УЗБЕКСКОЕ АГЕНТСТВО СВЯЗИ И ИНФОРМАТИЗАЦИИ

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Кафедра информатики и информационных технологии

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Алымбетова Рината студента 4-го курса факультета информационных технологий
по направлению информатики и информационные технологий

ТЕМА: Специальные режимы распознавания текста в программе FineReader

Научный руководитель:_____доц. Аламинов М.
ст. преп. Омарова Х.

Зав. кафедрой:_____доц. Бурханов Ш.А.

НУКУС - 2012 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время вместе с расширяющимся применением Internet и электронной почты остается широко распространенной такая форма обмена документами, как факсимильная связь. Она будет использоваться во всем мире еще долгое время, поскольку обладает следующими достоинствами: простота использования, очевидность, дешевизна, использование при передаче твердой копии (электронная версия документа не всегда есть в наличии). Однако отличительной чертой такого способа связи является передача изображения документа, сформированного с помощью сканирующего устройства факс-аппарата. Вследствие этого, применение факсимильной связи создает большие неудобства при учете, хранении и обработке входящей корреспонденции, особенно если велик ее объем. Очевидны также неудобства хранения электронных версий документа в виде изображений из-за большого объема файла и недоступности содержимого для автоматической обработки. Обработка документов может включать отбор документов по ключевым словам, определение тематики документа путем нахождения слов, характерных для какой-либо предметной области, автоматическое индексирование и перевод, а также классификацию документов согласно принадлежности организации-отправителю. Для решения всех этих задач необходим доступный текст документа.

Актуальность такой работы подтверждается последними публикациями. Даже общепризнанные лидеры среди пакетов оптического распознавания символов (optical character recognition, OCR), как раз и предназначенные для решения задач подобного рода, не справляются с распознаванием обычного факсимильного сообщения, несмотря на то, что текст можно легко прочесть визуально.

Существующие в настоящее время системы распознавания не всегда позволяют эффективно распознавать изображения печатных текстов низкого качества, характерные для документов, полученных по факсу.

Существует несколько причин, обуславливающих низкое качество факсимильных документов. Среди них — низкое разрешение факс-аппарата. В то время как распознаваемые документы обычно сканируются с разрешением не ниже 300 dpi, факсимильный документ обычно имеет разрешение 200x100 dpi (режим «Standard»), и лишь при улучшенном качестве передачи (режим «Fine») 200x200 dpi. Разное разрешение по вертикали и по горизонтали приводит к тому, что высота изображения документа в пикселях оказывается в два раза меньше ожидаемой, а символы оказываются «сплюснутыми» и в таком виде практически не распознаются. При низком разрешении символы имеют малую высоту в пикселях, поэтому случайные изменения нескольких пикселей приводят не только к значительным изменениям формы символа, но часто меняют его топологию. Значительные искажения изображений документов возникают вследствие низкого качества исходной твердой копии, причиной которого является старение, копирование, использование морально устаревших печатающих устройств. Все эти дефекты не позволяют применять для распознавания символов многие известные методы, в частности метод сравнения с эталоном, структурные методы. Не являются информативными топологические признаки, признаки формы и многие другие признаки, традиционно считающиеся эффективными при распознавании изображений. Низкое разрешение и плохое качество приводят к ошибкам на этапе предварительной обработки, в частности, к пропуску строк. Обычной является ситуация, когда вполне читаемый текст распознается с ошибками почти в каждом слове вследствие неэффективности процедур распознавания и орфографической коррекции.

Характерными при факсимильной передаче являются такие помехи, как перекос документа, а также появление тонкой вертикальной линии, вызванное дефектом факс-аппарата и приводящее к соединению символов соседних строк. Это делает не эффективными описанные в последних

публикациях методы сегментации строк и вызывает необходимость их усовершенствования или разработки новых.

Другим источником плохого качества документа являются помехи в линии связи. Для передачи через коммутируемую телефонную сеть графическая информация кодируется по строкам пикселей. Это приводит к тому, что кратковременная помеха искажает всю строку.

Для определения организации-отправителя система обработки и распознавания факсимильных документов должна использовать алгоритмы распознавания, адаптированные для обнаружения и идентификации уникальных для каждой организации признаков документа, а для обработки текста необходимы алгоритмы распознавания печатных символов низкого качества, что характерно для документов, полученных по факсу, а также алгоритмы определения тематики документа, использующие поиск по словарям. Кроме того, в условиях плохого качества электронной копии, для правильного распознавания слов не достаточно только лишь посимвольного распознавания, поскольку в этом случае ошибки неизбежны, не зависимо от алгоритма распознавания. Необходима подсистема проверки орфографии слов, взаимодействующая с распознающей системой с целью подбора наиболее близкого слова. Однако при анализе текста, полученного программой FineReader, видно, что в результате ошибок распознавания текст состоит из слов, представляющих собой бессмысленные наборы букв, хотя на изображении данного документа присутствуют разрешенные в языке слова.

Кроме того, в современных комплексных системах мониторинга и обработки информации возникает задача высокоскоростной обработки интенсивных информационных потоков и отбора конкретных факсимильных сообщений, что требует построения специализированных многопроцессорных вычислительных систем.

Одним из способов повышения производительности обработки может быть отбор факсимильных документов с помощью распознавания типа бланка по эмблеме или логотипу, без распознавания текстового

содержимого. Однако эта функция, как правило, вообще не поддерживается существующими средствами обработки изображений документов.

Существующие коммерческие пакеты оптического распознавания символов предназначены исключительно для работы на персональной ЭВМ. Они не доступны для усовершенствования, разработки новых систем и новых реализаций.

С учетом вышесказанного, имеется необходимость разработки методов и средств распознавания, пригодных для построения перспективных современных средств обработки факсимильных сообщений, используемых при создании новых программных пакетов, которые могут быть реализованы на различных платформах, и при построении специализированных комплексных систем мониторинга и обработки информации.

С задачей распознавания образов живые системы, в том числе и человек, сталкиваются постоянно с момента своего появления. В частности, информация, поступающая с органов чувств, обрабатывается мозгом, который в свою очередь сортирует информацию, обеспечивает принятие решения, а далее с помощью электрохимических импульсов передает необходимый сигнал далее, например, органам движения, которые реализуют необходимые действия. Затем происходит изменение окружающей обстановки, и вышеуказанные явления происходят заново. И если разобраться, то каждый этап сопровождается распознаванием.

С развитием вычислительной техники стало возможным решить ряд задач, возникающих в процессе жизнедеятельности, облегчить, ускорить, повысить качество результата. К примеру, работа различных систем жизнеобеспечения, взаимодействие человека с компьютером, появление роботизированных систем и др. Тем не менее, отметим, что обеспечить удовлетворительный результат в некоторых задачах (распознавание быстродвижущихся подобных объектов, рукописного текста) в настоящее время не удается.

В первом параграфе рассматриваются основные понятия теории распознавания образов, делается обзор методов и алгоритмов их распознавания, дается общая характеристика задач распознавания образов.

Во втором параграфе дается краткая история развития систем распознавания образов, рассматривается текущее состояние технологии оптического распознавания текста.

В третьем параграфе рассматривается система оптического распознавания текста ABBYY FineReader, в частности, задачи распознавания, проверка, редактирование и сохранение текста, описываются возможности программы.

В четвертом параграфе информация об автоматическом режиме распознавания текста в FineReader и режиме обучения с помощью эталона, о настройках программы под каждое определенное распознаваемое изображение.

В пятом параграфе приведен пример правильной настройки и подготовки программы для распознавания, а также пример конвертации документа формата PDF в DOC, формат программы MS Office Word.

§1. Что такое распознавание образов?

1.1 Определения. Методы распознавания образов

Первые исследования с вычислительной техникой в основном следовали классической схеме математического моделирования - математическая модель, алгоритм и расчет. Таковыми были задачи моделирования процессов происходящих при взрывах атомных бомб, расчета баллистических траекторий, экономических и прочих приложений. Однако помимо классических идей этого ряда возникали и методы основанные на совершенно иной природе, и как показывала практика решения некоторых задач, они зачастую давали лучший результат нежели решения, основанные на переусложненных математических моделях. Их идея заключалась в отказе от стремления создать исчерпывающую математическую модель изучаемого объекта. Причем зачастую адекватные модели было практически невозможно построить, а вместо этого удовлетвориться ответом лишь на конкретные интересующие нас вопросы, причем эти ответы искать из общих для широкого класса задач соображений. К исследованиям такого рода относились распознавание зрительных образов, прогнозирование урожайности, уровня рек, задача различения нефтеносных и водоносных пластов по косвенным геофизическим данным и т. д. Конкретный ответ в этих задачах требовался в довольно простой форме, как например, принадлежность объекта одному из заранее фиксированных классов. А исходные данные этих задач, как правило, задавались в виде обрывочных сведений об изучаемых объектах, например в виде набора заранее расклассифицированных объектов. С математической точки зрения это означает, что распознавание образов представляет собой далеко идущее обобщение идеи экстраполяции функции.

Важность такой постановки для технических наук не вызывает никаких сомнений и уже это само по себе оправдывает многочисленные исследования в этой области. Однако задача распознавания образов имеет и более широкий

аспект для естествознания. В контекст данной науки органично вошли и поставленные еще древними философами вопросы о природе нашего познания, нашей способности распознавать образы, закономерности, ситуации окружающего мира. В действительности, можно практически не сомневаться в том, что механизмы распознавания простейших образов, типа образов приближающегося опасного хищника или еды, сформировались значительно ранее, чем возник элементарный язык и формально-логический аппарат. И не вызывает никаких сомнений, что такие механизмы достаточно развиты и у высших животных, которым так же в жизнедеятельности крайне необходима способность различения достаточно сложной системы знаков природы. Таким образом, в природе мы видим, что феномен мышления и сознания явно базируется на способностях к распознаванию образов и дальнейший прогресс науки об интеллекте непосредственно связан с глубиной понимания фундаментальных законов распознавания. Понимая тот факт, что вышеперечисленные вопросы выходят далеко за рамки стандартного определения распознавания образов, необходимо так же понимать, что они имеют глубокие связи с этим относительно узким направлением.

Уже сейчас распознавание образов плотно вошло в повседневную жизнь и является одним из самых насущных знаний современного инженера. В медицине распознавание образов помогает врачам ставить более точные диагнозы, на заводах оно используется для прогноза брака в партиях товаров. Системы биометрической идентификации личности в качестве своего алгоритмического ядра так же основаны на результатах этой дисциплины. Дальнейшее развитие искусственного интеллекта, в частности проектирование компьютеров пятого поколения, способных к более непосредственному общению с человеком на естественных для людей языках и посредством речи, немыслимы без распознавания. Здесь рукой подать и до

робототехники, искусственных систем управления, содержащих в качестве жизненно важных подсистем системы распознавания .

Именно поэтому к развитию распознавания образов с самого начала было приковано немало внимания со стороны специалистов самого различного профиля - кибернетиков, нейрофизиологов, психологов, математиков, экономистов и т.д. Во многом именно по этой причине современное распознавание образов само питается идеями этих дисциплин.

Прежде, чем приступить к основным методам распознавания образов, приведем несколько необходимых определений.

Распознавание образов (объектов, сигналов, ситуаций, явлений или процессов) - задача идентификации объекта или определения каких-либо его свойств по его изображению (оптическое распознавание) или аудиозаписи (акустическое распознавание) и другим характеристикам.

Одним из базовых является не имеющее конкретной формулировки понятие множества. В компьютере множество представляется набором неповторяющихся однотипных элементов. Слово "неповторяющихся" означает, что какой-то элемент в множестве либо есть, либо его там нет. Универсальное множество включает все возможные для решаемой задачи элементы, пустое не содержит ни одного.

Образ - классификационная группировка в системе классификации, объединяющая (выделяющая) определенную группу объектов по некоторому признаку. Образы обладают характерным свойством, проявляющимся в том, что ознакомление с конечным числом явлений из одного и того же множества дает возможность узнавать сколь угодно большое число его представителей. Образы обладают характерными объективными свойствами в том смысле, что разные люди, обучающиеся на различном материале наблюдений, большей частью одинаково и независимо друг от друга классифицируют одни и те же объекты. В классической постановке задачи распознавания универсальное множество разбивается на части-образы.

Каждое отображение какого-либо объекта на воспринимающие органы распознающей системы, независимо от его положения относительно этих органов, принято называть изображением объекта, а множества таких изображений, объединенные какими-либо общими свойствами, представляют собой образы.

Методика отнесения элемента к какому-либо образу называется решающим правилом. Еще одно важное понятие - метрика, способ определения расстояния между элементами универсального множества. Чем меньше это расстояние, тем более похожими являются объекты (символы, звуки и др.) - то, что мы распознаем. Обычно элементы задаются в виде набора чисел, а метрика - в виде функции. От выбора представления образов и реализации метрики зависит эффективность программы, один алгоритм распознавания с разными метриками будет ошибаться с разной частотой.

Обучением обычно называют процесс выработки в некоторой системе той или иной реакции на группы внешних идентичных сигналов путем многократного воздействия на систему внешней корректировки. Такую внешнюю корректировку в обучении принято называть "поощрениями" и "наказаниями". Механизм генерации этой корректировки практически полностью определяет алгоритм обучения. Самообучение отличается от обучения тем, что здесь дополнительная информация о верности реакции системе не сообщается.

Адаптация - это процесс изменения параметров и структуры системы, а возможно - и управляющих воздействий, на основе текущей информации с целью достижения определенного состояния системы при начальной неопределенности и изменяющихся условиях работы.

Обучение - это процесс, в результате которого система постепенно приобретает способность отвечать нужными реакциями на определенные совокупности внешних воздействий, а адаптация - это подстройка параметров и структуры системы с целью достижения требуемого качества управления в условиях непрерывных изменений внешних условий.

