Вестник Сыктывкарского университета.
Сер.1.Вып.17.2013

УДК 004.932

БИНАРИЗАЦИЯ И СЕГМЕНТАЦИЯ ОТСКАНИРОВАННОГО ТЕКСТА

В. Л. Никитенков, А. А. Поберий

В данной статье рассматриваются подходы и методы бинаризации и сегментации применительно к отсканированному тексту. Рассматриваемые методы модифицируются для достижения большей эффективности в распознавании отсканированного текста.

Задачу распознавания любого текста можно разбить на несколько основных этапов: предобработка, сегментация, распознование, постобработка. Первый этап, как правило, включает в себя обработку изображений с целью уменьшения шумов и бинаризацию - преобразование цветного изображения в черно-белое (ЧБ). Сегментация заключается в определении текстовых областей и их последующем дроблении на строки, слова и символы. Распознавание сводится к непосредственному распознаванию символов. В процессе постобработки корректируются ошибочные результаты распознавания. В данной работе в первую очередь рассматриваются этапы предобработки и сегментации отсканированного текста.

1. Бинаризация изображения

Важным этапом предобработки является процесс бинаризации - процесс преобразования цветного изображения в ЧБ. Процесс бинаризации проводится в два этапа: преобразование цветного изображения в серое, серого изображения в ЧБ. Перевод цветного изображения в серое не является проблематичным, значение серого цвета каждого пикселя можно

выразить через три цветовых составляющих соответствующего цветного пикселя: красную, зеленую и синюю (цветовая схема RGB):

$$grey = \frac{(66 \times R + 129 \times G + 25 \times B + 128)}{256} + 16 \tag{1}$$

Где R, G, B - цветовые составляющие каждого пикселя - красная, зеленая, синяя соответственно. Второй этап (преобразования серого в ЧБ) является более сложным. Данный этап состоит в выработке такой функции, которая для каждого серого пикселя (принимающего значения от 0 до 255) ставила бы в соответствие либо 0 (белое) либо 1 (черное). Для определения данного значения используется некоторый порог t, при сравнении с которым выбирается выходное значение. Сложность задачи заключается в том, чтобы подобрать такое значение t, которое корректно отделит фон от текста и графики с наименьшей потерей информации и наименьшим количеством шума. До сих пор нет универсального алгоритма для решения данной задачи, но за последнее время появилось множество различных подходов в решении этой проблемы. Существующие подходы условно можно разделить на подход глобальный (для всех пикселей один и тот же порог t) и адаптивный (порог t меняется в зависимости от определенных условий). В данной работе для бинаризации изображения будет использоваться адаптивный, а не глобальный порог, изображение будет условно разбиваться на отдельные области и для каждой области будет применяется свой порог. Глобальный порог неприемлем для сканированных изображений, т. к. освещение распределено неравномерно, т. е. некоторые области текста могут быть темнее или светлее, следовательно, значение порога t общее для всех пикселей приводит к потере информации. Для поставленной задачи можно воспользоваться методом Отсу [2]. Основная идея этого метода заключается в том, что информация на изображении может быть только двух видов: фон и собственно информация. Фон представляется светлыми цветами, а информация темной. Составляется частотная гистограмма цветов (от 0 до 255 включительно) с последующим перебором цветов, чтобы отделить фон от текста (см. рис. 1).

