RasterID, или КАК ПОДНЯТЬ ЦЕЛИНУ

Аллегория

Разгребая завалы бумажных архивов, мы гордо вступили в новое тысячелетие — чуть прихрамывая, но все равно нога в ногу со всем остальным цивилизованным миром. Весь остальной мир, взглянув на восток, наконец понял, что у нас не пахано, и предложил... Все правильно, мы — гордые, мы не привыкли просить, у нас предложение порождает спрос, у нас уже все есть, или почти все; нас уже не пугают слова "сканер" или "плоттер", мы представляем, что такое "инженерная машина", у нас, в конце концов, есть электронный архив — правда, у каждого свой. А еще мы умеем строить грозные эсминцы и мощные электростанции, и, конечно, писать умные программы.

Москва... Softool-96... "Я купил у вас "Vidar". Дайте мне большую красную кнопку. Я хочу получить чертеж в Автокаде", — и молодой конструктор, разочарованный результатами векторизации, молча уходит. "Зачем вам векторный чертеж?" — но наш вопрос повисает в воздухе.

Мы уже ломаем стереотипы, мы уже знаем, что красный "Феррари" бессилен на бездорожье, а трактора не ездят по автомагистралям, но нам нужен сейчас именно трактор — простой и надежный, нам нужно пахать и сеять, нам нужно поднимать целину.

От разработчиков, или Чего мы хотели

Сканируем синьку, удаляем "мусор", выравниваем, печатаем — чертеж почти как новенький, сохраняем. Вот они, старые чертежи. В них очень редко вносятся изменения; в подавляющем большинстве случаев их всего лишь нужно быстро найти и распечатать — и всё.

А вот один из типовых примеров работы западного "репро-хауза", или сканирующего бюро. Крупная компания размещает заказ на сканирование 10 000 чертежей, результатом работы бюро является один или несколько компакт-дисков, содержащих файлы изображений, разложенные по различным папкам и поименованные определенным образом в соответствии с требованиями заказчика. В дополнение к файлам, как правило, приложен файл базы данных — например, в формате MS Access, содержащий ссылки на все отсканированные изображения и текстовую информацию, извлеченную из основной надписи чертежа (штампа). После получения этих CD компания-заказчик осуществляет слияние этих данных с данными своей архивной или EDM-системы. Работа оператора бюро заключается в том, чтобы подобрать правильное разрешение при сканировании, затем обработать полученные изображения, т.е. устранить геометрические искажения, "почистить" чертеж и сохранить его под нужным именем, а потом ввести содержимое одного или нескольких полей основной надписи в базу данных. Среднее время этих операций составляет 7-8 минут на один чертеж; при пяти рабочих местах нашего сканирующего бюро производительность составит около 300 чертежей в день, соответственно наш заказ может быть исполнен за 34 рабочих дня или полтора реальных месяца. Совершенно очевидно, что, сократив время обработки до 5 минут, мы увеличим производительность в полтора раза. "Вот за эти три минуты и будем бороться", — решили мы и взялись за дело...

Аксиома и три леммы, из которых мы исходили

Аксиома о растре: отсканированное растровое изображение, будучи поименованным, является полноценным электронным документом, так как его можно искать, смотреть,

редактировать и печатать на бумаге. Читателю, который не согласен с этой аксиомой, можем посоветовать для начала ознакомиться с материалами о гибридном редактировании в предыдущих номерах нашего журнала, а затем продолжить чтение этой статьи.

Лемма первая (исходная): при массовом вводе чертежей путем сканирования каждый из них проходит предварительную обработку по заранее заданному типовому сценарию. Сценариев может быть несколько, и это зависит в основном от качества бумажного оригинала. Каждый из сценариев представляет собой последовательность простых и, как правило, автоматических действий, задаваемых оператором в процессе обработки. Отсюда вывод первый: необходимо минимизировать вмешательство оператора в процесс обработки изображений.

Лемма вторая (спасибо ГОСТу): разновидностей основных надписей или штампов не так много; они, как правило, расположены в одном из углов и содержат основную информацию о чертеже, позволяющую однозначно его идентифицировать. Напрашивается вывод второй: нужно помочь оператору извлечь текстовую информацию из основной надписи чертежа.