1.2 Общая характеристика задач распознавания образов и их типы

В целом, можно выделить три метода распознавания образов:

Метод перебора. В этом случае производится сравнение с базой данных, где для каждого вида объектов представлены всевозможные модификации отображения. Например, для оптического распознавания образов можно применить метод перебора вида объекта под различными углами, масштабами, смещениями, деформациями и т. д. Для букв нужно перебирать шрифт, свойства шрифта и т. д.

Второй метод - производится более глубокий анализ характеристик образа. В случае оптического распознавания это может быть определение различных геометрических характеристик. Звуковой образец в этом случае подвергается частотному, амплитудному анализу и т. д.

Следующий метод - использование искусственных нейронных сетей (ИНС). Этот метод требует либо большого количества примеров задачи распознавания при обучении, либо специальной структуры нейронной сети, учитывающей специфику данной задачи. Тем не менее, его отличает более высокая эффективность и производительность.

Иску́сственные нейро́нные се́ти (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы. Первой такой попыткой были нейронные сети Маккалока и Питтса. Впоследствии, после разработки алгоритмов обучения, получаемые модели стали использовать в практических целях: в задачах прогнозирования, для распознавания образов, в задачах управления и др.

ИНС представляют собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты, особенно в сравнении с процессорами,

используемыми в персональных компьютерах. Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие локально простые процессоры вместе способны выполнять довольно сложные задачи.

С точки зрения машинного обучения, нейронная сеть представляет собой частный случай методов распознавания образов, дискриминантного анализа, методов кластеризации и т. п. С математической точки зрения, обучение нейронных сетей — это многопараметрическая задача нелинейной оптимизации. С точки зрения кибернетики, нейронная сеть используется в задачах адаптивного управления и как алгоритмы для робототехники. С точки зрения развития вычислительной техники и программирования, нейронная сеть — способ решения проблемы эффективного параллелизма. А с точки зрения искусственного интеллекта, ИНС является основой философского течения коннективизма и основным направлением в структурном подходе по изучению возможности построения естественного интеллекта с помощью компьютерных алгоритмов.

Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных или «зашумленных», частично искаженных данных.

1.3 Распознавание образов и классификация

Общая структура системы распознавания и этапы в процессе ее разработки показаны на рис. 4.

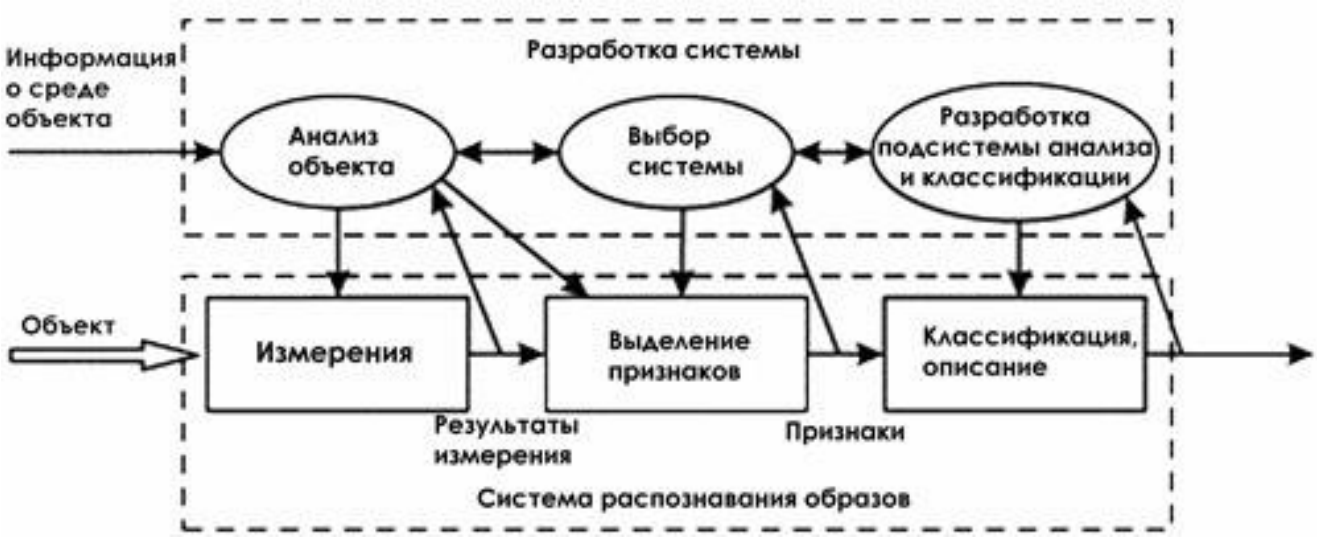


Рис. 4. Структура системы распознавания

Задачи распознавания имеют следующие характерные черты.

Это информационные задачи, состоящие из двух этапов: - преобразование исходных данных к виду, удобному для распознавания; - собственно распознавание (указание принадлежности объекта определенному классу).

В этих задачах можно вводить понятие аналогии или подобия объектов и формулировать правила, на основании которых объект зачисляется в один и тот же класс или в разные классы. В этих задачах можно оперировать набором прецедентов-примеров, классификация которых известна и которые в виде формализованных описаний могут быть предъявлены алгоритму распознавания для настройки на задачу в процессе обучения.

Для этих задач трудно строить формальные теории и применять классические математические методы (часто недоступна информация для точной математической модели или выигрыш от использования модели и математических методов несоизмерим с затратами).

Выделяют следующие типы задач распознавания: - Задача распознавания - отнесение предъявленного объекта по его описанию к одному из заданных классов (обучение с учителем); - Задача автоматической классификации - разбиение множества объектов, ситуаций, явлений по их описаниям на систему непересекающихся классов (таксономия, кластерный анализ, самообучение);

- Задача выбора информативного набора признаков при распознавании;
- Задача приведения исходных данных к виду, удобному для распознавания; - Динамическое распознавание и динамическая классификация - задачи 1 и 2 для динамических объектов;
- Задача прогнозирования - суть предыдущий тип, в котором решение должно относиться к некоторому моменту в будущем.

В качестве образов могут выступать различные по своей природе объекты: символы текста, изображения, образцы звуков и т. д. При обучении сети предлагаются различные образцы образов с указанием того, к какому классу они относятся. Образец, как правило, представляется как вектор значений признаков. При этом совокупность всех признаков должна однозначно определять класс, к которому относится образец. В случае, если признаков недостаточно, сеть может соотнести один и тот же образец с несколькими классами, что неверно. По окончании обучения сети ей можно предъявлять неизвестные ранее образы и получать ответ о принадлежности к определённому классу.

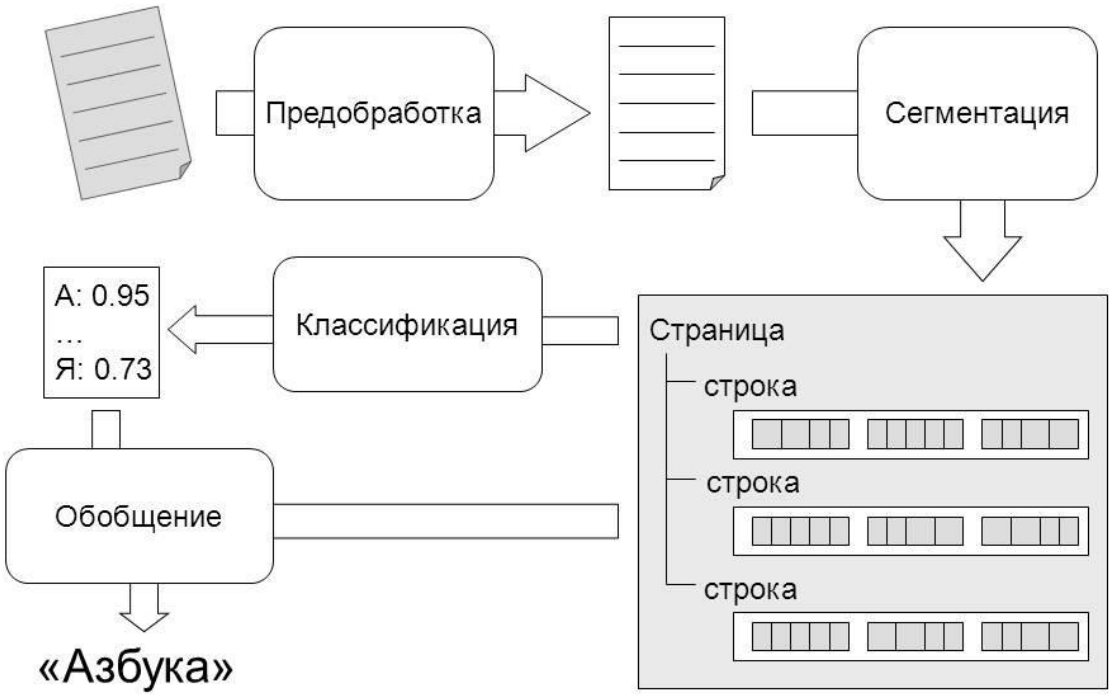
§2. Задачи распознавания текста

2.1 Оптическое распознавание символов. История

Несмотря на то, что в настоящее время большинство документов составляется на компьютерах, задача создания полностью электронного документооборота ещё далека до полной реализации. Как правило, существующие системы охватывают деятельность отдельных организаций, а обмен данными между организациями осуществляется с помощью традиционных бумажных документов.

Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид.

Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл. Более предпочтительным, по сравнению с графическим, является текстовое представление информации. Этот вариант позволяет существенно сократить затраты на хранение и передачу информации, а также позволяет реализовать все возможные сценарии использования и анализа электронных документов. Поэтому наибольший интерес с практической точки зрения представляет именно перевод бумажных носителей в текстовый электронный документ. На вход системы распознавания поступает растровое изображение страницы документа. Для работы алгоритмов распознавания желательно, чтобы поступающее на вход изображение было как можно более высокого качества. Если изображение зашумлено, нерезко, имеет низкую контрастность, то это усложнит задачу алгоритмов распознавания.



Поэтому перед обработкой изображения алгоритмами распознавания проводится его предварительная обработка, направленная на улучшение качества изображения. Она включает фильтрацию изображения от шумов, повышение резкости и контрастности изображения, выравнивание и преобразование в используемый системой формат (в нашем случае 8-битное изображение в градациях серого). Подготовленное изображение попадает на вход модуля сегментации. Задачей этого модуля является выявление структурных единиц текста – строк, слов и символов. Выделение фрагментов высоких уровней, таких как строки и слова, может быть осуществлено на основе анализа промежутков между тёмными областями. К сожалению, такой подход не может быть применён для выделения отдельных букв, поскольку, в силу особенностей начертания или искажений, изображения соседних букв могут объединяться в одну компоненту связанности (рис. 1) или наоборот — изображение одной буквы может распадаться на отдельные компоненты связанности (рис. 2). Во многих случаях для решения задачи

сегментации на уровне букв используются сложные эвристические алгоритмы.

Рисунок 1. Объединение нескольких букв в одну компоненту связанности.

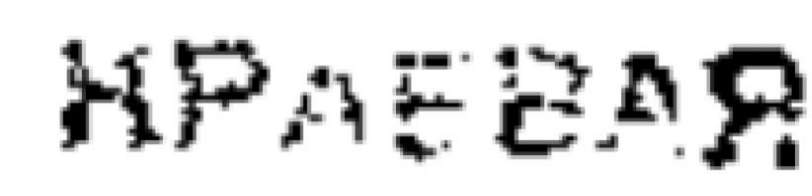


Рисунок 2. Распадение изображений букв на несвязанные компоненты вследствие низкого качества сканирования.

Полагаем, что для принятия окончательного решения о прохождении границы букв на таком раннем этапе обработки, системе распознавания недостаточно информации. Поэтому задачей модуля сегментации на уровне букв в разработанном алгоритме является нахождение возможных границ символов внутри буквы, а окончательное решение о разбиении слова принимается на последнем этапе обработки, с учётом идентификации отдельных фрагментов изображения как букв. Дополнительным преимуществом такого подхода является возможность работы с начертаниями букв, состоящих из нескольких компонент связанности без специальной обработки таких случаев.

Результатом работы модуля сегментации является дерево сегментации– структура данных, организация которой отражает структуру текста на странице. Самому верхнему уровню соответствует объект страница. Он содержит массив объектов, описывающих строки. Каждая строка в свою очередь включает набор объектов слов. Слова являются листьями этого дерева. Информация о возможных местах разделения слова на буквы

храниться в слове, однако отдельные объекты для букв не выделяются. В каждом объекте дерева хранится информация об области, занимаемой соответствующим объектом на изображении. Данная структура легко может быть расширена для поддержки других уровней разбиения, например колонок, таблиц.

Выявленные фрагменты изображения подаются на вход классификатора, выходом которого является вектор возможности принадлежности изображения к классу той или иной буквы. В разработанном алгоритме используется классификатор составной архитектуры, организованный в виде дерева, листьями которого являются простые классификаторы, а внутренние узлы соответствуют операциям комбинирования результатов низлежащих уровней (рис. 3).

