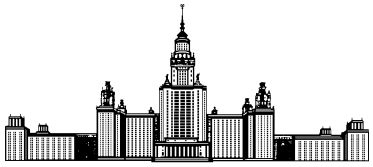


Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Захаров Егор Олегович

Сегментация текстовых блоков в изображениях рукописных документов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д.т.н., профессор
Местецкий Леонид Моисеевич

Москва, 2016

Содержание

Аннотация

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, которая возникает при организации навигации по большим массивам изображений текста. Задача сегментации строк состоит в нарезке изображений текста на фрагменты, включающие ровно одну текстовую строку. Сложность этой задачи определяется тем, что в рукописных документах (черновиках, дневниках, записных книжках) мы не можем опираться на предположения о структуре строк, справедливые для печатных документов: например, об обязательном наличии междустрочных интервалов, о параллельности строк и единой их ориентации на странице. В случае рукописных документов эти предположения либо не выполняются, либо выполняются лишь частично. Результатом данной работы является метод, позволяющий проводить эффективную сегментацию строк в данных изображениях.

1 Введение

1.1 Постановка задачи

В данной работе рассмотрена задача сегментации фрагментов строк в изображениях рукописных документов, которая является частью более общей проблемы организации навигации по большим массивам (архивам) с текстами.

Исходными данными задачи являются изображения, выходными данными также являются изображения, на которых выделены области, относящиеся к разным строкам. Данные области могут между собой пересекаться, но каждая должна целиком содержать одну текстовую строку. Пример подобной разметки можно увидеть на рис. 1.

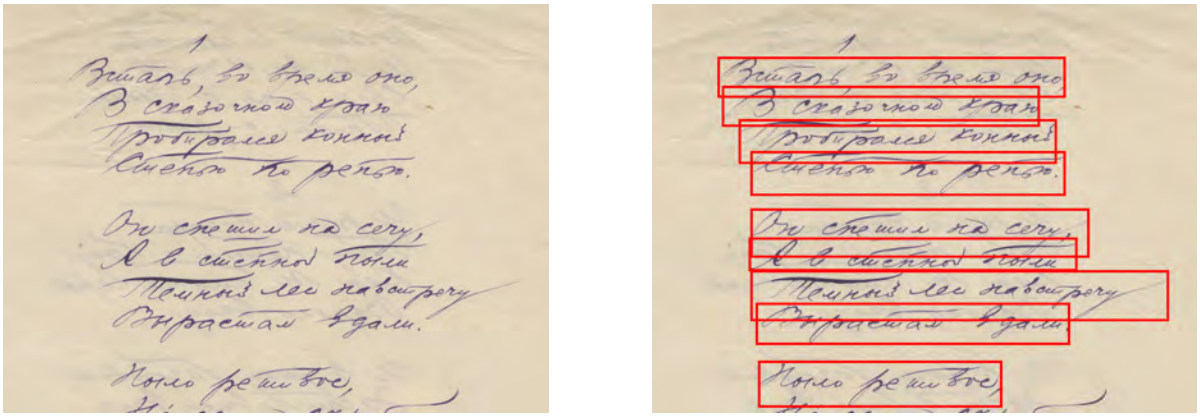


Рис. 1: Пример исходного изображения и изображения с размеченными строками

Основная сложность задачи заключается в том, что строчные сегменты имеют значительные области самопересечения, междустрочные интервалы могут быть слабо выражены, а также иметь разный угол наклона относительно документа. Предложенный метод сегментации решает поставленную задачу, основываясь на следующих априорных предположениях о структуре документов: каждая строка имеет линейную структуру (может быть хорошо приближена линией), локальное постоянство направлений строки (предполагается, что в соседних строках оно одинаково) и ограничение на угол направлений строк, поддающихся сегментации: $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$.

1.2 Обзор существующих методов

Все существующие методы делятся на две категории: методы, которые работют непосредственно с изображением: [3], [4], [5], и методы, которые работают с данными уменьшенной размерности: [6], [7].

В основе методов, описанных в [3], [4] и [5], лежит предположение о том, что всё изображение документа является непрерывным сегментом текста, то есть каждый пиксель исходного изображения обязательно относится к одной из строк. Данное предположение нарушается, например, на типичных изображениях разворота дневников, где существуют два отчётливых фрагмента текста (написанные на отдельных страницах), и строки из одной его части не должны содержать фрагментов строк из другой.

Методы [6] и [7] предполагают работу со связными компонентами бинарного изображения исходного текста и сведение задачи сегментации изображения к задаче кластеризации связных компонент по строкам, к которым они относятся. Метод [7] использует обучение метрики расстояний между связными компонентами и последующую их кластеризацию с учётом 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 компонент в строке по дискретной сетке (у каждой компоненты существует лишь конечное число направлений), и определение направления строки, исходя из направлений отдельных компонент. Недостатком данного метода является очень малое (в работе их 5) число возможных ориентаций строки, поэтому он неспособен работать с документами, наклоны строк в котором сильно варьируются.

2 Предложенный метод

Предлагаемый подход к решению связан с построением геометрического скелета для изображения текста, который представлен, как объект одного цвета на фоне другого цвета (бинарным изображением). Скелетом (или срединной осью) такого

Рис. 2: Схема предложенного метода

изображения мы будем называть множество центров максимальных вписанных кругов части изображения, относящейся к тексту. Это множество представляет собой планарный граф [2]. Задача сегментации строк рукописного текста сводится к выделению подграфов в полученном скелете, при этом подграфы должны быть: а) непересекающимися, б) каждый подграф соответствует одной строке. Данную задачу можно рассматривать как задачу кластеризации.

Рассмотрим схему кластеризации изображения документа, состоящую из следующих шести этапов: см. рис. 2. Идея данного метода основывается на нескольких наблюдениях.

1. Исключительно локальные методы сегментации строк работают плохо, потому что не учитывают глобальной структуры документа (наличия разрозненных абзацев текста и постоянства ориентации строк внутри абзацев).

2. Бинаризация с допустимым уровнем ошибок несёт в себе достаточно информации для качественной сегментации строк вне зависимости от них ориентации.

3. Построение скелета бинарного изображения не приводит к потере информации.

Данные предположения будут более подробно разобраны в соответствующих разделах, но являются предпосылками к созданию и отчасти обоснованием данного метода.

