Сегментация изображений страниц древних рукописей[1]

© Южиков В.С.

Казанский Государственный Университет
Y-Vladimir@yandex.ru

Аннотация

В статье описывается алгоритм для сегментации изображения страницы с текстом. Задача сегментации состоит в соотнесении каждого элемента страницы к одному из двух классов – текст или рисунок. Работа алгоритма начинается с разбиения всего изображения на небольшие области. Для классификации каждой области используются следующие критерии: 1). Доля черных пикселей во всей области. 2). Величина разброса толщины отдельных элементов области. 3). Наличие чередующихся строк и междустрочий.

1 Введение

Процесс создания электронных библиотек старинных книг и рукописей имеет свою специфику, обусловленную долгим хранением (часто в неподходящих условиях), а также ценностью многих оригиналов. Но при оцифровке приходится сталкиваться со многими характерными проблемами, связанными с дефектами страниц, таких как выцветание букв, неравномерный цвет бумаги, крупные и мелкие пятна и т.д. Это мешает дальнейшему использованию полученных изображений.

Кроме того, множество проблем возникает и при распознавании текста с таких изображений из-за наличия сильных дефектов и помех. Также на это влияет и отсутствие словарей древнерусской орфографии, играющих немалую роль в качественном распознавании. Поэтому, зачастую, единственным вариантом является создание библиотеки состоящей из отсканированных изображений страниц. В связи с этим возникает задача обработки и анализа таких изображений. Методы и алгоритмы для очистки изображений рукописей от пятен, помех, удаления просвечивания обратной стороны листа рассмотрены в [5].

Следующей важной задачей предварительной обработки является сегментация изображения страницы, т.е. необходимо определить границы текстовых блоков и блоков с рисунками – это позволяет описать макет страницы и использовать это описание при поиске и классификации изображений. Кроме того, выделенные блоки можно будет передать во внешнюю программу для дальнейшей обработки, например для распознавания символов в текстовых блоках, обработке и улучшению иллюстраций и т.д.

2 Обзор литературы

Данная проблема относится к задачам сегментации, то есть выделении на изображении однородных по какому-либо критерию областей. В случае фиксированного шаблона страницы сегментации, как таковой, не требуется – анализируется только положение опорных элементов оформления страницы. Далее, на их основе, вычисляются координаты блоков с информацией для ввода, используя заложенный макет [1, 2].

Если же данная информация отсутствует, то отделить на странице области с текстом от рисунков значительно сложнее, т.к. даже человек не всегда может однозначно определить, к какому классу отнести данный элемент. В качестве примера можно привести надписи на географических картах – возникает дилемма, то ли считать их частью рисунков, то ли отнести к тексту для последующего распознавания.

Алгоритмов сегментации достаточно много, но в основном применяются методы, основанные на одном из двух базовых свойств сигнала яркости: разрывности и однородности. В первом случае подход состоит в разбиении изображения на основании резких изменений сигнала, таких как перепад яркости на изображении. Вторая категория методов использует разбиение изображения на области, однородные в смысле заранее выбранных критериев. Примерами таких методов могут служить пороговая обработка, выращивание областей, слияние и разбиение областей [6]. Такие методы позволяет выделить на изображении страницы связные области, но этого недостаточно для разделения на текст и рисунки, что является весьма важным для точно сохранения макета страницы и качественного распознавания.

Для многих методов характерна неоднозначность разметки точек в реальных ситуациях из-за необходимости применения эвристик (выбор порогов совпадения яркостей,

[1] Работа выполнена при поддержке гранта РФФИ № 04-07-97501

выбор цифровых масок и т.д.). Заслуживает внимания в связи с этим предложенный метод многозначной разметки, основанный на комбинации различных приемов для снижения неопределенности. Важное практическое значение имеют допускающие параллельную обработку алгоритмы ускорения процесса разметки на основе логического анализа соседних элементов [8]

Недостатком текстурных методов сегментации для данной задачи является то, что рисунок на изображении страницы, как правило, представляет собой набор объектов, неоднородных по текстуре и регулярности. Поэтому такой подход можно использовать только для выделения областей с текстом и, частично, для выделения контурных рисунков.

В связи с этим, разработка метода, позволяющего надежно выделять на изображении страницы области с текстом, а также области с рисунками, является весьма актуальной.

3 Постановка задачи

На вход поступает растровое изображение H_{ij} страницы в градациях серого, где $i = [1..x]; j = [1..y]$, x, y – ширина и высота изображения в пикселях соответственно. Будет рассматриваться случай, когда все строки на странице параллельны друг другу и лежат горизонтально. Необходимо соотнести каждый элемент изображения с одним из двух классов: область с текстом и область с рисунками. А также выделить однородные непересекающиеся области из этих элементов. Граница каждой области должна представлять собой замкнутую ломаную, построенную из отрезков, параллельных осям координат. Рисунки могут быть как штриховые, так и полутоновые.

