Лабораторная работа 3.
Тема: Технология распознавания графических образов

Задание:

1. При помощи сканера ввести в компьютер документ, содержащий текст, таблицу и изображение;
2. Распознать полученные данные;
3. Исправить ошибки распознавания;
4. Сохранить полученную информацию в файл.

Используемое оборудование и программное обеспечение:

- Установочный файл ABBYY FineReader 6.0 Professional;
- Сканер.

Установка и запуск

Для установки FineReader используется специальная программа установки, входящая в состав дистрибутива. Всегда производите установку, используя дискету и CD–Rom только из дистрибутива, который Вы получили при покупке. Установочные файлы и файлы на CD-диске записаны в специальном виде, после копирования их на другой носитель программа инсталляции будет работать некорректно.

После того как программа установки проверит систему, наберите свое имя и укажите куда следует установить ABBYY FineReader. Установщик отобразит несколько опций установки. Выберите один из вариантов установки:

– Обычная (рекомендуется) – устанавливаются все компоненты дистрибутива, в том числе все языки распознавания, язык интерфейса (один) – выбранный при установке.

– Установка компонентов дистрибутива по выбору – из предложенного набора компонентов, входящих в дистрибутив, Вы можете выбрать только те, которые нужны Вам (в том числе доступные языки распознавания).

Чтобы запустить программу FineReader выберите пункт ABBYY FineReader Professional в меню «Пуск -> Программы».

Получение и обработка изображения программой FineReader

Качество распознавания во многом зависит от качества исходного изображения. FineReader работает со сканерами через TWAIN-интерфейс. Это единый международный стандарт, введенный в 1992 году для унификации взаимодействия устройств для ввода изображений в компьютер с внешними приложениями. При этом возможно два варианта взаимодействия программы со сканерами через TWAIN-драйвер:

– через интерфейс FineReader: в этом случае для настройки опций сканирования используется диалог программы FineReader «Настройки сканера».

– через интерфейс TWAIN-драйвера сканера: для настройки опций сканирования используется диалог TWAIN–драйвера сканера.

В режиме «Использовать интерфейс TWAIN–драйвера сканера», как правило, доступна функция предварительного просмотра изображения (preview), позволяющая точно задать размеры сканируемой области, подобрать яркость, тут же контролируя результаты этих изменений. К сожалению, диалог TWAIN–драйвера сканера у каждого сканера выглядит по-своему, в большинстве случаев все надписи на английском языке. В режиме «Использовать интерфейс FineReader» доступны такие опции, как возможность сканирования в цикле на сканерах без автоподатчика, сохранение опций сканирования в отдельный файл и возможность использования этих опций в других пакетах. Вы можете легко переключаться между этими режимами на закладке «Сканирование/Открытие» диалога «Опции» (меню Сервис>Опции) установите переключатель в одно из положений: «Использовать интерфейс TWAIN–драйвера сканера» или «Использовать интерфейс FineReader».

Чтобы запустить сканирование нажмите кнопку «Сканировать» или в меню «Файл» выберите пункт «Сканировать». Спустя некоторое время в «Главном окне» программы FineReader появится окно с изображением вставленного листа.

Если Вы хотите отсканировать несколько страниц, то нажмите на стрелке справа от кнопки «Сканировать» и в локальном меню выберите пункт «Сканировать несколько страниц».

Если сканирование не началось сразу, то откроется встроенный TWAIN–интерфейс сканера и диалог «Настройки сканера».

Если Вы хотите сразу запустить распознавание отсканированных страниц, воспользуйтесь опцией «Сканировать и распознать» или «Сканировать и распознать несколько страниц».

Программа отсканирует и распознает изображения. В главном окне программы появятся окно с "изображением" вставленного листа и окно «Текст с результатом распознавания». Распознанный текст можно сохранить во внешние редакторы и форматы.

Установка параметров сканирования

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.

Основными параметрами сканирования являются:

– Тип изображения – серый (256 градаций). Сканирование в сером цвете является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Черно-белый тип изображения обеспечивает более высокую скорость сканирования, но при этом теряется часть информации о буквах, что может привести к ухудшению качества распознавания на документах среднего и низкого качества печати. Если Вы хотите, чтобы содержащиеся в документе цветные элементы были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

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

– Яркость – в большинстве случаев подходит среднее значение яркости – 50%. На некоторых документах при сканировании в черно-белом режиме может понадобиться дополнительная настройка яркости.

