ РАБОЧИХ, ДОЛЖНОСТЯМ СЛУЖАЩИХ

По теме:
«Системы оптического распознавания информации»

230401. Информационные системы (в строительстве)

Братск, 2015

Методическое пособие по выполнению практических работ разработано в соответствии с рабочей программой профессионального модуля «Выполнение работ по одной или нескольким профессиям рабочих, должностям служащих», требованиями ФГОС СПО и адресованы студентам по специальности среднего профессионального образования 230401 *Информационные системы (в строительстве)*

Организация: Государственное бюджетное профессиональное образовательное учреждение Иркутской области «Братский промышленный техникум»

Авторы-составители:

Петрович А. В., преподаватель информационных дисциплин

Рецензент:

Методические указания одобрены на заседании цикловой комиссии

Протокол №_____ от «_____» _____ 20___ г.

Председатель ЦК _____ /Орлова Н.А./

ПРАКТИЧЕСКАЯ РАБОТА
СКАНИРОВАНИЕ ИЗОБРАЖЕНИЙ И РАСПОЗНАВАНИЕ ТЕКСТА

Цель работы: получить представление об OCR – программах распознавания текста, познакомиться с возможностями данных программы, научить распознавать отсканированный текст, передавать и редактировать его в Word; знать системы распознавания символов, форм и текста; уметь пользоваться программой распознавания текста.

Краткие теоретические сведения

1. Системы оптического распознания символов – преобразуют элементы графического изображения в последовательности символов (FineReader, CuneiForm).
2. Системы оптического распознавания форм – распознаются рукопечатные текст (данные вводятся в поля печатными буквами).
3. Системы распознавания рукописного текста – преобразуют текст, созданный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.

С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический файл — обычную картинку. Текст можно будет читать и распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

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 (именно так работали OCR первого поколения), 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.

Как видим, для того, чтобы получить электронную, готовую к редактированию копию любого печатного текста, программе OCR необходимо выполнить «цепочку» из множества отдельных операций. Сначала необходимо распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и так далее. Далее выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст. Если исходный документ имеет типографское качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений), то задача распознавания решается методом сравнения с растровым шаблоном. 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 отличных от входного изображения точек.

При распознавании документов с низким качеством печати (машинописный текст, факс и так далее) используется метод распознавания символов по наличию в них определенных структурных элементов (отрезков, колец, дуг и др.).

Любой символ можно описать через набор значений 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 данными буквами — в величине углов, которые образует третий отрезок

с двумя другими. 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 соответствует распознаваемому символу.

Наиболее распространенные системы оптического распознавания символов, например, ABBYY FineReader и CuneiForm от Cognitive, используют как растровый, так и структурный методы распознавания. Кроме того, эти системы являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов) и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

При заполнении налоговых деклараций, при проведении переписей населения и так далее используются различного вида бланки с полями. Рукопечатные 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 написанные от руки символы, довольно сильно различающиеся у разных людей. Кроме того, система должна определить, к какому полю относится распознаваемый текст.

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. Программы для распознавания текста вы можете приобрести отдельно или получить бесплатно вместе с купленным вами сканером.

Возможно, самая известная программа для распознавания текстов – это FineReader от компании ABBYY. Именно эту программу чаще всего вспоминают, когда речь заходит о системах распознавания.

FineReader — омнифонтовая система оптического распознавания текстов. Это означает, что она позволяет распознавать тексты, набранные практически любыми шрифтами, без предварительного обучения. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати, что достигается благодаря применению технологии "целостного целенаправленного адаптивного распознавания".

FineReader имеет массы дополнительных функций, которые простому пользователю, возможно, и без надобности, но зато производят впечатление на определенные группы покупателей. Так, одним из козырей FineReader является поддержка неимоверного количества языков распознавания — 176, в числе которых вы найдете экзотические и древние языки, и даже популярные языки программирования.

Но далеко не все возможности включены в самую простую модификацию программы, которую вы можете получить бесплатно вместе со сканером. Пакетное сканирование, грамотная обработка таблиц и изображений — для всего этого стоит приобрести профессиональную версию программы.

Все версии FineReader, от самой простой до самой мощной, объединяет удобный интерфейс. Для запуска процесса распознавания вам достаточно просто положить документ в сканер и нажать единственную кнопку (мастер Scan & Read) на панели инструментов программы. Все дальнейшие операции — сканирование, разбивку изображения на «блоки» и, наконец, собственно распознавание программа выполнит автоматически. Пользователю останется только установить нужные параметры сканирования.

