УДК 28.23.15

МЕТОДЫ И ЭТАПЫ РАСПОЗНАВАНИЯ РУКОПИСНОГО ТЕКСТА

Полюхин Д.А., Сальников И.И.

Пензенский государственный технологический университет, Пенза, e-mail: denis1564@yandex.ru

Статья посвящена актуальному направлению в информатике – распознаванию рукописного текста. Выполнен обзор систем распознавания, использующих различные методы: шаблонные, признаковые, структурные. Отмечаются достоинства и недостатки этих методов. Приводятся характерные особенности распознаваемого рукописного текста – шрифтовое и размерное разнообразие символов; искажения в изображениях символов, разрывы образов символов; слипание соседних символов; перекосы при сканировании; посторонние включения в изображениях; сочетание фрагментов текста на разных языках; большое разнообразие классов символов, которые могут быть распознаны только при наличии дополнительной контекстной информации. В обзорном плане рассмотрены существующие этапы распознавания рукописного текста. На этапе предобработки изображения выполняется повышение качества изображения за счет фильтрации, шумоподавления и других, имеющих своей целью повысить качество изображения. Используется пороговая бинаризация, которая позволяет резко разделить текст и фон, упрощает в дальнейшем применение многих алгоритмов, а также избавляет от некоторых шумов на изображении. На этапе сегментации текст разделяется на удобные для анализа составные части. Выполняется разделение текста на отдельные строки (сегментация строк) и разделение строк на слова (сегментация слов).

Ключевые слова: распознавание рукописного текста, шаблонные, признаковые, структурные, фильтрация изображения, бинаризация, сегментация текста, нормализация строк

METHODS AND STAGES OF A DISCERNMENT OF THE HAND-WRITTEN TEXT

Polukhin D.A., Salnikov I.I.

Penza State Technological University, Penza, e-mail: denis1564@yandex.ru

The paper is devoted to actual direction in computer science – discernment of the hand-written text. The review of systems of a discernment using various methods is carried out: sample, character, structural. The virtues and shortages of these methods are marked. The characteristics of the recognized hand-written text – font and dimensional variety of numerals are reduced; distortions in images of numerals, ruptures of images of numerals; слипание of adjacent numerals; distortions at scanning; outside inclusions in images; a combination of fragments of the text on different languages; large variety of classes of numerals, which can be recognized only at presence of an additional context-sensitive information. In a survey plan the existing stages of a discernment of the hand-written text are considered. At a stage предобработки of an image the image enhancement is fulfilled at the expense of a filtration, noise quieting and others, having by the purpose to increase quality of an image. Is used threshold бинаризация, which allows sharply to divide the text and hum noise, simplifies in further application of many algorithms, and also saves of some noise on an image. At a stage of segmentation the text is divided into the constituents, convenient for the analysis. The separation of the text on separate lines (segmentation of lines) and separation of lines on words (segmentation of words) is fulfilled.

Keywords: handwriting recognition, template, feature, structural, image filtering, binarization, text segmentation, string normalization

Широко исследуемой проблемой в настоящее время является распознавание рукописного текста. На данный момент достигнутая точность распознавания даже ниже, чем для рукописного «печатного» текста. Более высокие показатели могут быть достигнуты только с использованием контекстной и грамматической информации. Например, в процессе распознания искать целые слова в словаре легче, чем пытаться проанализировать отдельные символы из текста. Знание грамматики языка может также помочь определить, является ли слово глаголом или существительным. Формы отдельных рукописных символов иногда могут не содержать достаточно информации, чтобы точно (с уровнем более 98%) распознать весь рукописный текст. Методы автоматического распознавания образов и их реализация в системах оптического чтения текстов (*OCR*-системах – *Optical Character Recognition*) – одна из самых плодотворных технологий искусственного интеллекта (ИИ). В приведенной трактовке *OCR* понимается как автоматическое распознавание с помощью специальных программ изображений символов печатного или рукописного текста, например, введенного в компьютер с помощью сканера, и преобразование его в формат, пригодный для обработки текстовыми процессорами, редакторами текстов и т.д. Иногда под *OCR* понимают устройство оптического распознавания символов или автоматического чтения текста. В настоящее время такие устройства при промышленном использовании обрабатывают до 100 тыс. документов в сутки. Промышленное использование предполагает ввод документов хорошего и среднего качества – это бланки переписи населения, налоговых деклараций, бланки статистического учета и т. п. [1].

