УДК 004.352.243

А.А. Мозговой

ПРОБЛЕМЫ ИЗВЛЕЧЕНИЯ РУКОПИСНЫХ СЛОВ ИЗ СКАНИРОВАННОГО ИЗОБРАЖЕНИЯ

Воронежский институт высоких технологий

Рассматриваются основные проблемы извлечения рукописных слов из сканированного изображения для последующего распознавания текста. Предлагаются алгоритмы решения проблем взаимного пересечения и неравномерной сегментации областей рукописных слов с примерами. Предполагается наличие изображения, содержащего исключительно текстовую составляющую.

Ключевые слова: оптическое распознавание, рукописный.

Рукописные слова перед распознаванием необходимо сначала извлечь из изображения. Если для печатного текста эта задача является тривиальной, при условии хорошего качества изображения [1], то в случае с рукописным текстом появляются дополнительные сложности. Прямоугольные области слов могут пересекаться друг с другом, внутри слов могут находиться неравномерные разрывы, усложняющие идентификацию объекта как целого, возможна нестабильная высота символов внутри слова и т.д. Можно отметить и одно положительное свойство рукописного текста – постоянство толщины линии, что может быть использовано для поиска областей с текстом SWT методом [2].

В общем виде схема алгоритма извлечения рукописных слов из изображения показана на рис.1. Пунктиром выделены участки алгоритма, которые отвечают за визуализацию процесса.

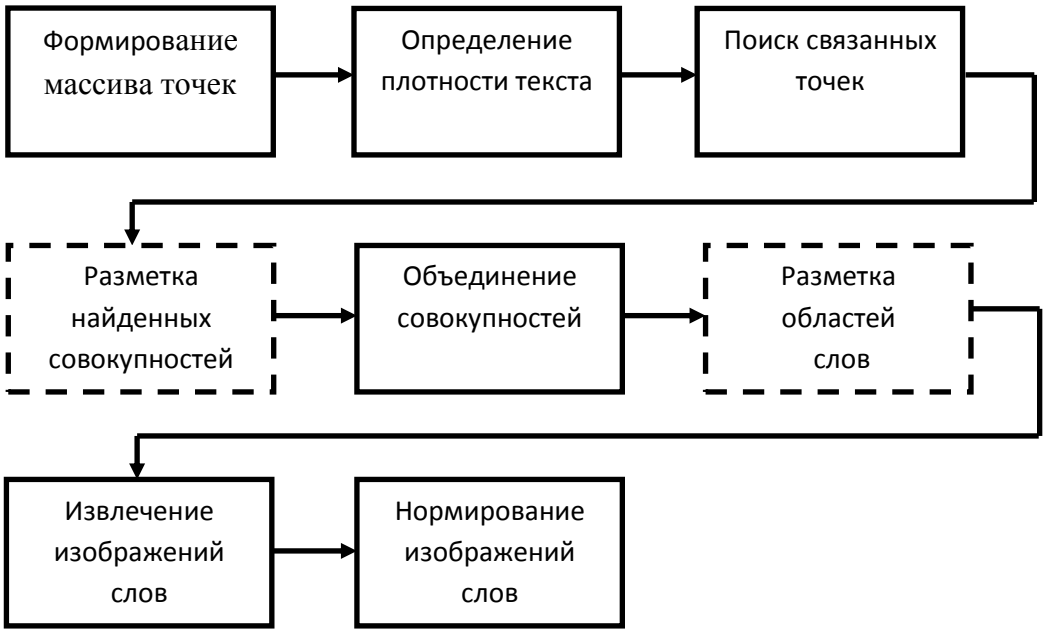

Рис. 1. Схема алгоритма извлечения рукописных слов из изображения

На вход алгоритма подаётся очищенное от шумов и посторонних объектов изображение, содержащее исключительно текстовую информацию. Сначала изображение преобразуется в однобитную матрицу для увеличения скорости обработки информации. Затем определяется плотность текста. Процедура определения плотности текста может быть выполнена в любое время до начала выполнения процедуры объединения совокупностей связанных точек, на втором месте в схеме она изображена лишь формально. Плотность текста $\rho_т$ используется в дальнейшем для расчёта предполагаемого расстояния между словами d:

$$d = \rho_{т\times} a = \frac{\sum_j (\sum_i d_i)}{N_d} \times a,\ d_i = \begin{cases} d_i, d_i < d_m \\ 0, d_i \geq d_m \end{cases}, \quad (1)$$

$$d_m = \frac{I_w}{N_d}. \quad (2)$$

где:

d_i – ширина промежутка i для j-ой строки в пикселях,

N_d – количество суммируемых промежутков,

I_w – ширина изображения в пикселях,

α – эмпирический коэффициент (α=1,2).

