Сегментация отсканированных документов

Алексей Вилькин, Марта Егорова
НИЯУ «МИФИ», Москва, Россия
{ aleksey.vilkin, marta.egorova}@gmail.com

Аннотация

Рассматривается восходящий подход сегментации отсканированных документов на области фона, текста и фотографий. На первом этапе изображение разбивается на блоки. Для каждого блока считается ряд текстурных характеристик. На основе этих характеристик построены два AdaBoost комитета классификаторов. Они объединяются в дерево решений, которое определяет тип блока. На втором этапе типы блоков итеративно корректируются на основе анализа соседних областей. На тестовой выборке достигнута точность сегментации более 80%.

Ключевые слова: *сегментация отсканированных документов, текстурные признаки*

1. ВВЕДЕНИЕ

Большая часть информации в современном мире хранится в электронном виде. Многие печатные документы также переводятся в электронный вид с помощью систем OCR (*optical character recognition*), которые требуют сегментации областей. Также анализ областей документа требуется при сохранении сканированных документов в PDF с MRC (*mixed raster content*) сжатием. При MRC сжатии для фона, текста, рисунков создаются собственные маски, и каждый тип информации сжимается с помощью собственного алгоритма сжатия.

Существует довольно много исследований по данной теме, в литературе можно выделить два основных подхода: восходящий (*bottom-up*) и нисходящий (*top-down*) подходы. В bottom-up [1,2,3] алгоритмах анализ начинается с объектов низкого уровня, таких как пиксели, зоны, соседние области; затем полученные объекты соединяются и классифицируются, как области документа. Такие алгоритмы хорошо обрабатывают области со сложной формой, но не учитывают при анализе объекты высокого уровня. Напротив, top-down алгоритмы[4] начинают со всего изображения в целом и пытаются разделить его на области конкретного типа. Такие алгоритмы не всегда способны обработать регионы сложной формы, например, непрямоугольные блоки текста или заголовки на несколько колонок текста. Многие подходы хорошо справляются со своей задачей для ограниченного набора изображений, так как для анализа используют специфичные признаки или предположения.

Мы предлагаем восходящий подход для работы с полутоновыми 8 bpp отсканированными изображениями документов, который на первом этапе классифицирует блоки, используя текстурные признаки, а затем итеративно уточняет класс блока за счет учета класса соседних блоков.

2. ОПИСАНИЕ АЛГОРИТМА

2.1 Количественные характеристики текстур

Изображение I первоначально разделяется на блоки размера NxN пикселей, где значение N зависит от разрешения отсканированного документа. Были рассмотрены размеры окна в интервале 10-80 пикселей и два варианта пересекающихся блоков и расположение блоков «плиткой». Размер окна должен быть достаточно маленьким, чтобы отделять различные области (текст, рисунок, фон) друг от друга, но и достаточно большим, для того чтобы полученные характеристики были адекватными.

В процессе экспериментов исследовалось более 20 количественных признаков текстур [1,3,5] среди которых при помощи AdaBoost toolbox [6] были выделены наиболее информативные:

- средняя яркость блока B_i: $\overline{B}_i = \dfrac{\sum_{r=1}^{N}\sum_{c=1}^{N} B_i(r,c)}{N^2}$ (1)

- средняя разница средних яркостей блоков B_k в 4-связной окрестности блока B_i: $\overline{dB}_i = \dfrac{\sum_{k=1}^{4} |\,\overline{B}_i - \overline{B}_k\,|}{4}$ (2)

- среднее вертикальной dB_y^i и горизонтальной dB_x^i производной блока : $\overline{d_{x,y}B}_i = \dfrac{\sum_{r=1}^{N}\sum_{c=1}^{N-1} dB_x^i(r,c) + \sum_{r=1}^{N-1}\sum_{c=1}^{N} dB_y^i(r,c)}{2N(N-1)}$ (3)

- однородность блока B_i: $H = \sum_{i,j} \dfrac{N_d(i,j)}{1+|\,i-j\,|}$, (4)

где N_d нормированная матрица совместной встречаемости (*co-occurrence matrix*), d задает пространственное отношение.