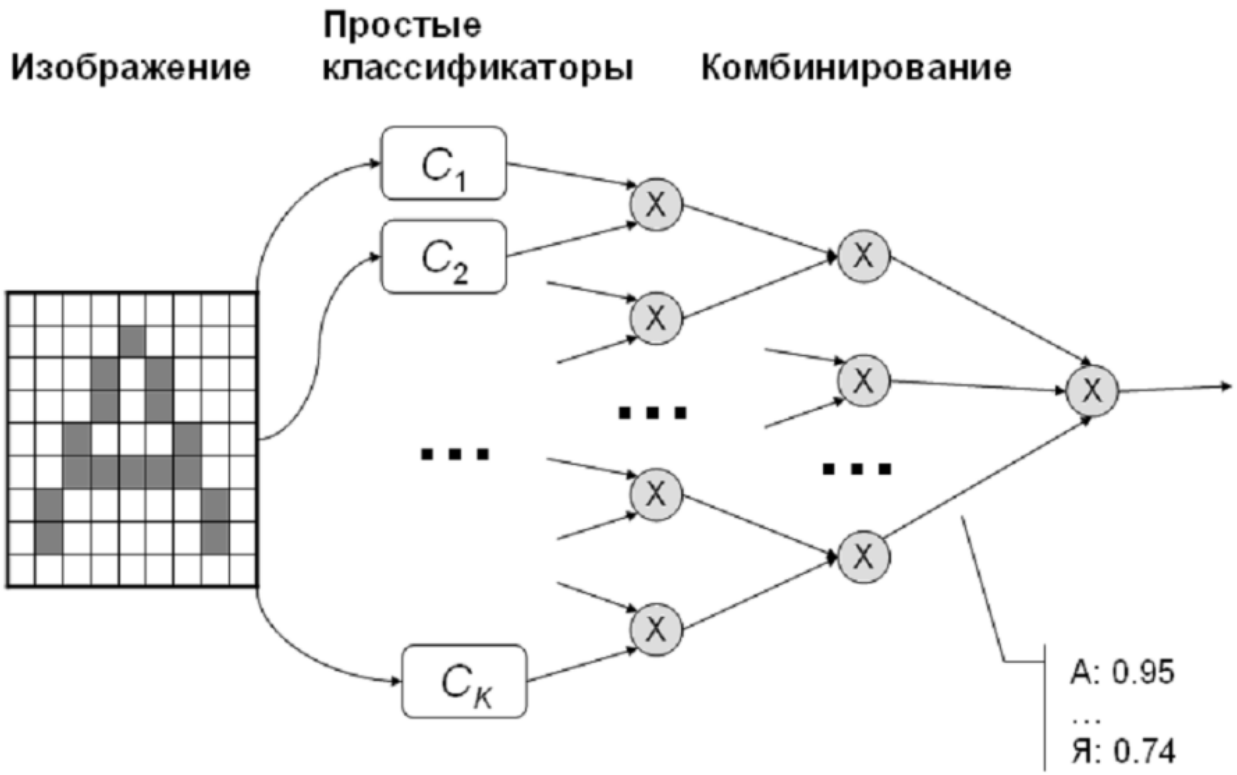


Рисунок 3. Архитектура классификатора

Работа простого классификатора осуществляется в два шага (рис. 4). Сначала по исходному изображению вычисляются признаки. Значение каждого признака является функцией от яркостей некоторого подмножества пикселей изображения. В результате получается вектор значений признаков, который поступает на вход нейронной сети. Каждый выход сети соответствует одной из букв алфавита, а получаемое на выходе значение рассматривается как уровень принадлежности буквы нечёткому множеству.

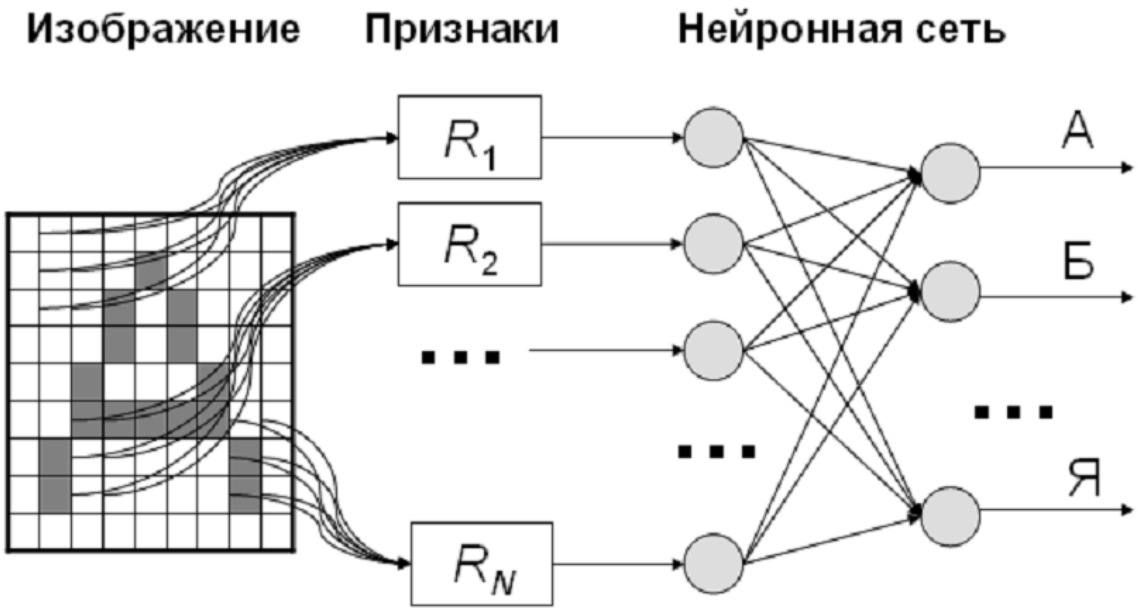


Рисунок 4. Простой классификатор

Задачей алгоритма комбинирования является обобщение информации, поступающей в виде входных нечётких множеств и вычисление на их основе выходного нечёткого подмножества множества распознаваемых символов. В качестве алгоритмов комбинирования используются операции теории нечётких множеств (такие как t-нормы и s-нормы), выбор наиболее уверенного эксперта.

Результатом работы классификатора является нечёткое множество, полученное в результате комбинирования на самом верхнем уровне. На последнем этапе принимается решение о наиболее правдоподобном варианте прочтения слова. Для этого используются уровни возможности прочтения

отдельных букв, меж-буквенной сегментации и частоты сочетаний букв в русском языке.

Оптическое распознавание символов (англ. optical character recognition, OCR) — это механический или электронный перевод изображений рукописного, машинописного или печатного текста в последовательность кодов, использующихся для представления в текстовом редакторе. Распознавание широко используется для конвертации книг и документов в электронный вид, для автоматизации систем учета в бизнесе или для публикации текста на веб-странице. Оптическое распознавание текста позволяет редактировать текст, осуществлять поиск слова или фразы, хранить его в более компактной форме, демонстрировать или распечатывать материал, не теряя качества, анализировать информацию, а также применять к тесту электронный перевод, форматирование или преобразование в речь. Оптическое распознавание текста является исследуемой проблемой в областях распознавания образов, искусственного интеллекта и компьютерного зрения.

Системы оптического распознавания текста требуют калибровки для работы с конкретным шрифтом; в ранних версиях для программирования было необходимо изображение каждого символа, программа одновременно могла работать только с одним шрифтом. В настоящее время больше всего распространены так называемые «интеллектуальные» системы, с высокой степенью точности распознающие большинство шрифтов. Некоторые системы оптического распознавания текста способны восстанавливать исходное форматирование текста, включая изображения, колонки и другие нетекстовые компоненты.

В 1929 году Густав Таушек получил патент на метод оптического распознавания текста в Германии, после чего за ним последовал Гендель, получив патент на свой метод в США в 1933. В 1935 году Таушек также

получил патент США на свой метод. Машина Таушека представляла собой механическое устройство, которое использовало шаблоны и фотодетектор.

В 1950 году Дэвид Х. Шепард, криптоаналитик из агентства безопасности вооружённых сил Соединённых Штатов, проанализировав задачу преобразования печатных сообщений в машинный язык для обработки компьютером, построил машину, решающую данную задачу. После того как он получил патент США, он сообщил об этом в «Вашингтон Дэйли Ньюз» (27 апреля 1951) и в «Нью-Йорк Таймс» (26 декабря 1953). Затем Шепард основал компанию, разрабатывающую интеллектуальные машины, которая вскоре выпустила первые в мире коммерческие системы оптического распознавания символов.

Первая коммерческая система была установлена на «Ридерс Дайджест» в 1955 году. Вторая система была продана компании «Стэндарт Ойл» для чтения кредитных карт для работы с чеками. Другие системы, поставляемые компанией Шепарда, были проданы в конце 1950-х годов, в том числе сканер страниц для национальных воздушных сил США, предназначенный для чтения и передачи по телетайпу машинописных сообщений. IBM позже получила лицензию на использование патентов Шепарда.

Примерно в 1965 году «Ридерс Дайджест» и «Ар-Си-Эй» начали сотрудничество с целью создать машину для чтения документов, использующую оптическое распознавание текста, предназначенную для оцифровки серийных номеров купонов «Ридерс Дайджест», вернувшихся из рекламных объявлений. Для печати на документах барабанным принтером «Ар-Си-Эй» был использован специальный шрифт OCR-A. Машина для чтения документов работала непосредственно с компьютером RCA 301 (один из первых массивных компьютеров). Скорость работы машины была 1500 документов в минуту: она проверяла каждый документ, исключая те, которые она не смогла обработать правильно.

Почтовая служба Соединённых Штатов с 1965 года для сортировки почты использует машины, работающие по принципу оптического распознавания текста, созданные на основе технологий, разработанных исследователем Яковом Рабиновым. В Европе первой организацией, использующей машины с оптическим распознаванием текста, был британский почтамт. Почта Канады использует системы оптического распознавания символов с 1971 года. На первом этапе в центре сортировки системы оптического распознавания символов считывают имя и адрес получателя и печатают на конверте штрих-код. Он наносится специальными чернилами, которые отчётливо видимы в ультрафиолетовом свете. Это делается, чтобы избежать путаницы с полем адреса, заполненным человеком, которое может быть в любом месте на конверте.

В 1974 году Рэй Курцвейл создал компанию «Курцвейл Компьютер Продактс», и начал работать над развитием первой системы оптического распознавания символов, способной распознать текст, напечатанный любым шрифтом. Курцвейл считал, что лучшее применение этой технологии — создание машины чтения для слепых, которая позволила бы слепым людям иметь компьютер, умеющий читать текст вслух. Данное устройство требовало изобретения сразу двух технологий — ПЗС планшетного сканера и синтезатора, преобразующего текст в речь. Конечный продукт был представлен 13 января 1976 во время пресс-конференции, возглавляемой Курцвейлом и руководителями национальной федерации слепых.

В 1978 году компания «Курцвейл Компьютер Продактс» начала продажи коммерческой версии компьютерной программы оптического распознавания символов. Два года спустя Курцвейл продал свою компанию корпорации «Ксерокс», которая были заинтересована в дальнейшей коммерциализации систем распознавания текста. «Курцвейл Компьютер Продактс» стала дочерней компанией «Ксерокс», известной как «Скансофт».

2.2 Текущее состояние технологии оптического распознавания текста

Точное распознавание латинских символов в печатном тексте в настоящее время возможно только если доступны чёткие изображения, такие как сканированные печатные документы. Точность при такой постановке задачи превышает 99%, абсолютная точность может быть достигнута только путем последующего редактирования человеком. Проблемы распознавания рукописного «печатного» и стандартного рукописного текста, а также печатных текстов других форматов (особенно с очень большим числом символов) в настоящее время являются предметом активных исследований.

Точность работы методов может быть измерена несколькими способами и поэтому может сильно варьироваться. К примеру, если встречается специализированное слово, не используемое для соответствующего программного обеспечения, при поиске несуществующих слов, ошибка может увеличиться.

Распознавание символов он-лайн иногда путают с оптическим распознавания символов. Последний — это офф-лайн метод, работающий со статической формой представления текста, в то время как он-лайн распознавание символов учитывает движения во время письма. Например, в он-лайн распознавании, использующем PenPoint OS или планшетный ПК, можно определить, с какой стороны пишется строка: справа налево или слева направо.

Он-лайн системы для распознавания рукописного текста «на лету» в последнее время стали широко известны в качестве коммерческих продуктов. Алгоритмы таких устройств используют тот факт, что порядок, скорость и направление отдельных участков линий ввода известны. Кроме того, пользователь научится использовать только конкретные формы письма. Эти методы не могут быть использованы в программном обеспечении, которое

использует сканированные бумажные документы, поэтому проблема распознавания рукописного «печатного» текста по-прежнему остается открытой. На изображениях с рукописным «печатным» текстом без артефактов может быть достигнута точность в 80 % — 90 %, но с такой точностью изображение будет преобразовано с десятками ошибок на странице. Такая технология может быть полезна лишь в очень ограниченном числе приложений.

Ещё одной широко исследуемой проблемой является распознавание рукописного текста. На данный момент достигнутая точность даже ниже, чем для рукописного «печатного» текста. Более высокие показатели могут быть достигнуты только с использованием контекстной и грамматической информации. Например, в процессе распознания искать целые слова в словаре легче, чем пытаться проанализировать отдельные символы из текста. Знание грамматики языка может также помочь определить, является ли слово глаголом или существительным. Формы отдельных рукописных символов иногда могут не содержать достаточно информации, чтобы точно (более 98%) распознать весь рукописный текст.

Для решения более сложных проблем в сфере распознавания используются как правило интеллектуальные системы распознавания, такие как искусственные нейронные сети.

§3. Программа FineReader.

3.1 Основные понятия. Распознавание, Проверка и редактирование, Сохранение полученного текста.

ABBYY FineReader — это система оптического распознавания текстов (OCR — Optical Character Recognition). Она предназначена для конвертирования в редактируемые форматы отсканированных документов, PDF-документов и файлов изображений, включая цифровые фотографии.
Преимущества программы ABBYY FineReader – это скорость и высокая точность распознавания.

Используемая в ABBYY FineReader система оптического распознавания быстро и точно распознает и максимально полно сохраняет исходное оформление документа, в том числе с текстом на фоне картинок, с цветным текстом на цветном фоне, с обтеканием картинок текстом и т.д.

Благодаря технологии адаптивного распознавания документов ADRT® (Adaptive Document Recognition Technology) ABBYY FineReader позволяет анализировать и обрабатывать документ целиком, а не постранично. В результате восстанавливается исходная структура документа, включая форматирование, гиперссылки, адреса электронной почты, а также колонтитулы, подписи к картинкам и диаграммам, номера страниц и сноски.