Лемма третья и последняя: у каждого предприятия свой электронный архив, система документооборота или хотя бы своя структура базы данных и используемая СУБД. Пытаться угодить всем мы не сможем и не будем. Логичен третий вывод: нужно обеспечить передачу данных о чертеже в произвольный приемник, т.е. предусмотреть достаточно простую возможность подключения внешнего потребителя извлеченной из чертежа информации.

Исходя из этих предпосылок, мы начали разработку программы с названием RasterID, поставив перед собой цель — превратить "обезличенный" растровый чертеж в полноценный электронный документ с уникальным идентификатором ID в вашей базе данных, отсюда и название программы. Мы также попытались создать удобное и понятное средство для пакетной обработки монохромных растровых изображений.

О главном, или Что у нас получилось

Конечно же RasterID поддерживает все сканеры, работающие через twain-драйверы, мы также умеем напрямую работать со всеми популярными моделями Contex'ов. Мало того, если ваш сканер поддерживает режим пакетного сканирования — как, например, Scamax, — мы со своей стороны гарантируем своевременный прием и сохранение каждого из изображений в отдельный файл либо как страницу внутри многостраничного TIFF-файла. Программа умеет читать и сохранять монохромные растровые изображения в большинстве популярных форматов. Это BMP, RLC, CAL, C4, TG4 и, конечно, все разновидности TIFF.

Теперь о командах обработки, которые могут быть использованы как в ручном, так и в пакетном режимах. Мы посчитали следующий набор команд более чем достаточным для предварительной обработки сканированных чертежей:

- Инвертировать;
- Отобразить зеркально по X/Y;
- Повернуть на 90/180/270;
- Устранить перекос;
- Обрезать различными способами;
- Вписать в ближайший формат;
- Удалить мусор;
- Сгладить;
- Корректировать по 4-м точкам;
- Изменить разрешение.

Мы не будем подробно описывать все доступные команды, для этих целей есть руководство пользователя, а остановимся лишь на новых и наиболее интересных. Это сглаживание и 4-точечная коррекция.

Сглаживание позволяет вам в ряде случаев восстановить качество плохо отсканированного чертежа. Как правило, это чертежи, нарисованные карандашом, или синьки с неправильно подобранным порогом бинаризации в процессе сканирования. При ближайшем рассмотрении на таких чертежах видны "рваные" тонкие линии либо линии, имеющие неровности на границе белого (см. рис. 1, 2).

Принцип работы сглаживания достаточно прост и заключается в конвертировании изображения в оттенки серого с последующим усреднением яркости каждой точки и пороговой бинаризацией. Побочный эффект работы команды — удаление "мусора" и малоразмерных объектов. Поэтому нужно достаточно аккуратно настраивать параметры сглаживания, особенно в режиме пакетной обработки, предварительно проведя эксперимент на выборке типовых чертежей. Данная рекомендация распространяется и на команды удаления "мусора".

Еще одна команда, на которую мы хотим обратить ваше внимание, — это 4-точечная коррекция. Она предназначена для устранения геометрических искажений сканирования. При использовании операции в ручном режиме вам необходимо указать на чертеже четыре угловых точки (как правило, это углы внешней или внутренней рамки) и реальные размеры прямоугольника, после чего RasterID устранит искажения в соответствии с заданными параметрами. А при наличии на чертеже внешней рамки команду можно запускать автоматически: программа попытается найти рамку и подобрать ближайший размер формата из заданного списка, к которому и будет приведен ваш чертеж. Перед этой опера-