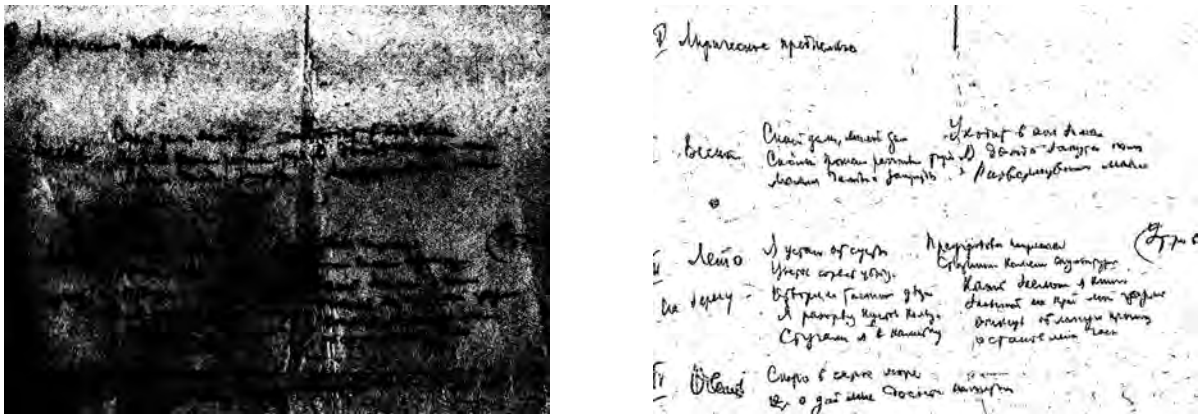


Рис. 3: Сравнение метода бинаризации Отсу (слева) и используемого метода (справа)

2.1 Бинаризация

2.1.1 Сравнение современных методов бинаризации

Работа большинства методов бинаризации основана на использовании порогов. В связи с этим их можно разделить на два класса: локальные и глобальные. К чисто глобальным методам относится бинаризация Отсу [8], к число локальным – метод Ниблака [9].

Глобальные методы хорошо подходят для бинаризации изображений, в которых пиксели фона могут быть отделены от пикселей текста, используя порог по гистограмме всего изображения. В случаях, когда яркости отличаются слабо, данный подход может привести к очень плохим результатам (рис. 3). Локальные методы в подобных ситуациях работают лучше, но приводят к большим шумам, потому что существенно зависят от выбора ширины окна, внутри которого устанавливается порог по яркости для каждого пикселя.

Современные методы пороговой бинаризации используют комбинации данных двух подходов, добавляя различные элементы пред- и пост-обработки получаемых разметок на текст и фон [10].

По результатам анализа выборки методов, предложенных на Document Image Binarization Contest (DIBCO) (проводился в рамках конференции по анализу текстов International Conference of Document Analysis and Recognition, ICDAR), были выбраны методы [11], [12], [13], [14], [15], показавшие наилучшие результаты.

Метод	F-score, %	Сложность
[11]	89.93	$o(1)$
[12]	92.03	$o(1)$
[13]	94.34	$o(n)$
[14]	90.03	$o(n)$
[15]	90.06	$o(n)$

Таблица 1: Сравнение методов бинаризации

Критерием лучшего метода является совокупная оценка качества его работы, а также скорости и возможности применения параллельных вычислений. При этом требования по скорости являются значительно более важными из-за больших объемов архивных данных, подлежащих разметке.

В качестве оценки качества работы берётся F-score, измеренная по выборке DIBCO-2010 (значения взяты из соответствующих статей):

$$F = \frac{2 * recall * precision}{recall + precision} \tag{1}$$

где $recall$ – отношение числа правильно выделенных текстовых пикселей к числу всех истинных текстовых пикселей, а $precision$ – отношение правильно выделенных текстовых пикселей к числу всех размеченных текстовых пикселей. Данные метрики измеряются по всему массиву текстов.

В качестве оценки скорости работы методов использовалась временная сложность алгоритма (с учётом возможности параллельного выполнения по отдельным пикселям). Приемлемой является сложность ниже линейной, т.к. архивные документы имеют размерность $n \sim 10^7$, и её уменьшение в общем случае приводит к потере информации.

Сравнение методов приведено в таблице 1, по его результатам наилучшим методом является метод [12], который и использован в дальнейшем. Приведём его краткое описание.

2.1.2 Бинаризация при помощи локального контраста

Было показано [12], что для анализа изображений сложной структуры, в которых текст не может быть качественно сегментирован глобальными пороговыми методами, хорошо подходит комбинация глобальных и локальных методов. Как было сказано ранее, использование локальных методов ограничено произволом в выборе ширины окна, внутри которого подбирается локальный порог. Поэтому идея предложенного метода заключается в том, что сначала глобальными методами находятся пиксели, гарантированно принадлежащие тексту, по ним оценивается средняя ширина рукописного штриха, а потом локальным методом находится более точная разметка изображения.

Ключевым в данном методе является использование нелинейной локальной контрастности изображения:

$$C(i,j) = I_{max}(i,j) - I_{min}(i,j) \tag{2}$$

где $C(i,j)$ – контрастность пикселя (i,j), $I_{max}(i,j)$ и $I_{min}(i,j)$ – максимальная и минимальная яркости в некоторой окрестности данного пикселя.

В [12] предложенно и обоснованно использование взвешенной и нормированной нелинейной контрастности:

$$C(i,j) = \alpha\frac{I_{max}(i,j) - I_{min}(i,j)}{I_{max}(i,j) + I_{min}(i,j) + \epsilon} + (1-\alpha)(I_{max}(i,j) - I_{min}(i,j)) \tag{3}$$

Первое слагаемое измеряет контрастность в изображениях с сильной дисперсией яркости пикселей, второе – с низкой. Коэффициент α предложено выбирать пропорционально этой дисперсии.

Пример получаемой карты контрастностей показан на рис. 4. Как видно, локальный контраст является одним из способов детектирования границ на изображении, и используется вместе с детектором Кэнни [16] для оценки ширины штриха и последующего применения одной из вариаций метода локального порога Ниблака (детали реализации описаны в соответствующей статье)

Существенным преимуществом данного метода является то, что он состоит из конечного числа шагов бинаризации и пост-обработки, внутри которых возможно

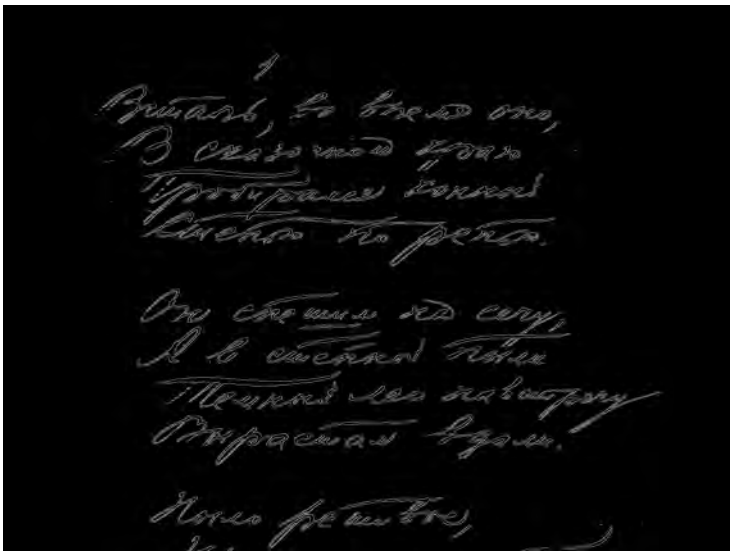


Рис. 4: Высококонтрастные пиксели изображения

параллельная обработка всех пикселей изображения, что даёт возможность эффективной реализации данного метода на графических процессорах, в отличие от методов [13], [14] и [15], где существенен последовательный обход пикселей изображения в строго заданном порядке.

2.2 Построение скелета

Для последующего анализа бинарного изображения используется геометрический скелет его части, размеченной на этапе бинаризации, как текст. Скелетом (или срединной осью) множества пикселей изображения мы будем называть множество центров максимальных вписанных в него кругов. Для растровой картинки это множество представляет собой также набор пикселей. В случае граничного представления бинарной картинки при помощи многогранника (в этом случае связные области текста приближаются непрерывными многоугольниками, уравнения границ которых задаются аналитически [2]), это множество представляет собой планарный граф.