- процент пикселей с градиентом больше порога:

$$P_g = \sum_{\forall (r,c)\in Bi} \{1 \,|\, \nabla\ B_i(r,c) > T\} / N^2 \text{, где } \nabla\ B_i(r,c)$$

вычислется как квадратный корень из суммы квадратов горизонатльной и вертикальной производных. (5)

- процент изменений значений пикселей после морфологической операции открытия B_i^o над бинарным изображением B_i^b, полученным бинаризацией по порогу 128:

$$P_m = \sum_{\forall (r,c)\in Bi} \{1 \,|\, B_i^o(r,c) \neq B_i^b(r,c)\} / N^2 \,. \quad (6)$$

а б в

Рисунок 1 Результаты сегментации документов: а) исходные документы; б) сегментация предложенным методом, в) сегментация программой FineReader 9.0

2.2 Классификация блоков

Количественные характеристики текстур позволяют отнести блок к одному из трех типов: рисунок, фон, текст. Для этого построено два AdaBoost классификатора: один для определения фотографий C_i , второй - для текста C_t. На рисунке 2 показано дерево решений, объединяющее оба классификатора для разделения на 3 класса. Каждый классификатор обучен на выборке из 57 документов разных типов, из разных источников. Если оба классификатора C_t, C_i дают положительный результат, то выбирается тот, что по модулю дальше от соответствующего порога T_t, T_i. Если оба отрицательные, то блок классифицируется как фон.

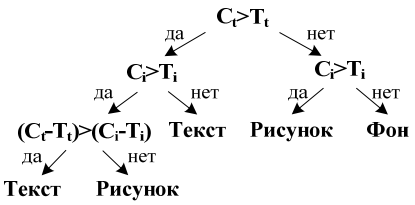

Рисунок 2 Дерево решений для классификации блока

Пороги полагаются равными сумме весов всех слабых классификаторов. Изменяя значение порога можно влиять на количество ошибок первого и второго рода.

2.3 Уточнение класса блоков

В результате классификации иногда возникают ошибки: отдельные блоки классифицируются неверно. Для получения однородных областей и подавления шумов для блока, класс которого отличается от класса соседних блоков мы ипользуем простую, но эффективную процедуру: определяем количество соседей каждого класса, и если число соседей какого-то класса больше 6, то блок реклассифицируется. Данная итерация выполняется несколько раз. Чем меньше окно, тем более разнородным, зашумленным получается результат.

3. РЕЗУЛЬТАТЫ

Работа алгоритма проверена на тестовой выборке, состоящей из 174 отсканированных изображений газет, документов, журналов разных типов. Результат классификации пикселов изображений представлен в таблице 1 в виде матрицы ошибок (*confusion matrix*) [7] усредненной по всем изображениям. Истинные значения посчитаны по вручную размеченным маскам. Для данного критерия наилучшие результаты с учетом производительности при разрешении сканирования 300 dpi были полученыдля размера окна в 30 пикселей и расположения блоков «плиткой».

Таблица 1 Относительная матрица ошибок, усредненная по всем изображениям тестовой выборки

		Истинные		
Классифицированные		Фон	Текст	Рисунок
	Фон	0.37	0.07	0.01
	Текст	0.04	0.33	0.02
	Фото	0.01	0.03	0.12

Предложенный алгоритм правильно классифицировал 82% пикселей. FineReader 9.0 правильно классифицировал 89% пикселей. Таким образом алгоритм даёт результаты, сравнимые с коммерческим ПО. Далее мы планируем улучшить алгоритм, в первую очередь, исследовать классификацию блоков с использованием SVM и, k-NN, а также применить более сложный анализ классифицированных блоков.

4. СПИСОК ЛИТЕРАТУРЫ

[1] J.J.Sauvola, M.Pietikäinen, *Page segmentation and classification using fast feature extraction and connectivity analysis. Int. Conf. on Document Analysis and Recognition, 1995*

[2] F. Wahl, K. Wong, R. Casey, *Block segmentation and text extraction in mixed text/image documents, Computer Graphics and Image Processing, 1982, Vol.20, pp. 375-390.*

[3] H. S. Baird, M. A. Moll, Chang An, *Document Image Content Inventories. Proc. of SPIE/IS&T Document Recognition & Retrieval, 2007.*

[4] F. Cesarini, S. Marinai, G. Soda, M. Gori, *Structured Document Segmentation and Representation by the Modified X-Y tree. International Conference on Document Analysis and Recognition, 1999, pp. 563.*

[5] Шапиро Л., Стокман Дж., *Компьютерное зрение, Бином. Лаборатория знаний, 2009.*

[6] *http://graphics.cs.msu.ru/ru/science/research/machinelearning/adaboosttoolbox*

[7] *http://en.wikipedia.org/wiki/Confusion_matrix*

Об авторах

Вилькин Алексей - студент 5 курса факультета Кибернетики НИЯУ «МИФИ». Его областью интереса являются распознавание образов, обработка изображений.

Марта Егорова – аспирантка НИЯУ «МИФИ». Ее областью интереса являются оценка и улучшение качества изображений, алгоритмы компьютерного зрения.