УДК 004.93:12

СЕГМЕНТАЦИЯ РУКОПИСНЫХ И МАШИНОПИСНЫХ ТЕКСТОВ МЕТОДОМ ДИАГРАММ ВОРОНОГО

С. А. Запрягаев, А. И. Сорокин

Воронежский государственный университет

Поступила в редакцию 1.03.2010 г.

Аннотация. В работе рассмотрена задача сегментации рукописных и машинописных текстов. Приводится описание широко распространённого метода сегментации на основе анализа «профилей» и обосновывается необходимость в разработке современных подходов к выделению строк, слов и символов текста. Предложен алгоритм сегментации текста, основанный на анализе диаграммы Вороного, построенной по центрам масс множества точек символов.
Ключевые слова: сегментация рукописного текста, диаграммы Вороного,алгоритм сегментации текста

Abstract. In the paper, the problem of hand and machine printed text segmentation is considered. The paper describes widely spread "profile"-based text segmentation approach and grounds essentiality of modern approaches to text line, word and character detection. A novel text segmentation algorithm based on analysis of Voronoi diagram generated for mass centers of character point sets is proposed.
Key words: handprinted text segmentation, Voronoi diagrams, text segmentation algorithm

ВВЕДЕНИЕ

В системах распознавания рукописных и машинописных текстов важным этапом в последовательности операций по распознаванию объектов является сегментация изображения. Роль сегментации сводится к отысканию на плоскости изображения элементов, подвергаемых распознаванию, таких как строки символов, отдельные слова или символы, рисунки, таблицы и другие объекты, содержащиеся в тексте.

Стандартные способы сегментации «хорошо» структурированных текстов (как правило, машинописных) обычно используют результаты вертикального и горизонтального сканирования изображения [1]. Так, на рис. 1 приведено изображение текста и так называемый «горизонтальный профиль» строк, заданный суммарным количеством «чёрных» точек, расположенных на горизонтальной прямой. Разбиение осуществляется на основе анализа локальных экстремумов «горизонтального профиля».

Однако для рукописного текста ситуация осложняется отсутствием линейного упорядочения. Метод профилей также не может быть использован в большинстве случаев ввиду отсутствия выраженных экстремумов профиля. Так, на рис. 2 приведено изображение горизонтального профиля рукописного текста.

Учитывая ограниченную применимость метода профилей, многие современные подходы к сегментации текста основаны на использовании метода диаграмм Вороного [2]. Выделение текстовых блоков в таких методах происходит в два этапа: первый заключается в построении диаграммы Вороного выделенной области. Второй этап использует полученную диаграмму как средство быстрого поиска так называемых «смежных» точек и производит постепенное укрупнение рассматриваемых блоков (т.е. выделение строк, затем их объединение в абзацы, параграфы и т.п.).

1. ОСНОВНЫЕ СВОЙСТВА ДИАГРАММ ВОРОНОВА

Диаграмма Вороного конечного множества точек S на плоскости представляет собой разбиение плоскости, при котором каждая область этого разбиения образует множество точек, максимально близких к одному из элементов множества S.

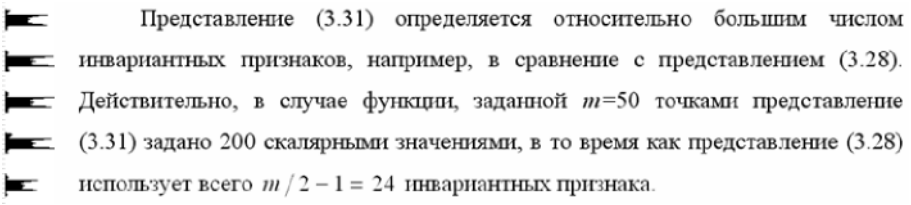


Рис. 1. Сегментация строк печатного текста на основе «профилей»