В работе использована быстрая реализация построения геометрического скелета по принципу, описанному в данной статье. Ключевой особенностью данной структуры является то, что скелет полностью сохраняет информацию о форме много-

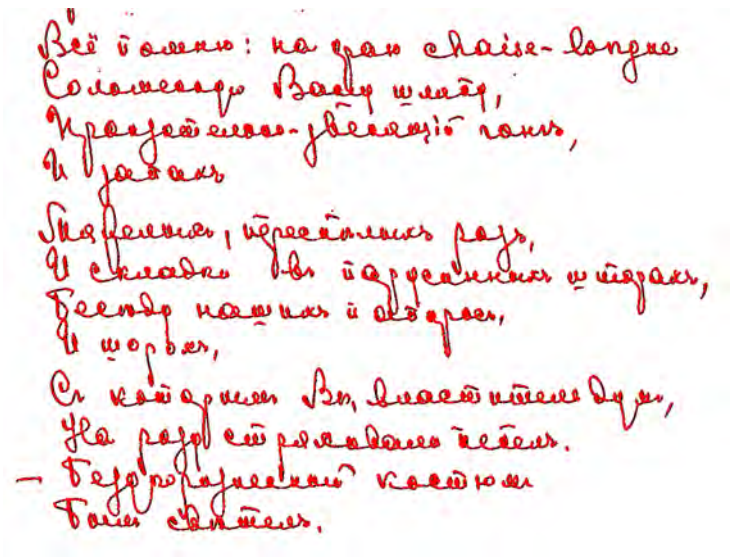


Рис. 5: Визуализация графа скелета для бинарного изображения

угольника, по которому он построен. Скелетизация многоугольника и построение многоугольника по скелету – обратимые преобразования [2].

В общем случае скелетом многоугольной фигуры является граф (V, E), где V – вершины скелета, т.е. центры максимальных вписанных в фигуру кругов, а E – кривые Безье первого и второго порядков [1], т.е. линии и параболы.

Последующий анализ изображения проводится с использованием графа (V, E), в котором для каждой вершины $v_i \in V$ задан вес, равный радиусу максимальной вписанной в фигуру окружности r_i. Визуализация данного графа для текста выглядит следующим образом: рис. 5.

2.3 Фильтрация скелета

Выделим в графе (V, E) связные компоненты: они соответствуют связным компонентам исходного бинарного изображения. Каждая компонента соответствует либо фрагменту слова на изображении, либо слову целиком, либо нескольким словам одновременно, либо является шумовым артефактом бинаризации (рис. 6).

На данном этапе предлагается произвести фильтрацию связных компонент скелета изображения для удаления шумов, оставшихся после бинаризации. Все ошибки бинаризации разделим на три рода:

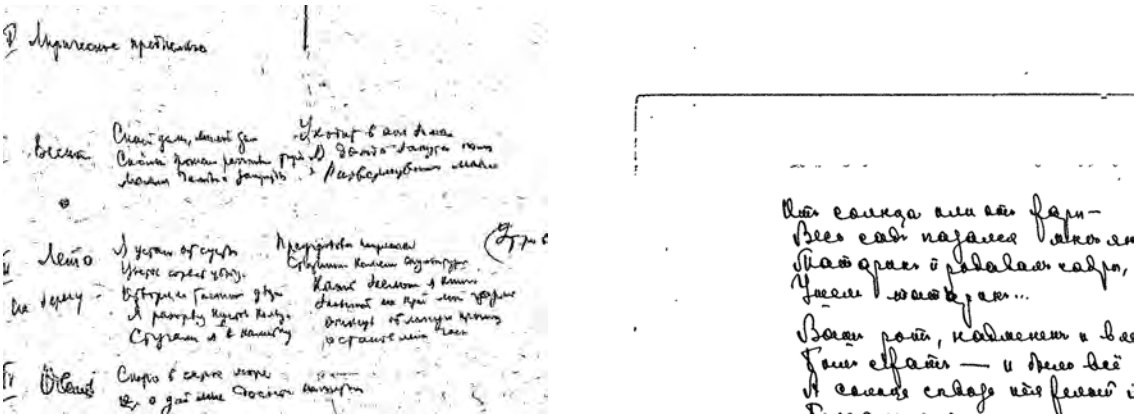


Рис. 6: Примеры бинаризации

1. Компоненты, отвечающие границам листа с документом и всему, что находится за границами.

2. Малые по площади компоненты внутри документа, которые отвечают реальным ошибкам бинаризации.

3. Компоненты, которые появляются вследствие наличия на листе с документом текстуры, неотличимой от текста по яркости (например, линованная бумага).

Ошибки первого и третьего рода не являются ошибками используемого метода бинаризации, т.к. первые появляются вследствие естественного различия между фоном докумена и фоном, на котором находится документ, а другие вследствие особенностей конкретной задачи. И локальные и глобальные методы бинаризации будут детектировать разность в яркости (или контрастности) на границе или в текстуре, и некорректно отнесут её к тексту (рис. 6).

При помощи скелетов возможно эффективное устранение всех трёх видов ошибок, при этом ошибки первого и третьего рода действительно нельзя исправить никак без привлечения априорной информации о структуре изображения документа.

2.3.1 Исправление ошибок бинаризации второго рода

Добавим в процедуру построения скелета порог на минимальную площадь связной компонены S_{min}, которую мы будем считать не шумовой. Площадью компоненты S бинарного изображения естественным образом будем считать число пикселей в ней.

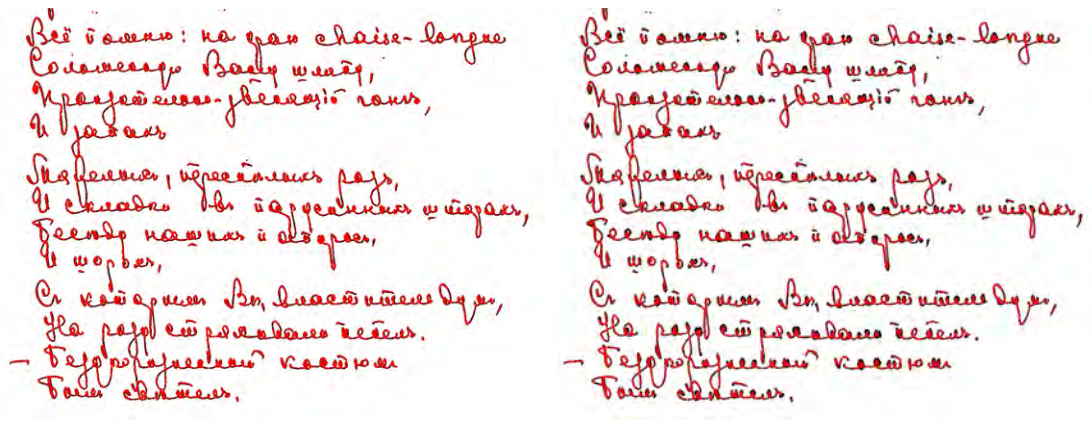


Рис. 7: Сравнение скелета (слева) и стриженного скелета с коэффициентом стрижки $n = 3$ (справа)

Также добавим ещё одну процедуру пост-обработки скелета, которая называется стрижкой (pruning) [2]. Основывается она на том факте, что граф скелета является деревом [2]:

1. Рассмотрим все ветви графа скелета (последовательности рёбер) длины n, которые заканчиваются на листовых вершинах.