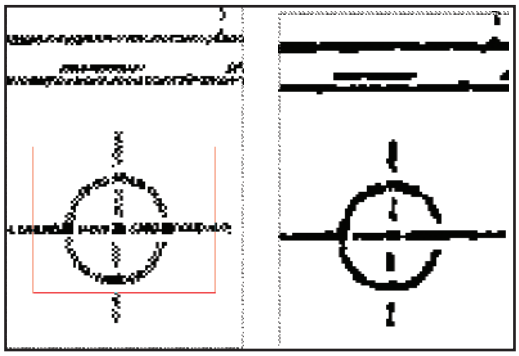

Рис. 1. Сглаживание на изображении с неправильно подобранным порогом бинаризации

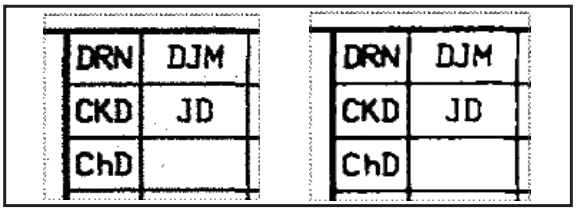

Рис. 2. Сглаживание на изображении с неровностями на границе с белым

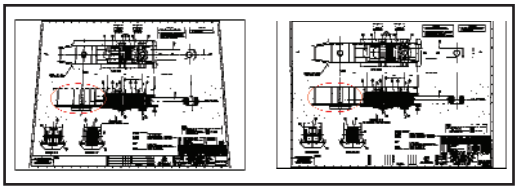

Рис. 3. Пример трапециевидных искажений и результаты работы 4-точечной коррекции

Рис. 5

цией рекомендуем устранить перекос изображения.

Любую из операций вы можете применить как ко всему изображению, так и к его части, программа позволяет вам задать прямоугольную рабочую область, с которой также можно производить операции копирования и вставки.

Вы пробовали писать сценарии?

Если нет, давайте попробуем вместе. Загрузите первое попавшееся изображение и посмотрите на него внимательно. Предположим, вы сканируете только A0; очевидно, что изображение лучше повернуть на 90 градусов. Сделайте это. Если изображение "грязное", запустите команду "Удалить мусор" в автоматическом режиме. Если при этом удалились и текстовые строки, лучше отменить операцию, задав вручную размеры "мусора" существенно меньшие, чем размеры текста.

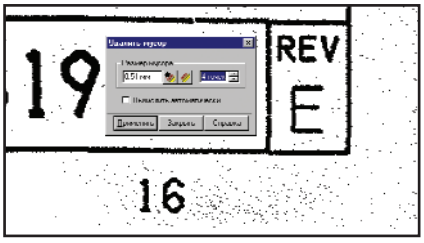

Рис. 4

Пойдем дальше. Взгляните на линии и попытайтесь оценить их качество — весьма вероятно, что придется использовать сглаживание. Будьте аккуратны с подбором параметров — особенно при наличии таких мелких объектов, как тексты, в которых могут залиться дырки в буквах "A" или "p" либо произойдет "слияние" соседних символов. Когда вас устроили результаты, запишите численные значения обоих параметров сглаживания, они нам еще пригодятся.

Смело запускайте "Устранить перекос" в автоматическом режиме: если программа не сможет оценить угол перекоса, операция просто не будет производиться. И наконец, если вас смущают поля вокруг внешней рамки чертежа, вызовите команду "Обрезать по рамке". Теперь можете сохранить или распечатать ваш чертеж, чтобы оценить плоды нашего совместного труда. Очень надеюсь, что вы будете удовлетворены результатом...

Теперь вспомним всё, что мы делали, и вызовем Мастер командных файлов. В левой половине вы видите список всех доступных команд. Вам достаточно мышью переместить все необходимые команды в правую часть диалога и расположить их в нужной последовательности. Так выглядит наш сценарий для пакетной обработки.

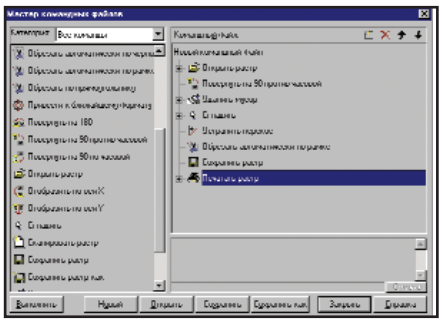

Рис. 6

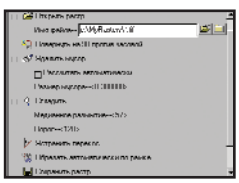

Рис. 7

После этого нужно задать параметры для команд, которые этого требуют (см. рис. 7). Здесь это команды "Открыть растр", "Удалить мусор" и "Сгладить". Обратите внимание, что в качестве имени файла мы можем указать все изображения из определенной папки, в нашем случае это "c:\MyRasters".