4 Описание методов

Предлагаемый подход основан на анализе бинарных изображений. Поэтому первым этапом обработки будет бинаризация полутонового изображения. Для этого можно использовать разные методы, например алгоритм, описанный в [5]. Полученное бинарное изображение разобьем на небольшие области H'_{ab} (*Рисунок 1*).

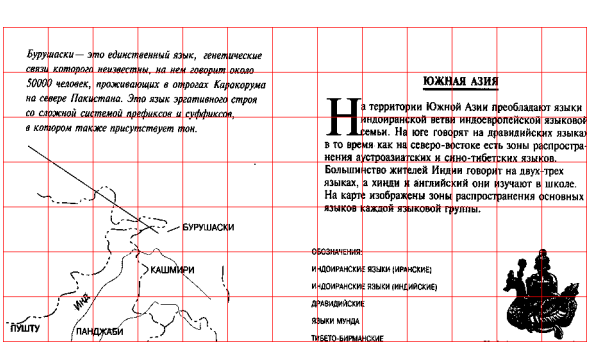

***Рисунок 1**. Разделение изображения на области.*

В качестве формы областей возьмем квадрат и равномерно поделим все изображение на такие области (при необходимости, границы изображения расширяются таким образом, чтобы изображение разделилось на целое число квадратов). Экспериментально было установлено, что хороший результат получается при выборе размера областей таким образом, чтобы ее пересекало 3-7 строк текста. При меньшем количестве строк возрастет ошибка классификации в дальнейшем, при большем же их количестве уменьшится точность определения границ блоков. Пронумеруем все полученные области - по горизонтали: a = [1..n], по вертикали: b = [1..m]. Дальнейшей задачей будет отнесение каждой области к одному из двух вышеназванных классов.

Предлагаются следующие метрики для классификации:

R_1: Доля черных пикселей во всей области.

R_2: Величина разброса толщины отдельных элементов области – отрезков, дуг, отдельных точки.

R_3: Наличие чередующихся строк (горизонтальные полосы с высокой долей черных пикселей) и междустрочий (горизонтальные полосы с низкой долей черных пикселей).

Для области с текстом построим критерии на основе метрик R_1, R_2, R_3 следующим образом:

$R_1^{текст}$ – Доля черных пикселей относительна невелика.

$R_2^{текст}$ – В пределах области содержатся однородные по толщине элементы, из которых состоят буквы.

$R_3^{текст}$ – Присутствуют строки и междустрочия.

Для области с полутоновым рисунком:

$R_1^{рисунок}$ – Доля черных пикселей во всей области относительна большая.

$R_2^{рисунок}$ – Толщина составляющих элементов имеет большой разброс.

$R_3^{рисунок}$ – Строки и междустрочия явно не выделяются.

Нетрудно заметить, что в последнем случае для штрихового или контурного рисунка критерии $R_1^{рисунок}$ и $R_2^{рисунок}$ могут, вообще говоря, не выполняться, что может привести к ошибочной классификации области. Чтобы уменьшить вероятность такой ошибки, предлагается ввести веса для каждого критерия и производить

классификацию на основе интегрального критерия. Так, например для $R_3^{текст}$ и $R_3^{рисунок}$ целесообразно увеличить вес, так как они является наиболее надежными критериями.

Метрика R_1 вычисляется по формуле (1):

$$R_1 = \frac{\sum_{i=1}^{\left[\frac{x}{a}\right]} \sum_{j=1}^{\left[\frac{y}{b}\right]} H'_{ab}(i,j)}{\left[\frac{x}{a}\right] \cdot \left[\frac{y}{b}\right]} \quad (1),$$

где $H'_{ab}(i,j)$ предполагается равным 0 или 1.

Для реализации метрики R_2 необходимо вычислить распределение толщины элементов изображения в области. В качестве вспомогательного инструмента была использована процедура скелетизации (утоньшения) изображения.

В данной работе был реализован метод, предложенный в [3]. Первым этапом алгоритма является выделение всех компонент связности на исходном изображении. В каждой компоненте для всех внешних и внутренних контуров ищутся начальные левые верхние точки, после чего происходит послойное удаление пикселей с контуров. Для очередной точки контура рассматривается конфигурация восьми соседних точек. Точка удаляется, если она не является концевой, (то есть не лежит на начальном или концевом интервале прямой или поворотной линии) и если после ее удаления все ее соседей будут по-прежнему образовывать связное множество. После анализа точки и ее соседей, а также возможности удаления осуществляется переход к следующей точке контура таким образом, чтобы остаться на границе изображения.