ABBYY FineReader распознает документы, написанные на одном или нескольких из 189 языков, включая арабский, вьетнамский, корейский, китайский, японский, тайский и иврит. В программу встроена функция автоматического определения языка документа.

Еще одной особенностью программы ABBYY FineReader является малая чувствительность к дефектам печати и способность распознавать тексты, набранные практически любыми шрифтами.

Программа включает широкий спектр работы с результатами распознавания — документы можно сохранять в различных форматах, отправлять по электронной почте, а также передавать в другие приложения для дальнейшей обработки.

ABBYY FineReader имеет простой и интуитивно понятный интерфейс, который позволяет работать с программой без дополнительной подготовки, освоив основные операции в самые короткие сроки. Поддерживаемые программой языки интерфейса можно переключать непосредственно из программы.

Встроенные задачи программы охватывают список наиболее часто используемых задач по конвертированию отсканированных документов, PDF и файлов изображений в редактируемые форматы и позволяют получить электронный документ одним нажатием кнопки.

Благодаря интеграции ABBYY FineReader с Microsoft Office и Проводником Windows, вы можете распознать документ непосредственно при работе с Microsoft Outlook, Microsoft Word, Microsoft Excel и Проводником Windows.

Программа имеет встроенную справку, содержащую примеры использования ABBYY FineReader для решения сложных задач конвертирования.

Если вам не нужно сохранять цветовое оформление документов, вы можете распознавать документы на 30% быстрее с помощью нового черно-белого режима. Кроме того, программа эффективно использует возможности многоядерных процессоров, что позволяет еще больше увеличить скорость обработки документов.

ABBYY FineReader позволяет сканировать бумажные книги и конвертировать их в форматы EPUB и FB2, которые широко используются для создания электронных книг. Вы сможете читать их на вашем iPad, планшете или другом портативном устройстве. Или отправьте результаты распознавания на свой адрес на сервере Kindle.com. Конвертируйте бумажные книги и статьи в нужный формат электронной книги, чтобы добавить в свою электронную библиотеку или архив.

Новая версия программы распознает и конвертирует изображения документов и PDF-файлы в формат OpenOffice.org Writer (ODT), точно

сохраняя исходное оформление и форматирование. Теперь вы без лишних усилий можно работать с документами в формате .odt или добавлять их в архив.

Усовершенствованный редактор стилей позволяет настраивать все параметры стилей в одном удобном диалоге. Все изменения происходят сразу во всем документе.

Вы можете рассортировать изображения страниц по нескольким документам FineReader для более точного сохранения оформления исходных документов.

Распознавать документы стало еще проще с помощью еще более легкого доступа ко всем базовым и пользовательским задачам распознавания. Улучшенное распознавание фотографий и новые инструменты для редактирования изображений.

ABBYY FineReader предлагает широкий диапазон новых мощных инструментов для редактирования изображений, включая настройку яркости, контрастности и уровней интенсивности света и тени, которая позволяет значительно улучшить исходное изображение и получить более точные результаты распознавания.

Улучшено определение стилей документа, текста на полях страницы, колонтитулов и заголовков, что позволяет существенно уменьшить время, необходимое для редактирования распознанных документов.

В нашей жизни становится все больше электронных документов. Тем не менее, деловые бумаги, журналы, книги по прежнему в ходу. Миллионы людей разных профессий используют ABBYY FineReader для перевода бумажных документов в электронный вид. Ведь в современном стремительном мире успеха добивается тот, кто умеет эффективно использовать свое время и может управлять информацией независимо от того, в каком виде она представлена.

Программа позволяет получить электронный документ одним нажатием, не вдаваясь в подробности работы программы. Встроенные сценарии предусматривают основные задачи по конвертированию PDF–документов, сканированию и распознаванию текстов и изображений

В программе ABBYY FineReader используются новейшая разработка компании ABBYY – технология Document OCR. Внедрение инновационной технологии Document OCR в программу позволило продвинуться далеко вперед в системах оптического распознавания. Теперь ABBYY FineReader проводит целостный анализ многостраничного документа. В результате сохраняется его логическая структура и восстанавливается не только основной текст документа, но и оформление: колонки, колонтитулы, шрифты, стили, сноски, нумерованные подписи к рисункам и таблицам. Полученный документ легко редактировать и использовать.

Существенные изменения внесены в технологию распознавания шрифтов. Теперь ABBYY FineReader определяет шрифт исходного документа и подбирает наиболее близкий к нему шрифт. Основываясь на новейших технологиях, ABBYY FineReader автоматически определяет языки, которые используются в документе, что существенно упрощает работу с программой. В наши дни в аппаратных средствах все чаще используются многоядерные процессоры. Все больше и больше компьютеров оснащены двух– или четырех ядерными процессорами. ABBYY FineReader, используя все возможности многоядерного процессора, позволяет без потери качества и времени одновременно выполнять различные шаги по обработке документов.

3.2 Характеристики исходного документа. Особенности получения изображения

Переданное в программу ABBYY FineReader изображение необходимо распознать и преобразовать в текст. Прежде, чем приступить к распознаванию, программа выделяет на изображении области с текстом, картинки, таблицы и штрих–кодами. Распознавание страниц, добавленных в

документ ABBYY FineReader, выполняется в автоматическом режиме с текущими настройками программы. Это позволяет работать с программой не дожидаясь распознавания всех страниц документа.

На качество полученного текста влияет правильно выбранный язык распознавания, режим распознавания, тип печати распознаваемого текста. Выполните распознавание вручную, если вы выделили области на изображении вручную или изменили следующие параметры в диалоге Опции

- язык распознавания на закладке *Документ*;
- тип печати на закладке *Документ*;
- настройки распознавания на закладке 2. *Распознать*;
- используемые шрифты на закладке *Дополнительные*;

Для запуска распознавания в ручную, нажимаем кнопку *Распознать* в окне *Изображение*, или в меню *Документ* выбираем пункт *Распознать* документ. Вы можете распознать все добавленные в документ ABBYY FineReader страницы. Для этого нажмите на стрелку справа от кнопки, в открывшемся меню выберите *Распознать документ*.

Результат распознавания отображается в окне *Текст*. В данном окне неуверенно распознанные символы выделяются цветом. Таким образом, вы легко заметите возможные ошибки, и их исправление не займет много времени. Вы можете редактировать полученный результат как в окне *Текст*, так и с помощью диалога *Проверка*.

Для того чтобы просмотреть неуверенно распознанное слово:

1. Щелкните на слово в окне *Текст*.

В окне *Изображение* показывается местоположение данного слова на странице, а в окне *Крупный план* под курсором показывается его увеличенное изображение.

2. Внесите изменения в случае необходимости.

Данный метод удобен для сравнения исходного и полученного документов. Программа ABBYY FineReader позволяет проверить неуверенно распознанные слова с помощью встроенного диалога проверки орфографии Используя данный диалог, вы можете просматривать неуверенно распознанные слова, находить орфографические ошибки, добавлять в словарь новые слова, изменять язык словаря.

ABBYY FineReader также позволяет редактировать оформление документа. Вы можете редактировать полученные результаты в окне *Текст* с помощью кнопок, расположенных на панели инструментов или панели *Свойства текста* (контекстное меню окна Текст>Свойства).

В процессе распознавания в документе выделяются стили. Все выделенные стили отображаются на панели *Свойства текста*. Редактируя стили, вы можете легко изменять форматирование, применяемое к тексту. При сохранении текста в формат RTF/DOC/WordML/DOCX все используемые стили сохраняются.

Результаты распознавания можно сохранить в файл, передать в указанное приложение, скопировать в буфер обмена или отправить по электронной почте в любом из поддерживаемых программой ABBYY FineReader форматов сохранения. Сохранить можно все страницы документа ABBYY FineReader или только выбранные. Для того чтобы результат максимально соответствовал вашим ожиданиям, следует внимательно отнестись к выбору опций сохранения.

Как сохранить распознанный текст:

1. В окне *Текст* нажмите стрелку справа от кнопки *Сохранить* и в выпадающем списке выберите необходимый пункт;

2. На панели инструментов окна *Текст* в выпадающих списках выберите:

- Формат сохранения документа;
- Режим сохранения оформления документа.
- Точная копия

Позволяет получить документ, оформление которого будет полностью соответствовать оригиналу. Рекомендуется использовать для документов сложного оформления, например, рекламных брошюр. Однако данный режим не предполагает внесение значительных правок в текст и оформление.

- Редактируемая копия

Позволяет получить документ, оформление которого может незначительно отличаться от оригинала. Документ, полученный с помощью данного режима, легко редактируется.

- Форматированный текст

В полученном документе сохранятся начертание и размер шрифта, разбиение на абзацы, но не сохраняется расположение объектов на странице и межстрочные интервалы. Таким образом, будет получен сплошной текст с выравниванием по левому краю.

- Простой текст

В данном режиме, в отличие от режима *Форматированный текст*, не сохраняется размер шрифта. В остальном оформление будет таким же.

- Опции…

Позволяет изменить настройки сохранения выбранного формата. В открывшемся диалоге *Опции* отметьте необходимые опции и нажмите кнопку *ОК*.

Список возможных режимов зависит от выбранного формата.

3. Нажмите к нопку *Сохранить*.

Качество распознавания во многом зависит от качества исходного изображения.

Тип печати

Документ может быть напечатан на различных устройствах, например, на пишущей машинке или матричном принтере. Качество распознавания таких документов может быть различным.

При распознавании текстов, напечатанных на матричном принтере в черновом режиме или на пишущей машинке, можно добиться более высокого качества распознавания, установив правильный *Тип печати*. Для большинства текстов тип печати определяется автоматически. Этому соответствует значение *Авто*, установленное в группе *Тип печати* документа в диалоге *Опции* (*меню Сервис>Опции…>закладка Документ*). При необходимости, вы можете выбрать другой тип печати в этой группе.

	Фрагмент страницы, напечатанной на пишущей машинке. Ширина букв одинакова (сравните, например, буквы "w" и "t"). Для таких текстов установите значение Пишущая машинка.
	Фрагмент страницы, напечатанной на матричном принтере. На картинке видно, что штрихи букв состоят из отдельно стоящих точек. Для таких текстов установите значение Матричный принтер.

Если вы распознаете программную распечатку, отметьте опцию *Распознать* как форматированный пробелами текст в группе *Тип печати* документа. В этом случае в распознанном тексте сохранится деление на строки; отступы от левого края будут переданы пробелами; каждая строка выделена в отдельный абзац, а расстояния между абзацами переданы пустыми строками. Все это позволит сохранить исходное форматирование текста при сохранении в формате txt.

Качество печати

Качество распознавания может существенно снизиться, если исходный документ напечатан в плохом качестве, то есть содержит много "мусора".

Для успешного распознавания документа, отпечатанного с плохим качеством, может потребоваться изменить настройки сканирования. Такой документ может содержать много "мусора", нечеткие границы букв, угловатые, неровные буквы с дефектами, перекос строк, смещение и неявные границы черных разделителей таблиц.

Подобные документы рекомендуется сканировать в сером режиме. При сканировании в сером режиме вам не нужно будет подбирать яркость сканирования, программа сделает это за вас автоматически. Серый тип изображения обеспечивает более высокую степень сохранения информации о буквах сканируемого текста. Это приводит к улучшению качества распознавания документов среднего и низкого качества печати. Вы также можете устранить некоторые дефекты вручную, используя инструменты по обработке изображения в окне *Редактировать изображение*.

Язык документа

Текст исходного документа может быть написан на нескольких языках. ABBYY FineReader поддерживает распознавание как одноязычных, так и многоязычных (например, англо–французских) документов. Для распознавания многоязычного документа необходимо выбрать несколько языков распознавания.

Чтобы выбрать язык для распознавания, в выпадающем списке *Язык* документа окна *Документ* выберите один из пунктов:

- Авто.

Язык будет выбираться автоматически из задаваемого списка словарных языков. Вы можете изменить состав данного списка. Для этого:

1. Нажмите с сылку *Выбор языков…* В результате откроется диалог *Редактор языков*.

2. Убедитесь, что включена опция *Автоматически выбирать язык распознавания* из списка включена.

3. Нажмите кнопку *Указать…*

4. В д иалоге Список языков отметьте необходимые языки.

- Язык или сочетание языков.

Выберите один из предложенных вариантов. Список языков включает в себя часто употребляемые языки на компьютере пользователя, а также английский, немецкий язык и французский языки.

- Выбор языков…

Выберите данный пункт, если вы хотите выбрать другие языки для распознавания. В открывшемся диалоге *Редактор языков* отметьте опцию Указать языки распознавания вручную и укажите один или несколько языков. Для этого отметьте пункты с соответствующими названиями языков. Если вы часто используете какую–либо комбинацию языков, то создайте новую группу, содержащую эти языки. Для того чтобы загрузить недостающие языки, в меню *Пуск>Программы>ABBYY FineReader* выберите команду *Загрузить* больше языков и следуйте инструкциям программы.

Выбор режима сканирования

В программе ABBYY FineReader возможны следующие варианты взаимодействия программы со сканерами:

- через интерфейс ABBYY FineReader;

В этом случае для настройки опций сканирования используется диалог программы ABBYY FineReader. Он позволяет устанавливать разрешение, яркость и тип изображения. Кроме этого здесь доступны такие функции как:

- сканирование многостраничных документов на сканерах без автоподатчика;

- автоматическое двустороннее сканирование (если данная возможность поддерживается сканером).

Для некоторых моделей сканеров опция *Использовать интерфейс ABBYY FineReader* может быть недоступна.

- через интерфейс TWAIN–драйвера сканера или WIA–драйвера сканера.

Для настройки опций сканирования используется диалог драйвера сканера.

По умолчанию сканирование выполняется через интерфейс драйвера сканера.

