
Сегментация 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 задачи 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 Сегментация строк · Рукописные документы

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 [1], [2], [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], [8] используют преобразованные данные, что позволяет существенно снизить размерность задач и повысить устойчивость к особенностям рукописных текстов. Метод [7] использует обучение метрики расстояний между связанными 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.

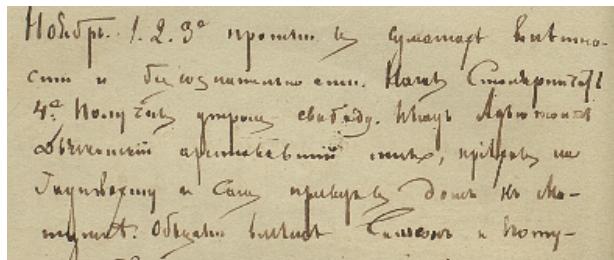


Рис. 1: Фрагмент исходных данных

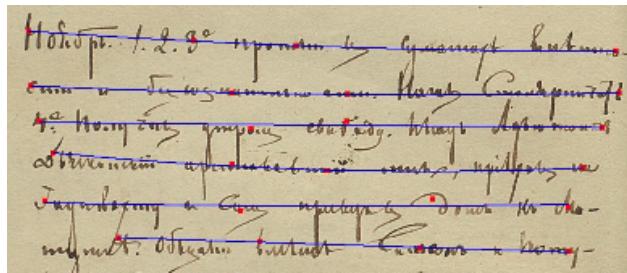


Рис. 2: Фрагмент выходных данных

Основная сложность задачи заключается в том, что строчные сегменты могут пересекаться, межстрочные 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 постоянство (соседние строки имеют схожую ориентацию)
- близкие компоненты внутри одной строки имеют схожую ориентацию
- угол наклона строк, поддающихся сегментации, ограничен диапазоном $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$

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 считаются в массив и к ним применяется алгоритм бинаризации для получения возможности последующего анализа. Таким образом, предварительная обработка текста состоит из следующих этапов:

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 изоб-

ражения:

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

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.

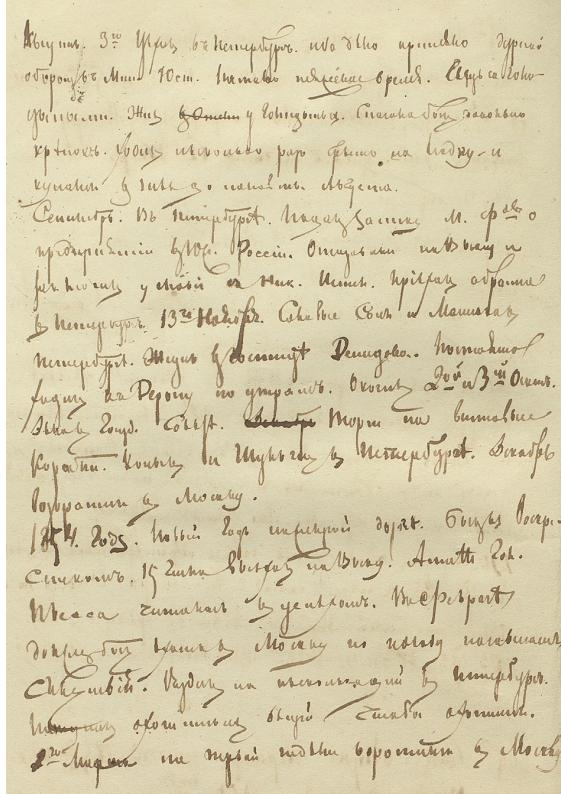


Рис. 3: Первоначальное изображение страницы.