За один проход снятие одного слоя точек проводится для каждой компоненты поочередно для каждого внешнего и внутреннего контуров. Процедура повторяется до тех пор, пока не останутся только неудаляемые точки.

Для вычисления толщины объекта во всех точках используется следующий способ: толщина элемента h в точке (i,j) равна сумме расстояния от данной точки до ближайшей белой точки (d) и расстояния от данной точки до ближайшей скелетной точки (s): $h_{ij} = d_{ij} + s_{ij}$.

В итоге у нас получается матрица $H_{ab}^{width}(i,j)$, в котором каждой черной точке сопоставляется значение толщины элемента, которому принадлежит эта точка. Тогда значение метрики R_2 для области H'_{ab} будет вычисляться по формуле (2):

$$R_2 = S^2 = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} (H_{ab}^{width}(i,j) - \overline{H}_{ab}^{width})^2 \quad (2).$$

Для построения метрики R_3 предлагается использовать следующий подход. Рассмотрим графики распределения количества черных пикселей вдоль горизонтальных направлений для текстовых областей и областей с рисунками (*Рисунок 2, 3*):

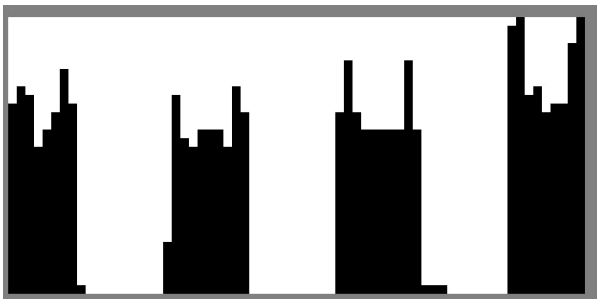

Рисунок 2. *Текстовая область.*

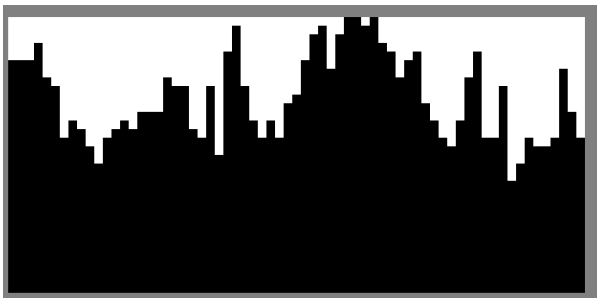

Рисунок 3. *Область с рисунком.*

Хорошо видно, что для текстовой области график имеет ярко выраженный периодический характер, обусловленный чередованием строк и междустрочий. На *Рисунке 3* такого не наблюдается. Для детектирования периодичности был использован метод на основе дискретного преобразования Фурье. В данном случае можно ограничиться частным случаем, когда присутствует только действительная компонента вектора входных данных, соответственно результат преобразования также будет представлен действительными числами. Формула для вычисления будет выглядеть следующим образом:

Обозначим – входной вектор (распределение количества черных пикселей вдоль горизонтальных направлений) $a = (a_0,...,a_{N-1})$, результирующий вектор $y = (y_0,...,y_{N-1})$, где $y_k = \sum_{j=0}^{N-1} a_j \cos\frac{2\pi}{N}$. Результат преобразования Фурье показан на *Рисунке 4,5*.

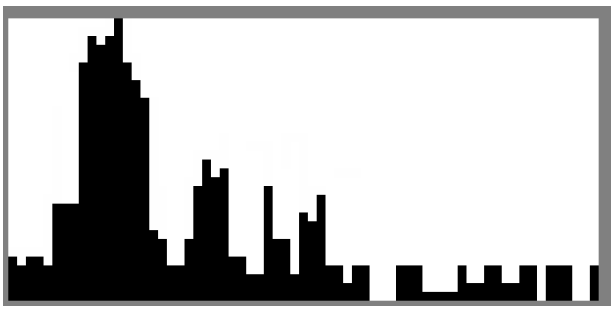

Рисунок 4. *Результат преобразование Фурье для Рисунка 2.*

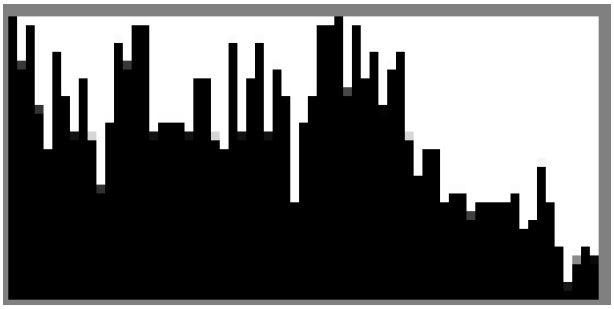

Рисунок 5. *Результат преобразование Фурье для Рисунка 3.*

На *Рисунке 4* четко выделяется доминирующая частота, соответствующая периоду графика на *Рисунке 2*. Для проверки этого будем использовать коэффициент эксцесса, который определяется как:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3, \quad \text{где} \quad \mu_4 = \mathrm{E}[(X - \mathrm{E}X)^4] \text{ - четвертый центральный момент, а } \sigma = \sqrt{D[X]} \text{ – дисперсия.}$$

Если расписать коэффициент эксцесса в полном виде [7], то получится выражение, представленной формулой (3)

$$\gamma_2 = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{j=0}^{N-1} \left(\frac{y_j - \bar{y}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (3),$$

где s – стандартное отклонение.