Вы можете легко переключаться между этими режимами:

1. Откройте диалог *Опции* на закладке 1. *Сканировать/Открыть* (*меню Сервис>Опции…*);

2. В группе *Сканер* установите переключатель в одно из положений: *Использовать интерфейс ABBYY FineReader* или *Использовать интерфейс сканера*.

Если в исходном изображении мелкий шрифт

Для успешного распознавания текста с мелким шрифтом необходимо отсканировать документ с более высоким разрешением.

1. Нажмите к нопку *Сканировать*;

2. В открывшемся диалоге укажите разрешение.

В зависимости от используемого режима сканирования откроется диалог *ABBYY FineReader* или диалог драйвера сканера.

3. Отсканируйте изображение.

Сравните результаты сканирования одного и того же документа с разными значениями разрешения. Для этого просмотрите полученные изображения в окне *Крупный план* в масштабе *С точностью до пикселя* (*меню Вид>Окно Крупный план>Масштаб*):

Особенности входного изображения	Рекомендуемое разрешение

	300 dpi – для обычных текстов (размер шрифта 10 и более пунктов).
	400–600 dpi – для текстов, набранных мелким шрифтом (9 и менее пунктов).

Настройка яркости сканирования

Если яркость сканирования была подобрана неверно, при распознавании возникнет сообщение о необходимости изменить яркость сканирования. Для сканирования некоторых документов в черно–белом режиме может понадобиться дополнительная настройка яркости.

Как изменить яркость:

1. Нажмите к нопку *Сканировать*;

2. В открывшемся диалоге укажите яркость.

В зависимости от используемого режима сканирования откроется диалог *ABBYY FineReader* или диалог *драйвера сканера*. В большинстве случаев подходит среднее значение яркости – 50%.

3. Отсканируйте изображение.

Если в полученном изображении вы обнаружили большое количество дефектов (разрывов или склеек букв), то обратитесь к таблице, приведенной ниже. В ней указаны возможные способы их устранения.

Особенности входного изображения	Рекомендации
	Пример хорошего (пригодного для распознавания) изображения
 "разорванные"; светлые, тонкие буквы	● Уменьшите яркость чтобы изображение стало темнее; ● Отсканируйте в сером. В этом случае осуществляется автоподбор яркости
 искаженные и залитые; склеенные символы; темные, толстые буквы	● Увеличьте яркость сделать изображение светлее; ● Отсканируйте в сером. В этом случае осуществляется автоподбор яркости.

§4. Методы распознавания программы FineReader

4.1 Автоматическое распознавание текстов

После сканирования документа получается графическое изображение исходного документа. Такое графическое изображение представляет собой набор разноцветных точек и редактированию в программах, предназначенных для обработки текстовых документов не подлежит. Программа FineReader решает проблему распознавания текста в составе точечного графического изображения.

Окно программы содержит строку меню, ряд панелей инструментов и рабочую область.

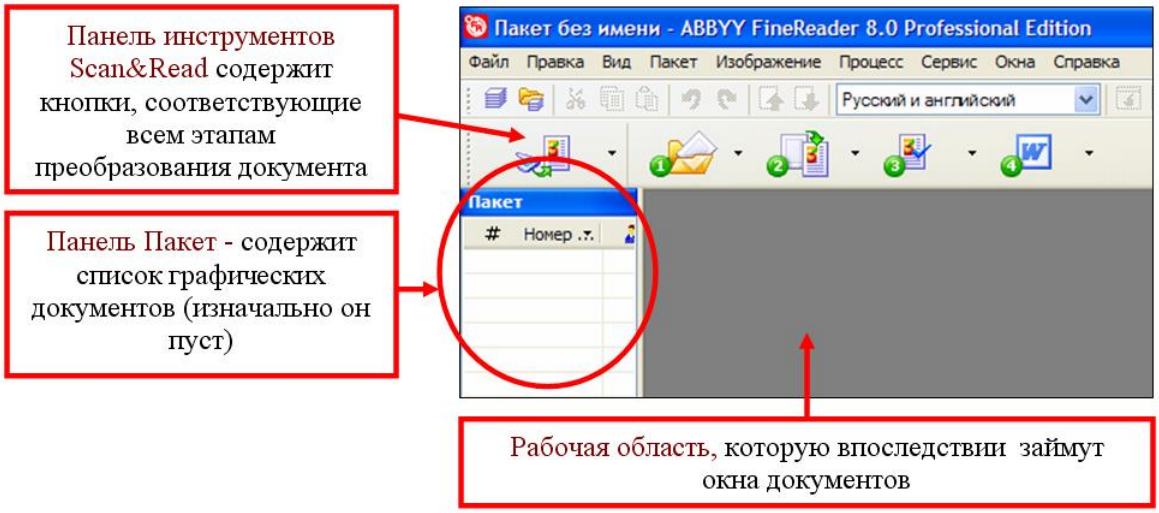


Рисунок 1 – Порядок распознавания текстовых документов

Преобразование бумажного документа в электронный происходит поэтапно или автоматически. Для автоматической работы используется инструмент Scan&Read.

- Первый этап работы – сканирование.

Если документ был уже отсканирован ранее, его открывают. Если изображение находится на бумажном носителе, то на первом этапе выбирают действие сканировать.

Программа FineReader использует для сканирования устройство, заданное по умолчанию. По завершении процесса сканирования полученное

графическое изображение автоматически выгружается в рабочую область программы FineReader.

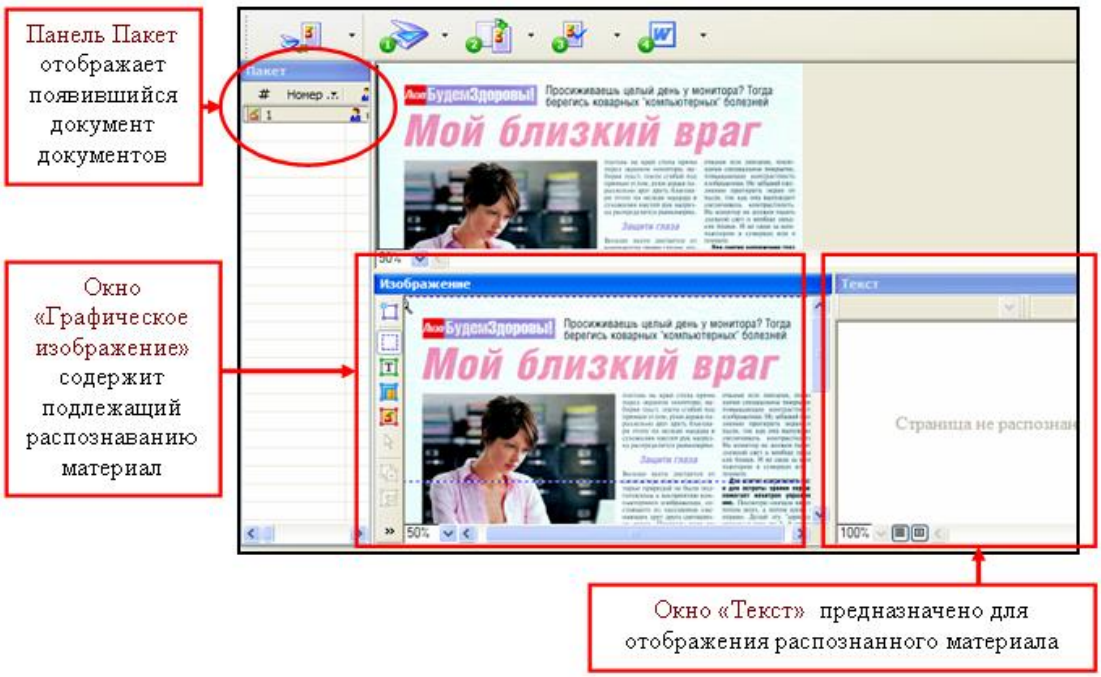


Рисунок 2 - Программа FineReader

Выполните первый этап – сканирование документа.

Второй этап – распознавание текста.

Прежде чем включать текст в документ, он разбивается на блоки, содержащее цельные фрагменты. Эту операцию программа может выполнить автоматически, хотя разбиение не всегда проходит удачно.

Границы и типы блоков можно устанавливать вручную.

Процесс распознавания отображается в специальном информационном окне:

- Полученный текст помещается в окно «Текст».

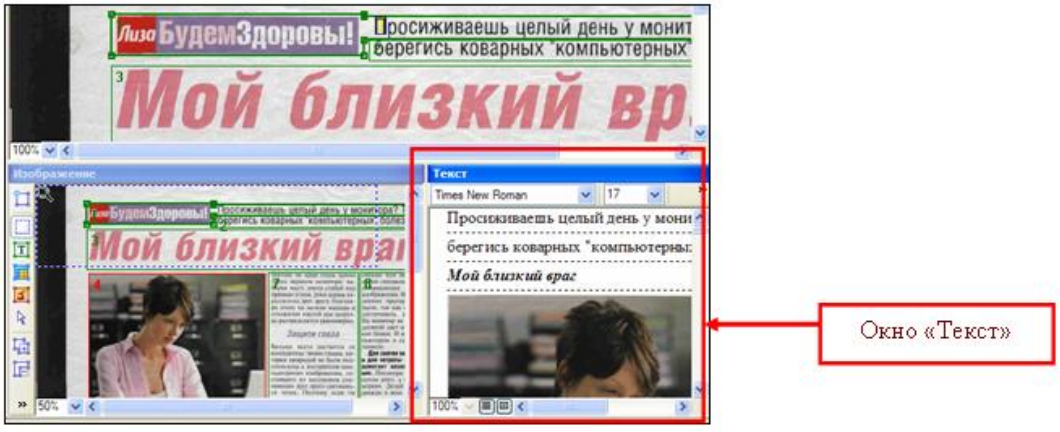


Рисунок 3 – Тестовое окно FineReader

Второй этап – автоматическое распознавание отсканированного документа.

Третий этап - проверка.

На данном этапе программа выполняет поиск ошибок распознавания. FineReader выделяет цветом те символы, которые она сама рассматривает как неоднозначно опознанные.

С помощью диалогового окна Проверка можно отредактировать нераспознанные символы.

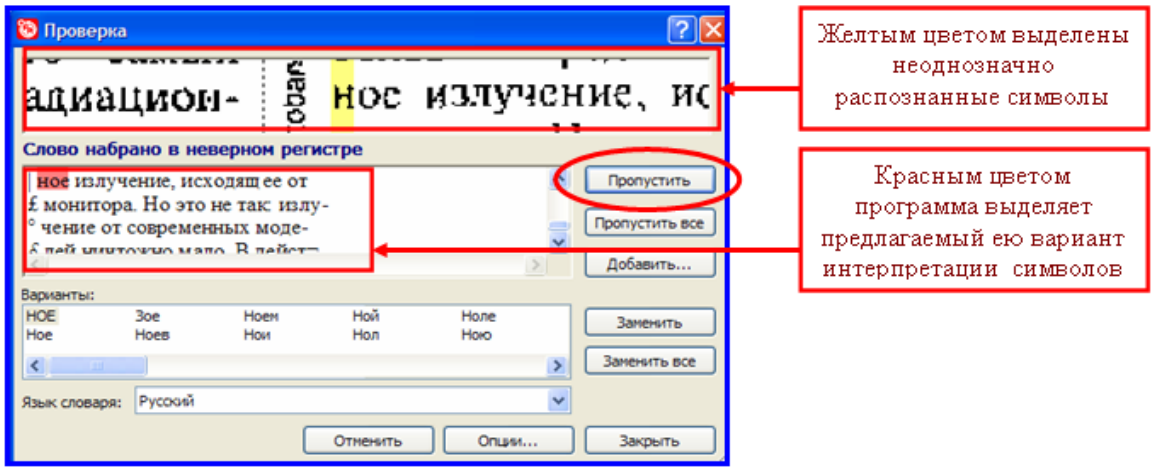


Рисунок 4 –Процесс распознавания

Если вариант интерпретации программы верный нажимаем кнопку Пропустить.

В случае обнаружения символов неверно распознанных программой ошибки исправляют вручную и фиксируют исправления нажатием кнопки Подтвердить.

Четвёртый этап – сохранение текста.

Программа FineReader предусматривает возможность прямой передачи полученного текста в Word:

Сохранение текстового документа выполняют в программе Word.

Сегментация текста на этапе распознавания.

При автоматической сегментации программа разбивает отсканированный документ на блоки различных типов: текстовые, графические и т. д.

Если исходный текст содержит рисунки, подрисуночные подписи, таблицы, примечания и другие элементы, автоматическое распознавание текста может пройти неудачно.

В таких случаях границы блоков указывают вручную. Для этого используют кнопки специальной панели инструментов Изображение.

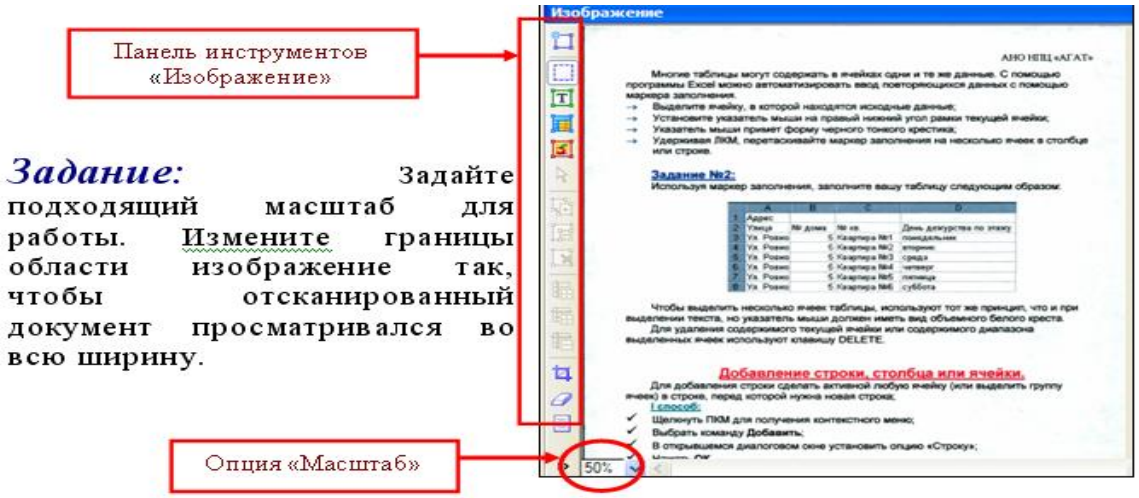


Рисунок 5 – Настройка изображения

Кнопки панели инструментов Изображение соответствуют различным типам блоков.