3.2 Кластеризация

В данной задачи под строками подразумевается кластеризованный набор компонент

$$\left\{ \bigcup_{k:a(C_k)=l} C_k \right\}_{l=1}^L,$$

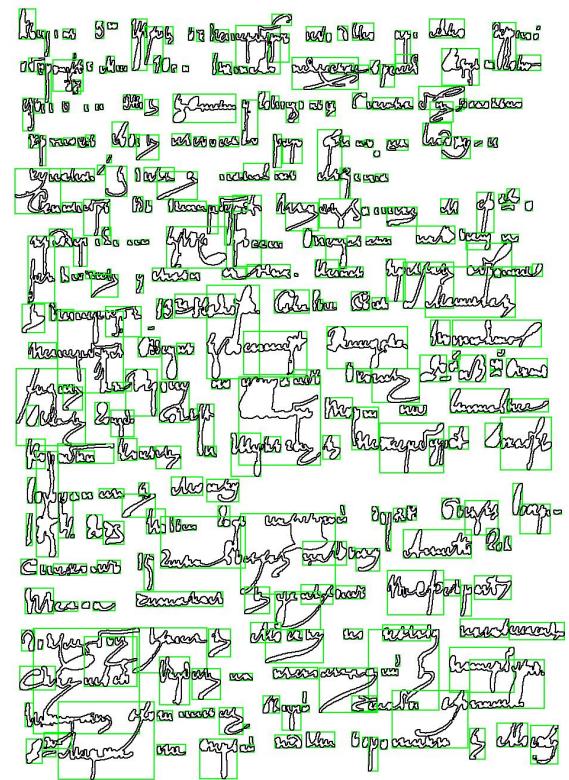


Рис. 4: Результат сегментации на связные компоненты.

где L — число кластеров (строк), а — алгоритм кластеризации, $C_k = (V_k, E_k)$ — связанные компоненты графа (V , E), полученные после бинаризации изображения и выделения контуров. Каждый кластер соответствует отдельной строке. Обозначим кластер, соответствующий строке l , как C_l , тогда по каждому кластеру

$$C_l = \left\{ \bigcup_{k:a(C_k)=l} C_k \right\}$$

можно однозначно построить соответствующую ему сегментацию исходного изображения. Тем самым задача сегментации строк в изображении однозначно сводится к задаче кластеризации компонент C_k .

3.2.1 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:

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

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

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 [9].

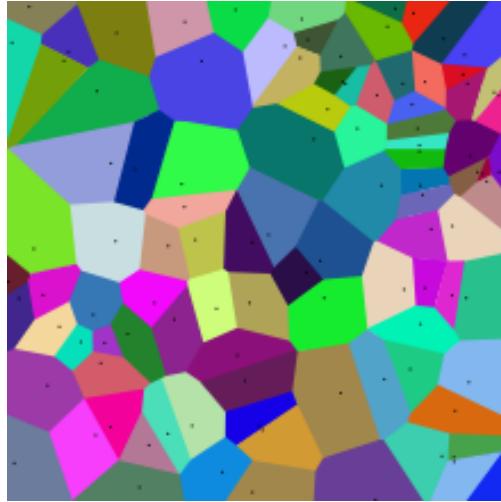


Рис. 5: Диаграмма Вороного для случайного множества точек (показаны чёрным). Разными цветами выделены области, все точки из которых ближе к соответствующему элементу множества точек.

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 ребра или ребра с небольшим наклоном (до 10°) в триангуляции Делоне, построенной на точках (X', Y') центров масс каждой компоненты связности страницы (рис. 6).

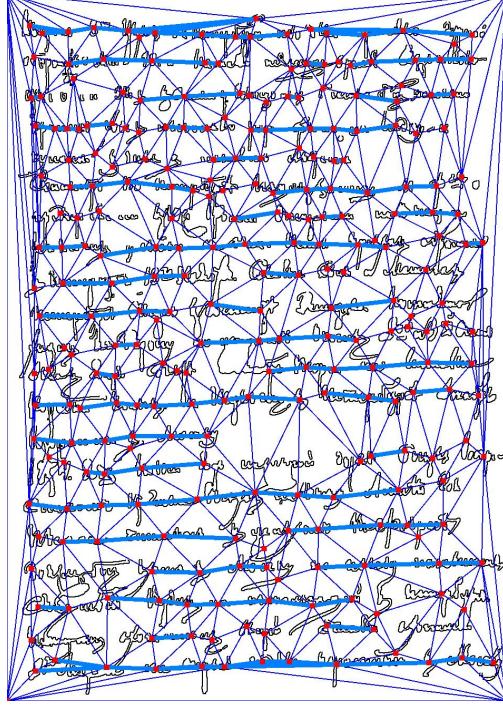


Рис. 6: Выделенные ребра триангуляции Делоне, обозначающие основные направления линий строк