И еще пара рекомендаций перед тем как вы нажмете кнопку "Выполнить". Будьте внимательны с командами "Обрезать по рамке": убедитесь, что она действительно есть, в противном случае у ваших валов останутся лишь фланцы. Если вы хотите получать изображения непосредственно со сканера, вам достаточно заменить команду "Открыть растр" на команду "Сканировать", указав при этом источник, включить сканер и дать ему прогреться. Если вы боитесь испортить исходные оригинальные файлы — просто замените команду "Сохранить" на "Сохранить как", указав другую директорию. Если вы укажете при этом другой формат, то получите отличный автоматический конвертор. Сохраните ваш сценарий для будущего использования. Теперь можете нажать кнопку "Выполнить" и заняться чем-нибудь полезным — например, посмотреть, с каким удовольствием прожорливый Contex поглощает ваши синьки, которые через полминуты в практически первозданном виде выплевывает ваш струйный принтер. Это и есть тот самый "репро-хауз", или новая жизнь старых чертежей.

Теперь о штампах или основных надписях

"Indexing" — это действительно просто. Не стоит пугаться иностранного слова, оно обозначает всего лишь процесс извлечения из чертежа уникальной текстовой информации, которая нужна ***вам*** в ***вашей*** системе документооборота. Как правило, такая информация содержится в угловом штампе. RasterID сможет найти основную надпись на чертеже и попытается распознать содержимое указанных вами полей штампа с последующей передачей в вашу базу данных. Для этого вам нужно один раз научить программу тем разновидностям штампов, которые вы используете, и сохранить их в отдельных файлах шаблонов.

Процесс обучения прост и понятен. Загрузите произвольный чертеж со штампом. Поверните его, если необходимо, устраните перекос и постарайтесь максимально улучшить качество чертежа. Затем вызовите команду "Обучить штампу" и

укажите прямоугольником зону, где расположена основная надпись.

RasterID постарается определить структуру штампа и отобразит ее на экране. Если топология штампа определена ошибочно, ее можно отредактировать, добавляя новые поля или удаляя некорректно распознанные. Включите анализ полей, содержимое которых вас интересует, и дайте им имена, соответствующие именам полей вашей базы данных. Если вы сканируете произвольно ориентированные чертежи, сообщите об этом нашей программе — RasterID будет искать штамп в каждом из четырех углов.

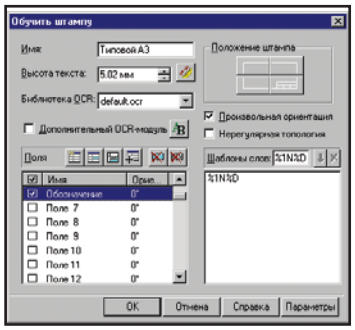

Рис. 8

Рис. 9

Программа умеет распознавать текстовые строки в каждом из заданных полей, используя встроенный либо внешний модули с возможностью задания шаблонов слов. Например, если вы знаете, что обозначение чертежа всегда представляет собой заглавную букву и несколько цифр, можно указать это последовательностью "%1N%D", тем самым улучшив качество распознавания. Вы также можете подключить как дополнительный свой собственный модуль распознавания текстов и использовать его. Правда, для этого вы должны знать хотя бы Бэйсик. Когда обучение закончено, сохраните шаблон штампа в отдельном файле для дальнейшего использования.

Теперь вы можете начать ***Indexing***. Вспомните тот сценарий обработки, который мы с вами создали, и снова позовите Мастер командных файлов. Типовой сценарий для извлечения полезной информации из чертежа может выглядеть так (рис. 10).

Обратите внимание: мы не включили в наш сценарий команды

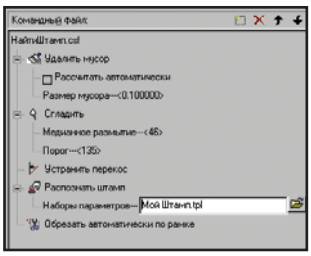

Рис. 10

открытия и сохранения файлов, а добавили в него только команды обработки и распознавания штампа. Не волнуйтесь, это сделано специально, так как список нужных файлов мы сможем задать непосредственно перед пакетной обработкой. В качестве параметра команды "Распознать штамп" мы указали имя созданного нами ранее шаблона "Мой Штамп.tpl". Теперь мы сохраним наш сценарий в файле с именем "Найти Штамп.csf" и будем его постоянно использовать.

Перед тем как начать процесс индексирования, давайте выберем один или несколько приемников ваших данных.