Если на изображении находится несколько областей с рукописным текстом разной плотности, то необходимо рассчитать расстояние между словами для каждой области.

Далее выполняется поиск связанных друг с другом точек рекурсивным методом. У каждой точки помечаются её соседи, координаты которых заносятся в стек. Сама точка присоединяется к текущей совокупности и удаляется из анализируемого бинарного множества. Из верхушки стека берётся следующая точка и операция повторяется. Признаком окончания рекурсии является достижение дна стека. В результате выполнения описанной процедуры получается некоторый набор совокупностей точек. Пример такого набора показан на рис.2.

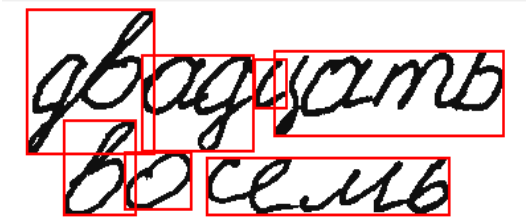

Рис. 2. Разметка найденных совокупностей точек

В случае с печатным текстом разметку областей слов на это можно было закончить, так как дальше всё очень просто: расстояния между буквами, словами и строками одинаковые; высота букв одинаковая; буквы

строго сегментированы и т.д. С рукописным текстом всё гораздо сложнее. На рис.2 можно видеть взаимные пересечения областей частей слов, «плавающие» размеры и нестабильную сегментацию. За объединение этих разрозненных совокупностей точек в слова отвечает процедура объединения совокупностей точек. Для её работы требуется определение предполагаемого расстояния между словами по формуле (1), а также определение относительного допуска ϭ, с помощью которого строится предположение о принадлежности элемента слова той или иной строке:

$$\sigma = \frac{(y1_{max} - y1_{min}) + (y2_{max} - y2_{min})}{4}, \quad (3)$$

где $y1_{max}$, $y2_{max}$, $y1_{min}$, $y2_{min}$ – вертикальные координаты верхних и нижних точек прямоугольных областей совокупностей точек, 4 – коэффициент найденным эмпирическим путём.

Признаком нахождения областей с точками на одной строке является взаимное поглощение их вертикальных размеров с допуском ±ϭ. Признаком принадлежности областей одному слову внутри одной общей строки является расстояние между ними ≤d. Результат работы описанной процедуры показан на рис.3.

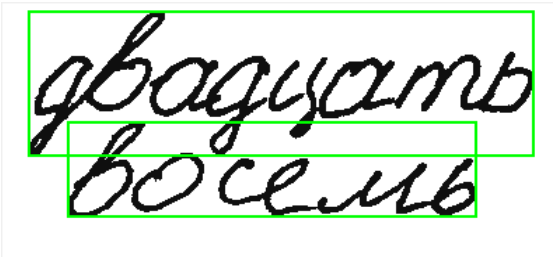

Рис. 3. Разметка найденных слов

Видимое пересечение областей слов не является препятствием для последующего их извлечения, так как совокупности точек объединялись лишь присвоением им одинакового маркера, обозначающего номер слова. Все точки по-прежнему хранятся с использованием массива с координатами этих точек. Сложный случай со «слипанием» слов не рассматривался. В дальнейшем предполагается либо морфологическая обработка «слипшихся» участков [3], либо их «разрезание».

Нормирование изображений извлечённых слов определяется алгоритмом их последующего распознавания. Чаще всего – это масштабирование до определённого размера по вертикали.

Описанная методика была успешно применена для извлечения слов из экспериментального набора чисел, написанных словами. Использованные сканированные изображения были высокого качества с разрешением 300ppi. В статье не рассматривались вопросы предварительной обработки изображения для увеличения качества операций извлечения из него

рукописных слов. Данная задача практически идентична подготовке изображения, содержащего печатный текст.

ЛИТЕРАТУРА

1. М.П. Кривенко Предварительная обработка при распознавании текстов по изображению низкого качества // Информатика и её применения, 2012. Т. 6. Вып. 4. С. 49-56.
2. B. Epshtein, E. Ofek, Y. Wexler Detecting text in natural scenes with stroke width transform // Computer Vision and Pattern Recognition (CVPR), 2010 IEEE, pp. 2963-2970.
3. А.Р. Гонсалес, Р. Вудс Цифровая обработка изображений/ Гл.9 Морфологическая обработка изображений. - М., 2006. - С. 747-811.

A.A. Mozgovoy

THE PROBLEM OF EXTRACTING HANDWRITTEN WORDS FROM THE SCANNED IMAGE

Voronezh Institute of High Technologies

The basic problem of extracting handwritten words from scanned images for subsequent recognition of the text. Algorithms for solving problems of mutual intersections and non-uniform segmentation of handwritten words with examples. Assumes the existence of an image containing only text component.

Keywords: optical recognition, handwriting.