ВВОД СКАНИРОВАННЫХ ДОКУМЕНТОВ В ЭЛЕКТРОННЫЙ АРХИВ ПРЕДПРИЯТИЯ

Давно известно, что при создании систем электронного архива и документооборота недостаточно регистрировать только те документы, которые изначально разрабатываются в электронном виде: несмотря на бурное развитие информационных технологий, на бумаге и сейчас хранятся огромные объемы технической, инженерной и технологической документации.

CADmaster уже не раз и достаточно подробно рассказывал о принципах работы с такими документами, позволяющих не отказываться от наработанного за многие годы, а использовать его, к примеру, в новых САПР-проектах. Всё просто: сканируем документы, повышаем их качество при помощи, скажем, программы Spotlight, а затем вводим эти документы в систему электронного архива и документооборота. Такой документ можно использовать в дальнейших разработках, вносить необходимые изменения, добавлять в архив новые версии и т.д. Не будем подробно останавливаться на всех тонкостях – кроме, пожалуй, одного момента...

Рассмотрим детальнее процесс ввода отсканированного документа в систему электронного архива (документооборота). Стандартная процедура здесь выглядит так: заполни поля электронной карточки сканированного документа, укажи файл изображения и нажми соответствующую кнопку. Но теперь представьте, что в вашем бумажном архиве сотни тысяч, а то и миллионы единиц хранения. Обычная технология регистрации документа, предоставляемая стандартной системой архива, сразу оказывается чрезвычайно громоздкой, требует невероятных временных затрат.

Единственный выход – автоматизировать процесс регистрации сканированного документа в электронном архиве. Например, в автоматическом режиме распознавать поля углового штампа, которые по сути являются полями учетной карточки документа, и регистрировать распознанную информацию в электронном архиве. При этом было бы совсем неплохо заодно и поднять качество сканированных изображений.

RasterID 2.0 предназначен именно для таких задач.

Теоретически для решения проблемы автоматизированного ввода нужно не так уж и много:

- система архива и документооборота должна поддерживать запись информации о файле документа в виде ссылки на "внешний" ресурс;
- необходим экспорт в систему документооборота всей информации о документах – либо напрямую при помощи дополнительных механизмов системы архива, других внешних приложений, СУБД или ODBC, либо через промежуточный формат *.xls, *.dbf и т.д.

Итак, распознаём в автоматическом режиме необходимые записи штампа (при помощи модуля распознавания текста) и экспортируем эту информацию в базу данных или систему документооборота.

Вот только... не всё так очевидно. Основная проблема в том, что надписи в штампах выполнены, мягко говоря, "не по ГОСТу". Значит, и вероятность безошибочного распознавания подобных надписей не так высока, как хотелось бы. Что делать?

Революционность RasterID 2.0 – в механизмах решения подобных проблем. Но об этом чуть позже. А сейчас – о возможностях программы.

Сканирование

В RasterID 2.0 реализован механизм работы с любым сканирующим оборудованием.

Модуль WiseScan – это всё, что необходимо для удобного, быстрого и интеллектуального сканирования на всех моделях сканеров Contex. Доступны автоопределение размера сканируемого оригинала, режим предварительного сканирования, точная настройка параметров цвет-