Блоки выделяются прямоугольными рамками различных цветов. Чтобы выделить блок необходимо:

1. Выбрать соответствующую кнопку панели инструментов;

2. Протягиванием определить границы блока.

Анализ макета страницы - выполняет автоматическое разбиение на блоки.

Выделить зону распознавания – позволяет выбрать щелчком мыши тот или иной блок, если автоматическое разбиение на блоки уже выполнено, и

определить зону для автоматического разбиения методом протягивания, если оно ещё не выполнено.

Выделить блок Картинка. Ластик - удаляет фрагмент отсканированного документа. Обрезка - позволяет вырезать любой фрагмент документа.

4.2 Распознавание с обучением

Режим "*Распознавание с обучением*" используется для:

- распознавания текстов, для набора которых использованы декоративные шрифты;
- распознавания текстов, в которых встречаются специальные символы (например, математические символы);
- распознавания большого объема (более 100 страниц) текста плохого качества.

В других случаях распознавание с обучением использовать не рекомендуется, так как затраты на обучение будут больше, чем полученный выигрыш в качестве распознавания.

Создание и обучение эталона

1. Откройте диалог *Опции* (*меню Сервис>Опции…*) на закладке 2. *Распознать*;

2. В группе *Обучение* установите переключатель в положение *Распознавание с обучением*;

3. Нажмите кнопку *Эталоны…*;

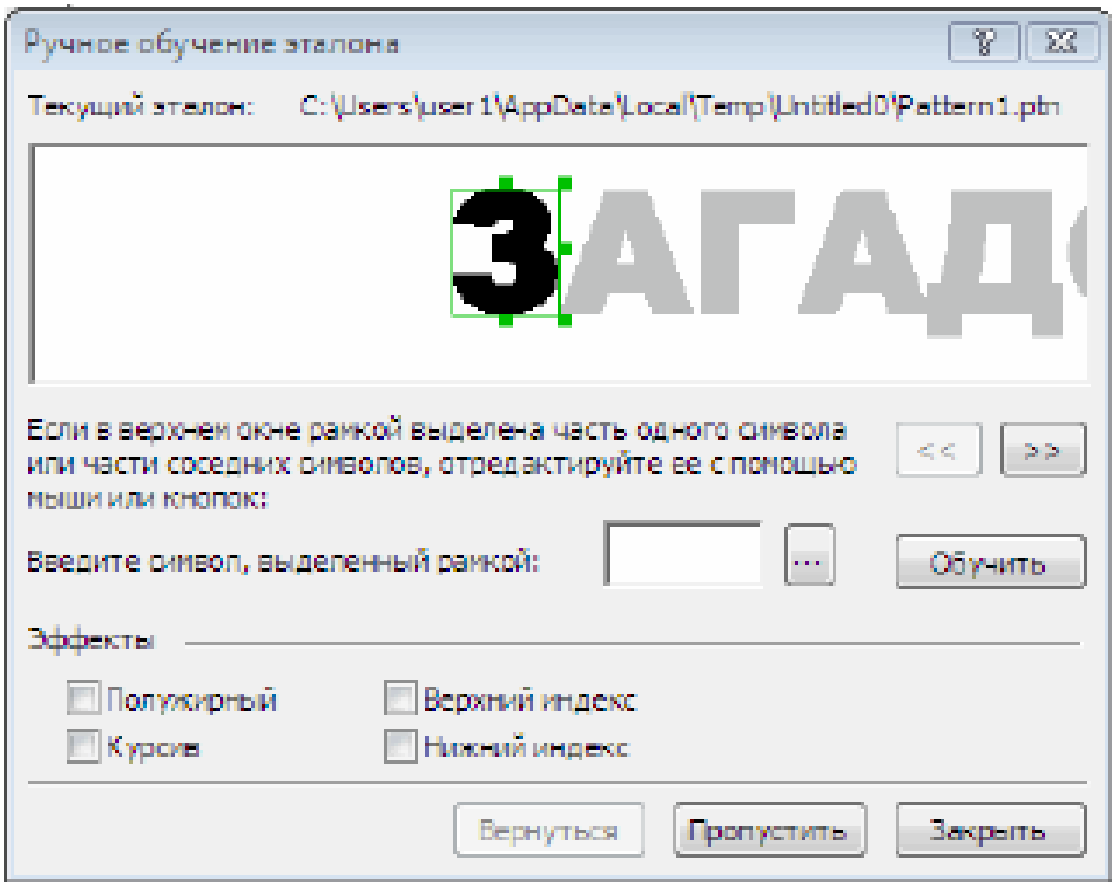
4. В открывшемся диалоге *Редактор эталонов* нажмите кнопку *Новый…*;

5. В открывшемся диалоге *Создать эталон* введите имя эталона и нажмите *ОК*;

6. Нажмите к нопку *Закрыть* в диалоге *Редактор эталонов*, затем кнопку *ОК* в диалоге *Опции*;

7. В окне *Изображение* нажмите кнопку *Распознать*;

Если в процессе распознавания встретится неизвестный символ, откроется диалог *Ручное обучение эталона* с изображением этого символа.



8. Обучите э талон *символам* или *лигатурам*.

Лигатуры – это сочетания двух или трех символов, которые из–за особенностей их начертания невозможно разделить при обучении и которые поэтому сразу обучаются как комбинации символов. Обучение лигатурам происходит так же, как и обучение отдельным символам.

Если вам важно в распознаваемом тексте сохранить начертание шрифта, верхний или нижний индексы, отметьте соответствующие опции в группе *Эффекты*. В процессе обучения вы можете вернуться к редактированию предыдущего символа. Для этого нажмите кнопку *Вернуться*. В этом случае охватывающий прямоугольник вернется на предыдущую позицию, а последняя обученная пара "изображение –символ" будет удалена из эталона. Кнопка *Вернуться* действует в пределах одного слова.

Выбор эталона для работы

Программа ABBYY FineReader позволяет использовать эталоны для более качественного распознавания документов.

1. В меню *Сервис* выберите пункт *Редактор эталонов*…

2. В открывшемся диалоге *Редактор эталонов* из списка существующих эталонов выберите нужный и нажмите кнопку *Выбрать*.

При работе с эталонами существуют следующие особенности:

1. Изображения некоторых символов не различаются системой распознавания и сопоставляются с каким–то одним символом. Например, прямой ('), левый (‘) и правый (’) апострофы хранятся в эталоне как изображение прямого апострофа. Таким образом, в результате распознавания в тексте никогда не появится правый или левый апостроф, хотя при обучении были указаны именно эти символы.

2. Для некоторых изображений решение о том, какому символу в распознанном тексте сопоставить встретившееся конкретное изображение, принимается на основе общего анализа распознанного текста. Так, например, решение о том, является ли символ, обозначаемый "кружком", буквой о или цифрой ноль, система принимает в зависимости от того, находятся ли рядом другие цифры или буквы.

3. Созданный эталон можно использовать только для распознавания текстов, использующих тот же шрифт и размер шрифта и отсканированных с тем же разрешением, что и документ, на котором данный эталон создавался.

4. Вы можете сохранить созданный эталон для работы с другими документами ABBYY FineReader. Для этого сохраните настройки документа ABBYY FineReader в файл набора опций (*.fbt).

5. При переходе к распознаванию текстов, напечатанных другим шрифтом, не забудьте отключить эталон. На закладке 2. *Распознать* диалога Опции установите переключатель в положение *Не использовать пользовательский эталон*.

Редактирование эталона

Прежде чем запускать распознавание с только что созданным эталоном, рекомендуется просмотреть эталон и, если потребуется, отредактировать его. Этим вы сведете к минимуму ошибки распознавания, которые могут возникнуть из–за неправильно обученного эталона. Эталон должен содержать только целые символы или лигатуры. Символы, обрезанные с краев, и символы с неправильными буквенными соответствиями следует удалить из эталона.

1. В меню *Сервис* выберите пункт *Редактор эталонов*….

2. В открывшемся диалоге *Редактор эталонов* выберите нужный эталон и нажмите кнопку *Редактировать*…

3. В открывшемся диалоге *Символы пользовательского эталона* выберите символ и нажмите кнопку *Свойства*…

В открывшемся диалоге:

- в поле *Символ* введите букву, которая соответствует символу;

- в поле *Эффекты* укажите правильное начертание: курсив, полужирный, верхний или нижний индексы.

Чтобы удалить неправильно обученные символы нажмите кнопку *Удалить* в диалоге *Символы пользовательского эталона.*

§5. Подготовка программы и распознавание текста

Что бы запустить программу ABBYY FineReader, нажимаем на «***Пуск - Все Программы - ABBYY FineReader 11 - ABBYY FineReader 11 Professional Edition***»

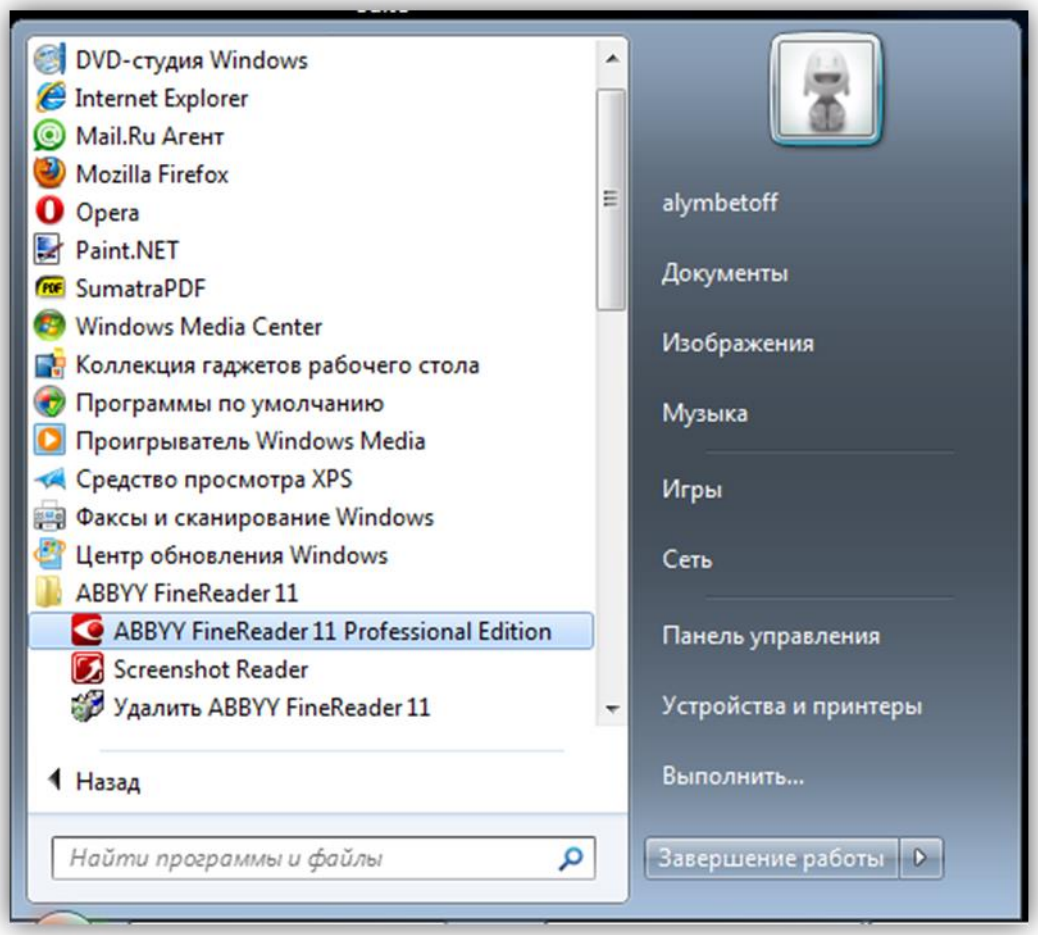


рис.1

Далее программа FineReader открывается и вы можете видеть интерфейс программы.

Перед тем как открыть файл, нужно выбрать «***Язык документа***» и «***Цветовой режим***». По умолчанию «***Язык документа***» будет стоять «***Автовыбор***». Язык можно изменить на тот в котором находится конвертируемый файл, так программа точно будет знать язык документа, что уменьшит и время обработки и вероятность ошибок некоторых слов в распознавании. «***Цветовой режим***» можно выбрать по необходимости, если

файл не содержит цветные оформления то можно оставить режим черно-белым.

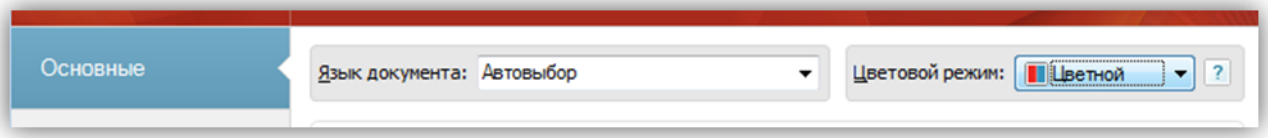


рис. 2

После настройки программы можно приступить к открытию файла. Что бы открыть файл который Вы хотите конвертировать в формат документа MS Office, нужно щелкнуть на кнопку «***Открыть***» в верхнем левом углу программы или нажать комбинацию клавиш «***Ctrl+O***».

