УДК 51-74

РАСПОЗНАВАНИЕ СТРОК В СТЕНОГРАФИЧЕСКИХ ДОКУМЕНТАХ

Гиппиев М.Б., Жуков А.В., Рогов А.А., Скабин А.В.

Петрозаводский государственный университет, Петрозаводск, Россия (185910,Россия, Республика Карелия, г. Петрозаводск, пр. Ленина, 33)

В настоящее время перевод изображений рукописного, машинного или печатного текста в текстовые данные стало одним из активно развивающимся направлением в распознавании образов. Существует большое количество программных средств, но они позволяют автоматизировать распознание печатного текста и текстовых форм, при условии хорошей оцифровки источника. В данной статье рассматривается проблема выделения строк на рукописных документах, которые были оцифрованы при помощи фотоаппарата. В печатных качественно оцифрованных документах выделение строк не является сложной задачей, которая решается распространёнными методами. В статье проводится сравнение двух новых методов выделения строк:метод ближайшего соседаи метод построения графа связей. Приводятся их алгоритмы, и также демонстрируются результаты работы на стенографических документах, и приводятся критерии оценки их работы. Кроме того,в статье приводится алгоритм распознавания надстрочных и подстрочных символов.

Ключевые слова: выделение строк, рукописные документы, метод ближайшего соседа, граф связей.

RECOGNITIONOF LINES INTHE HISTORICAL HANDWRITTEN DOCUMENTS

GippievM.B., ZhukovA.V., RogovA.A., SkabinA.V.

Petrozavodsk State University, Petrozavodsk, Russia (185910, Russia, Karelia, Petrozavodsk, street Lenina, 33)

Transformation of handwritten or printed text images into text data has currently become one of the most actively developing areas in pattern recognition. There are many software tools which allow automation ofprinted text and text forms recognition, provided with a good digitizing of the source. This article deals with string extraction on handwritten documents, whichhave been digitized using a camera. String selection in printed and qualitatively digitizeddocuments is not a difficult problem, which can be solved using widespread methods. Comparisonof two new string recognition methods is discussed: method of nearest neighbor and method of bond graph construction. Algorithms, as well as the results of their work on stenographic documents andcriteria of evaluation of their performance are described in the article. Moreover, the article presents an algorithm of subscript and superscript character recognition.

Key words: recognition of lines, handwritten document, method of nearest neighbor, bond graph.

Введение

Перевод изображений рукописного, машинописного или печатного текста в текстовые данные стал сегодня обыденным явлением. Существует большое количество программных средств, реализующих оптическое распознавание символов (opticalcharacterrecognition,OCR). Среди систем, поддерживающих русский язык, можно выделить «ABBYYFineReader», «CuneiForm», «GoogleTesseract» и другие. Указанные системы обладают богатым функционалом и позволяют автоматизировать распознавание и печатного текста и текстовых форм. Отметим, что некоторые из систем успешно борются с искажениями, характерными для сканированных печатных документов. Распознавание текстов с другими видами искажений, возникающих при фотографировании источника цифровыми фотоаппаратами, является серьезным препятствием для указанных систем. Одним из главных условий успешного распознавания печатного текста указанными системами, является необходимость

прямого горизонтального расположения строк в документе.

В современных системах OCR распознавание символа осуществляется с учетом предыдущих и последующих символов. Следовательно, возникает задача поиска соседних символов, а именно – разбитие текста на строки. В печатных текстах данная задача, как правило, легко решается, после определения направления текста. Наиболее распространённым является метод определения направления строк при помощи построения проекции символов на ось, перпендикулярную строкам текста (метод проекций). Пики в построенной проекции будут соответствовать строкам. Задача резко усложняется в случае фотографировании толстых книг или распознавании рукописного текста [1,2]. Любой рукописный документ несет в себе индивидуальные особенности, которые связаны с привычками автора, скоростью письма, аккуратностью и некоторыми другими факторами. На определение строк влияют: наклон текста в ту или иную сторону, заваливание, исправление и зачеркивание текста. Простейшие примеры такого текста можно увидеть на рис. 1. Кроме того, анализ рукописного текста, написанного с помощью стенографической записи,усложняет появление надстрочных и подстрочных символов.