Отметим следующие особенности предметной области, существенные с точки зрения *OCR*-систем: шрифтовое и размерное разнообразие символов; искажения в изображениях символов (разрывы образов символов, например, при увеличении изображения; слипание соседних символов и др.); перекосы при сканировании; посторонние включения в изображениях; сочетание фрагментов текста на разных языках; большое разнообразие классов символов, которые могут быть распознаны только при наличии дополнительной контекстной информации.

Методы распознавания**.** Системы распознавания реализуются как классификаторы, использующие различные методы: шаблонные (растровые); признаковые; структурные [2]. В классификаторе шаблонного типа с помощью критерия сравнения определяется, какой из шаблонов выбрать из базы. Самый простой критерий – минимум точек, отличающих шаблон от исследуемого изображения. К достоинствам шаблонного классификатора относятся хорошее распознавание дефектных символов («разорванных» или «склеенных»), простота и высокая скорость распознавания. Недостатком является необходимость настройки системы на типы и размеры шрифтов.

В признаковых классификаторах анализ проводится только по набору чисел или признаков, вычисляемых по изображению. Этот метод позволяет распознавать различные начертания символов, т.е. различные подчерки шрифты и т.д. Этот метод неизбежно вызывает некоторую потерю информации, так как используется топологическое представление, отражающее информацию о взаимном расположении структурных элементов символа. Эти данные могут быть представлены в графовой форме. При этом данный метод обеспечивает инвариантность относительно типов и размеров шрифтов. Недостатками являются трудность распознавания дефектных символов и медленная работа.

Основой структурно-пятенного метода является структурно-пятенный эталон [2]. Он имеет вид набора пятен с попарными отношениями между ними. Данное представление нечувствительно к различным начертаниям и дефектам символов. Алгоритм основан на сочетании шаблонного и структурного методов распознавания образов. При анализе образца выделяются ключевые точки объекта – так называемые «пятна». В качестве пятен, например, могут выступать: концы линий; узлы, где сходятся несколько линий; места изломов линий; места пересечения линий; крайние точки. После выделения характерных точек определяются связи между ними – отрезок, или дуга. Таким образом, итоговое описание представляет собой граф, который и служит объектом поиска в библиотеке «структурно-пятенных эталонов». При поиске устанавливается соответствие между ключевыми точками образца и эталона, после чего определяется степень деформации связей, необходимая, чтобы привести искомый объект к сравниваемому эталонному образцу. При этом, меньшая степень необходимой деформации предполагает большую вероятность правильного распознавания символа. Далее рассмотрим этапы обработки изображения.

1. Предобработка. На этом этапе выполняются следующие задачи: повышение качества изображения за счет фильтрации, шумоподавления и других, имеющих своей целью повысить качество изображения. На этом этапе происходит очистка изображения от дефектов сканирования. В частности, в самом начале работы к изображению в целях шумоподавления часто применяется фильтр Гаусса. Важную роль играет пороговая бинаризация, то есть перевод изображения в чёрно-белый формат из цветного или оттенков серого [3]. Это позволяет резко разделить текст и фон, упрощает в дальнейшем применение многих алгоритмов, а также избавляет от некоторых шумов на изображении. При этом используется гистограмма яркости изображения текста, на котором наблюдается два пика: высокий пик, соответствующий белому фону, то есть цвету бумаги, и пик в области тёмных пикселей, соответствующих яркости символов текста.