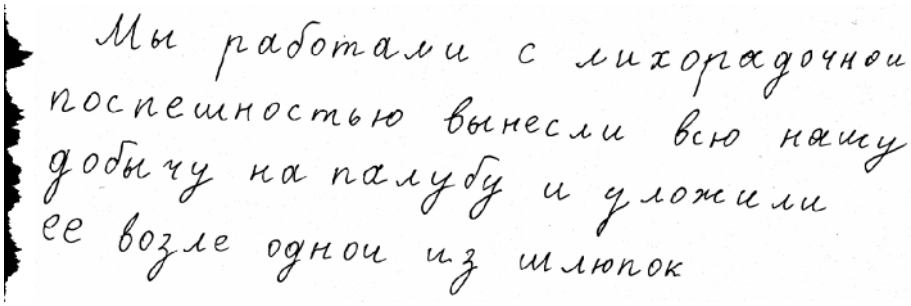


Рис. 2. Горизонтальный профиль рукописного текста

Другими словами, если S – конечное множество точек, то диаграмма Вороного $V(S)$ будет представлять собой разбиение плоскости на следующие подмножества:

$$V(s_i) = \{p : \forall j \neq i, d(p, s_i) < \\ < d(p, s_j)\}, \; s_i \in S. \\ V(S) = \bigcup_{i=1}^{N} V(s_i), \tag{1}$$

где $d(p,q)$ – евклидова метрика.

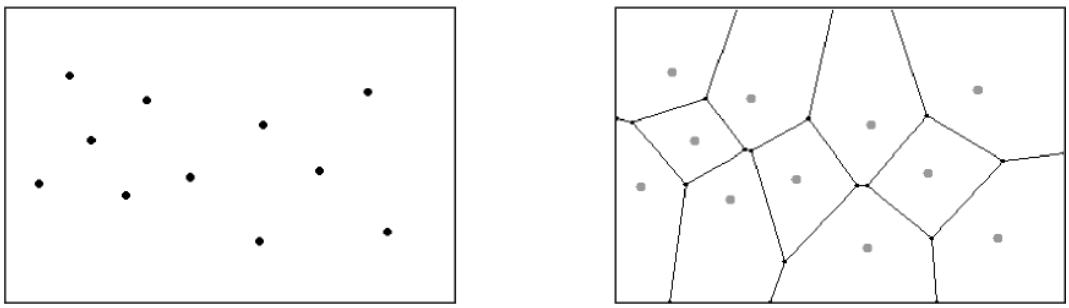
Множества $V(s_i)$ называют «ячейками Вороного», точки $s_i \in S, i = 1..N$ – «генераторами» [3]. На рис. 3 изображено множество генераторов и соответствующая им диаграмма Вороного.

Ячейка Вороного представляет собой выпуклый многоугольник [3], вершины многоугольников определяют вершины диаграммы Вороного, а соединяющие их отрезки – рёбра диаграммы Вороного. Таким образом, вся плоскость представляется объединением ячеек равноудалённых от точек-генераторов рёбер.

Известны следующие основные свойства диаграммы Вороного [3].

1. Каждая вершина диаграммы Вороного, полученной для множества N точек-генераторов, является точкой пересечения трёх рёбер диаграммы (при $N > 2$).

2. Каждый ближайший сосед $p \in S$ точки $q \in S$, определённый условием



а) множество генераторов

б) диаграмма Вороного

Рис. 3. Множество генераторов и диаграмма Вороного

$d(p,q) = \min\limits_{r \in S \setminus \{q\}} d(r,q)$, задаёт ребро ячейки Вороного. В этом случае ребро Вороного задано множеством точек $\{r : d(p,r) = d(q,r)\}$.

3. Многоугольник $V(s_i)$ является неограниченным тогда и только тогда, когда точка s_i лежит на границе выпуклой оболочки множества S.

4. Диаграмма Вороного построенная для множества N точек имеет не более $2N-5$ вершин и $3N-6$ рёбер.

Эффективный алгоритм построения диаграммы Вороного на плоскости был предложен в работе [4] Форчуном. Время работы алгоритма Форчуна для N генераторов составляет порядка $O(N\log(N))$. Подробное изложение деталей реализации и представления данных приводится в работе [5].