2. По очереди удалим каждую такую последовательность, восстановим по новому скелету контур фигуры.

3. Если площади фигур отличаются на величину, меньшую ϵ, то удалим данную ветвь.

В результате мы получим регуляризованный скелет (V_{pruned}, E_{pruned}), который наилучшим образом описывает форму исходного изображения, с учётом коэффициента стрижки. После стрижки скелета мы можем исключить компоненты, которые в итоге описываются малым числом вершин.

Объединение принципа стрижки и порога по минимальной площади компонент позволяет эффективно бороться с ошибками второго рода и сделать скелет более регулярным: рис. 7

2.3.2 Исправление ошибок бинаризации третьего рода

Ошибки третьего рода предлагается исправлять за счёт дополнительной информации, получаемой за счёт скелета: радиусов r_i в вершинах e_i. В изображениях текста, написанного на линованной или клетчатой бумаге, распределение радиусов имеет отличимую бимодальную структуру, что позволяет получить порог, который будет классифицировать вершины на принадлежащие фону или тексту. Получить порог можно, например, посредством минимизации внутриклассовой дисперсии (уже упоминавшийся метод Отсу).

Обозначим за $\{C_k\}_{k=1}^{K}$ множество всех связных компонент изображения, где C_k – множество вершин E_i из этих связных компонент, $n_k = |C_k|$. Тогда итоговая классификация компонент на текст и фон производится следующим образом:

$$C_k = \begin{cases} \text{text}, & \frac{\sum\limits_{i:E_i \in C_k} r_i}{n_k} > r_{thr}, \\ \text{background}, & \text{иначе.} \end{cases} \tag{4}$$

2.3.3 Исправление ошибок бинаризации первого рода

Ошибки первого рода эффективно исправляются детектированием границ документа и последующим удалением компонент, которые залезают за эту границу. Границы предлагается находить посредством анализа гистограмм горизонтальных и вертикальных проекций вершин скелета. Данные гистограммы можно перенормировать и получить вероятностные распределения. Проанализируем локальные максимумы данных распределений в обоих проекциях. Видно (рис. 8), что на изображениях с ошибками первого рода локальные максимумы проекций, отвечающие положению границ, имеют значительно более острые пики, чем остальные. Данная характеристика не должна является локальной, т.к. значительно зависит от формы распределения в достаточно большой окрестности, ограниченной ближайшим локальным минимумом (рис. 8). Одной из интегральных метрик остроты пика распределения является его куртозис:

$$\kappa = \frac{\mu_4}{\sigma_4} = \frac{\mathbb{E}[(X-\mu)^4]}{(\mathbb{E}[(X-\mu)^2])^2} \tag{5}$$

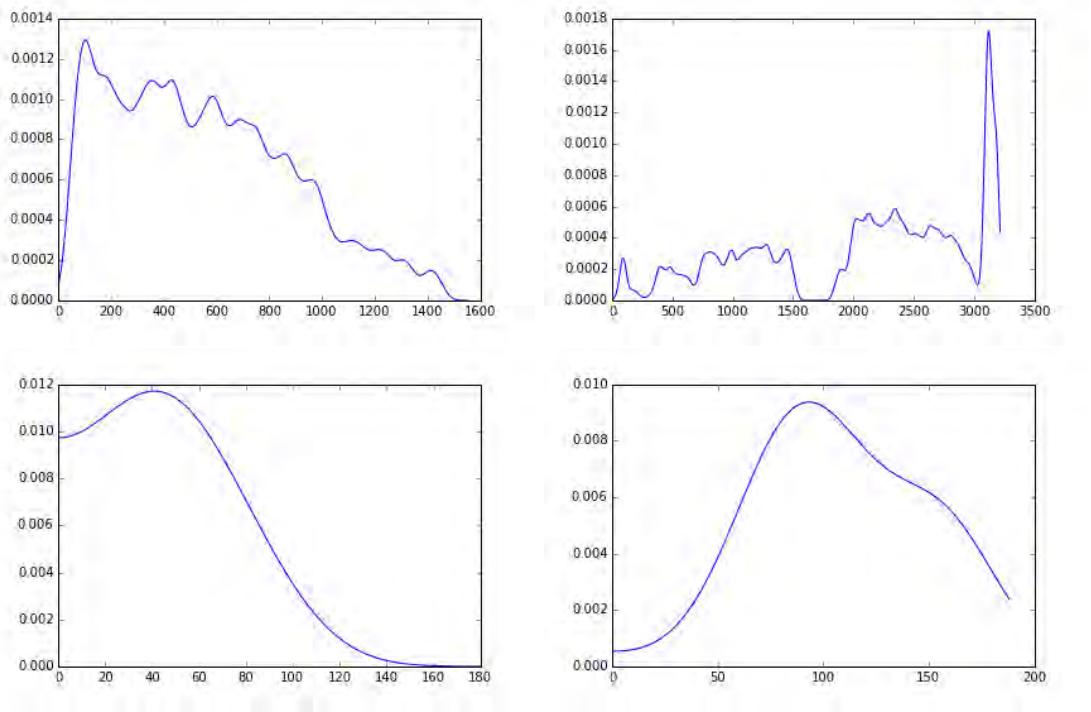


Рис. 8: Сравнение сглаженных гистограм проекций изображения без границ (слева) и изображения с чётко выраженной правой границей (справа). Во втором ряду показан нормированный правый пик этой проекции

Т.к. данная метрика является интегральной, то считать её предлагается по некоторой окрестности пика, предварительно перенормировав попавшее в неё распределение.

Тем самым, задача выделения границ сводится к задаче классификации крайних экстремумов гистограмм, как выбросов, по сравнению с куртозисами остальных экстремумов документа. Её можно решить, например, при помощи подсчёта квантилей этих распределений:

$$x_l^* = \begin{cases} \text{фон}, & \kappa_l > Q_3 + \alpha IQR, \\ \text{текст}, & \text{иначе}. \end{cases} \tag{6}$$

где x_l^* – координата углового экстремума в проекции, Q_3 – 25% квантиль выборки $\{\kappa_l\}$, IQR – её интерквантильный размах, а α – некоторый коэффициент.