Соответственно, значение метрики будет определяться как: $R_3 = \gamma_2$.

После построения метрик мы получили инструмент для классификации каждой области. Пример его работы приведен на *Рисунке 6* (светло-серые фрагменты – это области с текстом, темно-серые – области с рисунками):

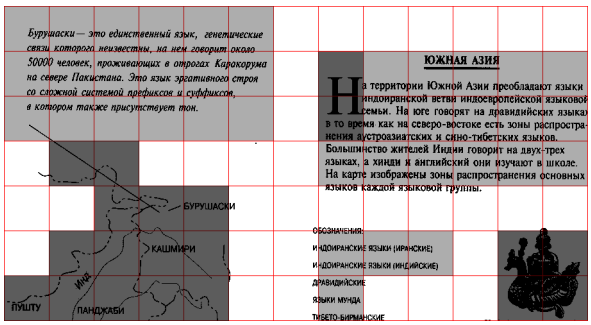

Рисунок 6. *Результат классификации областей.*

5 Заключение

Для проведения экспериментов была разработана программа, реализующая все вышеописанные алгоритмы. Суммарная метрика для критериев классификации вычислялась по следующим коэффициентам: $R = 0{,}12R_1 + 0{,}28R_2 + 0{,}6R_3$. Для критерия $R_1^{текст}$ текстовой области был выбран интервал доли черных пикселей вида: [0,05; 0,4], для $R_1^{рисунок}$: [0,2; 0,7].

Было обработано 127 изображений, взятых с различных источников. Вероятность ошибки первого рода при классификации областей составила 0,04, вероятность ошибки второго рода 0,11. Основной причиной ошибок является относительно высокая чувствительность метода к горизонтальности строк текста. Алгоритм может быть улучшен за счет введения инвариантности относительно наклона строк текста в пределах области, а также введением механизма уточнения границ найденных блоков.

В настоящее время эта система тестируется в библиотеке Казанского Государственного Университета, а также в Национальной библиотеке им. А. Навои (Республика Узбекистан, город Ташкент) при создании электронной коллекции старопечатных изданий.

Полученные результаты также могут быть использованы для систем автоматического ввода документов и систем распознавания печатного текста.

Литература

[1] Michael D. Garris, Darrin L. Dimmick. Form Design for High Accuracy Optical Character Recognition. IEEE Transactions PAMI, June 1996.

[2] J. Rocha, B. Sakoda, J. Zhou, and T. Pavlidis, "Deferred Interpretation of Grayscale Saddle Features for Recognition of Touching and Broken Characters," Proceedings of Document Recognition, SPIE, vol. 2181, San Jose, CA,pp. 342-350, February 1994.

[3] Е. В. Щепин, Г. М. Непомнящий. К топологическому подходу в анализе изображений. Геометрия, топология и приложения (Межвузовский сборник научных трудов). Москва, Мин. высшего и средн. спец. образ. РСФСР,

Московский институт приборостроения, 1990 г., с. 13-25.
[4] Ту Дж., Гонсалес Р. Принципы распознавания образов. Пер. с англ. – М.: Мир, 1978.
[5] В. С. Южиков. Об одном методе предварительной обработки изображений старопечатных текстов. Сборник «Исследования по информатике». Вып.9. – Казань: Отечество, 2005. – с. 125-132.
[6] Р. Гонсалес. Цифровая обработка изображений. Москва: Техносфера, 2005. – с. 812-836.
[7] И. Н. Володин. Лекции по теории вероятностей и математической статистике. — Казань: Центр инновационных технологий, 2001. – с. 60-61.
[8] Путятин Е.П., Аверин С.И. Обработка изображений в робототехнике. М: Машиностроение, 1990. 320 с.

Segmentation of the image of ancient manuscript page

Yuzhikov V.S.

In article the algorithm for segmentation of the image of text page was described. The problem of segmentation consists in correlation of each element of page to one of two classes - the text or figure. Work of algorithm begins with splitting the image into small areas. For classification of each area following criteria are used: 1). A share of black pixels in all area. 2). Value of disorder of width elements into area. 3). Presence of alternating lines and line spacing