FineReader работает со сканерами через TWAIN-интерфейс. Это единый международный стандарт, введенный в 1992 году для унификации взаимодействия устройств для ввода изображений в компьютер (например, сканера) с внешними приложениями. Качество распознавания во многом зависит от того, насколько хорошее изображение получено при сканировании. Качество изображения регулируется установкой основных параметров сканирования: типа изображения, разрешения и яркости.

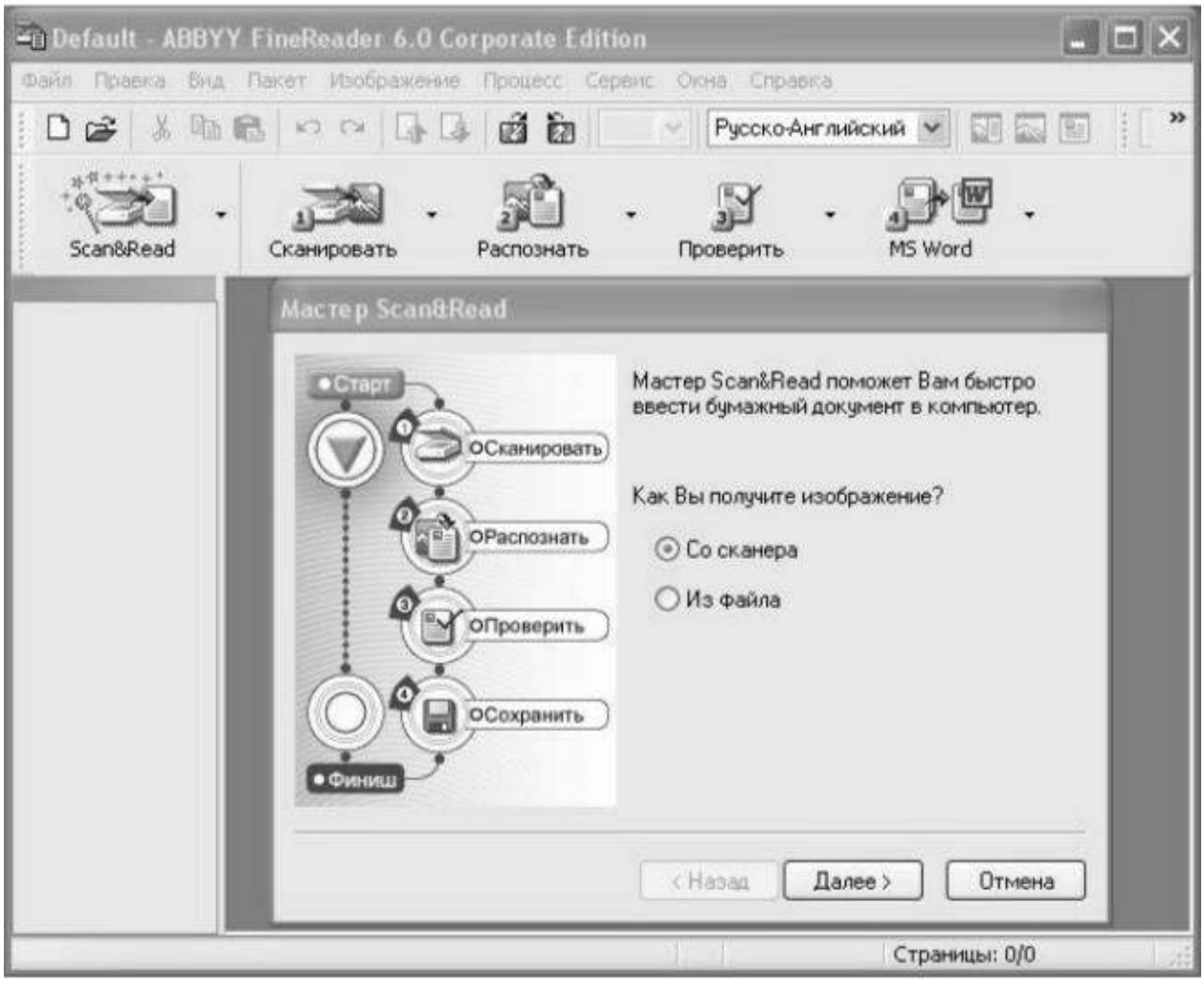
Сканирование в сером является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если Вы хотите, чтобы содержащиеся в документе цветные элементы (картинки, цвет букв и фона) были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

Оптимальным разрешением для обычных текстов является - 300 dpi и 400-600 dpi для текстов, набранных мелким шрифтом (9 и менее пунктов). После завершения распознавания страницы FineReader предложит пользователю выбор: сканировать и распознавать дальше (для многостраничного документа) или сохранить полученный текст в одном из множества популярных форматов — от документов Microsoft Office до HTML или PDF. Можно, впрочем, сразу же перебросить документ в Word или Excel, и уже там исправить все огрехи распознавания. При этом FineReader полностью сохраняет все особенности форматирования документа и его графическое оформление.