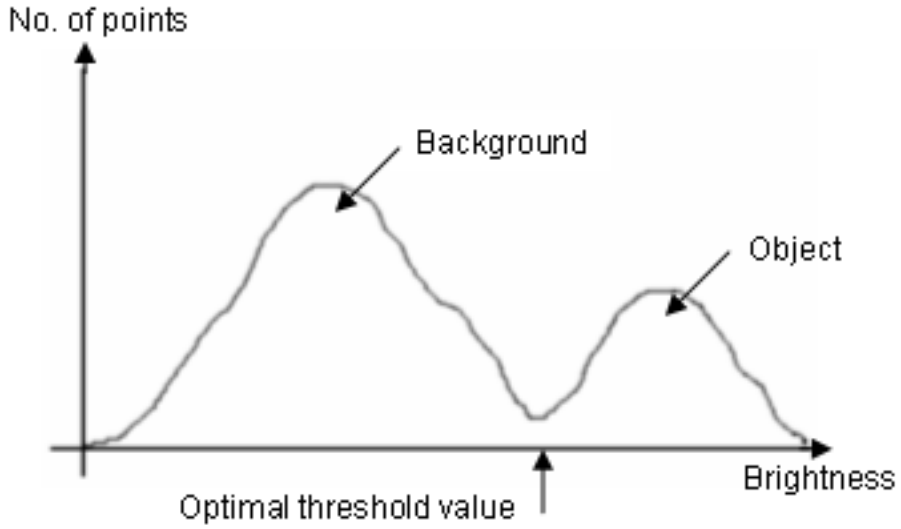

Рис. 1. Гистограмма цветов для нахождения разделяющего порога t.

Метод Отсу обычно используется для глобальной бинаризации. Применение метода Отсу при адаптивной бинаризации может давать шумовые значения для фоновых областей из-за наличия шумов в области изображения (см. рис. 2).

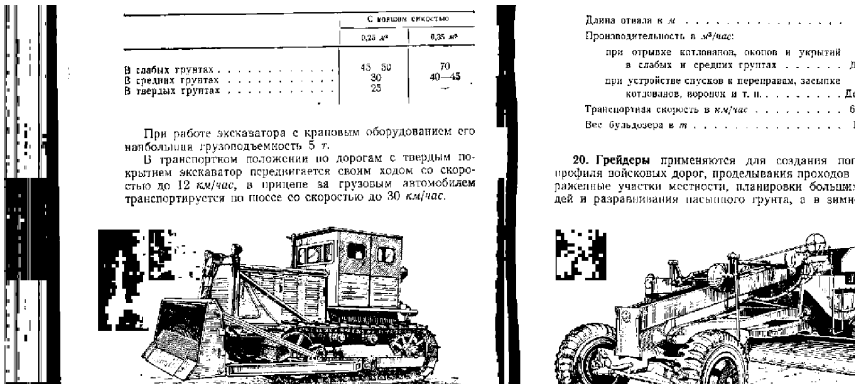

Рис. 2. Шумовые клетки при применении адаптивного метода Отсу.

Для исправления этого недостатка, после определения порога по методу Отсу можно проверить, является ли рассматриваемая область фоновой. Для этого достаточно вычислить дисперсию (сравнивая с некоторым фиксированным порогом), если она небольшая, то это скорее всего фон, в этом случае за порог t можно взять половину максимального значения $256 \times 0{,}5 = 128$. Описанная корректировка исправила полученные ранее ошибки (см. рис. 3).

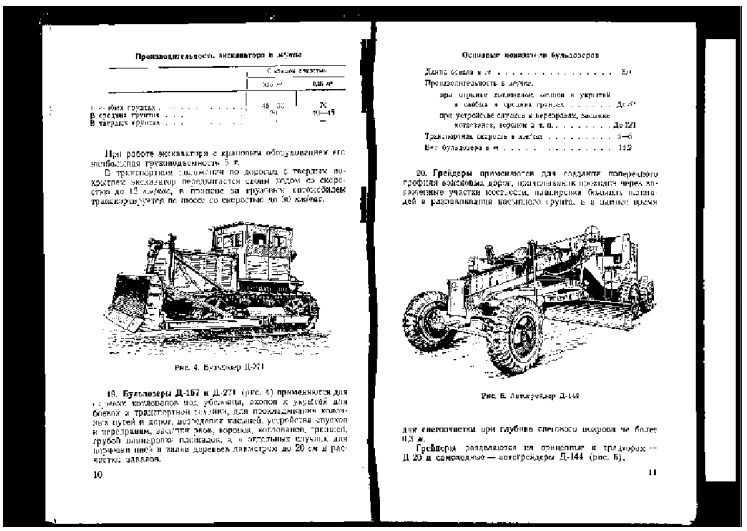

Рис. 3. Исправленный адаптивный порог Отсу.