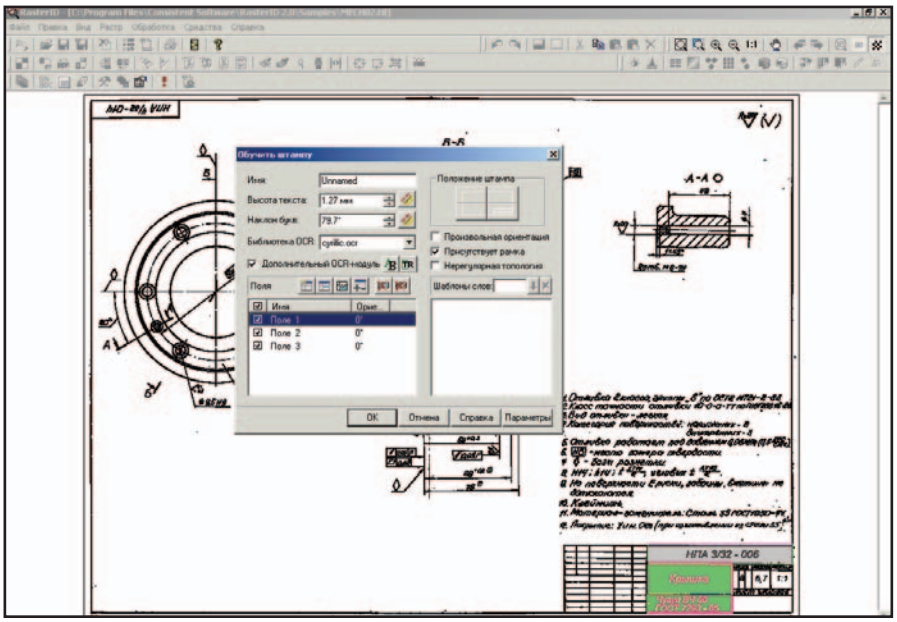

Рис. 1. Обучение штампу. При помощи соответствующих программных инструментов указано положение штампа, три поля (десятичный номер, название изделия, материал по ГОСТ), измерены высота текста и угол наклона, подключен дополнительный модуль OCR

ного и монохромного сканирования, пакетное сканирование, автоименование документов по заданной маске (с возможностью включения информации из распознанных полей).

К отсканированным изображениям можно применять предварительно заданную последовательность действий, включающую возможность распознавания штампа, проверки информации и ее экспорта во внешнюю базу данных.

"Виртуальный сканер" – при работе с некоторыми сканерами или инженерными системами (Xerox, KIP, Océ и др.) есть возможность "виртуального" сканирования. Требуется только указать папку, в которой сохраняются сканированные изображения, и составить сценарий обработки, после чего RasterID с установленной вами периодичностью обращается к данной папке и применяет указанную последовательность действий ко всем появляющимся там файлам.

RasterID также обеспечивает прямую поддержку сканеров с **TWAIN-интерфейсом**.

Повышение качества изображений

Думаю, если вы хоть раз видели отсканированную "синьку", вас не придется убеждать, что качество таких изображений далеко от идеального. Пакет RasterID 2.0 предоставляет различные возможности повышения качества сканированных документов: он позволяет удалять "мусор", заливать "дырки", сглаживать растровые линии, устранять возникающие при сканировании перекосы, обрезать пустые поля и т.д. С помощью операций коррекции изображений по четырем точкам рамки устраняются искажения и корректируется размер самого изображения. Эти и многие другие инструменты могут применяться в автоматическом (пакетном) режиме: или параллельно с процессом сканирования, или к указанному набору ранее отсканированных изображений.

Работа с цветными изображениями

Нужно ли работать с цветными изображениями в архиве технической документации? Опыт показывает, что да. К примеру, монохромное сканирование материалов невысокого качества (тех же традиционных российских "синек") не приносит желаемого результата: в подобных случаях лучше отсканировать изображение в режиме Gray Scale (256 градаций серого), а затем произвести обработку, повышающую качество. После обработки вы можете сохранить изображение как монохромное или оставить его как есть (Gray Scale).

Кроме того, все чаще появляются цветные чертежи, полученные при печати проектов, выполненных в САПР. Эти документы удобнее сканировать и нагляднее представлять в цвете!

RasterID располагает множеством средств обработки цветных изображений, среди которых инструменты коррекции яркости/контрастности, уровней, палитры, гамма-коррекции, фильтры размытия, контурной резкости, усреднения. Полноцветное изображение может быть приведено к индексированной палитре или преобразовано в градации серого. Есть возможность разделять цветное (серое) изображение на монохромные слои при помощи специальных процедур: бинаризации, уменьшения количества цветов, разделения по цветам.

Если все перечисленные возможности будут предоставлены людям квалифицированным и хорошо обученным, получение изображений отличного качества гарантировано.

Распознавание штампа. Индексация

Остановимся на функциях поиска штампа на чертеже, распознавания надписей в его полях и экспорта полученной информации.