3.2.2 Пространственная кластеризация

Для выполнения пространственной кластеризации мы используем упорядочивание ребер графа, построенного на основе бинаризованного изображения и выделенных контуров. Пусть на предыдущем этапе были выделены ребра триангуляции $E = \{e_1, e_2, \dots, e_n\}$. Алгоритм кластеризации состоит из следующих шагов:

1. Сортировка ребер На первом этапе ребра графа упорядочиваются по возрастанию координаты y , то есть:

$$E' = \{e_{(1)}, e_{(2)}, \dots, e_{(n)}\}, \quad \text{где } y_c(e_{(1)}) \leq y_c(e_{(2)}) \leq \dots \leq y_c(e_{(n)}),$$

где $y_c(e)$ — координата y центра ребра e .

2. Вычисление разницы высот Для каждой пары соседних ребер из упорядоченного множества E' вычисляется разница высот:

$$\Delta_i = y_c(e_{(i+1)}) - y_c(e_{(i)}), \quad \forall i \in \{1, 2, \dots, n-1\}.$$

3. Объединение ребер в группы Рёбра объединяются в группы $G_j, j \in \{1, 2, \dots, k\}$ по следующему критерию:

$$\Delta_i \leq P, \quad \forall e_{(i)}, e_{(i+1)} \in G_j,$$

где P — порог подобранный относительно распределения разниц высот Δ_i , описанный ниже.

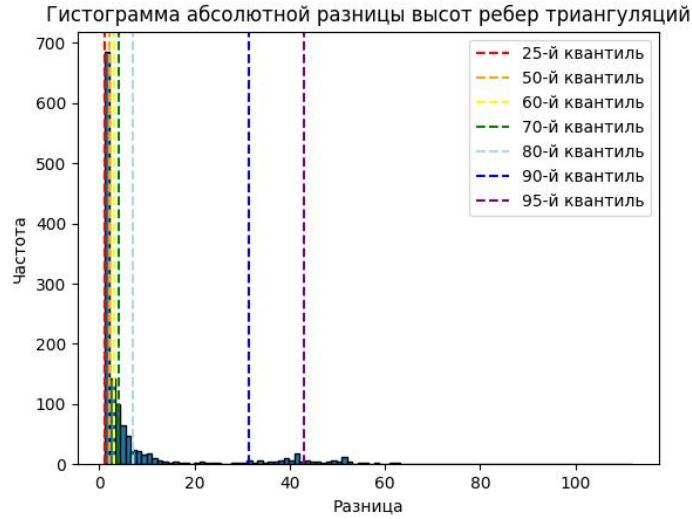


Рис. 7: Гистограмма абсолютной разницы высот рёбер триангуляции. На графике выделены значения различных квантилей, используемые для выбора порога кластеризации.

Для задачи оптимальный порог Р выбран равным 90-му перцентилю, так как видно из графика 7: в диапазоне с 25-го по 80-й перцентиль абсолютные разницы высот Δ_i остаются низкими и стабильными, что соответствует плотному расположению компонент внутри строк. Однако значения, превышающие 80-й перцентиль, начинают значительно увеличиваться, а разрыв между 80-м и 90-м перцентилями оказывается существенным, что указывает на появление выраженных вертикальных промежутков между сегментированными элементами. Это делает выбор 90-го перцентиля оптимальным, поскольку он корректно кластеризует строки, исключая объединение двух ближайших строк.