Чтобы установить параметры сканирования:

– При сканировании через TWAIN с использованием интерфейса FineReader: в меню «Сервис» выберите пункт «Настройки сканера» и в открывшемся диалоге «Настройки сканера» установите нужные опции.

– При сканировании через TWAIN с использованием интерфейса TWAIN–драйвера сканера для установки параметров сканирования используется диалог Вашего сканера, который открывается автоматически при нажатии на кнопку «Сканировать». Опции для установки параметров сканирования могут называться по-разному, в зависимости от модели сканера.

Сканирование многостраничных документов

Для удобства сканирования большого количества страниц в программе FineReader предусмотрен специальный режим сканирования: Сканировать несколько страниц. Он позволяет отсканировать несколько страниц в цикле. При этом:

– при сканировании через TWAIN с использованием интерфейса FineReader по окончании сканирования страницы сканер автоматически начинает сканирование следующей;

– при сканировании через TWAIN с использованием интерфейса TWAIN-драйвера сканера Twain-диалог сканера не закрывается после окончания сканирования первой страницы. Вы можете положить следующую страницу в сканер и отсканировать ее.

Для завершения сканирования в меню «Файл» выберите пункт «Остановить сканирование».

Открытие файлов с изображением

Если у Вас нет сканера, Вы можете распознавать графические файлы. Чтобы открыть изображение нажмите стрелку справа от кнопки «Сканировать» и в локальном меню выберите пункт «Открыть изображение». Внешний вид значка изменится; подпись «Сканировать» поменяется на «Открыть».

В меню «Файл» выберите пункт «Открыть изображение».

В Windows Explorer: щелкните правой кнопкой мыши на файле с изображением и в локальном меню выберите пункт «Открыть с помощью FineReader». Если на компьютере уже открыт FineReader, то изображение будет добавлено в текущий пакет, в противном случае перед добавлением изображения в пакет автоматически запустится FineReader с пакетом, с которым Вы работали в последний раз.

В диалоге «Открыть (Open)» выберите одно или несколько изображений. Выбранные изображения появятся в окне «Пакет», и

последнее из выбранных изображений откроется в окне «Изображение» и в окне «Крупный план» на экране FineReader, при этом копия изображения помещается в папку пакета.

Действия над изображением

Очистить от мусора. Распознаваемое изображение может быть сильно "замусорено", то есть содержать много лишних точек, возникших в результате сканирования документов среднего или плохого качества. Точки, близко расположенные к контурам букв, могут отрицательно сказаться на качестве распознанного текста. Чтобы уменьшить количество лишних точек, можно воспользоваться опцией «Очистить от мусора». Для этого в меню «Изображение» выберите пункт «Очистить изображение от мусора». Если Вы хотите очистить от "мусора" отдельный блок, то в меню «Изображение» выберите пункт «Очистить блок от мусора».

Инвертировать изображение. Некоторые сканеры инвертируют изображения при сканировании (черный цвет переводят в белый, а белый в черный). Чтобы получить стандартное представление документа (черный шрифт на белом фоне) необходимо в меню «Изображение» выбрать пункт «Инвертировать».

Повернуть или зеркально отразить изображение. При распознавании изображение должно иметь стандартную ориентацию: текст должен читаться сверху вниз, и строки должны быть горизонтальными. По умолчанию, программа при распознавании определяет и корректирует ориентацию изображения автоматически. Если ориентация изображения была определена ошибочно, то на закладке «Сканирование/Открытие» снимите отметку с пункта «Определять ориентацию страницы» и поверните изображение вручную. Чтобы повернуть изображение:

– на 90 градусов вправо – нажмите кнопку  или выберите в меню «Изображение» пункт «Повернуть по часовой стрелке»;

– на 90 градусов влево – нажмите кнопку  или выберите в меню «Изображение» пункт «Повернуть против часовой стрелки»;

– на 180 градусов – выберите в меню «Изображение» пункт «Повернуть на 180 градусов».

Чтобы зеркально отразить изображение относительно:

– горизонтальной прямой – выберите в меню «Изображение» пункт «Зеркально отразить относительно горизонтали»;

– вертикальной прямой – выберите в меню «Изображение» пункт «Зеркально отразить относительно вертикали».

Стереть участок изображения. Если Вы хотите исключить какой-то участок текста из распознавания или у Вас на изображении имеются большие участки мусора, то Вы можете стереть такие участки. Для этого выберите инструмент (на панели в окне «Изображение») и, нажав на левую кнопку мыши, выделите участок изображения, который Вы хотите удалить. Отпустите кнопку, выделенная часть изображения будет удалена.