После классификации крайних экстремумов, в случае, если какой-то из них был объявлен фоном, компоненты, чей центр тяжести $m_k = \frac{\sum\limits_{i:E_i \in C_k} p_i}{n_k}$ (p_k – координаты

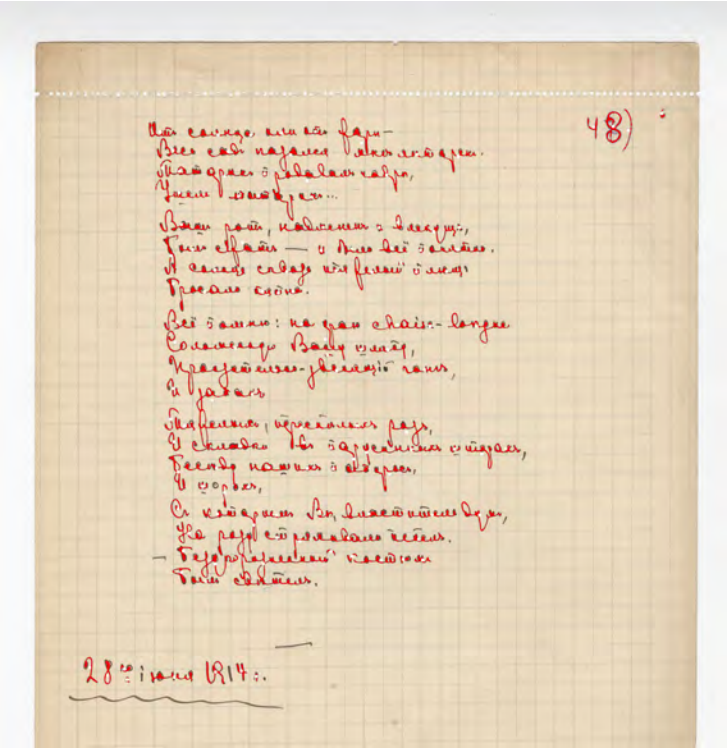


Рис. 9: Отфильтрованный скелет бинарного изображения, наложенный на оригинальное изображение

вершины E_k на изображении) находится за ближайшим локальным минимум гистограммы, объявляются фоновыми.

По итогам фильтрации мы получаем граф $(V_{filtered}, E_{filtered})$, все рёбра и вершины которого соответствуют сегментам текста на исходном изображении. Пример скелета изображения, полученный после фильтрации: рис. 9.

2.4 Начальная кластеризация

Предложенный метод сегментации строк основывается на нескольких предположениях о структуре текста в документе.

1. Отдельно каждая строка является линейной, т.е. может быть хорошо приближена прямой линией.

2. Близкие к друг другу строки имеют одинаковую ориентацию.

3. Ориентация строк, далёких друг от друга, может существенно отличаться.

4. Модуль угла любой строки по отношению к документу $\leq \frac{\pi}{4}$.

Под строками здесь понимается кластеризованный набор компонент $\left\{ \bigcup\limits_{k:\, a(C_k)=l} C_k \right\}_{l=1}^{L}$, где L – число кластеров, a – алгоритм кластеризации, $C_k = (V_p, E_p)$ – перенумерованные связные компоненты графа $(V_{filtered}, E_{filtered})$, к которому применён прунинг и все описанные выше фильтры. Каждый кластер соответствует отдельной строке. Так как каждой компоненте C_k соответствует её скелет, то по каждому кластеру $\left\{ \bigcup\limits_{k:\, a(C_k)=l} C_k \right\}$ можно однозначно восстановить соответствующую ему сегментацию исходного изображения. Тем самым задача сегментации строк в изображении однозначно сводится к задаче кластеризации компонент C_k. Рассмотрим более простой случай, когда каждую исходную компоненту можно отнести лишь к одной строке, т.е. когда слова в разных строках изображения текста не пересекаются.

Формально предположение 1 означает, что существует направление, вдоль которого дисперсия кластеризованных компонент максимальна и значительно больше дисперсии вдоль ортогонального ему направления. Семантически первое направление задаёт вектор, вдоль которого измеряется длина строки, а второе – вектор, вдоль которого измеряется ширина. Очевидно, что в любой строке, значимой с точки зрения рукописного текста, дисперсия вдоль первого направления будет значительно больше дисперсии вдоль второго. Этим обосновывается адекватность предложенного подхода постановке задачи.

При этом в русском и других европейских языках отдельные слова, которые соответствуют компонентам графа, также удовлетворяют данному предположению: их собственные направления наибольшей дисперсии близки к направлениям дисперсии в строке, которой они принадлежат (исключением являются служебные слова, но строки не могут состоять лишь из них). Это позволяет использовать в качестве приближения направления строки направления отдельных слов, входящих в неё.

С учётом данной терминологии и предположений метода введём набор признаков, которым мы будем характеризовать каждую компоненту C_k.

2.4.1 Набор признаков компонент

Поиск двух описанных выше ортогональных направлений формализуется в виде следующей задачи:

$$\begin{aligned} &w_{(1)} = \arg\max_{||w||=1} \Big[\frac{1}{n-1}\sum_i (p-\bar{p})\cdot w)^2\Big], \\ &w_{(2)} \cdot w_{(1)} = 0 \end{aligned} \tag{7}$$

где $w(i)$ – векторы первого и второго направления соответственно, $p = (x, y)$ – вектор с координатами точки из компоненты, $\bar{p}$ – вектор средних, $a \cdot b$ – скалярное произведение двух векторов.

Задача 9 полностью эквивалентна постановке задачи для метода главных компонент [17], поэтому будем называть полученные направления $w_{(1)}$ и $w_{(2)}$ главными направлениями компоненты.

Назовём углом ориентации компоненты в изображении следующую величину:

$$\theta_k = \arctan\Big(\min_i \Big[\frac{w^y_{k,(i)}}{w^x_{k,(i)}}\Big]\Big), \quad \theta_k \in (-\frac{\pi}{2}, \frac{\pi}{2}) \tag{8}$$

где $w_{k,(i)}$ – главные направления k-й компоненты. Будем считать, что характеристика θ_k является зашумлённым приближением угла $\hat{\theta}_l$ строки, в которой находится k-я компонента. Это является формализацией четвёртого предположения о структуре текста в документе, так как очевидным образом можно показать, что угол θ_k всегда будет лежать в более узких границах: $\theta_k \in (-\frac{\pi}{4}, \frac{\pi}{4})$. Делается оно для того, чтобы компоненты, отвечающие служебным словам, а также словам, которые на изображении или из-за бинаризации оказались разбиты на несколько связных компонент, имели начальный угол наклона, более близкий к истинному.

Также посчитаем ширину проекций вдоль каждой главной компоненты. Ширину проекции вдоль компоненты с меньшим углом назовём длинной, вдоль второй – шириной компоненты:

$$\begin{aligned} \text{length}_k &= \max_p[(p-\bar{p})\cdot w_{k,(i)}] - \min_p[(p-\bar{p})\cdot w_{k,(i)}] \\ \text{width}_k &= \max_p[(p-\bar{p})\cdot w_{k,(j)}] - \min_p[(p-\bar{p})\cdot w_{k,(j)}] \end{aligned} \tag{9}$$

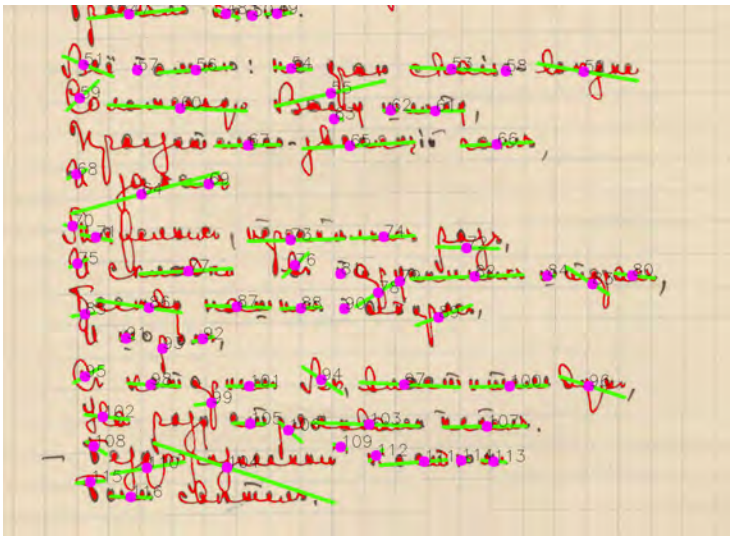


Рис. 10: Визуализация набора признаков. Линиями показаны проекции связных компонент на направление θ_k, они имеют длину length_k. Точки – центры масс соответствующих компонент.