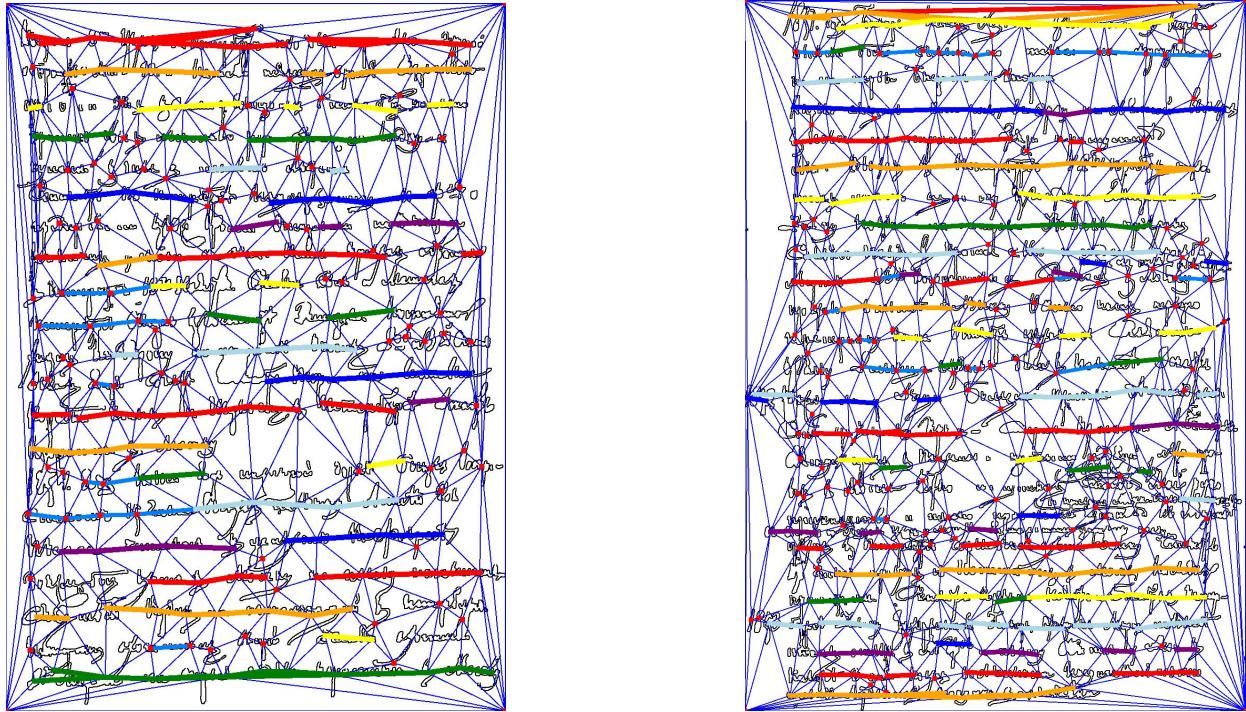


Рис. 8: Результаты первичной кластеризации.

В результате, удается выявить основную структуру страниц, учитывая различные межстрочные интервалы, однако в итоге могут выделяться излишние строки, которые объединяются при вторичной кластеризации.

3.2.3 Вторичная кластеризация

Повторная кластеризация проводится для уточнения границ строк текста на основе результатов предыдущего этапа. Множество групп рёбер после первичной кластеризации задаётся как $G = \{G_1, G_2, \dots, G_k\}$, где каждая группа G_i представляет кластер рёбер, соответствующий строке. Для каждой группы рассчитывается центр строки $C(G_i)$ как среднее значение координат y_c рёбер, входящих в неё.

Далее для соседних групп G_i и G_{i+1} вычисляется разница центров $\Delta_i = C(G_{i+1}) - C(G_i)$. Полученные значения Δ_i используются для определения необходимости объединения строк. Если Δ_i меньше выбранного порога P строки объединяются, иначе формируется новая строка. Таким образом, мелкие вертикальные разрывы внутри одной строки устраняются, а крупные разрывы между строками сохраняются.