Увеличить/Уменьшить масштаб изображения. На панели «Изображение» выберите инструмент /  и щелкните мышью на изображении. Изображение увеличится/уменьшится в два раза. Щелкните правой кнопкой мыши на изображение и в локальном меню выберите пункт Масштаб и нужный Вам масштаб.

Печать изображения. Вы можете напечатать одно изображение, открытое в окне «Изображение», несколько изображений, выделенных в окне «Пакет», или все изображения. Для этого в меню «Файл» выберите пункт «Печать изображения» и в открывшемся диалоге «Печать» установите параметры печати.

Отменить последнее действие. Для отмены последнего действия на панели «Стандартная» нажмите кнопку «Отменить» .

Общая информация по анализу макета страницы

Анализ макета страницы может проводиться как автоматически, так и вручную. В большинстве случаев FineReader сам успешно справляется с анализом сложных страниц. Автоматический анализ производится по кнопке «Распознать» одновременно с распознаванием текста.

Ручное выделение блоков может понадобиться, если

– хотите распознать часть страницы;

– в результате автоматического анализа блоки были выделены неправильно.

В случае если программа выделила неправильно некоторые блоки, часто оказывается быстрее исправить только их, воспользовавшись инструментами для редактирования блоков, а не выделять блоки на изображении заново вручную.

Блоки – это заключенные в рамку участки изображения. Блоки выделяют для того, чтобы указать системе, какие участки отсканированной страницы надо распознавать и в каком порядке. Также по ним воспроизводится исходное оформление страницы. Блоки разных типов имеют различные цвета рамок. Вы можете изменить цвета рамок блоков на закладке «Вид диалога» «Опции» (меню Сервис>Опции) в группе «Объекты». В поле «Объект» выберите нужный тип блока, а в поле Цвет – требуемый цвет.

При обработке изображений выделяют блоки следующих типов:

– Зона Распознавания – блок используется для распознавания и автоматического анализа части изображения. После нажатия на кнопку «Распознать» выделенный блок автоматически анализируется и распознается.

– Текст – блок используется для обозначения текста. Он должен содержать только одноколоночный текст. Если внутри текста содержатся картинки, выделите их в отдельные блоки.

– Таблица – этот блок используется для обозначения таблиц или текста, имеющего табличную структуру. При распознавании программа разбивает данный блок на строки и столбцы и формирует табличную структуру. В выходном тексте данный блок передается таблицей.

– Картинка – этот блок используется для обозначения картинок. Он может содержать картинку или любую другую часть текста, которую Вы хотите передать в распознанный текст в качестве картинки.

– Штрих-код – этот блок используется для распознавания штрих-кодов.

Опции анализа таблиц

В большинстве случаев программа делит таблицу на строки и столбцы автоматически. Дополнительная настройка опций анализа таблиц устанавливается на закладке «Распознавание» в группе «Таблицы». Эти опции рекомендуется использовать, если:

– в результате автоматического анализа макета страницы таблица была выделена и разделена на строки и столбцы неверно;

– документ содержит много однотипных таблиц, для которых известна дополнительная информация (например: таблица не содержит объединенных ячеек).

Опция «В каждой ячейке таблицы не более одной строки» текста используется для уверенного анализа таблиц с неполным количеством или без черных разделителей, с ячейками, содержащими не более одной строки текста.

Опция «Таблица не содержит объединенные ячейки» используется для уверенного анализа таблиц, не содержащих объединенных ячеек.



Рисунок 3.1 Группы инструментов по работе с изображением

Чтобы создать новый блок выберите один из инструментов:

– – выделить зону распознавания;

– – выделить текстовый блок;

– – выделить картинку;

– – выделить табличный блок.

Установите курсор мыши в угол предполагаемого блока, нажмите левую кнопку мыши и, не отпуская кнопки, потяните в противоположный по диагонали угол. Отпустите кнопку мыши, выделенная часть изображения будет заключена в рамку.

Вы можете поменять тип блока. Для этого щелкните на блоке правой кнопкой мыши и в локальном меню выберите «Тип блока», а затем – нужный пункт.

Чтобы добавить прямоугольную часть блока:

– Выберите инструмент ;

– Установите курсор мыши внутри блока, к которому Вы хотите добавить часть. Нажмите левую кнопку мыши и, не отпуская кнопки, потяните по диагонали. Выделив нужную часть изображения, отпустите кнопку мыши. Выделенный прямоугольник будет добавлен к блоку.