2. Выделение региона интереса. На этом этапе на бинаризованном изображении выделяется непосредственно область, на которой находится распознаваемый текст, и отбрасываются элементы, текстом не являющиеся [3,4]. К ним относятся такие объекты, как кляксы, пятна на бумаге, не удалённые в процессе бинаризации, картинки и др. Для их удаления можно, например, выделять компоненты связности на изображении, вычислять геометрические признаки и на их основе классифицировать компоненту связности как часть текста или дефект, используя методы машинного обучения или эвристики.

3. Сегментация и нормализация текста. На этом этапе текст разделяется, или сегментируется, на удобные для анализа составные части [5]. Наиболее естественными действиями на данном этапе является разделение текста на отдельные строки (сегментация строк) и разделение строк на слова (сегментация слов), а также, теоретически,

разделение слов на элементарные составные части. Кроме того, на данном этапе проводится нормализация текста приведение выделенных составных частей к некоторому стандартному виду для снижения вариативности и упрощения распознавания.

Сегментация строк. Задача сегментации (разделения) строк в машинопечатных документах на сегодняшний день считается полностью решённой. Но в задачах при разделении строк в общем случае возникают сложности, не позволяющие напрямую применять алгоритмы, пригодные для машинопечатных текстов:

– строки не только могут не являться параллельными, но и могут изгибаться;

– различные строки могут быть слишком близки, а элементы текста, принадлежащего различным строкам, могут налагаться друг на друга.

Например, если коэффициент формы области (отношения квадрата её периметра к площади) меньше некоторого значения, а площадь больше некоторого значения, то это с большой вероятностью дефект (т.к. рукописный текст обычно является некоторой кривой).

Выделение базовых линий. Эти методы основаны на идее, что человек пишет либо по, либо поверх некоторой воображаемой линии. Данные методы пытаются аппроксимировать эту линию, а затем восстановить по ней строку. В преобразовании Хафа выделяются прямые, если они не слишком искривлены. Преобразование Хафа применяется к центрам компонент связности пикселей текста. Такой подход требует, чтобы строки текста были близки к прямым, но зато позволяет выделять строки, расположенные в произвольном месте и идущие под произвольными углами.

Пересечение элементов различных строк представляет собой проблему не только сегментации строк, но и распознавания текста, так как отнесение элемента к неправильной строке очевидно ухудшает его распознаваемость. Пересекающиеся компоненты являются проблемой для методов горизонтальной проекции (так как они увеличивают значение профиля проекции в тех местах, где должен быть его минимум) группировочных методов (так как они используют связные компоненты пикселей текста для построения строк), но слабо влияют на некоторые методы выделения базовых линий. Для поиска пересекающихся элементов из различных строк можно использовать такие признаки, как размер компонент связности текста, факт отнесения одной компоненты к нескольким строкам или, напротив, не относящимся ни к какой строке. После нахождения таких сомнительных компонент нужно определить, относятся ли они к какой-то строке или же их нужно декомпозировать на элементы, относящиеся к разным строкам. Такая вертикальная декомпозиция компонент сложная задача. Простое решение заключается в разрезании компоненты на части горизонтальными линиями, но можно применить и более тонкие подходы, например, выделение отдельных штрихов.

3. Сегментация слов. На этом этапе работы системы распознавания выделенные строки текста разделяются на отдельные слова. В отличие от машинописного текста, в котором расстояние между словами более-менее постоянно, а интервалы между символами внутри слова гораздо меньше, чем интервалы между словами, в рукописном тексте размер интервалов между словами может варьироваться в очень широких пределах. Компоненты связности текста, отнесённые к одной строке на предыдущем этапе работы системы распознавания, объединяются в слова на этом этапе.