2. СЕГМЕНТАЦИЯ ТЕКСТА НА ОСНОВЕ ОБОБЩЁННЫХ ДИАГРАММ ВОРОНОГО

В работе [2] предлагается использовать обобщение диаграммы Вороного для выделения текстовых блоков (т.е. абзацев, параграфов) и строк внутри этих блоков. Генераторы обобщённой диаграммы Вороного представляют собой конечные множества точек $S_1 \ldots S_N$. В этом случае обобщённая диаграмма Вороного $A(S_1,\ldots,S_N)$ области будет задана разбиением пространства на множества вида:

$$A(S_i) = \{p : \forall j \neq i, d(p,S_i) < d(p,S_j)\}, \quad (2)$$

где $d(p,S_i) = \min\limits_{s \in S_i} d(p,s)$,

$$A(S_1,\ldots,S_N) = \bigcup_{i=1}^{N} A(S_i),$$

Определение в точности совпадает с в случае, когда генераторы $S_1 \ldots S_N$ представляют собой одноточечные множества. Рассмотрим структуру обобщённой диаграммы области $A(S_1,\ldots,S_N)$ и её связь с диаграммой Вороного $V(S)$ множества $S = \bigcup\limits_{i=1}^{N} S_i$.

В работе [6] предлагается способ выделения слов и символов текста, основанный на использовании обобщённой диаграммы Вороного . Каждый символ трактуется как объект, состоящий из множества связных точек. Таким образом, весь текст представляется набором связных множеств точек $C_1,\ldots,C_N$.

При этом вводятся два типа расстояний, используемых для сегментации.

Если множество отрезков прямых $\{l_1, l_2, \ldots, l_m\}$ представляют собой границу между двумя связанными компонентами C_i и C_j, то минимальное расстояние от каждого из компонентов до границы между компонентами C_i, C_j определяется как

$$md_s(C_i,C_j) = \min_{1 \le k \le m} \left(g(l_k,C_i), g(l_k,C_j)\right),$$

где $g(l_k,\cdot)$ – расстояние L_1 от прямой, содержащей l_k, до соответствующего компонента.

Аналогично вводится расстояние от компонента C_i до его границы, заданной рёбрами ячейки Вороного, т.е. отрезков прямых $\{e_1, e_2, \ldots, e_{k_i}\}$

$$md_C(C_i) = \min_{1 \le k \le k_i} g(e_k, C_i).$$

На основе расстояний $md_s(\cdot,\cdot)$ и $md_C(\cdot)$ задаются правила объединения символов в слово и поиска промежутков между словами.

Правило 1. Если условие $md_S(C_i,C_j) \le 2 \times \min\left(md_C(C_i), md_C(C_j)\right)$ выполнено для смежных множеств C_i и C_j, то символы, соответствующие этим множествам объединяются в одно слово.

Правило 2. Если условие правила 1 не выполняется, то граница ячейки Вороного, отделяющая C_i и C_j является также и границей слов, к которым принадлежат символы, соответствующие этим множествам.

На рис. 4 приведено изображение обобщённой диаграммы Вороного, построенной по связным компонентам, отвечающим символам текста. Для примера введём два множества-генератора C_1 и C_2, которым соответствуют символам «ь» и «е». Пунктиром обозначены расстояния $g(l_k,C_1)$ и $g(l_k,C_2)$, доставляющие минимум расстояния $md_S(C_i,C_j)$. Правило 1 для примера, изображённого на рис. 4 не выполнено, так как расстояние $md_S(C_i,C_j)$ от каждого из компонентов C_1, C_2 больше, чем удвоенное расстояние $md_C(C_1)$ или $md_C(C_2)$, т.е. минимальное от компонентов до их собственных границ. Примером использования правила 1 можно считать символы «ш» и «ь», так как расстояние от множеств, представляющих эти символы, до общей границы совпадает с минимальным расстоянием каждого из символов до собственной границы.

3. СЕГМЕНТАЦИЯ ТЕКСТА НА ОСНОВЕ ТОЧЕЧНЫХ ДИАГРАММ ВОРОНОГО

Метод, описанный в параграфе 1, предполагает использование множества всех точек изображения для построения обобщённой диаграммы

Вороного. Так, в случае изображения с высоким разрешением, построение диаграммы Вороного для множества всех точек является достаточно затратным (например, при разрешении 1637×1481 точек и 10% заполнении изображения чёрными точками потребуется построение порядка 200000 ячеек диаграммы Вороного).