FineReader накладывает на допустимую форму блоков некоторые ограничения, обусловленные необходимостью передавать на распознавание неразрывные текстовые строки. Для соблюдения этих ограничений FineReader либо автоматически корректирует границы блока после операций добавления или удаления прямоугольной части блока, либо запрещает некоторые операции по передвижению отрезков, образующих границу блока.

Чтобы выделить один или несколько блоков, выберите инструмент и щелкните мышкой по нужному блоку или, удерживая кнопку мыши, нарисуйте прямоугольник, охватывающий нужные блоки. Вы можете выделить один или несколько блоков, используя стандартные инструменты выделения блоков. Чтобы выбрать несколько блоков, нажмите клавишу SHIFT или CTRL (при этом должен быть выбран один из инструментов: , , или ) и мышью щелкните на требуемых блоках.

Чтобы передвинуть блок, нажмите клавишу ALT (при этом должен быть выбран один из инструментов: , , , или ) и мышью переместите блоки.

Чтобы удалить блок, выберите инструмент и выделите блок, который Вы хотите удалить.

Редактирование таблицы

Для редактирования таблицы выберите на панели «Изображение» один из инструментов:

–  – чтобы добавить вертикальную линию;

–  – чтобы добавить горизонтальную линию;

–  – чтобы удалить линию.

Чтобы объединить ячейки таблицы, в меню «Правка» выберите пункт «Объединить ячейки таблицы».

Чтобы разбить ранее объединенные ячейки таблицы, в меню «Правка» выберите пункт «Разбить ячейки таблицы».

Чтобы объединить строки таблицы (в этом случае деление на столбцы остается), в меню «Правка» выберите пункт «Объединить строки таблицы».

Если в результате автоматического анализа таблицы разделение на строки и столбцы произошло неверно, прежде чем анализировать таблицу вручную заново, попробуйте сначала отредактировать результаты автоматического анализа.

Чтобы отредактировать таблицу вручную, отредактируйте таблицу, используя инструменты (панель в окне «Изображение»):

–  – Добавить вертикальную линию;
– Добавить горизонтальную линию;
– Удалить линию.

Чтобы объединить ячейки или строки таблицы, в меню «Правка» выберите пункты «Объединить ячейки таблицы» или «Объединить строки таблицы». При объединении строк таблицы деление на столбцы остается.

Объединенные ячейки можно снова разделить, воспользовавшись командой «Разбить ячейки таблицы» (меню «Правка»).

Распознавание

Задача распознавания состоит в том, чтобы преобразовать отсканированное изображение в текст, сохранив при этом оформление страницы. Прежде чем приступить к распознаванию текста, необходимо установить основные параметры распознавания: язык распознавания, тип печати распознанного текста и тип страницы.

Перед запуском распознавания проверьте установленные опции: язык распознавания, тип печати распознаваемого текста и тип страницы.

Чтобы запустить распознавание, нажмите кнопку «Распознать» на панели «Scan&Read». В меню «Процесс» выберите нужный Вам пункт:

– «Распознать» – чтобы распознать открытую страницу или все страницы, выделенные в окне «Пакет»;

– «Распознать все» – чтобы распознать все нераспознанные страницы пакета;

– «Распознать Блок» – чтобы распознать блок или несколько блоков, выделенные на изображении;

– «Запустить фоновое распознавание» – чтобы запустить распознавание в фоновом режиме.

Язык распознавания

FineReader поддерживает распознавание как одноязычных, так и многоязычных (например, английско-французских) документов.

Чтобы указать язык распознаваемого текста, выберите соответствующую строку в списке на панели «Распознавание».

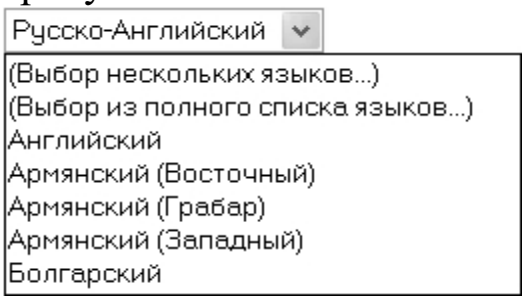

Рисунок 3.2 Выбор языка распознавания

Если Вы хотите распознать документ, написанный на нескольких языках, то в списке языков на панели «Стандартная» выберите пункт «Выбор нескольких языков…».

Тип печати входного текста

Для большинства текстов тип печати определяется автоматически. Этому соответствует значение «Авто» (группа «Тип печат» и, меню «Сервис->Опции», закладка «Распознавание»).

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 текстов, напечатанных на пишущей машинке – значение «Пишущая машинка»;

– для текстов, напечатанных на матричном принтере – значение «Матричный принтер».