Данную проблему адаптивного метода Отсу также удалось решить разбиением изображения на вертикальные области. Преимущество данного подхода перед адаптивным заключается в том, что разбиение происходит только по одному из измерений (по ширине) и не требуется корректировка в каждой области для исправления описаной ранее ошибки адаптивной бинаризации Отсу (см. рис. 4). Идея данного подхода вытекает из наблюдения, что освещенность сканированного изображения меняется по горизонтали, темнее по краям и на стыке страниц и ярче посередине каждой страницы.

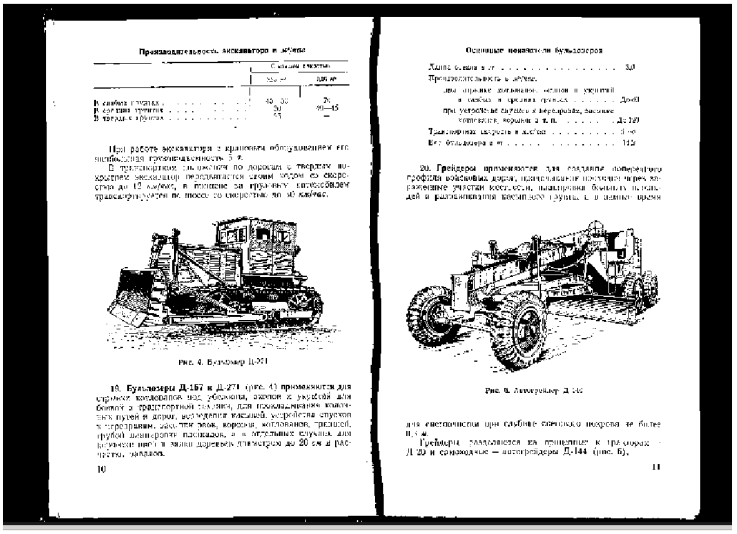

Рис. 4. Порог Отсу при разбиении на вертикальные области.

2. Сегментация

2.1. Нахождение областей. Существует множество методов и подходов к проблеме сегментации. В данной работе рассматривается подход, основанный на разбиении изображения на клетки с последующим анализом взаимной корреляции строк и математического ожидания каждой клетки [3]. Данный метод можно дополнить еще одним параметром - среднеквадратическим отклонением, который также способен характеризовать каждую клетку. Для сегментации серое изображение разбивается на клетки со сторонами, составляющими 1% от ширины изображения. Затем вычисляются статистические характеристики каждой клетки: среднеквадратическое отклонение σ, математическое ожидание μ и средняя взаимная корреляция всех строк клетки $corr$. По полученным характеристикам происходит классификация каждой клетки на фоновую или текстовую.

$$corr < 0,97; 0,03 < \mu \leq 0,85; 3 < \sigma < 9 \Longrightarrow текст \tag{2}$$

Данные значения подбирались эмпирически [3]. Далее необходимо определить границы областей, для этого обходим изображение, применяя маски для каждого пикселя:

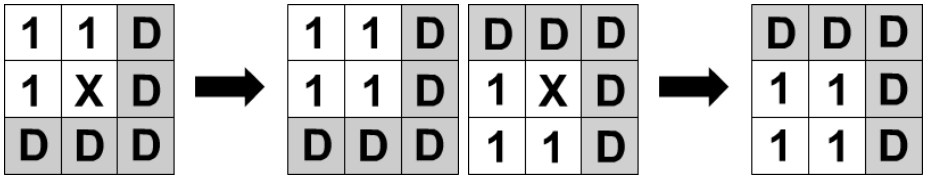

Рис. 5. Маски выделения текстовых областей для пикселей («0» - фон, «1» - текст, «D» - любое значение, «X» - обрабатываемый пиксель).

Обходим изображение в обратную сторону, применяя следующие маски:

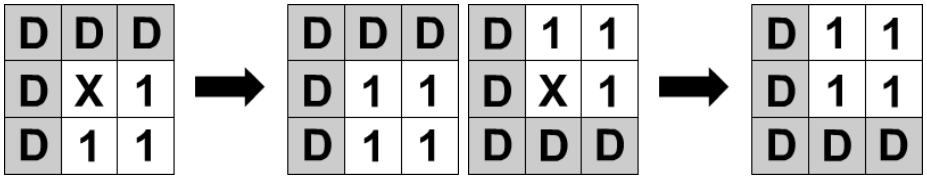

Рис. 6. Маски выделения текстовых областей для пикселей («0» - фон, «1» - текст, «D» - любое значение, «X» - обрабатываемый пиксель).