Учитывая вычислительную сложность построения диаграммы Вороного области, существует возможность получения новых, более эффективных алгоритмов сегментации текста, основанные на анализе взаимного расположения центров масс символов. Так, изображение текста на листе формата А4 содержит в среднем 20 00–3000 символов, центры масс которых используются при построении диаграммы Вороного.

Для выделения строк, слов и символов текста разработан алгоритм, основанный на использовании диаграмм Вороного. Каждое связное множество точек S_i, соответствующее некоторому символу текста, заменено центром его масс:

$$c_i = \frac{1}{|S_i|} \sum_{s \in S_i} s, \;\; i = 1..N. \qquad (3)$$

Множество точек $C = \{c_i, i = 1..N\}$ рассматривается как множество генераторов диаграммы Вороного . На рис. 5 изображен исходный текст и соответствующая ему диаграмма Вороного, построенная по центрам масс символов. Прямолинейные границы ячеек на рис. 5 являются достаточно грубым приближением диаграммы Вороного области (рис. 4), тем не менее, они сохраняют информацию о расстоянии между буквами, словами и строками. Например, расстояние от букв «ь» и «е» до границы, разделяющей их, указывает на то, что эти буквы относятся к разным словам.

Пусть c – точка-генератор диаграммы Вороного, а соответствующая ей ячейка ограничена множеством отрезков прямых $E = \{e_1 ... e_m\}$. Обозначим наименьшее расстояние от точки c до границы ячейки $V(c)$ как

$$md_C(c) = \min_{e \in E} d(c, e),$$

где $d(c, e)$ – евклидово расстояние от точки c до отрезка e.

Обозначим как $md_s(c, p)$ расстояние от двух смежных точек–генераторов c и p до общей границы ячеек Вороного $V(c)$ и $V(p)$, по построению – $md_s(c, p)$ совпадает с половиной евклидова расстояния между точками c и p.

Сформулируем алгоритм поиска «соседних» символов, принадлежащих одному слову.

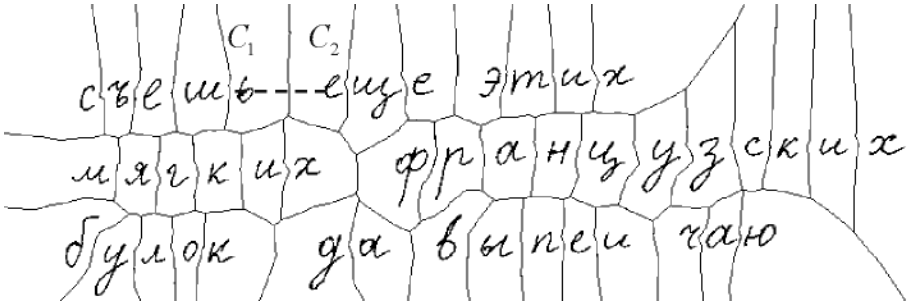


Рис. 4. Диаграмма Вороного области

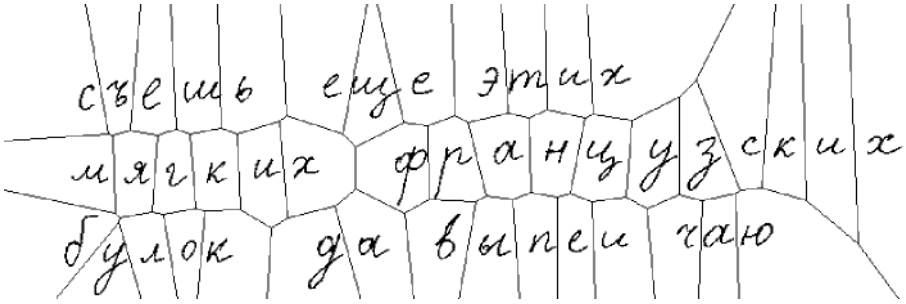


Рис. 5. Диаграмма Вороного множества центров масс символов

В качестве критерия принадлежности символов, заданных центрами масс $c = (c_x, c_y)$ и $p = (p_x, p_y)$ одному слову используются следующие условия.

