Сегментация рукописных строк в изображениях рукописных документов

A Preprint

Смирнова В.С.
ВМК МГУ
svictoriast@yandex.ru

Местецкий Леонид Моисеевич
ВМК МГУ
mestlm@mail.ru

Abstract

Понятие строки является ключевым при работе с электронными архивами сканированных текстовых документов как печатных, так и рукописных. В данной работе рассматривается задача сегментации строк в изображениях рукописных документов, необходимая для задача распознавания текста и навигации по большим массивам текстовых изображений. Цель сегментации — аппроксимировать линии текста кривой. Основная сложность заключается в том, что для рукописных документов (черновиков, дневников, записных книжек) недопустимо полагаться на предположения о структуре страниц, справедливых для печатных документов, таких как наличие межстрочных интервалов, параллельность строк и их единая ориентация. В результате работы предложен метод, обеспечивающий эффективную сегментацию строк в таких изображениях.

Keywords First keyword · Second keyword · More

1 Введение

Автоматизация обработки рукописных документов играет важную роль в современных задачах цифровизации исторических архивов и текстовых данных. Одной из ключевых задач в этом контексте является сегментация строк, которая позволяет структурировать текстовые данные для последующего анализа и поиска. Однако рукописные тексты существенно отличаются от печатных документов из-за отсутствия структурности страниц: строки могут быть наклонены, перекрывать друг друга или иметь непостоянные интервалы.

Существующие методы сегментации можно разделить на две основные категории. Методы первой категории [3], [4], [5] обрабатывают изображение текста целиком, предполагая, что текстовые фрагменты представляют собой непрерывные строки, где каждый пиксель принадлежит одной из них. Однако это предположение часто нарушается, например, на изображениях разворотов блокнотов или дневников, где текст на разных страницах не связан.

Методы второй категории [6], [7] используют преобразованные данные, что позволяет существенно снизить размерность задач и повысить устойчивость к особенностям рукописных текстов. Метод [7] использует обучение метрики расстояний между связными компонентами и последующую их кластеризацию с учетом данной метрики. В качестве объекта исследования выступали тексты, составленные из иероглифов, поэтому данный метод слабо подходит для сегментации документов на европейских языках. Метод [6] предполагает детектирование ориентации компонентов в строке по дискретной сетке (у каждой компоненты существует лишь конечное число направлений) и определение направления строки, исходя из направлений отдельных компонентов. Недостатком данного метода является очень малое (в работе их 5) число возможных ориентаций строки, поэтому он не подходит для работы с документами, наклоны строк в которых сильно варьируются.

В данной работе ставится задача разработки метода, который эффективно решает проблему сегментации строк для рукописных документов.

2 Постановка задачи

Исходными данными(рис. 3) задачи являются изображения рукописного текста, а выходными(рис. 4) — изображения с выделенными линиями строк, аппроксимированными квадратичными кривыми. Такой подход позволяет учитывать локальные изгибы и наклоны строк, что значительно повышает точность сегментации.

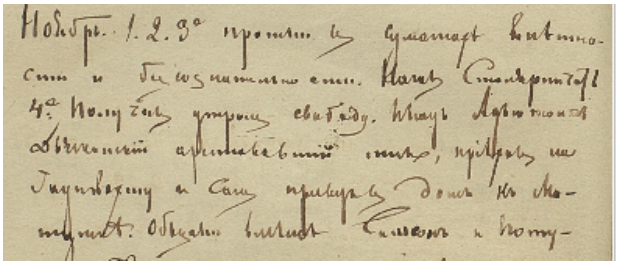

Рис. 1: Фрагмент исходных данных

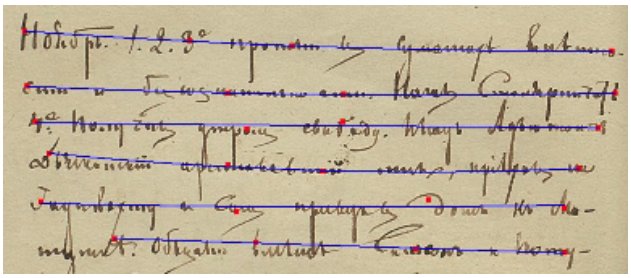

Рис. 2: Фрагмент выходных данных

Основная сложность задачи заключается в том, что строчные сегменты могут пересекаться, междустрочные интервалы часто выражены слабо или вовсе отсутствуют, а угол наклона строк может значительно варьироваться относительно документа. Для решения этой проблемы предложенный метод сегментации опирается на следующие априорные предположения о структуре документа: каждая строка обладает линейной структурой (с возможностью приближения линией), направления строк сохраняют локальную постоянность (соседние строки имеют схожую ориентацию), а угол наклона строк, поддающихся сегментации, ограничен диапазоном $\theta \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$.