Программу нужно "обучить" распознаванию штампа. Процедура проста: достаточно обвести штамп прямоугольником и, если необходимо, отредактировать распознанную топологию. Далее следует указать те поля, которые содержат необходимую информацию, задать им имена и сохранить темплет (шаблон). Темплеты нужно создать для всех различных по топологии штампов.

При распознавании указанные поля углового штампа записываются в ячейки таблиц приемников данных. Кроме того, и это очень важно, в одно из полей может передаваться графический фрагмент (картинка) углового штампа, что в дальнейшем значительно упрощает проверку достоверности распознанных данных.

О распознавании надписей: пакет RasterID может работать как с внутренним OCR-модулем, так и с OCR-модулем FineReader Pro 5.0. Разумеется, процесс распознавания нуждается в контроле.

Приемники данных

RasterID является открытой системой: при помощи стандартных способов программирования может быть создан модуль экспорта, обеспечивающий прямую передачу распознанных данных из полей углового штампа в архивную систему пользователя.

С RasterID поставляются модули экспорта данных в MS Excel, MS Access, текстовый файл с разделителями; поддерживается передача данных при помощи механизма ODBC. Большинство современных СУБД имеет встроенные механизмы экспорта и импорта, позволяющие "пакетно" импортировать файлы перечисленных стандартных форматов в формат таблиц СУБД (например, Data Transformation Services Wizard, поставляемый с MS SQL Server).

Промежуточный формат также может быть экспортирован при помощи механизмов самой системы архива или других внешних приложений.

Поскольку одним из полей приемника данных является ссылка на файл документа, необходимо, чтобы архивная система работала по ссылке с файлами документов, хранящимися на "внешнем" ресурсе.

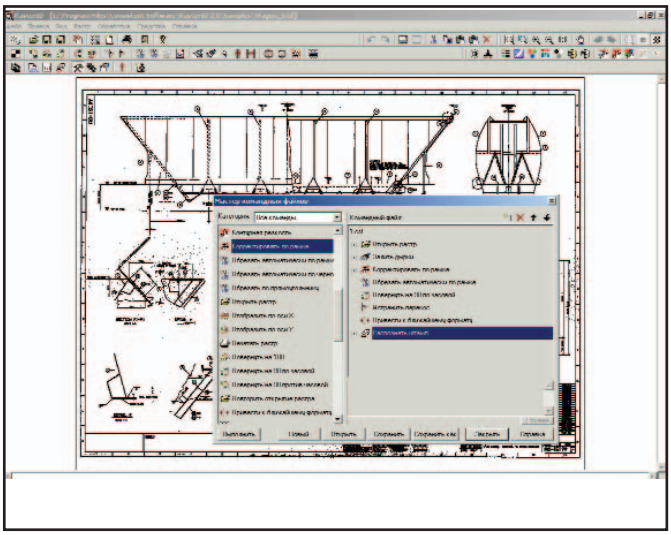

Рис. 2. Формирование командного файла с использованием Мастера командных файлов

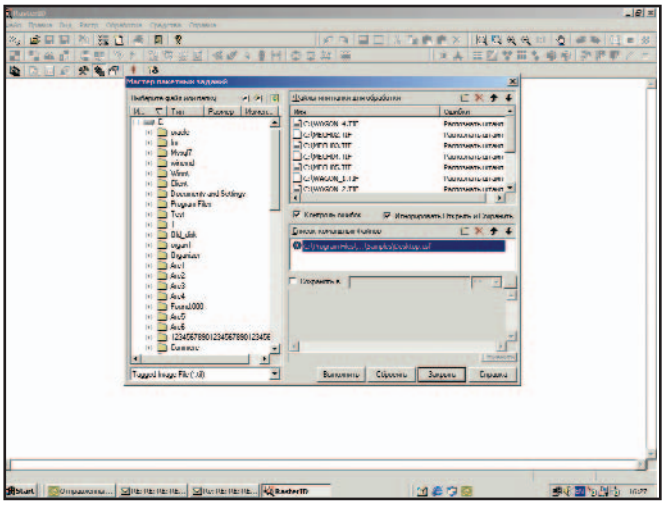

Рис. 3. Работа с командными файлами с использованием Мастера пакетных заданий

Контроль качества распознавания

Не будем строить иллюзий и утверждать, что "распознать можно всё". Качество распознавания в первую очередь зависит от подлинника: распознать печатный текст проще, чем написанный пусть даже строго по ГОСТу, но от руки. А если текст, мягко говоря, "не по ГОСТу" или само изображение – крайне низкого качества? Для таких ситуаций в новой версии пакета предусмотрены встроенные механизмы контроля качества распознавания.

