

Распрямление текстовых строк на основе непрерывного гранично-скелетного представления изображений

Антон Масалович, Леонид Местецкий
Московский государственный университет,
Москва, Россия
anton_m@abbyy.com, l.mest@ru.net

Аннотация

Рассматривается задача распрямления текстовых строк на изображениях, предназначенных для распознавания. Предлагается подход к решению, включающий построение непрерывного внешнего скелета изображения сканированного текста, выделение в нем цепей, расположенных в пробелах между строчками, и в построении преобразования изображения, переводящего эти цепи в горизонтальные прямые линии. Приводятся результаты вычислительных экспериментов на искривленных изображениях, показывающие возможности предлагаемого подхода.

Ключевые слова: распознавание текста, преобработка изображений, распрямление строк, непрерывное гранично-скелетное представление, внешний скелет.

1. ВВЕДЕНИЕ

Все современные системы распознавания текстов устроены так, что качество распознавания изображений напрямую зависит от качества самого распознаваемого изображения. Одним из факторов, существенно влияющих на качество изображений, является деформация строк текста на изображении. При сканировании толстой книги такая деформация выражается в искажениях строк (загибании их краёв) вблизи сгиба. Если для получения изображений документов используют цифровые камеры вместо сканеров, то искривление строчек происходит вследствие характерных дефектов, связанных с искажением изображения за счёт используемой оптики, к примеру, закругление изображений по краям в результате так называемого «эффекта бочки» (бочкообразные дисторсии). На одном изображении могут сочетаться несколько деформаций, из-за чего невозможно построить точную модель искривления страницы.

Поскольку все современные системы распознавания текста ориентированы на работу с прямыми горизонтальными строками, возникает задача распрямления деформированных строк сканированного документа, которая должна быть решена на этапе преобработки.

Известно несколько решений задачи распрямления строк [3,4,5]. Некоторые решения основываются на восстановлении искривления строки по прямоугольникам символов в строке, некоторые основываются на построении искажения документа с использованием третьего измерения. Существуют алгоритмы, основанные на численном подборе изгиба строки (с помощью построения функции адекватности изгиба). Все эти методы имеют ряд недостатков. Прямоугольники символов на размытой фотографии, к примеру, очень непредставительны для задачи определения

строк. Трёхмерное моделирование требует твердых условий на начальное искажение строк. Численные методы работают слишком медленно и не очень стабильно.

Предлагаемый подход к решению задачи основывается на построении внешнего скелета изображения текста в сканированном документе. Идея решения состоит в том, что в таком скелете образуются длинные ветви, ориентированные вдоль междустрочных пробелов. Предлагается метод нахождения таких ветвей и последующего их преобразования в горизонтали. Распрямление междустрочных ветвей скелета индуцирует общее преобразование строк текста, распрямляющее эти строки.

Работа выполнена при поддержке РФФИ, грант 05-01-00542.

2. ФОРМУЛИРОВКА ЗАДАЧИ

Пусть у нас имеется изображение текстового документа. Изображение должно быть черно-белым, причем текст должен быть черным, а фон белым. На изображении не должно присутствовать инвертированных областей. На изображении не должно быть шума. В современных системах распознавания текста имеются эффективные алгоритмы, приводящие практически любое изображение текстовых документов к указанному виду.

Алгоритм распрямления строк должен работать следующим образом. Получая на вход изображение текстового документа, алгоритм должен преобразовать его так, чтобы на выходе получилось изображение, в котором все строки текста расположены строго горизонтально.

Введем еще два естественных требования к исходным изображениям, необходимые для разрабатываемого алгоритма:

1. На изображении должны присутствовать достаточно большие текстовые блоки. Это связано с тем, что алгоритм работает не со строками, а с междустрочными промежутками, поэтому на изображении должно быть достаточное количество строк текста, расположенных друг под другом.
2. Искривление строк изображения можно аппроксимировать достаточно гладкой кривой. Это связано с тем, что для аппроксимации искривления изображения в алгоритме применяются гладкие кривые.

3. НЕПРЕРЫВНОЕ ГРАНИЧНО-СКЕЛЕТНОЕ ПРЕДСТАВЛЕНИЕ ИЗОБРАЖЕНИЯ

Для решения поставленной задачи воспользуемся непрерывным гранично-скелетным представлением изображения [1,2].

Многоугольная фигура – это замкнутая область на плоскости, граница которой состоит из непересекающихся простых многоугольников.