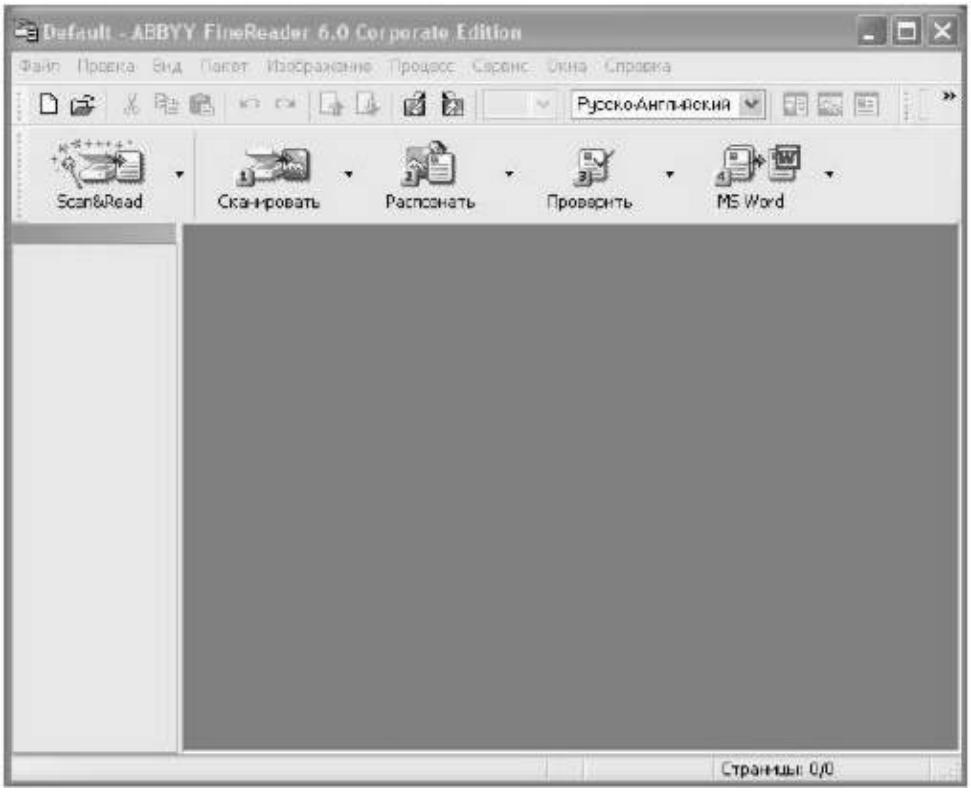
Задание 1

Отсканировать и преобразовать в электронный текстовый документ страницу текста.