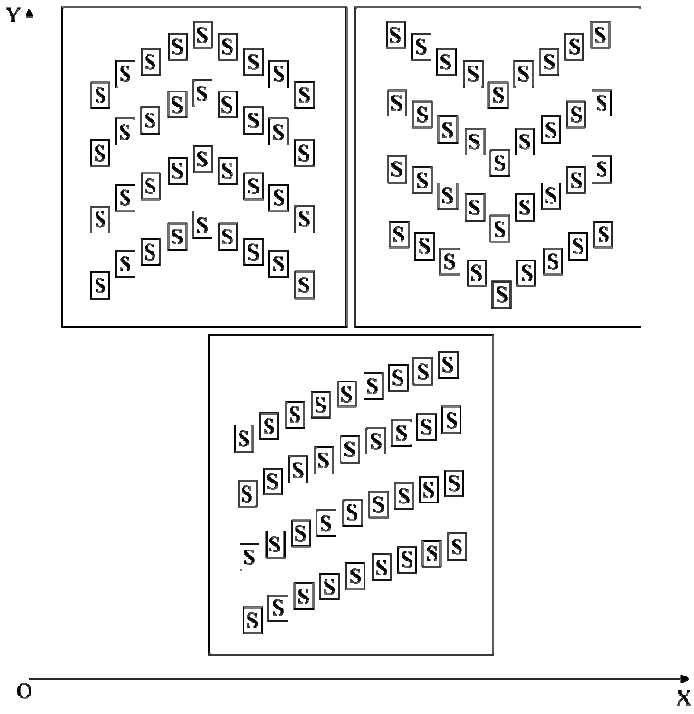

Рисунок 1. Примеры расположения текста на бумаге

Имеющиеся методы определения строк графического представленного текста существенно зависит от вида текста [1]. В работе [4] описывается алгоритм распознавания строк методом ближайшего соседа, основанный на предварительном выделении на исторической стенограмме отдельных символов. В данной статье предлагается модификация этого алгоритма, улучшающая его качество.

Алгоритм распознавания строк методом ближайшего соседа

Основной идеей алгоритма распознавания строк, описанного в работе [4], является гипотеза о том, что расстояние между соседними символами в строке меньше, чем расстояние между соседними строками. Он состоит из следующих шагов:

1. Определяются направления координатных осей (ось OX направлена параллельно предполагаемому направлению строк);
2. Для каждого символа ищется ближайший символ – сосед, который удовлетворяет критериям поиска;
3. Формируются множества соседей, которыесоответствуют строкам в стенограмме.

Опишем критерий поиска. Для этого введем обозначения:

s_i – i-ый символ в стенограмме $i = 1..N$;

d_{ij} – расстояние между центрамиi-го и j-го символов;

α_{ij} – угол между линией, соединяющей центры i-ого и j-ого символов, и линией параллельной оси OX (рис. 2);

k – произвольный коэффициент, который подбирается в зависимости от стенограммы;

l_{ij} – проекция на ось OX расстоянияd_{ij};

α_{max} – максимальный угол между линией, соединяющей центры символов, и линией параллельной оси OX;

α_{min} – минимальный угол между линией, соединяющей центры символов, и линией параллельной оси OX;

l_{max} – максимальное расстояние по оси OX между центрами соседних символов;

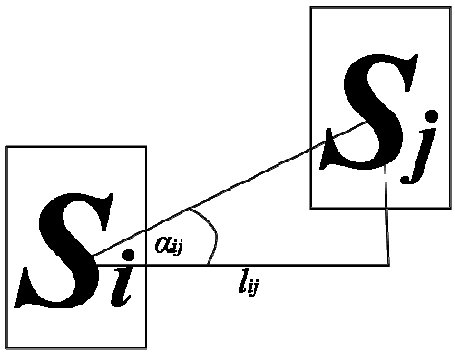

Рисунок 2. Поиск соседнего символа

NS_i – массив соседей, где i – номер текущего символа, а значение элемента NS_iсоответствует номеру соседа, либо -1, если соседа нет.

Расстояния между символами с номерами i и j, вычисляются по следующей формуле:

$$d_{ij} = \alpha_{ij}k + l_{ij} \quad (1)$$

Рассмотрим этапы алгоритма более формально:

1. В начале работы алгоритма для каждого символа происходит поиск соседа, то есть символа, до которого расстояние является минимальным, и который удовлетворяет условиям: угол между символами $\alpha_{min} \leq \alpha_{ij} \leq \alpha_{max}$ и расстояние по горизонтали $l_{ij} \leq l_{max}$. Таким образом, формируется массив NS_i.

2. Затем для каждого символа из этого массива выполняются следующие действия:

2.1. Определяется строка из одного элемента – текущего символа (s_i).