Для выбора порога P аналогично предыдущему пункту рассматривалось распределение абсолютной разницы высот строк (Δ_i).

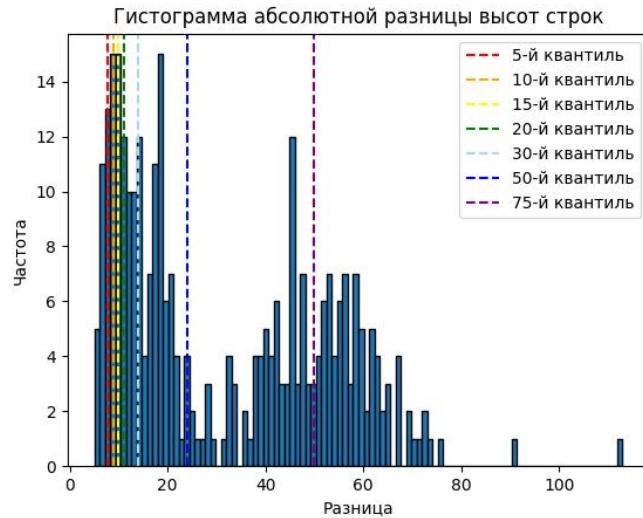


Рис. 9: Гистограмма абсолютной разницы высот строк. На графике выделены значения различных квантилей, используемые для выбора порога кластеризации.

Соответствующая гистограмма 9 демонстрирует, что низкие значения Δ_i (до 30-го перцентиля P_{30}) соответствуют небольшим разрывам внутри одной строки, отвечающим ошибкам первичной кластеризации из-за наклона текста. После P_{30} значения начинают резко увеличиваться, отражая крупные разрывы между строками. До 30-го перцентиля плотность распределения высока, указывая на компактное расположение сегментов внутри строк, тогда как после 50-го перцентиля частота значений заметно падает, что свидетельствует о межстрочных промежутках. Выбор P_{30} как порога является оптимальным, так как он позволяет объединить близко расположенные сегменты в пределах одной строки, сохраняя разделение между строками.