3 Описание предложенного метода

3.1 Предварительная обработка изображений

Входные данные представляют собой изображения страниц архивных дневников XIX века, представленных в формате .jpg. В первую очередь, из фотографий выделяется страница текста, затем полученные изображения считываются в массив и к ним применяется алгоритм бинаризации для получения файла в формате .bmp. Таким образом, предварительная обработка текста состоит из следующих этапов(рис.??):

1. Медианное размытие для сглаживания изображения и удаления шума.
2. Так как фотографии имеют неравномерное освещение, то для бинаризации используется адаптивный метод порогового преобразования, вычисляющий порог для небольших областей изображения:

$$T(x, y) = \frac{1}{W} \sum_{(i,j)\in\mathcal{N}(x,y)} I(i, j) \cdot G(i, j) - C$$

Где:

- $T(x, y)$ — пороговое значение для пикселя в координатах (x, y).

- W — сумма всех весов в гауссовом окне.
- $\mathcal{N}(x, y)$ — окрестность пикселя (x, y), размером в 9.
- $I(i, j)$ — интенсивность (яркость) пикселя в координатах (i, j) в окрестности $\mathcal{N}(x, y)$.
- $G(i, j)$ — гауссов вес для пикселя в координатах (i, j) в окрестности $\mathcal{N}(x, y)$.
- $C = 6$ — константа, которая вычитается из средней взвешенной суммы.

3. Морфологические операции, помогающие улучшить форму и четкость объектов на изображении (дилатация, эрозия, открытие и закрытие с различными размерами ядер).
4. Извлечение контуров из бинаризованного изображения.

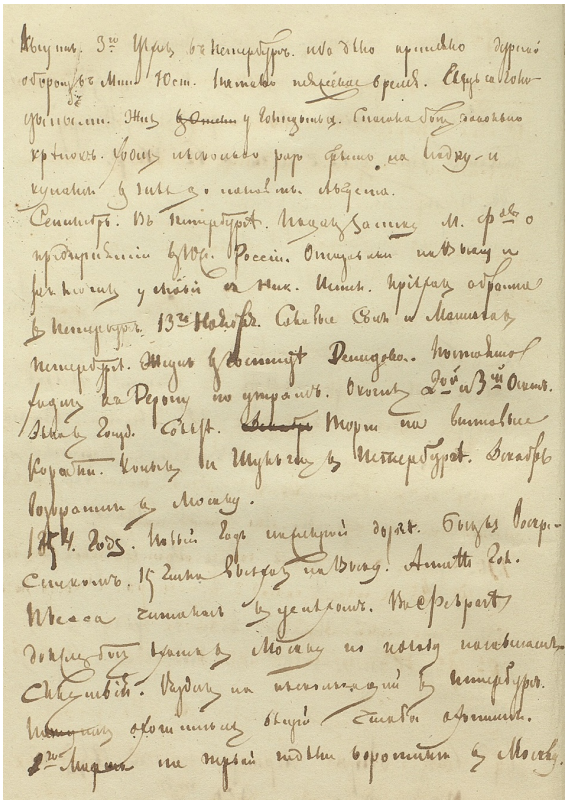

Рис. 3: Первоначальное изображение страницы.

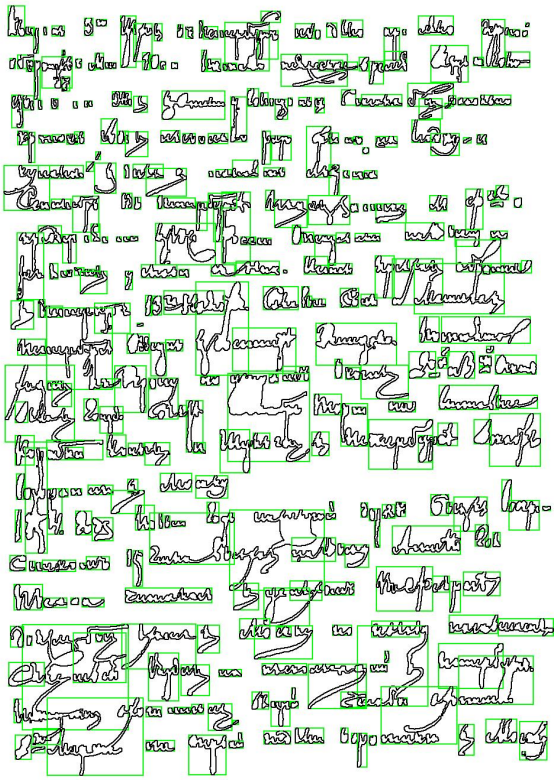

Рис. 4: Результат сегментации на связные компоненты.