2.2. Формируется конец строки. Если у символа есть сосед ($NS_i \neq -1$), тогда к концу строки добавляется символ с номером NS_i, и затем для последнего символа, операция повторяется до тех пор, пока для вновь добавленного символа s_l значение соответствующего ему элемента из массива NS_l, не будет равно -1 ($NS_l \neq -1$).

2.3. Аналогичным образом формируется начало строки. Если есть в массиве NS элемент, значение которого равняется номеру текущего элемента $NS_p = i$, тогда в начало строки добавляется символ с номером s_p.

Таким образом, у нас формируется строка, то есть множество символов, относящихся к строке: $(s_0, \dots, s_{i-1}, s_i, s_{i+1}, \dots, s_n)$, где $s_0, \dots, s_{i-1}$ – номера символов, добавленных на 2.3 шаге и $s_{i+1}, \dots, s_n$ – номера символов, добавленных на 2.2 шаге.

Для выбора оптимальных параметров работы алгоритма была разработана система, которая позволяет вручную выделять строки в стенограмме – построить эталон, сгенерировать наборы входных параметров алгоритма, и для каждого набора параметров получить количественные оценки (полнота, точность, F-мера) путем сравнения результата работы алгоритма с построенным эталоном. Наилучшие результаты на основании F-меры алгоритм распознавания строк методом ближайшего соседа показал при следующих параметрах, представленных в табл.1:

Таблица 1. Значение критериев при различных параметрах

α_{max}	α_{min}	k	l_{max}	F-мера	Точность	Полнота
40	**-40**	**2**	**200**	**0,8343**	**0,9446**	**0,7877**
32	-40	2	190	0,8342	0,9476	0,7867
29	-34	2	190	0,8323	0,9477	0,7837
26	-34	2	190	0,825	0,9453	0,7736
29	-40	2	170	0,822	0,9357	0,7774

На рис. 3 представлен фрагмент лучшего результата работы данного алгоритма.

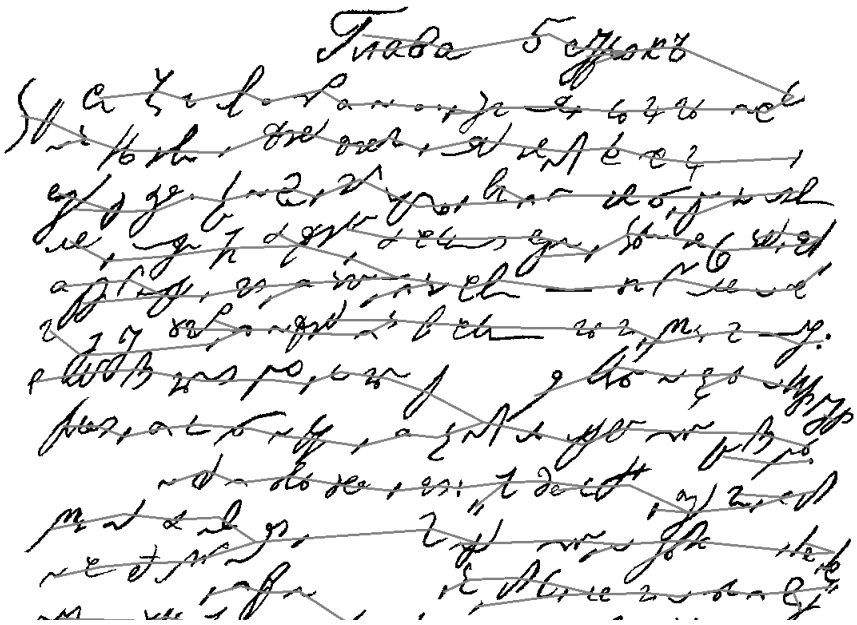

Рисунок 3. Фрагмент результата работы алгоритма распознавания строк методом ближайшего соседа

Проведенные исследования показали недостаточно хорошее качество распознавания строк. Нормировка формулы для расчета расстояния не улучшила результаты:

$$d_{ij} = \frac{\alpha_{ij}}{|\alpha_{max} - \alpha_{min}|}k + \frac{l_{ij}}{l_{max}} \qquad (2)$$

Алгоритм распознавания строк методом построения графа связей

Предлагаемый алгоритм является модификацией выше описанного. Его идея заключается в построении всех возможных связей между символами и последовательном объединении символов, входящих в связи, начиная со связей с наименьшими расстояниями.

В данном алгоритме используются следующие обозначения:

– R_r– список связей, где для каждой связи задан номер первого символа –$R_r.FS$, номер второго символа –$R_r.SS$, расстояние между символами –$R_r.D$.