Скелетом многоугольной фигуры называется геометрическое место точек плоскости, имеющих не менее двух ближайших точек на границе фигуры. Различают внешний и внутренний скелеты многоугольной фигуры.

Непрерывное гранично-скелетное представление изображения – это совокупность многоугольников минимального периметра, разделяющих все черные и белые точки изображения, и скелета многоугольной фигуры, образованной этими многоугольниками (рис. 1).



Рис. 1. Непрерывное гранично-скелетное представление изображения.

Для решения задачи распрямления строчек нас будет интересовать **внешний скелет изображения** – та часть скелета, которая лежит в белой фоновой области изображения.

Скелет многоугольной фигуры состоит из линий, которые представляют собой сегменты прямых и парабол. Скелет можно рассматривать как плоский граф, у которого рёбрами являются сами сегменты прямых и парабол, а вершинами – концевые точки этих сегментов, т.е. точки соединения сегментов друг с другом. Алгоритм вычисления скелета через диаграммы Вороного многоугольной фигуры [1,2] дает возможность в явном виде вычислять этот граф.

4. ОСНОВНАЯ ИДЕЯ АЛГОРИТМА

Если рассмотреть ту часть внешнего скелета изображения сканированного текстового документа, которая лежит между двумя горизонтальными строками, то нетрудно увидеть, что в нем можно выделить четкую горизонтальную составляющую (рис. 2).

Для двух искривленных строк разделяющая их линия скелета будет не горизонтальной, но также вполне заметной (рис. 2).

Задача состоит в том, чтобы выявить такие междустрочные ветви и затем построить на их основе преобразование изображения, распрямляющее строки.



Рис. 2. Скелетная ветвь между горизонтальными и искривленными линиями текста.

Идея алгоритма заключается в следующем:

- Строится непрерывное гранично-скелетное представление изображения.
- На внешнем скелете изображения выделяются достаточно длинные и гладкие участки.
- Полученные ломаные линии аппроксимируются гладкими кривыми.
- Строится преобразование пространства, которое каждую полученную кривую преобразует в горизонтальную прямую.
- Полученное преобразование применяется к изображению.

5. ПРЕДОБРАБОТКА СКЕЛЕТА

Во внешнем скелете изображения в области текста очень много ненужных элементов, появившихся из-за неровностей букв. Такие элементы не представляют никакой ценности для задачи определения строк и достаточно сильно усложняют алгоритмы нахождения междустрочных просветов и увеличивают время работы алгоритма.

Фактически, нам нужно оставить только те ребра скелета, которые проходят между двумя различными черными объектами (рис. 3). Есть несколько путей для выделения таких ребер. При реализации нашего подхода мы воспользовались тем, что нужная нам часть скелета представляет собой циклические подграфы скелетного графа. Таким образом, можно определить существенные части скелета, выделив в нём циклические компоненты. **Циклическая компонента графа** – это подграф такой, что из каждой его вершины можно проложить путь по подграфу без самопересечений, который возвращается в ту же вершину. Выделение циклического подграфа осуществляется путём последовательного удаления из скелета всех вершин степени один и инцидентных им рёбер до тех пор, пока не останутся только вершины степени больше единицы.

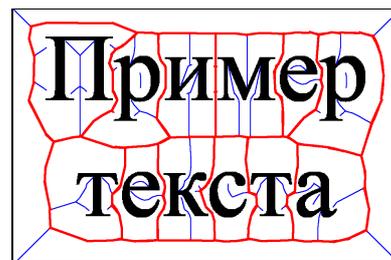


Рис. 3. Скелет изображения, с выделенными значащими ребрами.

6. КЛАСТЕРИЗАЦИЯ РЕБЕР СКЕЛЕТА

Так как задача заключается в выделении междустрочных ветвей скелета, ребра, разделяющие буквы текста в одной строке, нас не интересуют. Соответственно, возникает задача выделения «горизонтальной» составляющей скелета, образованной ребрами, которые лежат между строками текста. Выделение этой составляющей осуществляется методом кластеризации ребер скелета на междустрочные и межбуквенные ребра.

Если параллельным переносом сдвинуть все ребра скелета так, чтобы один конец каждого ребра попал в начало координат, то оставшиеся концы ребер образуют достаточно четко выраженную крестообразную конфигурацию. На рис.4 приведен пример изображения и его скелета, а на рис.5 – «крест», полученный в результате сдвига его ребер. Соответственно, все ребра скелета с легкостью разбиваются на два класса по углу наклона ребра к горизонтали. При этом даже необязательно реализовать автоматические системы определения порога кластеризации. Выставление порога кластеризации в 60 градусов от горизонтали дает вполне удовлетворительные результаты такого разбиения.