Также введём ещё одну метрику, которая будет измерять степень вытянутности компоненты вдоль направления, отвечающего углу θ_k: отношение дисперсии вдоль него к сумме дисперсий вдоль двух направлений, обозначим его за γ_k.

Объединение данного набора признаков с признаком, характеризующим пространственное положение компоненты ($\bar{p}$, координата центра тяжести) является достаточным для построения требуемой сегментации.

Визуализация набора признаков выглядит следующим образом: рис. 10.

2.4.2 Первичная кластеризация соседних компонент

Следующим этапом кластеризации компонент является выделение тех из них, которые гарантированно принадлежат одной строке. Этот этап направлен на то, чтобы сагрегировать маленькие компоненты C_k, чей угол θ_k плохо приближает истинный угол направления строки $\hat{\theta}_l$ из-за маленькой дисперсии вдоль главного направления (т.е., компоненты с низким коэффициентом γ_k) в объединения $\left\{ \bigcup\limits_{k:\, a(C_k)=l} C_k \right\}_{l'=1}^{L'}$, которые мы будем считать едиными компонентам на второй итерации кластеризации.

На этом же этапе требуется разрешить более глобальную задачу группировки компонент на кластеры, в которых ориентация строк будет одинаковой.

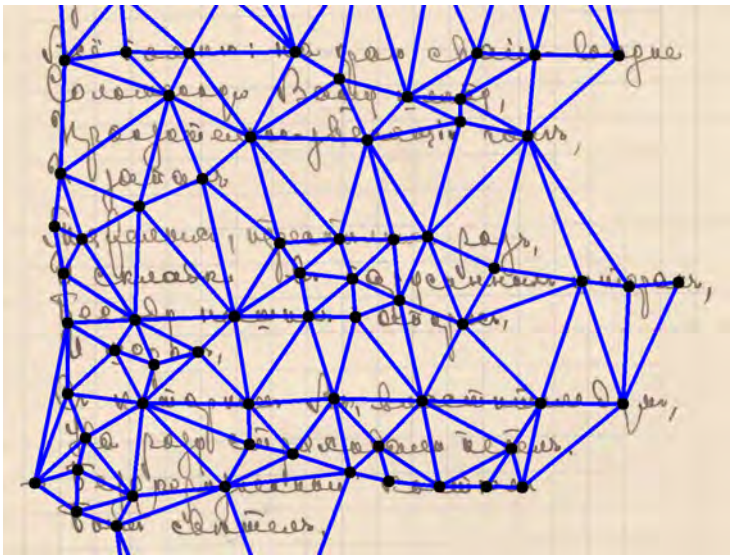


Рис. 11: Фильтрованная по длине триангуляция Делоне на центрах масс связных компонент скелета

Для решение данных задач задействуем второе и третье предположения метода и введём на компонентах C_k систему соседства, т.е. построим ещё один граф (P, E^*), вершинами которого будут являться центры компонент $\bar{p}_k$, а рёбра задавать систему соседства. Естественным графом, отвечающим данным критериям, является триангуляция Делоне [2]. Это планарный граф, максимальной кликой которого является клика размера 3 (треугольник). Пример полученного графа: рис. 11.

Для каждой компоненты он задаёт систему связности. Зададим метрику, которая будет измерять расстояние между двумя компонентами:

$$d_{kj} \equiv d(C_k, C_j) = \begin{cases} ||\bar{p}_k - \bar{p}_j||, & (\bar{p}_k, \bar{p}_j) \in E^* \\ \inf, & \text{иначе.} \end{cases} \tag{10}$$

Далее хотелось бы с учётом этой системы связности скорректировать полученные значения углов θ_k и заменить на средневзвешенные по системе соседства. Но изначально триангуляция включает рёбра, имеющие большую длину и соединяющие противоположные части изображения. Исходя из предположения 3, эти вершины не должны влиять на ориентацию друг друга. Рёбра полученного графа требуется отфильтровать.

Составим из конечных значений данной метрики выборку и проверим её на наличие выбросов при помощи подсчёта квантилей. Удалим из триангуляции рёбра,

которые соответствуют выбросам:

$$(\bar{p}_k, \bar{p}_j) = \begin{cases} \text{оставить}, & d_{kj} < Q_3 + \alpha IQR, \\ \text{удалить}, & \text{иначе.} \end{cases} \tag{11}$$

Получаем новую систему соседства $(P, E^*_{\text{filtered}})$, которая уже не является связным графом. Исходя из предположений 1 и 3, компоненты, которые попали в разные связные участки полученного графа не могут являться частями одних строк.

Взвесим углы смежных компонент по полученной системе соседства $(P, E^*_{\text{filtered}})$ следующим образом:

$$\begin{gathered} \theta_k^{\text{new}} = \alpha\theta_k + (1-\alpha) \sum_{j:\ (k,j)\in E^*_{\text{filtered}}} \beta_j \theta_j, \\ \sum_j \beta_j = 1 \end{gathered} \tag{12}$$

Определим веса в 13 так, чтобы компоненты с большей длиной и большей относительной дисперсией вдоль главного направления, определяющего длину, давали больший вклад в итоговый угол θ_k^{new}. Тем самым мы даём тем компонентам, которые наиболее похожи на строки (протяжённые и узкие), корректировать направления менее стабильных компонент. Положим $\beta_j = \frac{\gamma_j \text{length}_j}{\sum_j \gamma_j \text{length}_j}$, $\alpha = \frac{\gamma_k \text{length}_k}{\text{med}[\gamma_j \text{length}_j]}$, где med – медиана.

Скорректированные θ_k^{new} позволяют нам сделать оценку доверительного интервала на истинные углы θ_l, которые лежат внутри одной области связности, определённой графом $(P, E^*_{\text{filtered}})$. Учитывая предположение 2, мы считаем, что все θ_l между собой похожи, поэтому построение единого интервала для всех близких строк также имеет смысл. Будем использовать робастный интервал, полученный при помощи квантилей распределения, так как внутри введённой выше области связности даже после коррекции могут присутствовать компоненты с зашумлённой ориентацией:

$$\{\hat{\theta}_l\} \in [\text{med}\theta_k + IQR, \text{med}\theta_k - IQR] \tag{13}$$

где IQR – интерквантильный размах выборки θ_k.

Тем самым для каждой области связности получен достаточно робастный и узкий интервал, который включает в себя предполагаемые истинными ориентации строк.