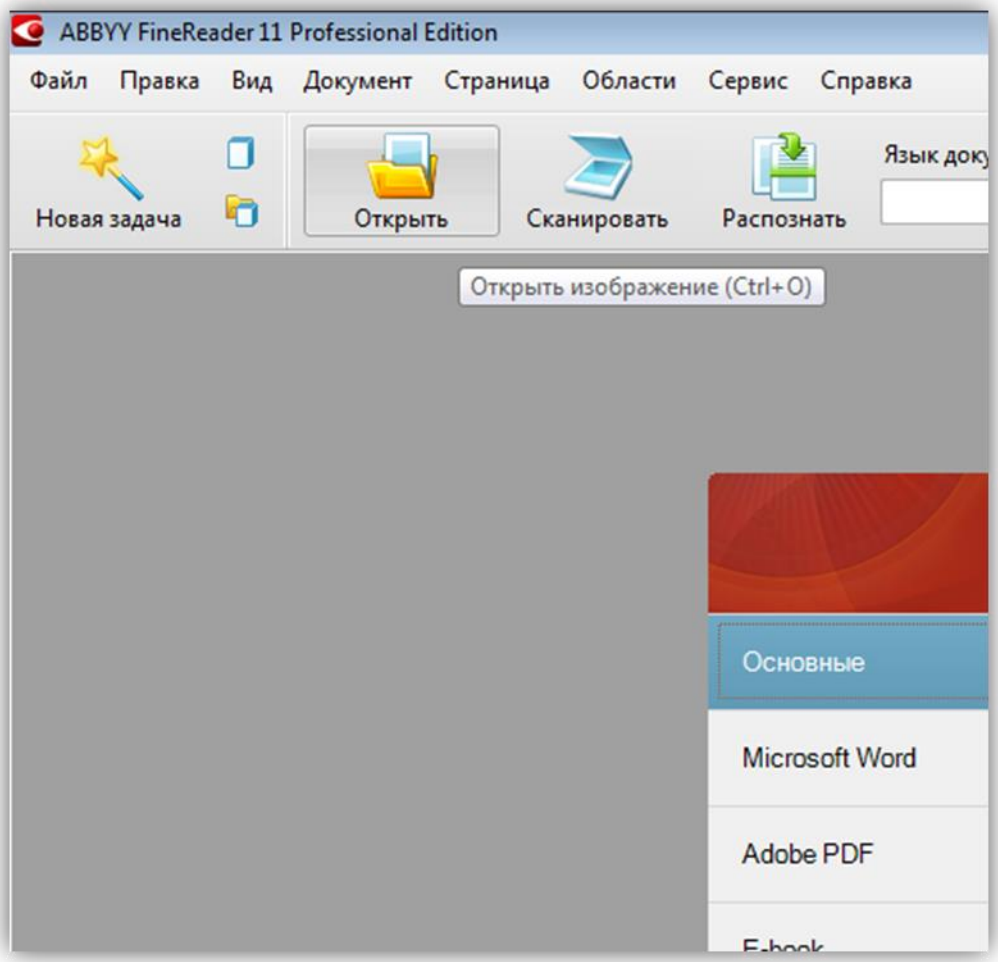


рис. 3

После нажатия на кнопку «***Открыть***», появляется окно выбора, находим и открываем нужный нам файл. Это могут быть как файлы изображений формата BMP, PNG, TIFF, GIF, JPEG, PCX так и документы формата PDF, DJVU.

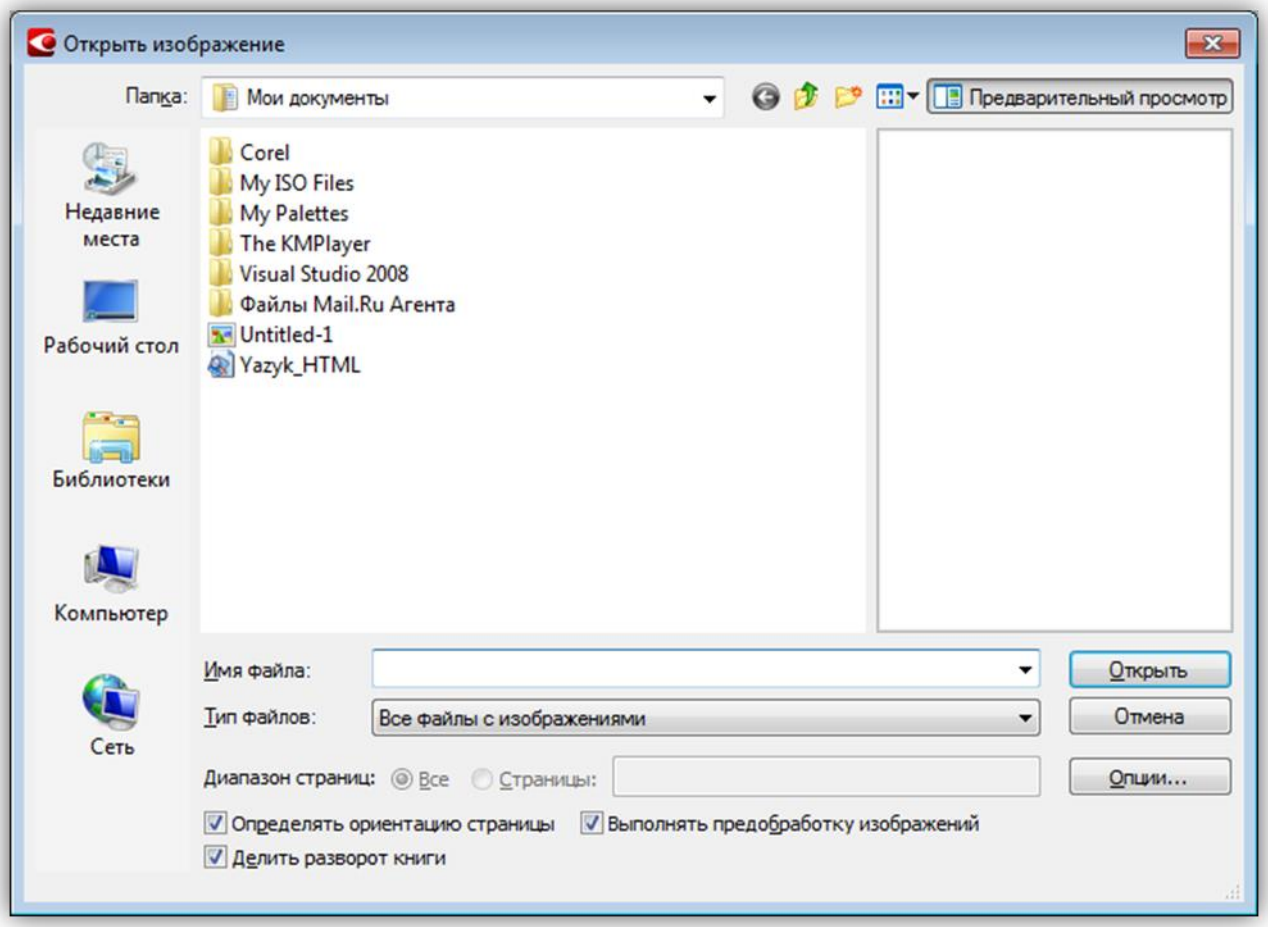


рис.4

Выбираем необходимый нам файл и открываем. После открытия программа автоматически приступит к распознаванию текста из документа или изображения.

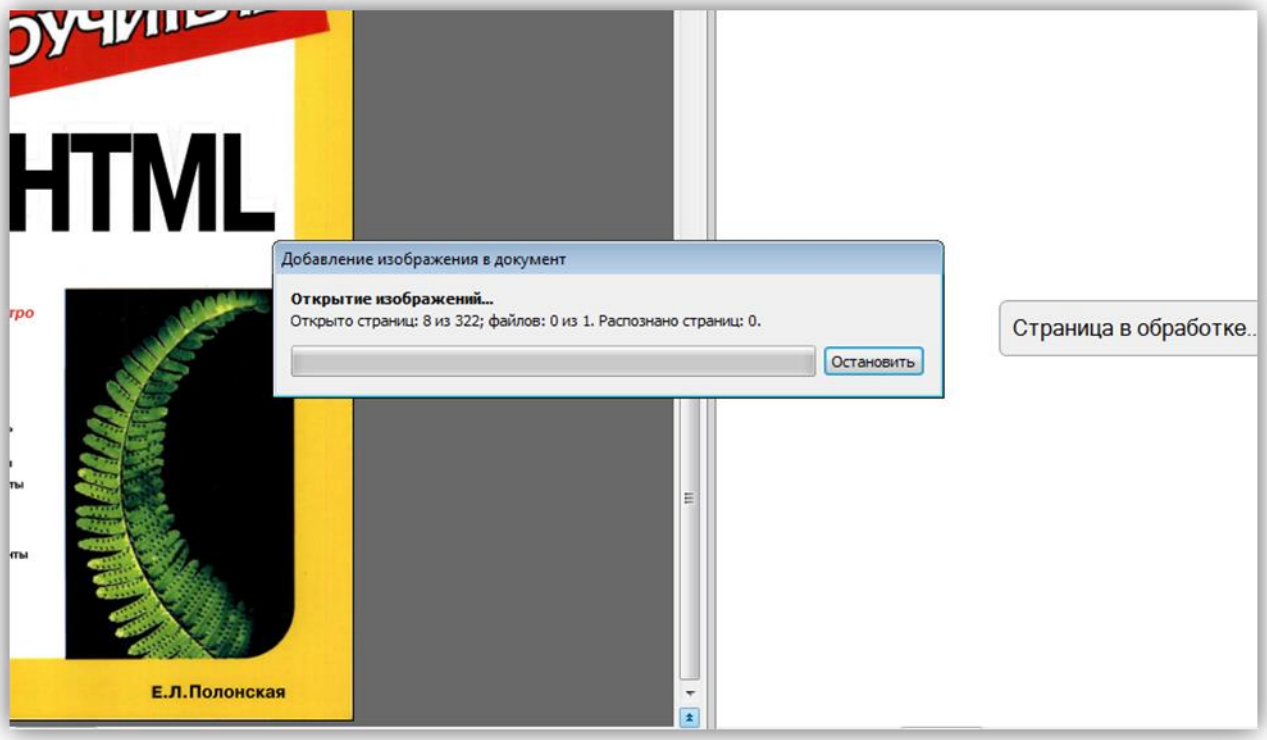


рис. 5

При этом программа будет сообщать об ошибках распознавания.

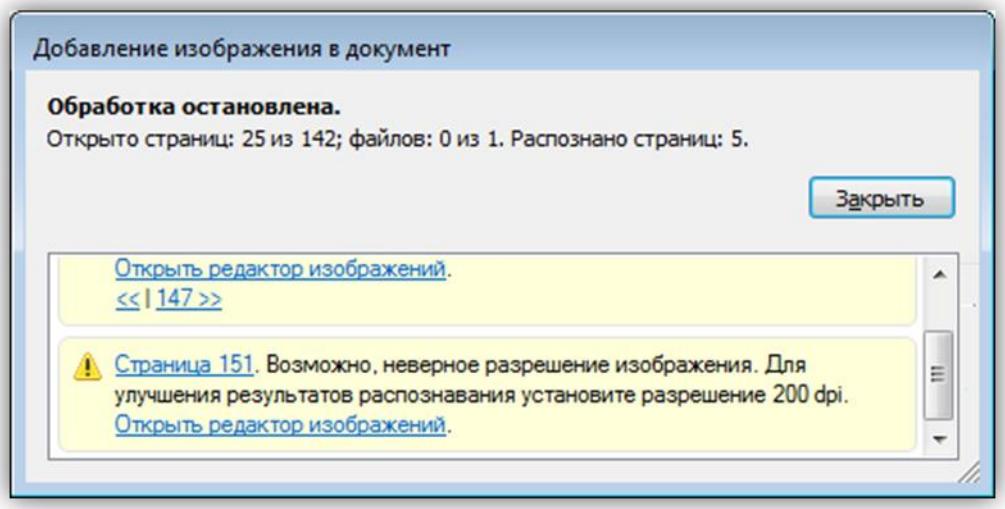


рис. 6

Если появляется сообщение «***Возможно, неверное разрешение изображения***» то следует «***Открыть редактор изображений***»

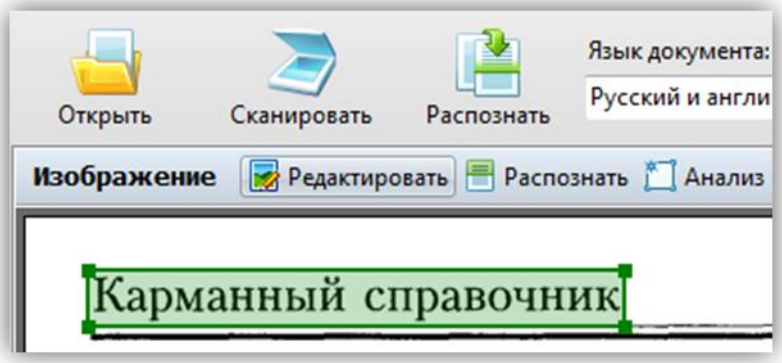


рис. 7

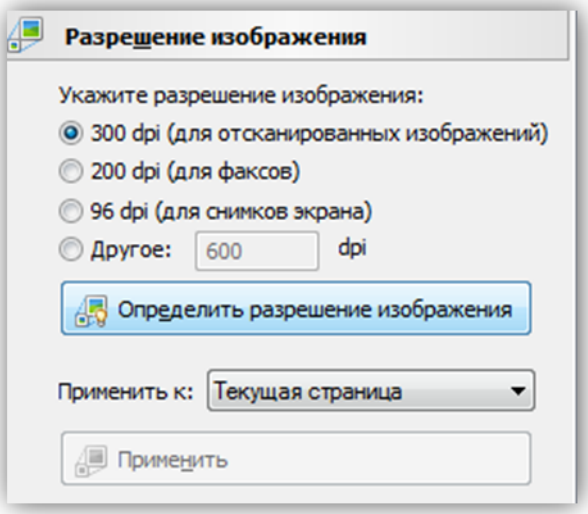


рис. 8

После перехода в режим редактирования, в правом углу нажмите на кнопку «***Определить разрешение изображение***». Программа автоматически определит необходимое для распознавания разрешение. Нажимаем на кнопку «***Применить***» и выходим из режима редактирования изображения.