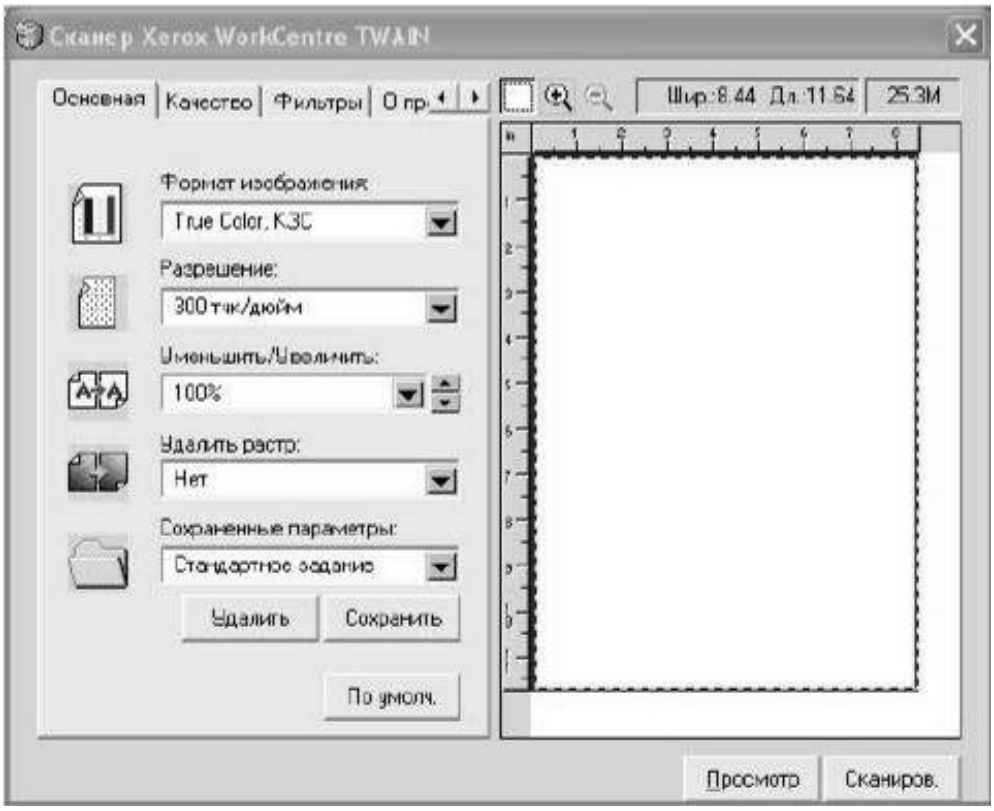
1. В операционной системе Windows запустить программу ABBYY FineReader-6 Corporate Edition (Пуск – Все программы – ABBYY FineReader-6 Corporate Edition).



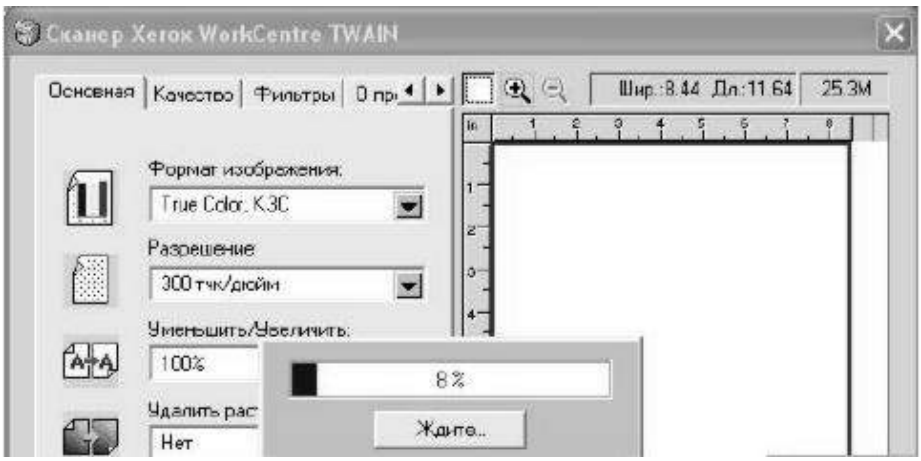
2. Поместить бумажную страницу с текстом в окно сканирования сканера (МФУ).
3. Можно использовать «Мастер Scan&Read».
4. Мы будем выполнять работу без помощи «Мастер Scan&Read», поэтому нажимаем кнопку Отмена.