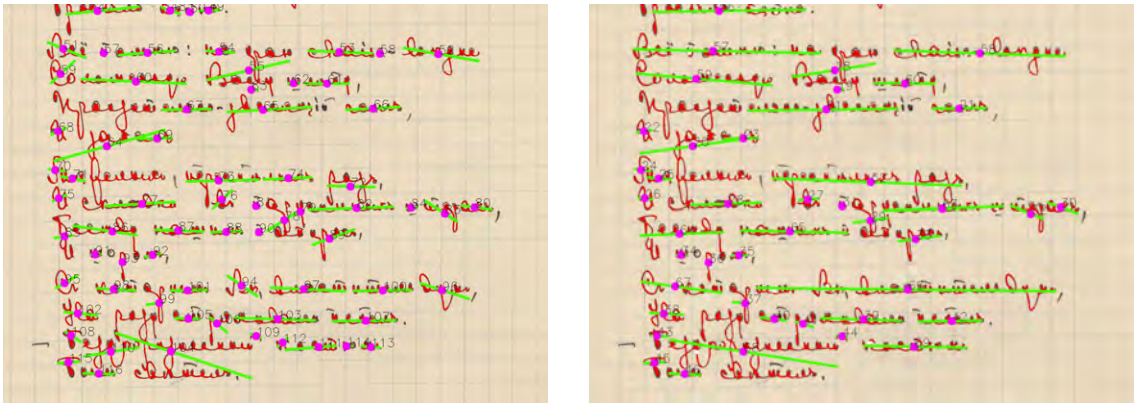


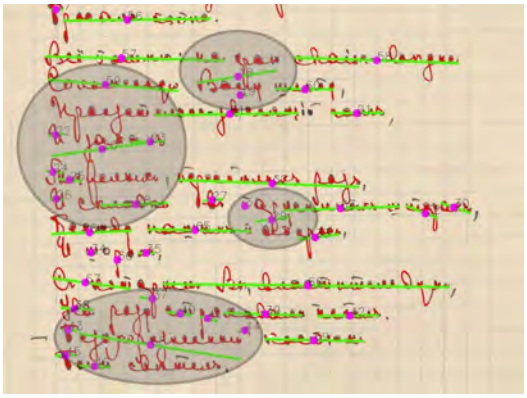
Рис. 12: Слева показаны главные направления компонент до начальной кластеризации, а справа – после)

Теперь с его помощью решим задачу аггрегации как можно большего числа компонент в фрагменты строк $\left\{ \bigcup\limits_{k:\, a(C_k)=l} C_k \right\}_{l'=1}^{L'}$. Для этого мы всё ещё не можем использовать отдельные компоненты, так как их протяжённость не сопоставима с длиной строки. Заметим, что отфильтрованная система соседства триангуляции Делоне теперь включает в себя лишь компоненты, которые относятся либо к одной, либо к разным строкам с одинаковой ориентацией. При этом, исходя из предположения о линейности строк, компоненты из одной строки также должны хорошо приближаться одной линией с углом $\hat{\theta}_l$. Это позволяет нам рассматривать рёбра триангуляции, как кандидаты на задание систему соседства внутри одной строки, и произвести первичную сегментацию, используя их углы θ_{kj}.

Удалим из графа $(P, E^*_{\text{filtered}})$ рёбра, угол θ_{kj} которых не входит в интервал 13. Полученные связные компоненты и будут являться искомыми фрагментами строк $\left\{ \bigcup\limits_{k:\, a(C_k)=l} C_k \right\}_{l'=1}^{L'}$, и, как видно на рис. 12, объединять компоненты в кластеры, которые гарантированно принадлежат лишь одной строке.

2.5 Разрез компонент, содержащих фрагменты слов из разных строк

Выше был рассмотрен случай, в котором предполагается, что каждая из компонент принадлежит только одной строке. Данное предположение очень часто нарушается (см. 13).



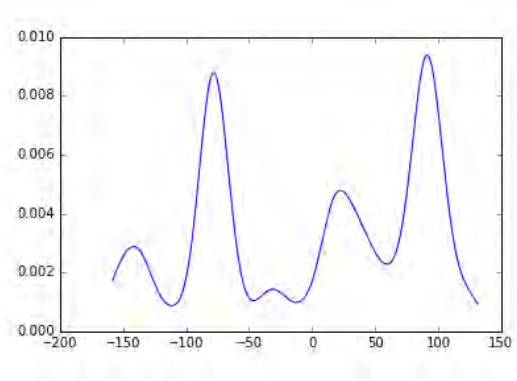


Рис. 13: Компоненты, содержащие в себе фрагменты слов, относящиеся к разным строкам (слева), а также сглаженная проекция одной из них (n 20) на направление, ортогональное установленному направлению текста в документе

Для детектирования подобных компонент рассмотрим множество ширин width'_k. Компоненты, образованные склейкой междустрочных фрагментов, с точки зрения этой выборки являются выбросами. Их можно детектировать при помощи порога, построенного на квантилях (аналогично порогу 6).

После этого оценим при помощи углов компонент θ'_k, полученных при начальной кластеризации, угол θ_l направления текста в данном фрагменте документа:

$$\hat{\theta}_l = \text{med}\ \theta'_k \tag{14}$$

Спроецируем каждую из широких компонент на это направление и построим сглаженную гистограмму проекции (рис. 13). Разрежем каждую такую компоненту вдоль локальных минимумов гистрограммы, результат можно видеть на рис. 14.

2.6 Построение метрики близости

Примем связные компоненты, полученные на предыдущем этапе, за новые компоненты $C'_k = \left\{ \bigcup_{k:\, a(C_k)=l} C_k \right\}_{l'=1}^{L'}$ и пересчитаем для них все используемые признаки. Дополнительно расширим на них введённую ранее систему соседства: C'_k будет лежать в той же системе связности компонент, что и все входящие в него компоненты.

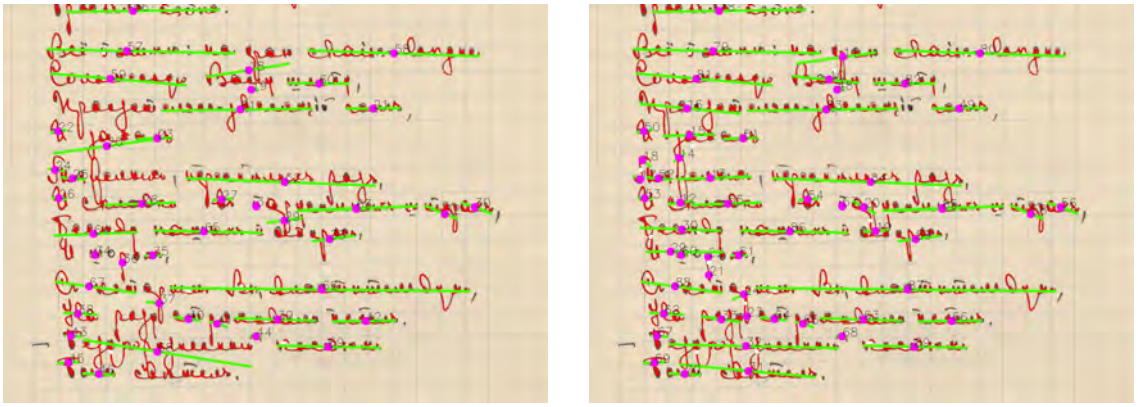


Рис. 14: Сравнение компонент до и после разреза

После этого введём на C_k следующую метрику:

$$d(C_i, C_j) = \begin{cases} \left\| \left(\frac{n_x(L_i, L_j)}{\max\limits_k [\gamma_k \text{length}_k]}, n_y(L_i, L_j) \right) \right\|, & C_i \text{ и } C_j \text{ лежат в одной системе соседства,} \\ \inf, & \text{иначе} \end{cases} \tag{15}$$

где $n(L_i, L_j)$ – кратчайший вектор между линейными сегментами L_i и L_j, заданные в системе главных направлений компоненты с наибольшей длиной $[\gamma_k \text{length}_k]$. Линейные сегменты задаются главными осями метода главных компонент и равны по длине проекции данных на них: рис. 12.