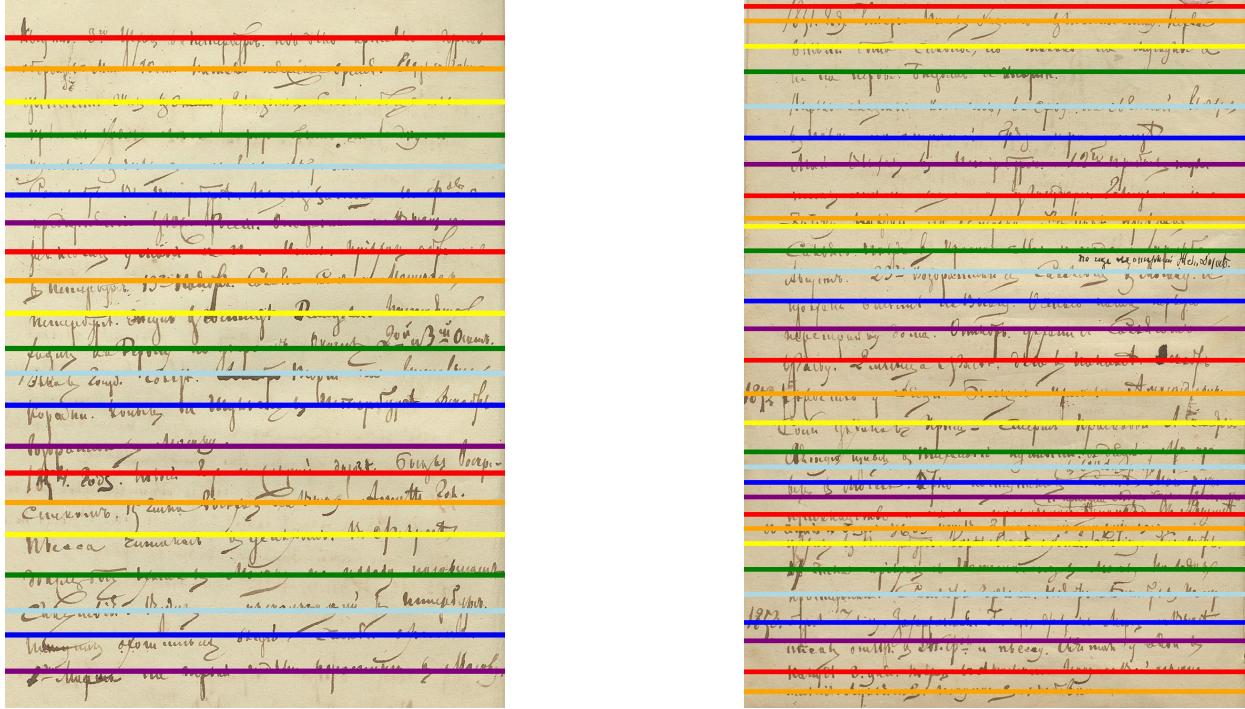


Рис. 10: Результаты кластеризации.

3.3 Сегментация строк

Для аппроксимации строки, представленной кластером $G_i = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, выбирается подмножество ключевых точек, равномерно распределённых по длине строки. Этот выбор направлен на обеспечение устойчивости аппроксимации и корректного описания геометрии строки. Выбор точек осуществляется следующим образом:

1. Крайние точки: Выбираются крайние левые и правые точки строки, соответствующие минимальному и максимальному значениям координаты x : Левая точка: (x_{\min}, y_{\min}) , Правая точка: (x_{\max}, y_{\max}) .
2. Квантильные точки: Для равномерного распределения точек вдоль строки используются значения 25-го и 75-го процентилей по координате x . Эти процентили определяют x -координаты точек, которые характеризуют центральные части строки
3. Если в строке недостаточно уникальных точек для определения всех четырёх ключевых точек (например, строка состоит из двух или трёх точек), алгоритм адаптируется. В этом случае:

- При наличии двух точек: добавляется средняя точка между крайними левыми и правыми точками:

$$(x_{\text{middle}}, y_{\text{middle}}) = \frac{(x_{\min}, y_{\min}) + (x_{\max}, y_{\max})}{2}.$$

- При наличии трёх точек: используется медианная точка по координате x .

Такой подход гарантирует, что аппроксимация строк корректно описывает линию их прочтения, учитывая искривления её компонент, вызванные наклоном текста, локальными дефектами сегментации и разброс в расположении символов.

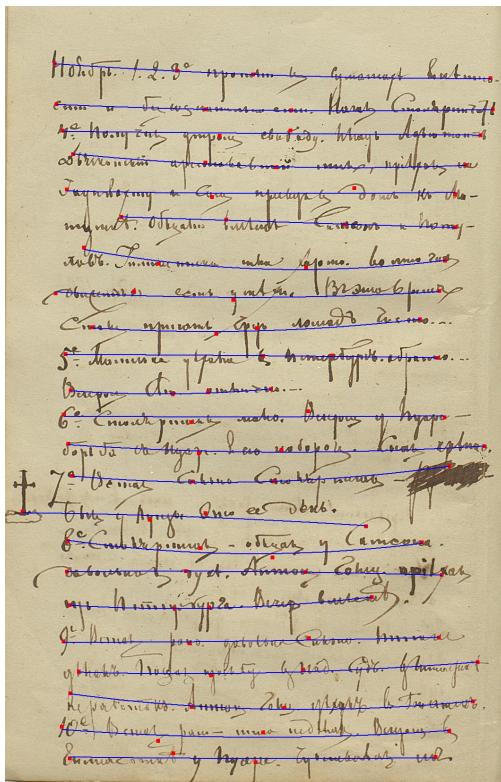


Рис. 11: Страница с сегментированными строками

3.4 Эксперименты

Качество работы метода проверялось по набору изображений 60 документов из архива дневников одного автора. Изображения из этого набора были использованы в качестве иллюстраций в данной работе.