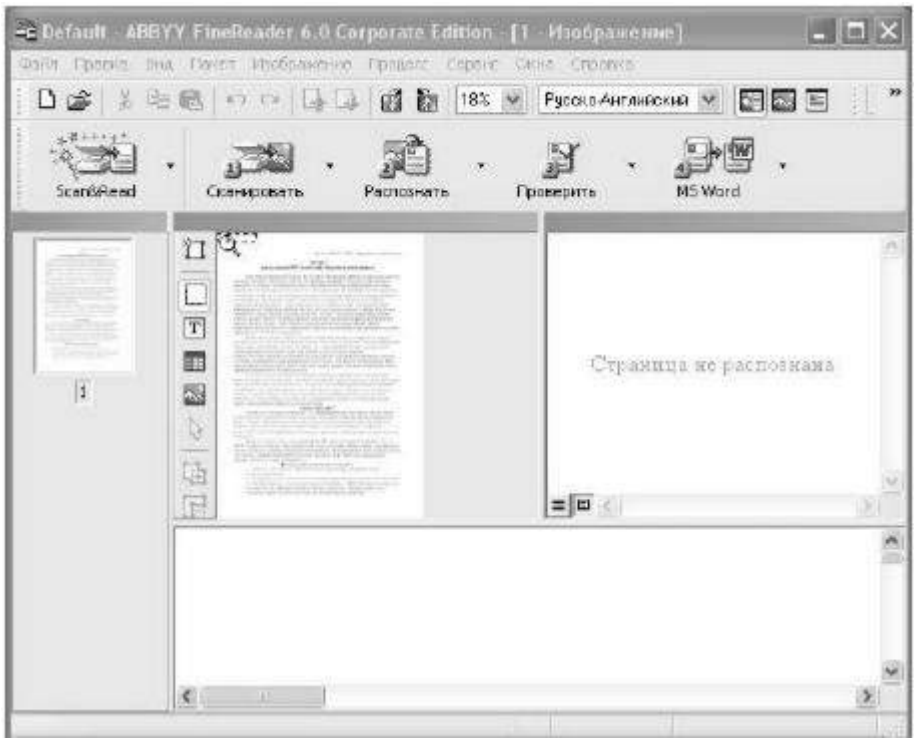
5. Вызываем вкладку с кнопки Сканировать – Сканировать изображение. Откроется окно «Сканер Xerox WorkCentre TWAIN»



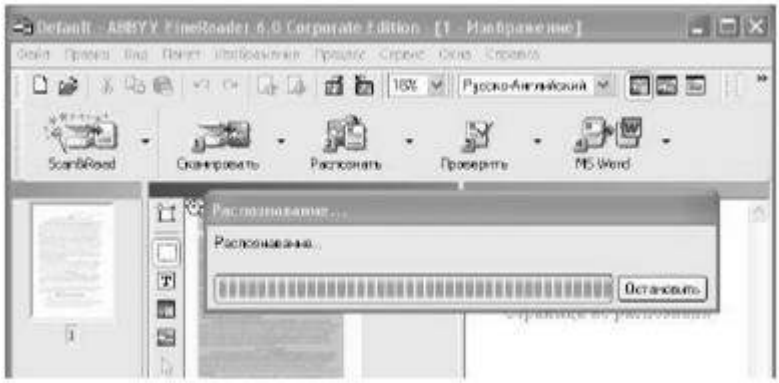
6. Запускаем режим сканирования кнопкой Сканиров.



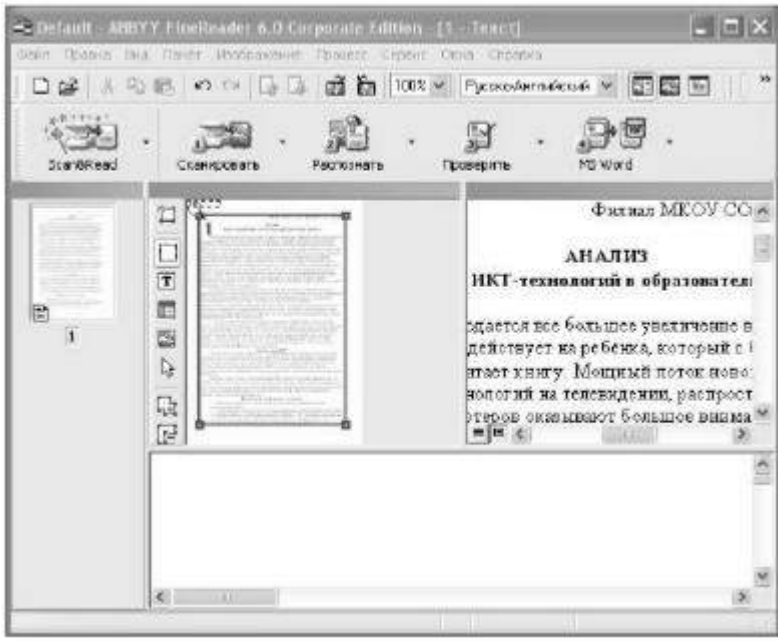
7. После завершения сканирования открывается окно с изображением документа.