Данное расстояние является частным случаем расстояния Махолонобиса в двумерном пространстве (для случая, когда обе компоненты лежат в одной системе соседства). В нашем случае расстоянием 15 задано явное предпочтение расстоянию вдоль направления, которое мы считаем аппроксимацией направления строки. Т.е. линейные сегменты, находящиеся в внутри одной строки, по этой метрике будут ближе к друг другу, чем линейные сегменты в разных строках, и степень их близости будет пропорциональна вытянутости наибольшего из сегментов.

При обучении метрики можно также воспользоваться и автоматическими методами обучения метрик с учителем. Для этого достаточно разметить данные, получаемые на предыдущем этапе, и запустить произвольный алгоритм обучения по всему набору полученных признаков. В данной работе метрика подобрана вручную, что, однако, не ограничивает общность подхода.

2.7 Итоговая кластеризация

После обучения или подсчёта метрики 15 требуется с её помощью кластеризовать компоненты по мере близости между ними. Для того чтобы сделать алгоритм кластеризации более робастным, предлагается производить кластеризацию жадным образом, начиная с компонент наибольшей длины. Процесс кластеризации является итеративным, на каждом этапе мы объединяем между собой компоненты, расстояние между которыми меньше оценки на ширину компонент EW, которую можно получить посредством анализа распределения ширин width_k. Также порог на расстояния можно подбирать и автоматически.

После сходимости процесса (когда число итоговых компонент перестало уменьшаться), оставшиеся некластеризованные компоненты относятся к ближайшему кластеру. Пример получаемого результата на рис. 15.

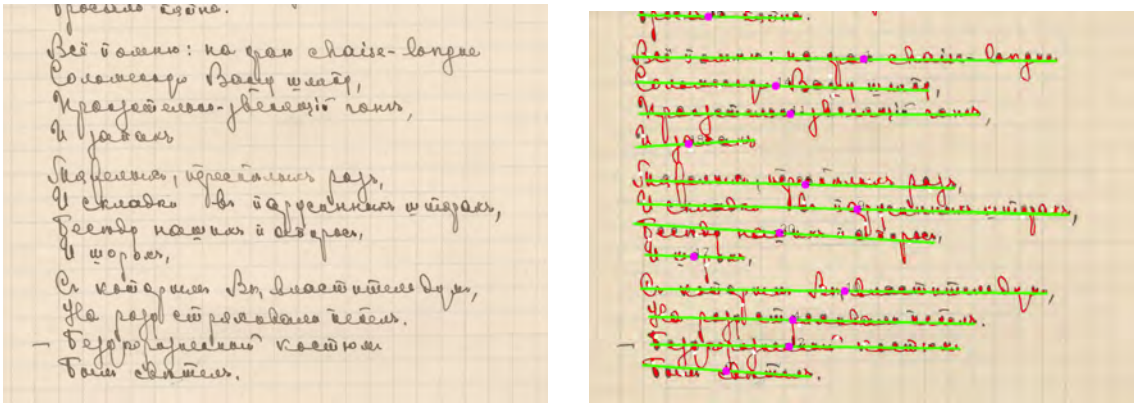


Рис. 15: Исходное изображение и изображение с сегментированными строками

3 Заключение

Основным результатом данной работы является формулировка и обоснование нового подхода к решению задачи сегментации строк в изображениях рукописных документов, который основан на построении, фильтрации и кластеризации скелетов бинарного изображения.

Ключевым результатом является сформулированный общий подход к кластеризации, производящийся в два этапа: этап, связанный с уменьшением шума в данных,

и этап, связанный с построением метрики, которая впоследствии используется в алгоритме автоматической кластеризации. Общность подхода заключается в том, что нигде явно не используется привязка к конкретному почерку, стилю письма или языку (кроме самых общих предположений), но при этом имплементация этих особенностей возможна тривиальным образом посредством ручной разметки компонент на принадлежащие/не принадлежащие одной строке и обучения соответствующей метрики при помощи методов distance metric learning.

По итогам исследования данный метод сегментации будет реализован для массива документов из архива рукописных документов русских писателей.

Список литературы

[1] Леонид Местецкий. *Скелет многоугольной фигуры – представление плоским прямолинейным графом*. Труды 20 международной конф. ГРАФИКОН-2010, СПб, ИТМО, с. 222-229, 2010.

[2] Леонид Местецкий. *Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры*. Физматлит, 2009.

[3] A. Sanchez et al. *Text Line Segmentation in Images of Handwritten Historical Documents*. First Workshops on Image Processing Theory, Tools and Applications, 2008.

[4] A. Nicolaou et al. *Handwritten Text Line Segmentation by Shredding Text into its Lines*. International Conference on Document Analysis and Recognition, 2009.

[5] Alireza Alaei et al. *A new scheme for unconstrained handwritten text-line segmentation*. Pattern Recognition Volume 44, Issue 4, 2011.

[6] Jayant Kumar et al. *Handwritten Arabic Text Line Segmentation using Affinity Propagation*. 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*. Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, 2008.

[8] Nobuyuki Otsu. *A Tlreshold Selection Method from Gray-Level Histograms*. TRANSACTIONS ON SYSTREMS, MAN, AND CYBERNETICS, VOL. SMC-9, NO. 1, 1979.

[9] W. Niblack. *An introduction to digital image processing*. Englewood Cliffs, Prentice Hall, N.J., pp. 115-116, 1986.

[10] N. Chaki et al. *Exploring Image Binarization Techniques,*. Studies in Computational Intelligence, 2014.

[11] Bolan Su, Shijian Lu and Chew Lim Tan. *Binarization of Historical Document Images Using the Local Maximum and Minimum.* International Conference of Document Analysis and Recognition, 2010.

[12] Bolan Su, Shijian Lu and Chew Lim Tan. *A Robust Document Image Binarization Technique for Degraded Document Images.* IEEE Transaction on Image Processing, 2012.

[13] K. Ntirogiannis, B. Gatos, I. Pratikakis. *A combined approach for the binarization of handwritten document images.* Pattern Recognition Letters, 2012.

[14] Morteza Valizadeh, Ehsanollah Kabir. *Binarization of degraded document image based on feature space partitioning and classification.* Springer Verlag, 2010.

[15] Morteza Valizadeh, Ehsanollah Kabir. *An adaptive water flow model for binarization of degraded document images.* Springer Verlag, 2012.

[16] J. Canny. *A Computational Approach to Edge Detection.* TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-8, NO. 6, 1986.

[17] Lindsay I Smith. *A tutorial on Principal Components Analysis.* COSC 453, 2002.