– Для каждого символа задан номер строки, к которой он относится – $s_i.L$, номер следующего за ним символа в строке – $s_i.NS$ (если не задан, то $s_i.NS = -1$) и номер символа, для которого следующим символом в строке является текущий, – $s_i.PS$ (если не задан, то $s_i.PS = -1$).

l_k – *k*-ая строка.

– $VerticalDistances(l_{k1}, l_{k2})$ – расстояние вдоль оси OY между центрами символов*k1*-ой строки и символами *k2*-ой строки.

– h_{max} – максимальное вертикальное расстояние между символами строк.

Алгоритм состоит из следующих этапов:

1. Строятся все возможные связи между символами (R_r), то есть находятся пары символов такие, что угол между линией, соединяющий первый символ ($First = R_r.FS$) со вторым ($Second = R_r.SS$), и линией, параллельной оси OX, не больше максимального и не меньше минимального углов ($\alpha_{min} \leq \alpha_{FirstSecond} \leq \alpha_{max}$), а расстояние по горизонтали между символами не превышает максимального ($l_{FirstSecond} \leq l_{max}$). $R_r.D = d_{FirstSecond}$.
2. Связи упорядочиваются в порядке возрастания расстояний между символами ($R_r.D$).
3. Каждый символ помещается в отдельную строку.
4. Для каждой связи R_r из списка R_r выполняются следующие действия:

4.1. Если символы, входящие в связь R_r, можно соединить (рис. 4), то есть для первого символа ($First = R_r.FS$) не задан следующий за ним символ ($s_{First}.NS = -1$), а для второго символа ($Second = R_r.SS$) не задан предшествующий ($s_{Second}.PS = -1$), тогда объединяем строки, содержащие эти символы ($l_{s_{First}.L}$ и $l_{s_{Second}.L}$).

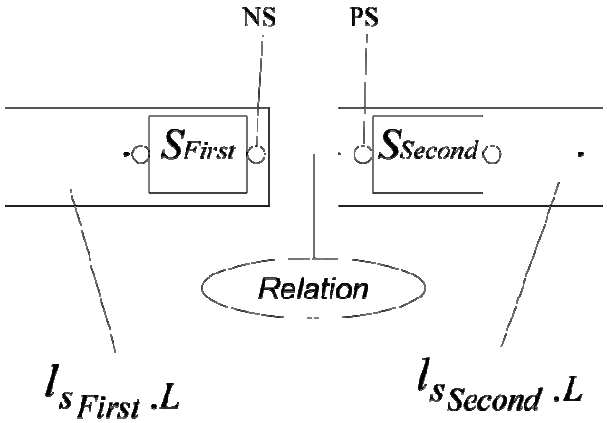

Рисунок4. Символы, которые можно соединить

4.2. Иначе, если символы соединить нельзя, то находим вертикальные расстояния между символами строк $l_{s_{First}.L}$ и $l_{s_{Second}.L}$ ($VerticalDistances\left(l_{s_{First}.L}, l_{s_{Second}.L}\right)$), и если для любого $h \in VerticalDistances\left(l_{s_{First}.L}, l_{s_{Second}.L}\right)$ выполняется условие $h \leq h_{max}$, тогда объединяем строки $l_{s_{First}.L}$ и $l_{s_{Second}.L}$.

4.3. Символы в строке упорядочиваются по горизонтальным координатам центров символов.

Расстояния между символами находятся таким же образом, как и в предыдущем алгоритме.

Наилучшие результаты алгоритм распознавания строк методом построения графа связей показал при следующих параметрах, представленных в табл. 2:

Таблица 2. Значения критериев при различных параметрах

α_{max}	α_{min}	k	l_{max}	h_{max}	F-мера	Точность	Полнота
20	**-28**	**1**	**200**	**10**	**0,9806**	**0,9809**	**0,9808**
17	-22	1	190	10	0,9773	0,9775	0,9775
11	-25	4	190	15	0,9771	0,9775	0,9773
11	-40	3	190	30	0,9745	0,9749	0,9748
14	-28	4	190	20	0,9678	0,9685	0,9683

На рис.5 представлен фрагмент лучшего результата работы данного алгоритма.