Рис. 4. Пример кластеризации ребер скелета.

В результате в одном кластере оказываются «вертикальные» составляющие скелета (ребра, разделяющие буквы в одной строке), а в другом кластере – «горизонтальные» ребра, разделяющие объекты в разных строках. Для решения задачи распрямления строк существенными являются ребра из второго кластера.

Как видно из примера кластеризации ребер скелета на рис. 4, кластеризация ребер не идеальная, однако ни одного ребра между строками текста не было классифицировано неправильно, а короткие неправильно кластеризованные ребра между буквами текста не представляют существенных трудностей для алгоритма и будут из-за своей незначительной длины отсеяны на следующем шаге.

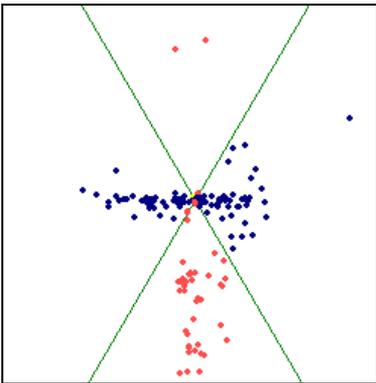


Рис. 5. Концы ребер скелета изображения при их начале в одной точке.

7. ВЫДЕЛЕНИЕ И АППРОКСИМАЦИЯ МЕЖДУСТРОЧНЫХ ВЕТВЕЙ СКЕЛЕТА

Предлагается следующий алгоритм для выделения и аппроксимации межстрочных ветвей скелета.

- В качестве начального ребра возьмём произвольное длинное ребро из кластера междустрочных ребер.
- Пусть у нас уже есть выделенная ветвь скелета, тогда для этой ветви построим аппроксимирующую кривую.
- На конце ветви в силу особенностей скелета будет максимум два возможных ребра – продолжения ветви. Проверим, конец какого из ребер имеет меньшее расстояние до построенной кривой.
- То ребро, для которого расстояние меньше, добавим к ветви и скорректируем уравнение кривой с учётом этого ребра.
- Прекратим строить ветвь, если оба рассматриваемых ребра совсем не подходят, т.е. слишком удалены от построенной кривой.
- Удалим аппроксимирующую линию из рассмотрения, если она получилась слишком короткой.
- Выберем следующее начальное ребро из тех ребер, что мы еще не рассматривали.

В качестве варианта аппроксимирующей кривой при реализации данного метода мы использовали квадратичную регрессию множества вершин построенной ветви.

Каждую ветвь аппроксимируем кривой

$$B = \left\{ \begin{cases} x(t) = a_1 * t^2 + a_2 * t + a_3 \\ y(t) = b_1 * t^2 + b_2 * t + b_3 \end{cases} \right\} t \in [0,1]$$

Будем считать, что первая вершина ветви соответствует параметру $t=0$, последняя вершина – $t=1$. Для каждой из n вершин ветви определим значение параметра t_i как отношение длины ветви от начала до этой вершины к длине всей ветви. Параметры кривой будем подбирать так, чтобы минимизировать суммарное отклонение точек ломанной от кривой:

$$\sum_{i=0}^n (a_1 t_i^2 + a_2 t_i + a_3 - x_i)^2 + (b_1 t_i^2 + b_2 t_i + b_3 - y_i)^2 \rightarrow \min_{\substack{a_1, a_2, a_3 \\ b_1, b_2, b_3}}$$

Приравняв нулю частные производные функционала по параметрам $a_1, a_2, a_3, b_1, b_2, b_3$, и решив соответствующую систему линейных уравнений, получаем искомые коэффициенты для аппроксимации ветви скелета квадратичной кривой.

Пример выделения строк на изогнутом тексте приведен на рис. 6.

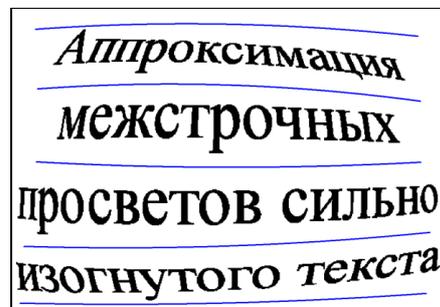


Рис. 6. Пример аппроксимации межстрочных просветов на основе скелета.

