УДК 004

РАЗРАБОТКА ПРОГРАММНОГО МОДУЛЯ ДЛЯ РАСПОЗНАВАНИЯ И ИДЕНТИФИКАЦИИ БЛАНКОВ ДОКУМЕНТОВ ПО ЗАДАННЫМ ШАБЛОНАМ, С ПОСЛЕДУЮЩЕЙ РЕГИСТРАЦИЕЙ В БАЗЕ ДАННЫХ

Т.Б. Каримкулов
Научный руководитель: О.В. Марухина к.т.н., доцент ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: tantay@tpu.ru

Abstracsts: *This particular solution is designated for stream input forms, which automatically extracts the identifying information from paper documents and saves it in the information system of the company. Search and pattern recognition or text to graphic images occurs along the contours, using filters of OpenCV and threshold segmentation by method of (OTSU).*

Keywords: UmguCV, Barcode, image.
Ключевые слова: Компьютерное зрение, изображение, распознавание объектов.

Поиск и распознавание образов или текста на графических изображениях является актуальным научным и практическим направлением в связи с нарастающими потоками документооборота в современных деловых и научных центрах. Центр обеспечения качества образования (ЦОКО) является структурным подразделением федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский Томский политехнический университет». Одной из основных задач центра является организация работ по созданию и модификации банков оценочных материалов (контрольных задач, тестовых заданий, программно-дидактических тестов и т. п.) в различных областях знаний и оценки качества обучения, разработанный Центром обеспечения качества образования, направленный на развитие механизмов мониторинга системы образования и уже внедренный в работу вуза.

Одной из наиболее важных задач, стоящих перед специалистами информационно-программного обеспечения Центра тестирования ТПУ, – обработка бланков для тестирования. Задача, сформулированная в рамках данного исследования, такова: имеется множество бланков документов формата А4, которые переводятся в цифровой вид (в формате многостраничных .tiff-файлов) посредством сканирования. Бланк имеет уникальный идентификационный номер (штрих-код) – ID (в формате XX – XXXXXXXXX), в котором зашифрована серия (классификатор бланка) и номер бланка соответственно. Цель – распознать ID бланка, прикрепить его к файлу (либо в название, либо в метаданные файла, в зависимости от возможности обработки СУБД) и загрузить на сервер с последующей регистрацией в базе данных.

Таким образом, сформируем алгоритм формализации работы с системой распознавания бланков:

1. Снять изображение бланка с устройства ввода.
2. Привести изображение в стандартный вид.
3. Найти на изображении объект (штрих-код), который необходимо распознать.
4. Распознать объект.
5. Сохранить штрих-код.
6. Выгрузить изображение со штрих кодом на сервер.

Разработка программного обеспечения для распознавания объектов производилась на основе открытой библиотеки компьютерного зрения OpenCV, с использованием кроссплатформенной «обертки» (wrapper) для .NET библиотеки – Emgu CV [1]. Данное решение было принято, исходя из того, что OpenCV имеет открытый исходный код и обладает необходимым функционалом для работы с машинным зрением. Описываемая программа разрабаты-

валась на языке C#, который продолжает набирать популярность в программировании под MS Windows. Загрузка и масштабирование изображения выполняются путём создания объекта класса System.Drawing.Bitmap и передачи нужных параметров в его конструктор.

Основываясь на экспериментальном опыте, был выбран метод нахождения объектов по контурам. Перед тем как начать поиск контуров, изображение необходимо предварительно обработать различными фильтрами. Приведение изображения к универсальному виду разделена на этапы:

1. Бинаризация изображения – преобразование в чёрно-белое изображение, где черный будет отвечать наличию «краски», а белый – ее отсутствию. Делается это потому, что ряд алгоритмов, например, получение контура объекта, конструктивно не работают с полутонами. Одним из самых простых способов бинаризации является пороговая фильтрация (выбираем *t* в качестве порогового значения, все пиксели с интенсивностью больше *t* – фон, меньше – «краска»), но в силу ее низкой адаптивности будет использоваться метод порогового преобразования «Otsu Threshold» [2].
2. Размытие (выполняется для уменьшения шума).
3. Нахождение границ по алгоритму Кенни [3].
4. Нахождение маркеров, самая важная деталь в бланке – черные квадраты (я назвал их маркеры). Верхний центральный маркер нужен для определения где верх, а где низ бланка.
5. Поворот изображения.

После предварительной обработки, на изображении находятся контуры и записываются в специальные переменные в виде последовательности точек с определёнными координатами. Затем отфильтровываются мелкие и крупные контуры, по заданным условиям (размерам), остается контур максимально приближенный к искомому, с учетом погрешности. Найденный контур (в нем содержится ID бланка) вырезается для распознавания. ID бланка будет упакован в штрих код PDF-417 (в планах и другие типы штрих кодов, а также серийный номер из последовательности символов).

В настоящее время в мире существует множество разновидностей баркодов. Распознавание и генерация большинства из них реализована в библиотеках, доступных для разработчиков. При работе со штрих-кодами в .NET используется библиотека Zxing. Библиотека умеет генерировать и распознавать всевозможные 1D и 2D баркоды: QR-Code, Codabar, EAN, UPC, Aztec, Data Matrix. И главное, она умеет работать с PDF 417. Тем не менее, иногда можно наткнуться на оригинальный тип баркода, распознать который сходу не получится. И тогда метод тщательного всматривания и использования хорошо спроектированной библиотеки с открытым исходным кодом помогает быстро получить результат.

Данное решение разработано для Центра обеспечения качества образования при ТПУ и предназначено для потокового ввода бланков, которое автоматически извлекает идентифицирующую информацию из бумажных документов и сохраняет ее в базу данных корпоративной информационной системы.

Список литературы

1. EmguCV Tutorial [Электронный ресурс], URL: http://www.emgu.com/wiki/index.php/MainPage (свободный, дата доступа 10.03.15).
2. Сегментация изображения. Метод Отса [Электронный ресурс]. – URL: http://habrahabr.ru/post/128768/ (свободный, дата доступа 10.03.15).
3. Оператор Кэнни [Электронный ресурс]. – URL: https://ru.wikipedia.org/wiki/Оператор_Кэнни (свободный, дата доступа 10.03.15).