Чтобы поменять тип печати, на закладке «Распознавание» диалога «Опции» в группе «Тип печати» выберите требуемый пункт.

Если программа некорректно распознала блоки с вертикальным направлением текста, то щелкните правой кнопкой мыши по блоку с вертикальным направлением текста и в локальном меню выберите пункт «Свойства». В открывшемся диалоге «Свойства» выберите нужный вариант в списке «Направление текста» и распознайте изображение заново.

Общая информация по сохранению распознанного текста

Кнопка «Сохранить» позволяет передать результаты распознавания в выбранное приложение или сохранить их в файл. Внешний вид значка меняется в зависимости от выбранного режима сохранения; подпись «Сохранить» меняется на название выбранного приложения.

Чтобы сохранить распознанный текст, нажмите стрелку справа от кнопки «Сохранить» и в локальном меню выберите необходимый пункт.

После того как Вы экспортировали распознанный текст в выбранное Вами приложение, отправили его по электронной почте, передали в буфер или сохранили в файл, "информация" об этом действии отразится на значке кнопки «Сохранить». Поэтому для того, чтобы повторить ту же операцию для другого изображения, Вам достаточно нажать на этот значок.

Опции сохранения распознанного текста устанавливаются на закладке «Форматирование» диалога «Опции». Также часть этих опций можно установить в диалогах «Мастер сохранения» результатов и «Сохранить текст как».

Режим сохранения форматирования и раскладки распознанного текста (при сохранении в форматах RTF, DOC или HTML):

– Сохранять полное оформление документа – сохраняется полное оформление документа: разбиение на абзацы, гарнитура и размер шрифта, колонки, направление текста, цвет букв и фон текста, а так же сохраняется структура таблиц;

– Сохранять начертание и размер шрифта – сохраняется структура таблиц, разбиение на абзацы, начертание и размер шрифта;

– Не сохранять оформление – в этом случае сохраняется разбиение на абзацы и структура таблиц;

– Сохранять (не сохранять) картинки – позволяет сохранить картинки в распознанном тексте. Эта опция доступна при сохранении в форматы RTF, DOC, PDF или HTML.

При сохранении в PDF и HTML-файлы картинки сохраняются в формате JPEG. В этом формате используется метод сжатия изображения с "потерей качества". Значение, установленное в поле «Качество JPEG», позволяет регулировать качество сохраненного изображения. Возможно установить одно из значений в диапазоне от 1 до 100 (по умолчанию выставлено среднее значение – 50).

Режимы передачи страниц пакета (при сохранении нескольких страниц пакета)

Каждая страница в отдельный файл – позволяет сохранить каждую страницу пакета в отдельный файл. В этом случае к названию файла в конец автоматически добавляется порядковый номер страницы в пакете.

Сохранять с именами исходных изображений – позволяет сохранить распознанные страницы в отдельные файлы выбранного формата, причем имена этих файлов будут совпадать с именами соответствующих файлов изображений.

Пофайловое деление по пустым страницам – весь пакет рассматривается как совокупность групп страниц, разделенных пустыми страницами. Для каждой такой группы при сохранении образуется отдельный файл, в который последовательно записываются страницы группы; имена формируемых в этом режиме файлов образуются добавлением к указанному пользователем имени файла индексов следующего вида: –1, –2, –3 и т.д.

Все страницы в один файл – позволяет сохранить все (или все выделенные) страницы пакета в один файл.

Сохранение распознанного текста в формате PDF.

Опции сохранения оформления документа:

– Только текст и картинки – сохраняется распознанный текст и имеющиеся в тексте картинки;

– Только изображение – сохраняется только изображение;

– Текст поверх изображения картинки – изображение передается картинкой. Текстовые области сохраняются как текст поверх картинки;

– Текст под изображением картинки – все изображение сохраняется как картинка. Под нее "записывается" распознанный

текст. В этом случае сохраняется полный дизайн документа: цветовые выделения шрифта, фона, оформление бумаги и т.д.;

Чтобы установить опции при сохранении в формате PDF, необходимо в меню «Сервис» выберите пункт «Форматы», затем на закладке PDF диалога «Форматы» установите требуемые опции. При сохранении в режиме «Только текст и картинки» и в режиме «Текст поверх изображения страницы» Вы можете указать, чтобы при сохранении текста в формат PDF неуверенно распознанные слова заменялись их изображением. Для этого на закладке PDF диалога «Форматы» отметьте опцию «Заменять неуверенно распознанные слова их изображениями».