8. ПРЕОБРАЗОВАНИЕ ИЗОБРАЖЕНИЯ

У алгоритма непосредственного преобразования изображений должны быть следующие ограничения:

- Каждую выделенную кривую он должен преобразовывать в отрезок прямой, той же длины, что и исходная кривая, и минимально отклоняющийся от этой прямой.
- Каждая точка, не лежащая ни на одной выделенной кривой, должна преобразовываться в соответствии с близлежащими кривыми.

Общее преобразование плоскости ищется в виде $(x(u, t), y(u, t))$. Задача состоит в построении таких зависимостей, чтобы при некоторых значениях параметра $u = u_0, u_1, \dots, u_k$ уравнения

$$(x_j(t) = x(u_j, t), y_j(t) = y(u_j, t)), j = 0, \dots, k$$

соответствовали ранее построенным k аппроксимирующим линиям междустрочных пробелов (рис.6). Такое преобразование предполагает аппроксимацию коэффициентов междустрочных кривых функциями $a_1(u), a_2(u), a_3(u), b_1(u), b_2(u), b_3(u)$.

Непосредственное распрямление состоит в вычислении нового изображения в растровой сетке в прямоугольных декартовых координатах (u, t) , в котором точка с координатами (u, t) раскрашивается в цвет точки с координатами $(x(u, t), y(u, t))$ исходного (искривленного) изображения.

Предварительные эксперименты показали реализуемость такого подхода. В настоящее время разрабатывается алгоритм построения такого преобразования в виде кубического патча Безье.

9. ЗАКЛЮЧЕНИЕ

Проведенные вычислительные эксперименты показали корректность и реализуемость предложенного подхода к решению задачи распрямления текстовых строк на основе непрерывного гранично-скелетного представления. Полученные результаты свидетельствуют о приемлемой точности и высокой скорости работы алгоритма.

В настоящее время разрабатывается полномасштабная программная реализация алгоритма.

10. БИБЛИОГРАФИЯ

- [1] Местецкий Л.М. *Непрерывный скелет бинарного растрового изображения. Труды международной конференции "Графикон-98", Москва, 1998.*
- [2] Местецкий Л.М. *Скелет многосвязной многоугольной фигуры. Труды международной конференции "Графикон-2005", Новосибирск, 2005.*
- [3] Li Zhang, Chew Lim Tan. *Warped Image Restoration with Application to Digital Libraries. Proceedings of 8th international*

conference on Document Analysis and Recognition "ICDAR-2005", Seoul, Korea, 2005

[4] Hironori Ezaki, Seiichi Uchida, Akira Asano, Hiroaki Sakoe. *De-warping of document images by global optimization. Proceedings of 8th international conference on Document Analysis and Recognition "ICDAR-2005", Seoul, Korea, 2005*

[5] Udrian Ulges, Christoph H. Lampert, Thomas M. Breuel. *A Fast and Stable Approach for Restoration of Warped Document Images. Proceedings of 8th international conference on Document Analysis and Recognition "ICDAR-2005", Seoul, Korea, 2005*

Об авторах

Леонид Моисеевич Местецкий – доктор технических наук, профессор факультета вычислительной математики и кибернетики Московского Государственного Университета им. М.В. Ломоносова.

Адрес: Москва, 119899, Воробьевы горы, МГУ, 2-й учебный корпус, факультет ВМиК, кафедра математических методов прогнозирования.

E-mail: l.mest@ru.net

Антон Андреевич Масалович – аспирант факультета вычислительной математики и кибернетики Московского Государственного Университета им. М.В. Ломоносова

Адрес: Москва, 119899, Воробьевы горы, МГУ, 2-й учебный корпус, факультет ВМиК, кафедра математических методов прогнозирования.

E-mail: anton_m@abbyy.com

Warped image restoration based on continuous skeletal-border representation

Abstract

This paper is devoted to a problem of de-warping of document images intended to recognition. Novel technique is proposed that based on construction of continuous border-skeletal representation of image. Smooth curves related to gaps between text lines are extracted from outer skeleton of the image. Image is transformed to de-warp these curves into straight horizontal lines. Some experimental results is listed, which indicates the possibilities of proposed algorithm.

Keywords: text recognition, OCR, image preprocessing, warped image restoration, continuous border-skeletal representation, outer skeleton.

About the authors

Leonid Mestetskiy is a professor at Moscow State University, Department of Mathematical methods of Forecast. His contact email is l.mest@ru.net.

Anton Masalovitch is a Ph.D. student at Moscow State University, Faculty of Computational Mathematics and Cybernetics. His contact email is anton_m@abbyy.com.