1. Символы принадлежат одной строке:

$$|p_y - c_y| < \alpha \min\left(md_C(p), md_C(c)\right), \qquad (4)$$

где $\alpha > 0$ – заранее заданная константа.

2. Символы расположены достаточно «близко» друг к другу и принадлежат одному слову:

$$md_S(p,c) < \beta \min\left(md_C(p), md_C(c)\right), \qquad (5)$$

где $\beta > 0$ – заранее заданная константа.

Для сегментации большинства как машинописных, так и рукописных текстов использовались следующие параметры, полученные эмпирическим путём: $\alpha = 1,3$, $\beta = 1,5$.

АЛГОРИТМ ПОИСКА «СОСЕДНИХ» СИМВОЛОВ СЛОВА

АЛГОРИТМ 1

1. Пусть заданы константы α и β.

2. Пусть c – точка генератора, для которого необходимо найти соседний символ, U – множество точек генераторов, которые допускаются в решении.

3. Перебором рёбер, ограничивающих ячейку Вороного, порождённую точкой $c = (c_x, c_y)$, найти точку $p = (p_x, p_y)$, смежную с ней, такую, что выполняются следующие условия:

$$\begin{cases} p \in U \\ |p_y - c_y| < \alpha \min\left(md_C(p), md_C(c)\right) \\ md_S(p,c) < \beta \min\left(md_C(p), md_C(c)\right) \end{cases}.$$

4. Если p не найдена, вернуть $\varnothing$.

5. Обновить множество генераторов $U = U \setminus \{p\}$.

6. Вернуть p.

Сформулируем алгоритм выделения слова на основе алгоритма поиска «соседних» символов. Обозначим использование алгоритма поиска соседних символов как НАЙТИ_СОСЕДНИЙ_СИМВОЛ (c; U), где c – точка генератора, для которого необходимо найти соседний символ, U – множество точек генераторов, которые допускаются в решении.

АЛГОРИТМ ВЫДЕЛЕНИЯ СЛОВА

АЛГОРИТМ 2

1. Пусть c – точка генератора, соответствующая некоторому символу слова, U – множество точек генераторов, которые допускаются в решении.

2. Инициализировать список w, задающий порядок букв слова.

3. Положить $p = c$.

4. p = НАЙТИ_СОСЕДНИЙ_СИМВОЛ (p; U) (**Алгоритм 1)**

5. Если $p \neq \varnothing$, добавить p в **конец** списка w. (соседи «справа»), к 3.

6. Положить $p = c$.

7. p = НАЙТИ_СОСЕДНИЙ_СИМВОЛ (p; U) (**Алгоритм 1)**

8. Если $p \neq \varnothing$, добавить p в **начало** списка w. (соседи «слева»), к 6.

9. Список w содержит упорядоченные центры элементов арифметических выражений

10. Вернуть w

Сформулируем алгоритм сегментации текста, использующий алгоритмы поиска «соседних» символов и выделения слов на изображении. Аналогично алгоритму поиска «соседних» символов, обозначим использование алгоритма выделения слова как НАЙТИ_СЛОВО(c; U), где c – точка генератора, для которого необходимо найти соседний символ, U – множество точек генераторов, которые допускаются в решении.

АЛГОРИТМ СЕГМЕНТАЦИИ ТЕКСТА

АЛГОРИТМ 3

1. Заполнить список U упорядоченными по возрастанию координаты y точками-генераторами диаграммы Вороного.

2. Инициализировать список строк текста *L_List*.

3. Инициализировать текущую строку $L \leftarrow \varnothing$.

4. Если список U пуст, перейти к шагу 10.

5. Выбрать следующую точку $c = (c_x, c_y)$ из списка U, положить $U \leftarrow U \setminus \{c\}$.

6. Положить $W \leftarrow$ НАЙТИ_СЛОВО (c; U) (**Алгоритм 2)**

7. Если текущая строка L пуста, добавить слово W к строке L, к п. 5.

8. Если найденное слово W не принадлежит строке L, добавить строку L к списку *L_List*, положить $L \leftarrow \varnothing$, к п. 4.

9. Добавить слово W к строке L, к п. 4.