Метрика качества была подсчитана посредством анализа сегментированных изображений. За TP обозначается число сегментированных строчных фрагментов, которые действительно являются строками, за FP — количество фрагментов, которые распознаны как строки, но ими не являются, а за FN — нераспознанные строки. Тогда можно ввести две метрики:

$$\text{recall} = \frac{TP}{TP + FN} \quad \text{и} \quad \text{precision} = \frac{TP}{TP + FP}.$$

Первая метрика показывает то, насколько предложенный метод вообще способен детектировать строки, а вторая — насколько точно он это делает.

Запишем значения, полученные для данных метрик, в таблицу: (см. рис. 20).

TP	FP	FN	precision	recall
1208	32	24	0.97	0.98

Таблица 1: Значения метрик, подсчитанных на блоке изображений. Всего в изображениях было найдено 1232 строки, из них корректно были определены 1208. 32 строк были найдены ошибочно, 24 не выделены.

Предложенный метод обладает большой чувствительностью (recall), и точностью (precision). Большинство ошибок, которые входят в число FN , связаны с тем, что метод распознаёт две подряд идущие строки как одну, а ошибки FP связаны с излишней сегментацией строк, то есть выделения шума или многократным выделением одной и той же строки.



Рис. 12: Основные примеры ошибок работы предложенного метода.

4 Заключение

В данной работе был предложен метод сегментации строк в изображениях рукописных документов. Метод ориентирован на сложные случаи анализа исторических рукописей, таких как архивные дневники XIX века, которые характеризуются неоднородным освещением, отсутствием строгой структуры страниц и вариативностью почерка.

Основными этапами метода являются предварительная обработка изображений для выделения контуров и связных компонент. На основе связных компонент была реализована кластеризация объектов в строки на основе триангуляции Делоне на странице и пространственного анализа ребер получившегося графа. Метод обеспечивает устойчивую аппроксимацию линий строк, учитывая искривления, наклон текста и локальные дефекты сегментации.

Эксперименты, проведенные на архивных рукописях, показали высокую точность работы метода, демонстрируя показатели precision 0.97 и recall 0.98. Это подтверждает его эффективность в задаче выделения строк текста в сложных условиях.

Несмотря на полученные высокие результаты, метод может быть усовершенствован за счет внедрения более точной кластеризации, основанной на анализе не только статистических данных о пространственном расположении объектов, но и их текстурных, геометрических и контекстных характеристик. Это позволит повысить точность сегментации, особенно в сложных случаях, таких как близкорасположенные или пересекающиеся строки.

Список литературы

- [1] A. Sanchez et al. Text line segmentation in images of handwritten historical documents. In First Workshops on Image Processing Theory, Tools and Applications, 2008.
- [2] A. Nicolaou et al. Handwritten text line segmentation by shredding text into its lines. In International Conference on Document Analysis and Recognition, 2009.
- [3] Alireza Alaei et al. A new scheme for unconstrained handwritten text-line segmentation. Pattern Recognition, 44(4):917–928, 2011.
- [4] А.А. Масалович and Л.М. Местецкий. Численные методы детектирования и исправления геометрических искажений в изображениях текстовых документов. Dissertation, 2010. URL <https://istina.msu.ru/dissertations/4704243/>.
- [5] Nilar Phy Wai and Nu War. Text line segmentation on myanmar handwritten documents using directional gaussian filter. In IEEE Xplore, 2024. URL <https://ieeexplore.ieee.org/abstract/document/10533060/authors#authors>.
- [6] Jayant Kumar et al. Handwritten arabic text line segmentation using affinity propagation. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pages 135–142, 2010.
- [7] Fei Yin et al. Handwritten text line segmentation by clustering with distance metric learning. In Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, 2008.
- [8] G. Renton, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet. Handwritten text line segmentation using fully convolutional network. In International Conference on Document Analysis and Recognition, 2017. URL <https://ieeexplore.ieee.org/document/8270267>.
- [9] Леонид Местецкий. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. Физматлит, 2009.