

Технология сканирования и распознавания документов SCANIFY

Один из атрибутов нашей повседневной жизни — регистрация персональных документов (паспорт, водительское удостоверение, страховое свидетельство и т. п.), паспортных данных, их верификация и ввод в информационные системы и базы данных различных организаций. Однако до сих пор подобные операции, ставшие обязательными элементами многих информационных систем, чаще всего выполняются вручную, несмотря на практически всеобщую компьютеризацию и широкое применение методов сканирования и распознавания при работе с бумажными документами.

Причин тому несколько, и одна из основных — трудности автоматического распознавания информации, зафиксированной на подобных официальных удостоверениях личности. Это связано с наличием цветного гербового фона на многих документах, а также с большим разбросом в яркости и контрастности отдельных полей. Проблема усугубляется широким использованием пластика для изготовления удостоверений и ламинирования бумажных документов. Это хорошо видно на примере ксерокопий, которые для таких носителей информации часто получаются явно недостаточного качества.

Второй важный момент заключается в том, что успех автоматизации обработки бумажных документов во многом определяется возможностью разработки такого дизайна документа, чтобы он максимально облегчал задачу распознавания. Но в задаче распознавания государственных документов такой прием не работает: вопрос о создании, например, нового образца паспорта в форме, рассчитанной на автоматическую обработку записанной в нем информации, пока находится лишь в стадии рассмотрения.

Учитывая это, компания Cognitive Technologies (www.cognitive.ru) создала технологию Scanify, которая позволяет автоматизировать процесс ввода документов, достигая при этом гораздо более высокого (по сравнению со стандартным ксерокопированием) качества копии, а также автоматически извлекать и экспортировать из документов не только текстовую, но и графическую информацию (фотографию, подпись владельца),

операции, вести автоматический контроль и проверку вводимых данных.

Один из ключевых этапов распознавания — отделение собственно распознаваемых символов от фона. Документы государственного образца, как правило, имеют фоном гербовую бумагу. Это фактически делает невозможным стандартную пороговую фильтрацию фона — ведь нередко контрастность деталей фона выше, чем контрастность заполнения. Таким образом, задача распознавания сводится к классической задаче цвето-текстурного анализа.

В последнее время появились достаточно мощные математические инструменты для анализа и фильтрации текстур, и работоспособность технологии Scanify обеспечивается одним из самых перспективных из них — Wavelet-фильтрацией. Вэйвлеты (wavelet) позволяют исследовать смеси сигналов и выделять отдельные компоненты даже в тех случаях, когда их вклад меняется от точки к точке текстуры. В силу того, что несущие в себе смысловую информацию символы присутствуют на фоне вездесущего гербового рисунка далеко не везде, это свойство вэйвлетов позволяет выделять на сложном изображении значимую информацию.

Дополнительные сложности при распознавании создают печати, «наползающие» на значимый текст и графическую информацию. Для решения данной проблемы в системе используется метод обобщенного преобразования Хафа (для поиска печатей различных форм) и линейная спектральная модель цвета (для разделения изображений печати и текста).

В результате цвето-текстурного анализа и обработки получается «очищенное» изображение документа с подавленным гербовым фоном, качество которого существенно выше, чем у самых лучших ксерокопий. Это позволяет избавиться от ксерокопирования документов, заменив этот процесс распечаткой на принтере обработанного технологией Scanify изображения документа. Кроме того, за счет подавления фона сложность графического образа документа снижается, и появляется возможность значительного сжатия объема графических образов.

В классических системах ввода документы поступают на сканирование по од-

ному. Это оправдано при потоковом сканировании анкет или форм, однако при построении системы для обработки документов более естественной представляется возможность ввода нескольких документов за один проход сканера. Преимущество такого подхода — простота и скорость ввода, а также возможность перекрестных проверок данных между различными документами.

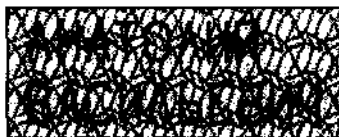
Для реализации такой функциональности система должна найти, разделить и классифицировать различные документы, лежащие на стекле сканера. Эта задача решается введением так называемых профилей — наборов признаков документов, создаваемых в процессе обучения системы. Текущая версия продукта поддерживает распознавание и обработку следующих типов документов: паспорт гражданина РФ, заграничный паспорт, водительское удостоверение, военный билет, свидетельство о регистрации ТС, свидетельство обязательного пенсионного страхования.

Дальнейшая обработка документа в Scanify (собственно распознавание, верификация и экспорт данных) происходит с использованием технологий Cognitive Forms. В результате обеспечивается не только высокое качество автоматического распознавания, но и полный контроль вводимых данных оператором, в том числе перекрестная проверка взаимного соответствия данных из нескольких документов, а также соответствия документа записи в базе данных.

Scanify предоставляет возможность экспорта введенных данных в различных форматах: XML, DBF, Business VCard, чистый текст и т. д. Набор функций Scanify API позволяет создавать приложения в области автоматической обработки документов, а также встраивать весь функционал технологии (в виде объектов ActiveX и низкоуровневого интерфейса C++) непосредственно в информационную систему предприятия. Тем самым обеспечивается возможность простого подключения функций обработки документов в приложения, использующие скриптовые языки с поддержкой ActiveX, а также возможность интеграции на уровне исходных кодов приложения.



Исходный документ



Расознаваемое поле



Результат обработки стандартным методом адаптивной бинаризации



Результат вэйвлет-обработки по технологии Scanify