Способ контроля зависит от приемника, в который были экспортированы данные. При экспорте данных в MS Excel для контроля передается фрагмент растра, содержащий штамп. В MS Access используют специальный модуль проверки качества. Если штамп распознается из программы напрямую, для контроля служит диалог *Редактирование распознанных данных*: он появляется после того как команда *Распознать штамп* завершила распознавание.

В любом случае контролировать распознавание удается достаточно эффективно.

Пакетная обработка

Все упомянутые выше возможности RasterID могут применяться в автоматическом (пакетном) режиме: программа предлагает мощный и простой в использовании механизм создания командных файлов.

Метод Drag and Drop позволяет сформировать код командного файла, просто перетаскивая мышкой нужные команды.

Затем в Мастере пакетных заданий указываются каталог, папки или отдельные файлы отсканированных изображений, назначается соответствующий командный файл и... можно идти отдыхать: остальное RasterID сделает сам.

Распознаваемое и нераспознаваемое. Что делать?

Поговорим о вещах банальных, но неизбежных. Что делать, если, как уже сказано, текст в угловом штампе выполнен явно "не по ГОСТу"? Можно ли повысить вероятность безошибочного распознавания, к примеру, десятичного номера документа? Или расширить функционал RasterID в других случаях? На этих вопросах стоит остановиться подробнее.

Компонент CSRaster

Самые "продвинутые" читатели, думаю, давно знакомы с технологией ActiveX. Для тех же, у кого это знакомство еще впереди, приведем неоспоримый факт: существуют компоненты ActiveX, которые могут инсталлироваться в любую среду объектно-ориентированного программирования – например VB, Visual C++, Delphi.

В дистрибутив пакета RasrerID входят ActiveX-компонент CSRaster и Руководство разработчика.

Зачем он нужен

Революционность подобного решения в том, что ActiveX-компонент располагает всеми свойствами, методами и функциями, необходимыми для создания любых приложений на основе функционала RasterID. Другими словами, если вам недостаточно тех функций, которые предоставляет стандартная

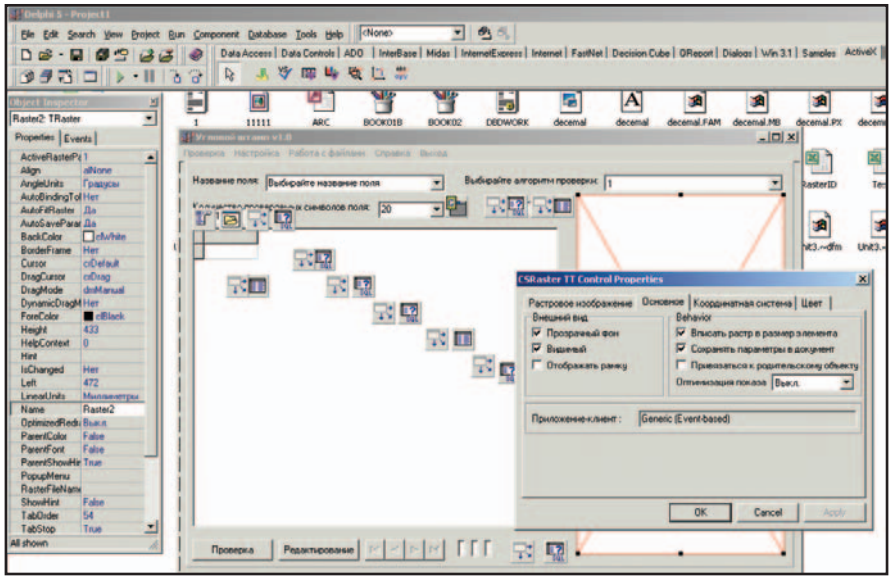

Рис. 4. Приложение, использующее компонент CSRaster, создаваемое в среде объектно-ориентированного программирования Delphi

версия пакета RasterID, и при этом вы имеете навыки программирования, вам по силам самому создать требуемый функционал.