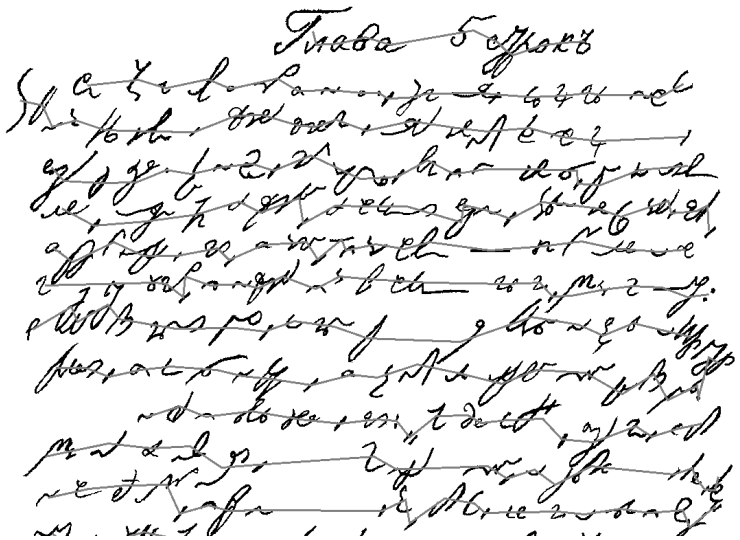

Рисунок 5. Фрагмент результата работы алгоритма распознавания строк методом построения графа связей

Предлагаемый алгоритм распознавания строк методом построения графа связей показал лучший результат, нежели алгоритм распознавания строк методом ближайшего соседа, что подтверждается как оценками, так и визуальным представлением результатов работы алгоритмов.

Алгоритм распознавания подстрочных и надстрочных символов

При распознавании строк возникает задача определения, к какому типу (основной, подстрочный и надстрочный) относится каждый символ, входящий в строку. В качестве возможного варианта решения данной задачи можно предложить следующий алгоритм. После распознавания строки строим линию аппроксимации по центрам ее символов. Анализ показал, что, как правило, строки в стенограммах имеют форму, которую можноаппроксимировать полиномом второй степени.Определим расстояния между центрами символов и линией аппроксимации. Для некоторого символа обозначим это расстояние за ε, тогда вероятность того, что данный символ является основным равна$P_{\text{осн}} = e^{-\alpha\varepsilon}$, а вероятность того, что символ является либо подстрочным (символ расположен под

линией аппроксимации), либо надстрочным (символ расположен над линией аппроксимации),$P_{\text{неосн}} = 1 - e^{-\alpha\varepsilon}$, где α – некоторый коэффициент, который выбирается в зависимости от стенограммы, при этом $P_{\text{осн}} + P_{\text{неосн}} = 1$.

Заключение

Рассмотренный в статье алгоритм будет реализован в создаваемой компьютерной программе для распознавания исторических стенограмм [3,5,6].

Работа выполнена при поддержке Программы стратегического развития (ПСР) ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

Список литературы

1. МасаловичА.А., МестецкийЛ.М.Распрямление текстовых строк на основе непрерывного гранично-скелетного представления изображений// Графикон-2006: Труды международной конференции. –Новосибирск, 2006.

2. МестецкийЛ.М.Скелет многосвязной многоугольной фигуры// Графикон-2005: Труды международной конференции. – Новосибирск, 2005.

3. Рогов А.А., Скабин А.В., Штеркель И.А. Автоматизированная информационная система распознавания исторических рукописных документов. Информационная среда ВУЗА XXI века. Материалы международной научной конференции. 4–10 декабря 2012 г. Купио (Финляндия). – С. 127-130.

4. Рогов А.А., Скабин А.В., Штеркель И.А. О дешифровке рукописных исторических документов //Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL 2012. – С. 111-117.

5. Рогов А.А., Скабин А.В., Штеркель И.А. О дешифровке исторических рукописных документов //Информационные технологии и письменное наследие El’Manuscript 2012: материалы IV Международной научной конференции. –Петрозаводск, 3 – 8 сентября 2012.– С. 230–233.

6. Скабин А.В., Рогов А.А. Бинаризация и выделение символов исторической стенограммы // Ученые записки Петрозаводского государственного университета. Серия «Естественные и технические науки». –2013. –№ 4 (133).

Рецензенты:

Печников Андрей Анатольевич, доктор технических наук, доцент, ведущий научный сотрудник Лаборатории телекоммуникационных систем Института прикладных математических исследований, ИПМИ КарНЦ РАН, Петрозаводск.

Питухин Евгений Александрович, доктор технических наук, профессор, доцент кафедры математического моделирования систем управления, Петрозаводский государственный университет, г.Петрозаводск.