10. Если строка L не пуста, добавить её к списку строк *L_List*.

Шаг 8 описанного данного алгоритма является проверкой того, что найденное слово W не принадлежит строке L и заключается в исследовании двух множеств точек-генераторов S_L и S_w, составленных, соответственно, из точек,

принадлежащих сформированной строке L и точек найденного слова W.

Пусть точка $l \in S_L$ – точка строки L с наибольшей координатой y:

$$l = (x_l, y_l) : \forall s = (x, y) \in S_L, y_l \geq y.$$

Аналогично, $w \in S_w$, $w = (x_w, y_w)$ – точка найденного слова с наименьшей координатой y_w, $\forall s = (x, y) \in S_W, y_w \leq y$.

Если $y_w - y_l > \gamma d_C(w)$, где γ – некоторая заранее заданная константа, то считается, что слово W не принадлежит строке L. Для сегментации рукописных и машинописных символов константа γ выбиралась равной 1, 3.

4. АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Вычислительные эксперименты по сегментации текстов, проведённые для набора образцов рукописных и машинописных текстов, свидетельствуют о достаточной точности сегментации с использованием представленного подхода. Для рукописных текстов ошибки определения принадлежности символа слову, т.е. отношение числа ошибочно классифицированных символов к общему числу символов текста, составляет порядка 2,7%. На рис. 6 приведено графическое представление результатов сегментации текста.

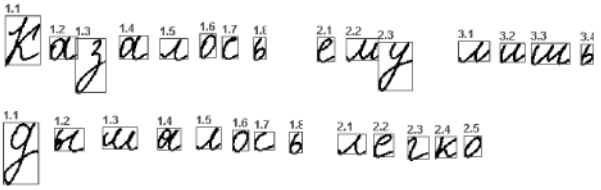


Рис. 6. Результат работы алгоритма сегментации текста

Каждому символу поставлена в соответствие метка, обозначающая соответственно порядковый номер слова в строке и порядковый номер символа в слове.

ЗАКЛЮЧЕНИЕ

Таким образом, на основе экспериментальных данных можно сделать вывод о том, что описанный метод сегментации текста является новым эффективным методом к выделению строк, слов и изолированных символов текста на основе использования диаграмм Вороного. Представленный метод используется в системе распознавания рукописных и печатных текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Shafait F.* Performance Comparison of Six Algorithms for Page Segmentation / F. Shafait, D. Keysers, T. Breuel // Image Understanding and Pattern Recognition (IUPR) research group. – 2006. – 12 pp.

2. *Kise K.* Segmentation of page images using the area voronoi diagram. / K.Kise, A.Sato, M. Iwata // Computer Vision and Image Understanding. – 1998. – Vol. 3, no. 70. – P. 370–382.

3. *Препарата Ф.* Вычислительная геометрия. / Ф. Препарата, М. Шеймос – М.: Мир, 1989. – 295 с.

4. *Fortune S.* A sweepline algorithm for Voronoi diagrams / S. Fortune // Proceedings of the second annual symposium on Computational geometry. – 1986. – P. 313 – 322.

5. *Computational Geometry Algorithms and Applications* / [Edited by M. Berg]. – 3rd Edition. – Berlin.: Springer-Verlag, 2008 – 386 pp.

6. *Wang Z.* Word Extraction Using Area Voronoi Diagram. / Z. Wang, Y. Lu, C. Lim // CVPRW '03. – 2003. – P. 31 – 36.

Запрягаев Сергей Александрович – д. ф.-м.н, проф. каф. цифровых технологий Воронежского государственного университета. Тел.(4732)208-257. E-mail:zsa@main.vsu.ru.

Сорокин Андрей Игоревич – аспирант, кафедра цифровых технологий, Воронежский государственный университет. Тел. (4732) 208-257. E-mail: sorokinai@narod.ru.

Zapryagaev S. A. – Doctor of Physics-math. Sciences, Professor of the dept. of digital technologies Voronezh State University. Tel.(4732)208-257. E-mail:zsa@main.vsu.ru.

Sorokin Andrey Igorevich – Post-graduate student of the dept. of digital technologies Voronezh State University. Tel.(4732)208-257.