В качестве примера покажем, как работают внешние механизмы контроля качества распознавания штампа, созданные автором при помощи программирования в среде Delphi.

Решения с использованием компонента CSRaster

Условие задачи просто, но актуально для большинства предприятий: необходимо создать часть электронного архива на основе 600 000 сканированных документов. Все документы имеют угловой штамп и сделанные вручную надписи. Ключевым полем является децимальный номер документа. Условия хранения и состояние исходных документов, мягко говоря, плохие.

Требуется максимально повысить вероятность правильного распознавания децимального номера при автоматическом вводе информации в СУБД.

Использование экспорта в ODBC без проверки результатов распознавания недопустимо: велика вероятность внесения неверной информации. С другой стороны, "вручную" проверять все 600 000 записей не слишком-то продуктивно...

Итак, возможное решение.

Используем механизмы среды программирования для инсталляции ActiveX-компонента.

Применяя указанные в Руководстве разработчика свойства и методы, несложно создать приложение с необходимым функционалом. Из всего функционала требуется выбрать только то, что необходимо для решения конкретной задачи – распознавания штампов и создания файла, содержащего результаты распознавания. На этой стадии совсем не важно, какой именно приемник данных выбрать: можно использовать MS Excel, а можно, например, ODBC. В первом случае при дальнейшем решении задачи будет использован механизм экспорта/импорта в СУБД MS SQL Server (Data Transformation Services Wizard). Каким путем идти зависит, наверное, от более конкретно сформулированной задачи и еще от технологии обработки, выбирать которую только вам.

Следующий шаг – написание "недостающего" функционала. Не будем подробно рассматривать строки кода на Паскале. Опишем только логику. Создаваемое приложение должно распознать поля углового штампа и проверить качество распознавания. Я применил следующую схему: задаются критерии (алгоритмы) проверки, включающие правила нахождения символов в различных позициях полей штампа. Почему именно так? Во-первых, эта логика наиболее приемлема при проверке правильности распознавания децимальных номеров, которые по сути являются ключевым полем в описании документа. Во-вторых, существует вполне определенная логика составления децимального номера документа, порядка расположения в нем символов. В-третьих, язык Transact SQL позволяет описывать эту логику в тексте запроса.

Помимо запроса, содержащего логику нахождения того или иного символа в той или иной позиции, "дописываемый" функционал конечно же должен содержать средства вывода запроса (тех записей, которые не удовлетворяют описанному алгоритму проверки). Плюс к тому необходимо создать средство сравнения информации, хранящейся в поле СУБД (результат распознавания), и реальной информации, которая содержится на растровом изображении. Здесь, конечно, сравнение придется проводить "вручную". Также необходимо средство редактирования выявленных в полях СУБД "несоответствий".

А теперь вернемся к условию задачи. Мы получили средство распознавания углового штампа 600 000 растровых изображений. Кроме того, созданная программа позволяет выводить и редактировать все записи о неверных результатах распознавания. Полностью избежать "ручной" работы не удалось, но, поскольку число "неправильных" результатов гораздо меньше общего количества документов, можно утверждать, что эта работа сведена к минимуму.

Конечно, это всего лишь один пример (правда, достаточно неплохо работающий). Возможно, перед вами будут стоять совсем другие задачи, связанные с распознаванием угловых штампов, пакетной обработкой растровых изображений, созданием электронных архивов. Наш пример доказывает, что при использовании пакета RasterID эти проблемы могут быть решены. Вид приложения, созданного в среде Delphi с использованием компонента CSRaster и описанного выше, приведен на рис. 4.

Алексей Рындин
Consistent Software/Бюро ESG
(Санкт-Петербург)
Тел.: (812) 430-3434
E-mail: aryndin@esg.spb.ru