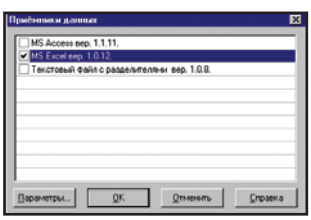

Рис. 11

В стандартную поставку программы входят три наиболее популярных: текстовый файл с разделителями, файл MS Excel и файл MS Access. Если у вас нет навыков программирования, вы можете использовать готовые приемники данных, а затем производить импорт этих файлов в свою систему документооборота. Сейчас мы так и поступим, выбрав в качестве приемника MS Excel, а чуть позже объясним, как же все-таки принять извлеченные данные непосредственно в свою систему.

Не буду больше испытывать ваше терпение — вызовем Мастер пакетных заданий, где вы сможете выбрать нужные файлы, наблюдать за процессом обработки и проанализировать его результаты.

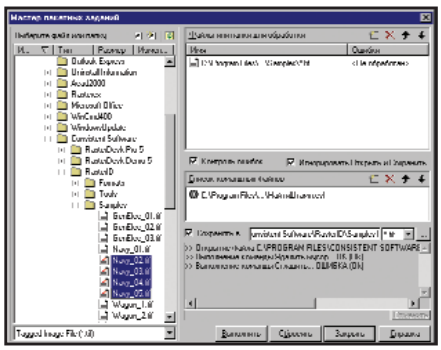

Рис. 12

Когда вы создали один или несколько типовых сценариев обработки ваших растровых изображений, определились с приемником данных, отсканировали тысячи чертежей, вам остается только выбрать группу файлов, требуемый сценарий и нажать кнопку "Выполнить". Если вы включили при этом контроль ошибок, все необработанные чертежи, и только они, останутся в списке для обработки. Придя завтра на работу, вы сможете просмотреть их и завершить обработку вручную.

А пока предлагаю взглянуть на результаты индексирования, экспортированные в MS Excel.

Созданная программой таблица содержит имя файла, угол поворота изображения, вырезанный из него штамп, имена и содержимое распознанных полей. Вы можете, глядя на оригинальный штамп, проверить и отредактировать распознанные текстовые строки. Если количество чертежей не слишком велико и у вас еще нет своего архива, это одно из самых простых решений.

Похожее решение мы предлагаем и для MS Access, где вы можете просмотреть все обработанные файлы, организовать поиск и редактирование записей и воспользоваться прочими возможностями, предоставляемыми этой незатейливой СУБД. Непосредственно отсюда вы можете

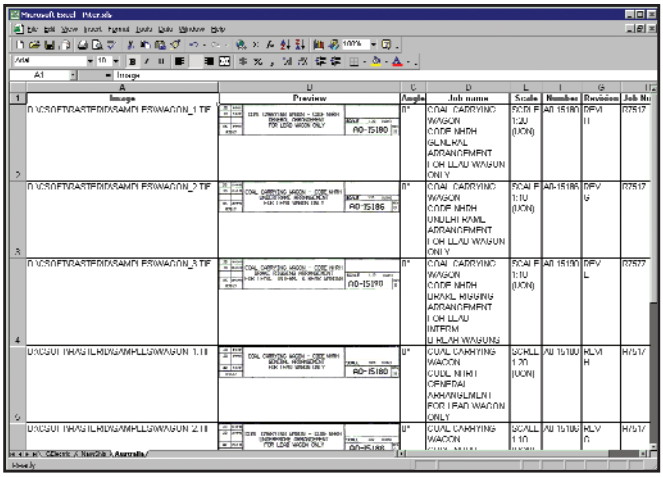

Рис. 13

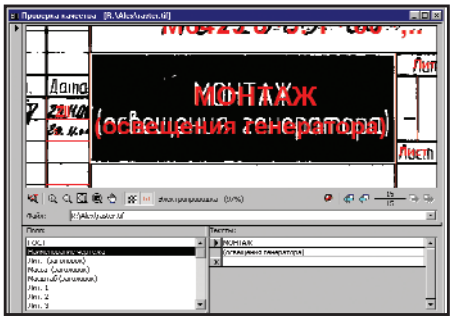

Рис. 14

вызвать RasterID, чтобы произвести необходимые операции над изображением вручную или просто его распечатать (см. рис. 14). Все, что видите на картинке, представляет собой программу на Бэйсике, работающую внутри MS Access и поставляемую в исходных текстах, которые можно изменять, дополнять и использовать по своему усмотрению. Но обо всем по порядку...