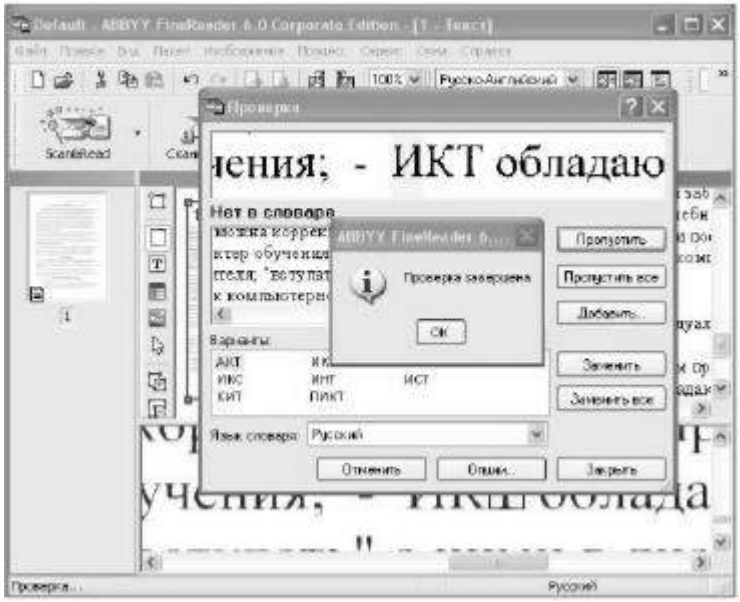
8. Изображение текстового документа на бумажной странице можно сохранить как графический файл командой Файл – Сохранить изображение как.
9. Для преобразования элементов графического изображения в последовательности текстовых символов нужно нажать кнопку Распознать.



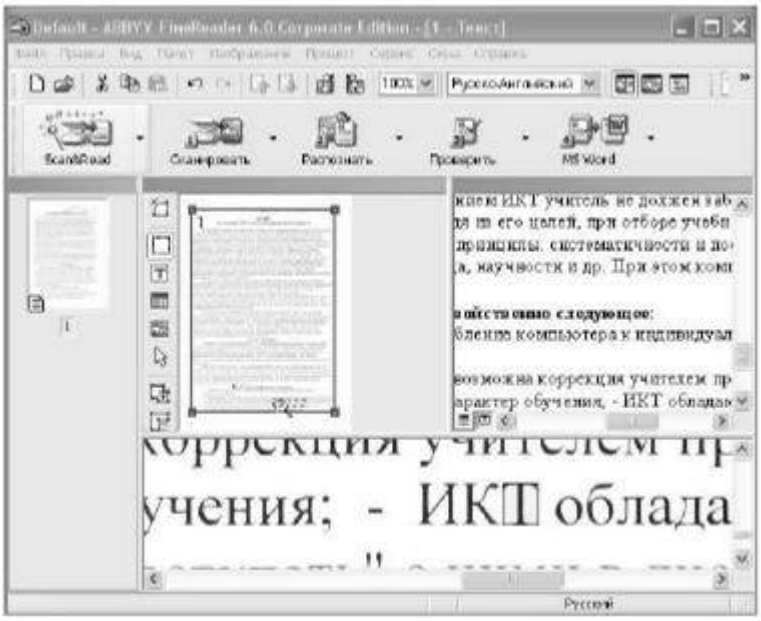
10. После завершения распознавания в окне распознавания программы появится текстовый фрагмент документа.



11. Для редактирования текстового документа необходимо запустить процесс проверки нажатием кнопки Проверить. (В этом режиме с помощью данной программы можно исправить ошибки, допущенные при распознавании).



12. Результаты распознавания и проверки можно сохранить в файл, передать во внешнее приложение, не сохраняя на диск или скопировать в буфер обмена. Распознанный текст можно отправить в Microsoft Word. Для этого щелкните кнопку Передать в MS Word. Запуститься программа Microsoft Word и откроется распознанный текст, который вы можете редактировать и форматировать, сохранить в файл.



Задание 2

1. Откройте документ FineReader «Страница из газеты» в папке своей сетевой папке. Распознайте отсканированный документ в FineReader, выделив блоки. Сохраните распознанный документ в Microsoft Office Word и отредактируйте его.
2. Окройте в FineReaderPDF-файл«Правила внутреннего распорядка»из своей сетевой папки распознайте его, сохраните вMicrosoft Office Word. Отредактируйте документ.

Вопросы для защиты работы:

1. Зачем нужны программы распознавания текста?
2. Как происходит распознавание текста?
3. Какие программы распознания текста вы знаете? Какими пользовались?
4. Какое разрешение является оптимальным для сканирования текста, изображений?

Список используемых источников:

1. Файловый архив студентов StudFiles [Электронный документ] — URL: http://www.studfiles.ru
2. Информационный портал Студопедия [Электронный документ] — URL: http://studopedia.ru