Так же стоит обратить внимание на изображения в документе. Если программа видит их как «***Текст»*** то следует изменить их на «***Картинка***».

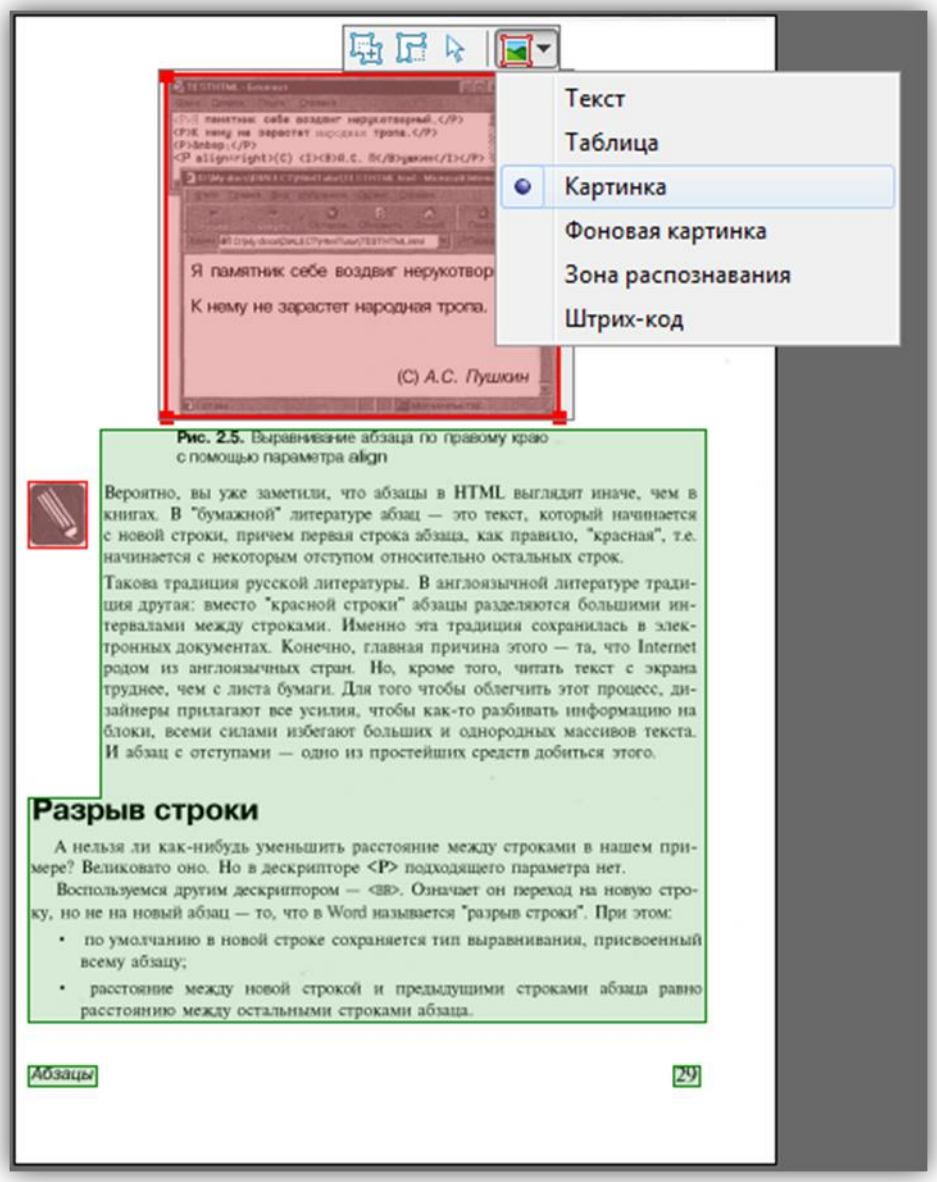


рис. 9

Что бы сохранить документ нужно нажать на кнопку «***Сохранить***»

Выбираем и нажимаем нужный тип документа

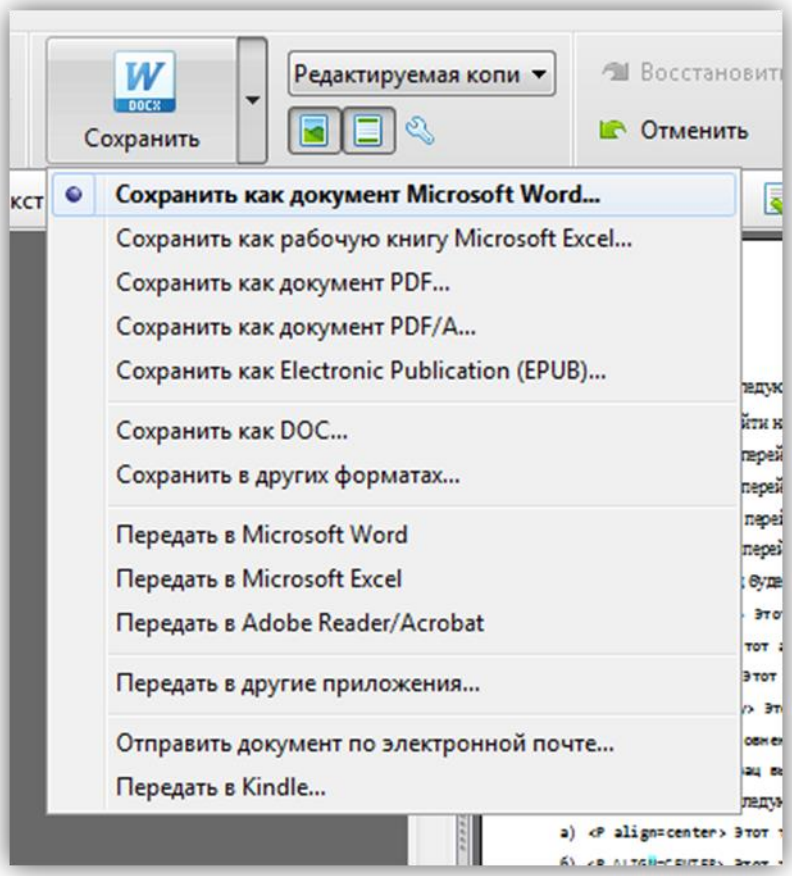


рис. 10

На этом можно считать работу по настройке программы и по преобразованию PDF файла в текст, поддающийся обработке текстовыми процессорами законченной.

Как видно далее в отсканированной странице то есть в странице формата PDF, некоторые символы в словах а в некоторых местах и целые слова были с нечеткими, размазанными границами и на рисунках присутствовали тексты. Благодаря правильной настройки программы в ходе преобразования FineReadery удалось без ошибок распознать тексты с нечеткими границами в тоже время программа не стала распознавать тексты в рисунке на странице.

Страница документа формата DOC

Запас никогда не помешает. Например; для того, чтобы использовать заголовки не подряд, а через один. Скажем, если нам нужна двухуровневая вложенность, то мы вовсе не обязаны использовать обязательно заголовки <H1> и <H2>. Вместо этого можно воспользоваться, например, <H2> и <H4> или любыми другими, и таким образом подобрать нужное соотношение размеров.

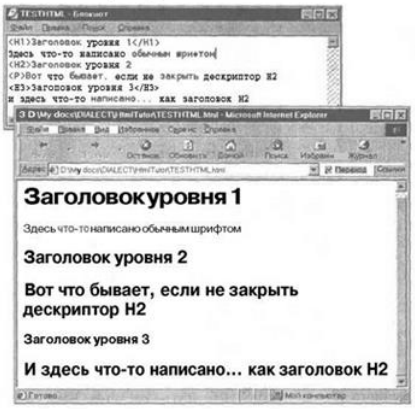


Рис. 5.2. Если не закрыть дескриптор заголовка, возможны самые неожиданные эффекты. Но если он закрыт, следующий текст всегда начинается с новой строки

Параметры заголовка

Итак, у нас есть средство для разметки заголовков и подзаголовков страницы. Но не слишком ли оно однообразно? В конце концов, это всего лишь вариации на тему размера шрифта в пределах одного абзаца.

Есть ли у дескрипторов <H4> параметры?

Разумеется. И это те же параметры, что и у дескриптора <P>. Точнее, один параметр абзаца — выравнивание (align). Как и обычные абзацы, заголовки по 'молчанию выравниваются по левому краю. Но с помощью параметра align их южно выровнять по правому краю или по центру (рис. 5.3).

А как же с другими параметрами? Ведь заголовки часто отличаются не только разгром шрифта и выравниванием, но и цветом, и гарнитурой, да мало ли еще чем?

Заголовки

57

Страница документа формата PDF

Запас никогда не помешает. Например; для того, чтобы использовать заголовки не подряд, а через один. Скажем, если нам нужна двухуровневая вложенность, то мы вовсе не обязаны использовать обязательно заголовки <H1> и <H2>. Вместо этого можно воспользоваться, например, <H2> и <H4> или любыми другими, и таким образом подобрать нужное соотношение размеров.

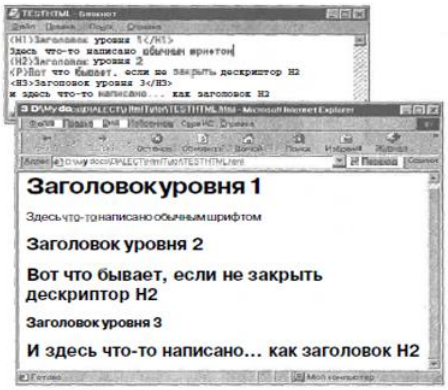


Рис. 5.2. Если не закрыть дескриптор заголовка, возможны самые неожиданные эффекты. Но если он закрыт, следующий текст всегда начинается с новой строки

Параметры заголовка

Итак, у нас есть средство для разметки заголовков и подзаголовков страницы. Но не слишком ли оно однообразно? В конце концов, это всего лишь вариации на тему размера шрифта в пределах одного абзаца.

Есть ли у дескрипторов <H4> параметры?

Разумеется. И это те же параметры, что и у дескриптора <P>. Точнее, один параметр абзаца — выравнивание (align). Как и обычные абзацы, заголовки по молчанию выравниваются по левому краю. Но с помощью параметра align их южно выровнять по правому краю или по центру (рис. 5.3).

А как же с другими параметрами? Ведь заголовки часто отличаются не только размером шрифта и выравниванием, но и цветом, и гарнитурой, да мало ли еще чем?

Заголовки 57

ЗАКЛЮЧЕНИЕ

В данной работе мы рассмотрели систему оптического распознавания символов ABBY Fine Reader. Применение систем оптического распознавания символов играет огромную роль в работе по оцифровке бумажных документов, различной технической и художественной литературы.

В процессе написания данной работы, были достигнуты следующие цели:

1. Дана характеристика основным понятиям теории распознавания изображений, дана характеристика задач распознавания образов и их типов.

2. Были рассмотрены задачи распознавания текстов, дан обзор текущему состоянию систем оптического распознавания символов.

3. На конкретном примере была рассмотрена система оптического распознавания символов ABBY Fine Reader, был рассмотрен интерфейс программы и основные задачи и функции, выполняемые в данной программе.

4. Рассмотрены методы распознавания текста в программе Fine Reader, дана качественная характеристика этих методов, описаны их преимущества и недостатки.

5. В пятом параграфе приведен конкретный пример, по настройке программы, рассмотрена преобразование документа или изображения в текст, поддающийся обработке текстовыми процессорами.

ЛИТЕРАТУРЫ

1. **Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen.** *Character Recognition Systems: A Guide for Students and Practitioners. - Wiley-Interscience: Ноябрь, 2007*

2. **Kunihiko Fukushima.** *Neocognitron for handwritten digit recognition. Neurocomputing*

3. **Богданов В., Ахметов К.** *Системы распознавания текстов в офисе. // Компьютер-пресс — 1999 №3, с.40-42*

4. **Павлидис Т.** *Алгоритмы машинной графики и обработки изображений. М:, Радио и связь, 1986*

5. **Nagao M., Matsuyama T.** *Edge Preserving Smoothing. // Proc. Fourth Intern. Joint Conf. on Pattern Recognition (November, 1978), pp. 518-520.*

6. **Багрова И. А., Грицай А. А., Сорокин С. В., Пономарев С. А., Сытник Д. А.** *Выбор признаков для распознавания печатных кириллических символов // Вестник Тверского Государственного Университета 2010 г., 28, стр. 59-73*

7. **Abdou I.E., Pratt W.K.** *Quantitative Design and Evaluation of Enhancement/Thresholding Edge Detectors. // IEEE Proceedings, 67 (1979), pp 753-763.*

8. **Квасников В.П., Дзюбаненко А.В.** *Улучшение визуального качества цифрового изображения путем поэлементного преобразования // Авиационно-космическая техника и технология 2009 г., 8, стр. 200-204*

9. **Арлазаров В.Л., Куратов П.А., Славин О.А.** *Распознавание строк печатных текстов // Сб. трудов ИСА РАН «Методы и средства работы с документами». — М.: Эдиториал УРСС, 2000. — С. 31-51.*

10. *http://www.intuit.ru/department/economics/manstats/6/4.html*

11. *http://www.abbyy.ru*