Бартош Жигадло
Politechnika Wrocławska
Вроцлав, Польша
Ольга Романова
Одесский национальный политехнический университет
Одесса, Украина

ПОДХОДЫ К РЕШЕНИЮ ЗАДАЧИ СЕГМЕНТАЦИИ ТЕКСТА

Статья посвящена актуальному вопросу сегментации текстовых изображений. Автор делает анализ различных подходов к решению задачи сегментации, представленных в современной литературе.

Ключевые слова: *сегментация изображения, тесселяция, бинарная маска.*

В системах распознавания рукописных и машинописных текстов важным этапом в последовательности операций по распознаванию объектов является сегментация изображения. Сегментация изображения – это разделение изображения на области, однородные по некоторому критерию. Сегментация, при которой области разбиения не пересекаются, называется тесселяцией. Цель сегментации состоит в упрощении или изменении представления изображения, чтобы его было легче анализировать в дальнейшем. Результатом сегментации является множество сегментов, которые покрывают все изображение. Иначе говоря, каждый пиксель отмечен некоторой меткой некоторого класса. На сегодняшний день известно большое количество алгоритмов сегментации изображений, использующих разные признаки и подходы. Кроме того, при исследовании сегментации изображений возникает задача оценки качества на некоторой заранее определенной выборке. Роль сегментации изображений текста 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.

Задача сегментации текстовых изображений является очень актуальной. На текущий момент огромная часть информации хранится в электронном виде. Поиск и извлечение необходимых знаний происходят намного проще благодаря полуавтоматическим системам навигации по различным корпусам тестов, изображений и видео. Для навигации по рукописным документам нужно выделить объекты, присутствующие на изображении, такие как текст, иллюстрации, печатные вставки. Архивы старых документов, литературы или

искусства содержат огромное количество рукописей разных периодов и авторов. Поиск и навигация по архиву происходит вручную, что отнимает большое количество времени. У работников возникают разные сценарии использования архива. Например, найти все зарисовки в черновиках определенного автора. Для такого случая достаточно классификации на текст, изображения и фон. В более сложных случаях класс «текст» может подразделяться на прозу, поэзию, ремарки, исправления и т.д.

Множество подходов к решению задачи сегментации текста описано в литературе. Большинство алгоритмов основаны на предположении о горизонтальной ориентации страницы, а исходными данными служат печатные материалы.

В первую группу алгоритмов можно отнести методы, основанные на дроблении страницы на однородные прямоугольные блоки. Далее из блоков выделяются различные признаки и производится кластеризация выбранным алгоритмом. Например, в [1] документ разрезается на однородные прямоугольные блоки фиксированного размера, для каждого блока считается дискретное преобразование Фурье, затем блоки кластеризуются с помощью метода k-средних. На выходе алгоритм возвращает две бинарных маски: для текста и для изображений. В [5] к документам также предъявляется требование о горизонтальном расположении строк. Сначала происходит цветовая коррекция документа, затем разбиение на блоки, вычисление совместного распределения яркости для блока по двум координатам, кластеризация методом k-средних. В [6] на вход алгоритму подаются отсканированные документы, весь текст расположен горизонтально, допустимы изображения любых размеров. Происходит перевод цветного изображения в серое, разбиение на блоки, дискретное вейвлет-преобразование, излечение границ текста, удаление не текстовых регионов. В [7] для блоков вычисляется вейвлет-преобразование и на основе их обучается скрытая марковская модель. Проблема этой группы алгоритмов в строгом требовании к горизонтальной ориентации страниц. Во вторую группу алгоритмов относятся алгоритмы, основанные на анализе границ бинаризованных компонент. В [2] на вход подается документ, текст в котором расположен под одним и тем же углом. Производится определение ориентации текста с помощью проекции интенсивности изображения. Затем происходит размытие исходного документа, разделение на неоднородные блоки и обход в глубину по границам блоков, на основании которого происходит слияние блоков в различные классы. На выходе имеем бинарное изображение с границами текста в документе. В [3] после пороговой бинаризации изображения

происходит вертикальное и горизонтальное сглаживание, поиск границ, объединение разных компонент из эвристических соображений и разделение на компоненты. На выходе алгоритм возвращает 2 бинарных изображения: с текстом и с иллюстрациями. Основным результатом работы ныне существующих алгоритмов сегментации является бинарная маска, выделяющая текст на изображении.

Список использованной литературы

1. D. Sasirekha, Dr. Chandra, Enhanced Techniques for PDF Image Segmentation and Text Extraction. 2012.
2. A.Antonacopoulos, Page Segmentation Using the Description of the Background. 1998.
3. Priti P. Rege, Chanchal A. Chandrakar, Text-Image Separation in Document Images Using Boundary/Perimeter Detection. 2012.
4. C. Strouthopoulos, N. Papamarkovs, A.E. Atsalakis, Text extraction in complex color documents. 2001.
5. M-W Lin, J-R Tapamo, B Ndovie, A Texture-based Method for Document Segmentation and Classification. 2006.
6. Neha Gupta, V.K. Banga, Image Segmentation for Text Extraction. 2012.
7. Sunil Kumar, Rajat Gupta, Nitin Khanna, Student Member, IEEE, Santanu Chaudhury, and Shiv Dutt Joshi, Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model. 2007.