3.2 Построение макета страницы

Первым этапом кластеризации компонент связности текста в строки является выделение основных направлений строк на основе системы соседства точек (X', Y'), являющихся центрами масс тяжести выделенных компонент связности. Центры масс для каждой компоненты рассчитываются по формулам:

$$X' = \frac{\sum_{i=0}^{Width} X_i}{No.\ of\ White\ pixels} \quad \forall X_i = 1$$

$$Y' = \frac{\sum_{j=0}^{Height} Y_j}{No.\ of\ White\ pixels} \quad \forall Y_j = 1$$

Здесь X_i и Y_j — координаты белых пикселей в бинаризированном изображении, а $No.\ of\ White\ pixels$ — общее количество белых пикселей в компоненте связности. Наиболее очевидной моделью для определения пространственной близости точечных объектов на плоскости является триангуляция Делоне.

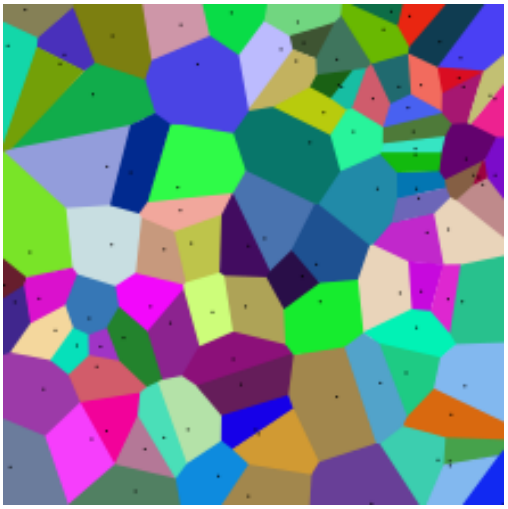

Рис. 5: Диаграмма Вороного для случайного множества точек (показаны чёрным). Разными цветами выделены области, все точки из которых ближе к соответствующему элементу множества точек.

Триангуляция Делоне однозначно соответствует диаграмме Вороного для того же множества точек. Диаграмма Вороного представляет собой разбиение плоскости, где каждая область соответствует одному элементу множества точек и содержит все точки, которые ближе к этому элементу, чем к любому другому. В качестве метрики используется евклидово расстояние.

Триангуляция Делоне строит граф, где рёбра соединяют только те точки, которые являются соседями в диаграмме Вороного. Таким образом, она служит системой соседства для компонент связности.

Так как близкие друг к другу компоненты внутри строки имеют схожую ориентацию, строкам соответствуют горизонтальные рёбра или рёбра с небольшим наклоном (до 10°) в триангуляции Делоне, построенной на точках (X', Y') центров масс каждой компоненты связности страницы(рис. 6).

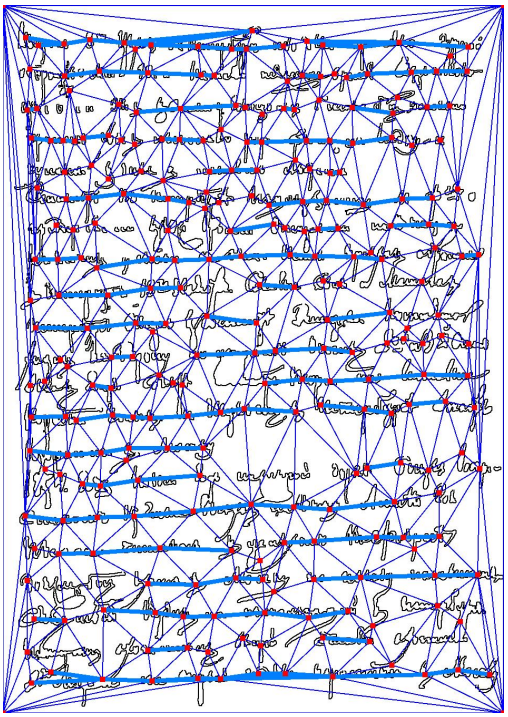

Рис. 6: Выделенные ребра триангуляции Делоне, обозначающие основные направления линий строк

3.3 Пространственная кластеризация

Список литературы