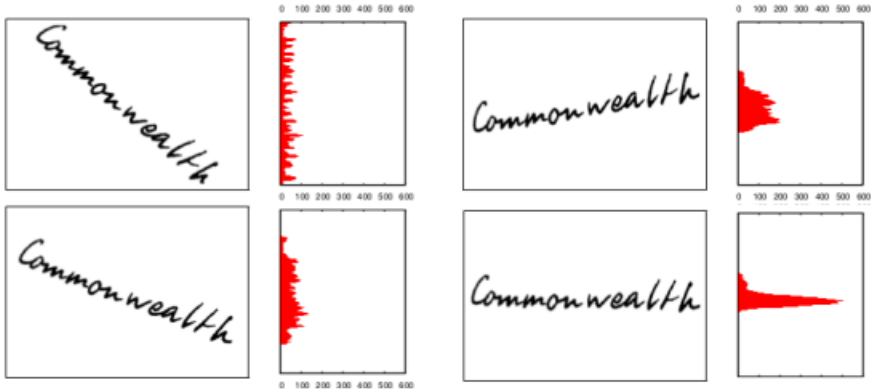

Рис. 1. Коррекция строк по горизонтали с использованием гистограммы профиля

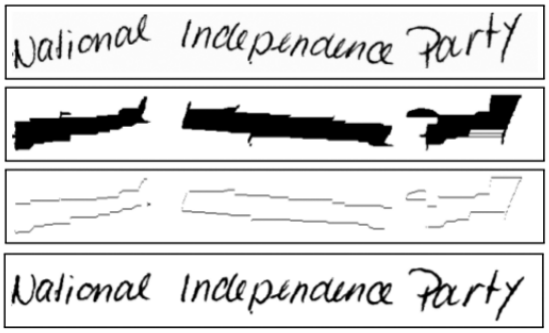

Рис. 2. Коррекция строк по горизонтали с использованием алгоритма линейной регрессии

4. Нормализация. В силу высокой вариативности начертания слов их распознавание является очень сложным процессом [6]. Нормализация служит для приведения слова к некоторому стандартному виду без значительной потери информации, необходимой для распознавания. Одними из наиболее часто используемых методов нормализации является метод коррекции наклона слова от горизонтальной и вертикальной линии [7]. Простейший метод коррекции по горизонтали состоит в выполнении максимизации его на некотором диапазоне (рис. 1). Существуют и другие методы нормализации, например, коррекция размера и выделение скелета текста, но они применяются реже.

Существуют и другие методы, например, основанные на сглаживании и линейной регрессии (рис. 2).

Список литературы

1. Предварительная обработка изображений – Национальная библиотека им. Н.Э. Баумана [Электронный ресурс] // Предварительная обработка изображений. – URL: https://ru.bmstu.wiki/Предварительная_обработка_изображений. (дата обращения: 02.03.2019).

2. Выделение областей интереса на основе классификации изолиний – тема научной статьи по медицине и здравоохранению читайте бесплатно текст научно-исследовательской работы в электронной библиотеке КиберЛенинка [Электронный ресурс] // Выделение областей интереса на основе классификации изолиний. – URL: https://cyberleninka.ru/article/n/vydelenie-oblastey-interesa-na-osnove-klassifikatsii-izoliniy. (дата обращения: 04.03.2019).

3. Прэтт У. Цифровая обработка изображений. Ч.2. – М.: Мир, 1982. – 480 с.

4. 868.pdf [Электронный ресурс] // Автоматическая сегментация текста с учетом его семантической структуры. – URL: https://openbooks.ifmo.ru/ru/file/868/868.pdf (дата обращения: 02.03.2019).

5. Сальников И.И. Поэлементный анализ растровых изображений. – Пенза: Приволжский Дом знаний, 2015. – 180 с.

6. Методы оффлайн-распознавания рукописного текста – pdf [Электронный ресурс] // Методы оффлайн-распознавания рукописного текста. – URL: https://docplayer.ru/25970063-Metody-offlayn-raspoznavaniya-rukopisnogo-teksta.html. (дата обращения: 03.03.2019).

7. GK.pdf [Электронный ресурс] // Методы оффлайн-распознавания рукописного текста. – URL: http://www.textolog-rgali.ru/userfiles/articles/article2/GK.pdf. (дата обращения: 05.03.2019).