Взгляд изнутри, или Как добавить свое...

У вас уже есть свой электронный архив и вы привыкли к своей системе документооборота, а это значит, что у вас есть люди, которые все это создали или хотя бы осуществляют техническую поддержку. Эта глава посвящена им.

Я надеюсь, все представляют себе детский конструктор "Lego". Эдакая куча красивых разноцветных кубиков, стеклышек и колесиков в яркой коробке с буклетом, показывающим, что из них можно собрать. Так вот, наша программа — это красивая игрушка-трактор, собранная из таких кубиков. Вы можете разобрать его, собрать свой и прицепить к нему тележку...

А теперь попробую серьезно.

Уже больше десяти лет мы профессионально занимаемся обработкой и распознаванием растровых изображений и накопили немалый опыт. Все лучшее, что мы умеем и используем в наших программах гибридного редактирования, теперь открыто для вас. Давайте проведем маленький эксперимент: запустите MS Word и позовите команду "Вставить/Объект". Вы увидите список хорошо знакомых названий, среди которых будет CSRasterTT. Это и есть наш объект, который открывает, показывает растры, хранит внутри себя полный набор операций с растровым изображением, умеет искать штамп и передавать распознанные текстовые строки во внешний приемник данных — ваш приемник.

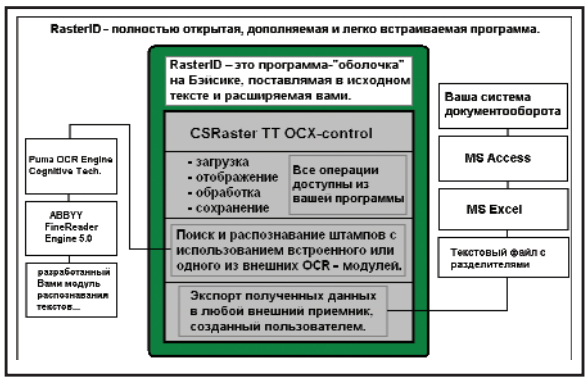

Рис. 15

RasterID — это всего лишь красивая оболочка, написанная на языке VisualBasic и использующая значительную часть функций нашего объекта. Получив программу, вы получаете полный набор исходных текстов оболочки и всех приемников данных, включая подробную инструкцию по написанию собственного. Вы можете легко и просто расширять нашу программу, менять пользовательский интерфейс, встраивать наш объект в свою архивную систему для просмотра и обработки растровых изображений. Все стандартные приемники данных прекрасно модифицируются и адаптируются к вашим нуждам. И, конечно, мы обеспечиваем вас профессиональными консультациями и подробной информацией обо всех возможностях нашей программы.

В заключение – пара историй...

Голландская компания Eneco после покупки RasterID в течение одной недели адаптировала приемник данных для MS Access. Задача состояла в обработке 80 000 растровых изображений и заполнении пяти полей базы данных, значения трех из которых содержатся в имени файла, а два извлекаются из штампа чертежа. В файлах содержались двенадцать разновидностей штампов. После двух недель работы с программой был проведен временной тест на выборке из 500 различных чертежей. Эксперимент показал, что время обработки одного чертежа сократилось с восьми до трех с половиной минут, после чего было принято решение о внедрении программы.

Или вот другой пример. Питерское отделение нашей компании разработало технологию наполнения электронного архива для ЦКБ "Рубин". Для нормальной работы оператора необходимо было добавить простую команду, которая обрезает изображение по нужному формату бумаги. После небольших изменений в RasterID была добавлена отдельная кнопка, которая позволяла выбрать нужный формат из списка. В поле чертежа появлялась рамка заданного размера, и оператор, мышью позиционируя рамку, производил обрезку изображения.

Мы можем привести еще массу примеров адаптации и использования нашей программы в различных западных и российских компаниях и очень надеемся, что она поможет вам в создании собственного электронного архива.

Александр Крылов
Consistent Software
Тел.: (095) 360-1524
E-mail: alex@csoft.sitek.net