В результате, текстовые области получают углы, которые можно выделить как начало и конец прямоугольной области:

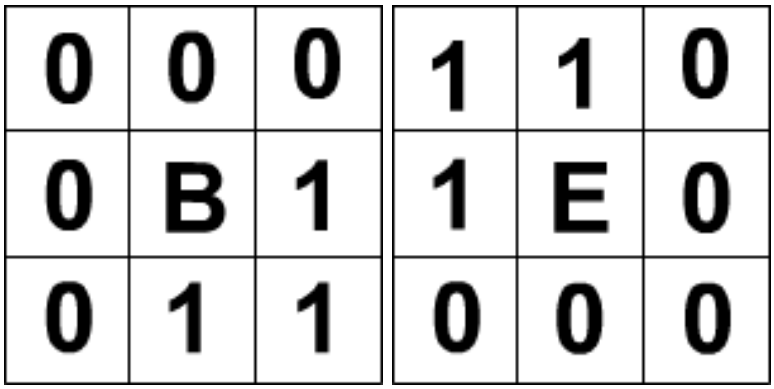

Рис. 7. Маски для обозначения начала и конца области («B» - начало, «E» - конец).

Данный метод помогает быстро находить границы текстовых областей, после чего можно переходить к следующему шагу сегментации - разделению строк. После нахождения текстовой области, ее необходимо разбить на строки, для этого можно воспользоваться статистическим анализом строк изображения [1]. Определяя долю черных пикселей каждой строки текстовой области, можно наблюдать колебания: возрастание при входе в текстовую область и убывание при выходе из нее. Опираясь на это наблюдение можно получить строки.

2.2. Разделение слов. Каждую строку необходимо разбивать на слова. Для этого можно оценивать долю черных пикселей, но уже каждого столбца. Для разделения слов нужно находить интервалы с наименьшей долей черных пикселей, ширина которых соизмерима с шириной одного символа, т. к. интервалы меньшей ширины – это разделители символов, а не слов. При реализации данного метода лучше пробегать столбцы скользящим окном шириной больше одного пикселя, т. к. в одном столбце возможно получить значение меньше порогового из-за разрывов внутри символов, вызванных ошибками бинаризации или низким качеством изображения.

2.3. Разделение символов. Чтобы отделить символы друг от друга, можно просматривать долю черных пикселей в скользящем окне пробегающем по ширине слова, и если эта доля меньше порога, следовательно, мы попали на разделитель символов (пробел). Но у этих методов есть существенный недостаток – они не инвариантны к повороту изображения, так как все символы будут располагаться под углом и просмотр скользящим окном по ширине может выдать ложные результаты. Чтобы не зависесть от небольших углов наклона изображения, пороговое значение, с которым сравнивается доля черных пикселей, должно

вычисляться для каждого случая отдельно, т. к. колебание доли черных пикселей при входе и выходе из текстовых областей сохраняется при изменении угла наклона. После сегментации можно переходить к непосредственному распознанию символов.

Литература

1. Papavassiliou V., Stafylakis T., Katsouros V., Garayannis G. Handwritten image segmentation into text lines and words // *Pattern recognition. – 2010. – №43. – С.369 –377.*

2. Otsu N. A threshold selection method from grey-level histograms // *IEEE Trans., Man., Cyber. 9(1): 62-66. doi:10.1109/TSMC. 1979.4310076*

3. Sauvola J., Piettkainen M. *Page segmentation and classification using fast feature extraction and connectivity analysis. Int. Conf. on Document Analysis and Recognition, 1995*

Summary

Nikitenkov V. L., Pobery A. A. Scanned text binarization and segmentation

In this article methods and approaches for binarization and segmentation for scanned text are considered. For more efficient results for scanned text recognition considered methods are modified.

Сыктывкарский университет *Поступила 11.05.2013*

Содержание

Contents