

АНАЛИЗ СТРУКТУРЫ ТАБЛИЦ НА ИЗОБРАЖЕНИЯХ ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ

Шолохова Татьяна Николаевна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: tanja200596@mail.ru

Потребность в преобразовании документов в электронный вид возникла из-за огромного количества проблем, связанных с хранением материальных форм данных: отсутствие возможности быстрого доступа, сложность создания копий, неизбежные повреждения с течением времени.

Целью данной работы является исследование применимости непрерывно-морфологического подхода к решению задачи распознавания структуры таблиц на изображениях документов. Непрерывно-морфологический подход подразумевает, что метод распознавания основывается на геометрических свойствах объектов, полученных путём аппроксимации растрового представления изображения.

Предложенный алгоритм содержит следующие стадии: предобработка изображения документа; построение и обработка внутреннего скелета изображения документа для выделения прямых линий; построение и обработка внешнего скелета изображения документа; поиск компонент связности внешнего скелета для выделения ячеек таблицы; построение графа связности ячеек; классификация документа на основе структуры таблиц.

Первые шаги выполнения работы заключаются в построении внутреннего и внешнего скелетов изображения документа. В геометрии и компьютерной графике под термином «скелет фигуры» понимают множество точек (пикселей, вокселей), являющихся центрами максимальных вписанных кругов. Алгоритм построения скелета позволяет получить его в виде плоского геометрического графа [1].

Внутренний скелет используется для извлечения прямых линий из изображения документа. Предобработка изображения для построения внутреннего скелета заключается в применении морфологической операции дилатации. В построенном скелете толщина рёбер, образующих скелет текста заметно больше толщины рёбер, образующих скелет таблицы. (Рис. 1.) Смежные горизонтальные рёбра скелета соединяются в горизонтальные линии, а смежные вертикальные рёбра в вертикальные линии.

После извлечения прямых линий на изображении удаляется весь

текст, далее производится построение внешнего скелета. Все отдельные компоненты связности скелета являются отдельными ячейками таблицы. (Рис. 2.) Для найденных ячеек вводятся отношения: горизонтальное соседство (левая граница одной ячейки и правая граница другой лежат на одной вертикальной прямой и пересекаются) и вертикальное соседство (аналогично). (Рис. 3.)

Для классификации документа из структуры графа соседства ячеек извлекаются признаки, описывающие структуру таблицы, такие как: количество ячеек; количества строк и столбцов; количество рёбер в графе соседства; количество компонент связности в графе соседства и другие.

Реализация предложенного алгоритма протестирована как на изображениях высокого разрешения, так и на нечётких фотографиях документов. Проведённые исследования показывают, что рассматриваемый подход является перспективным и его дальнейшее развитие может привести к разработке эффективных алгоритмов классификации документов.

Иллюстрации

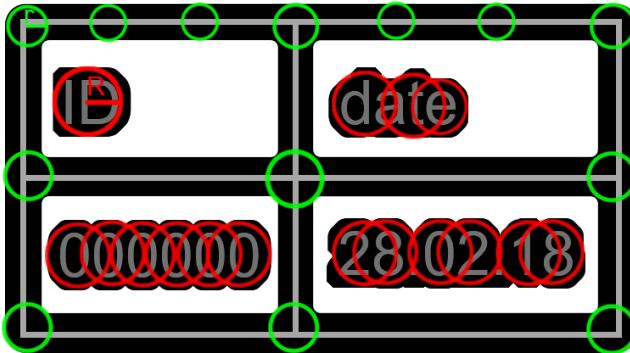


Рис. 1. Стадия поиска прямых линий. (изображение до дилатации отмечено серым цветом, чёрным отмечено изображение после дилатации)

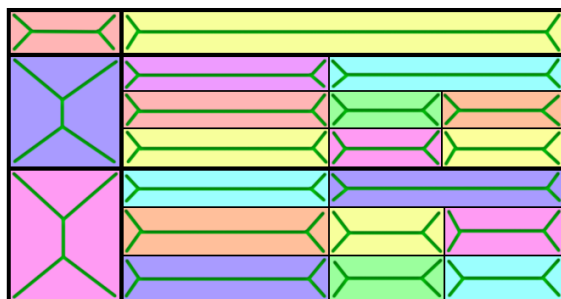


Рис. 2. Внешний скелет изображения таблицы: выделение ячеек.

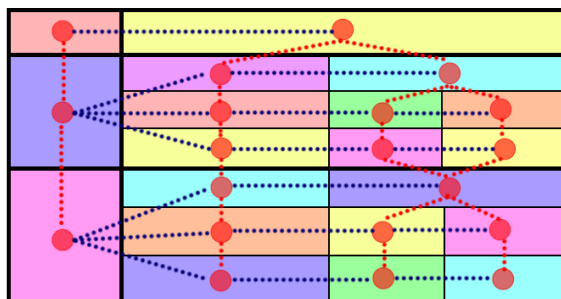


Рис. 3. Структура таблицы: граф смежности ячеек.

Литература

1. Местецкий Л. М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. // Москва, Физматлит, 2009.
2. Местецкий Л. М. Скелет многоугольной фигуры — представление плоским прямолинейным графом. // Санкт-Петербург, ГРАФИКОН-2010.