Генерация признаков для распознавания почерка в рукописных документах

Кадченко Иван Евгеньевич
ВМК МГУ
kadchenko.ivan@mail.ru

Местецкий Леонид Моисеевич
ВМК МГУ
mestlm@mail.ru

2023

Аннотация

В данной работе предлагается метод распознавания почерка человека по его уникальному набору шаблонов. Пусть даны изображения рукописного текста и требуется определить, какое количество различных авторов представлено на входных изображениях и к какому автору принадлежит каждый отдельный фрагмент текста. Решение данной задачи позволит устанавливать авторства различных документов, в том числе и исторических, а также упростить работу с архивами в целом. Предлагаемое решение опирается на штриховое представление рукописного текста. Гипотеза заключается в том, что каждый человек при написании различных букв использует примерно похожие, свойственнные именно ему движения руки, а также, что у большинства людей присутствуют какие-то штрихи, выделяющие конкретного автора.

Ключевые слова: Распознавание почерка · Шаблоны писателя · Рукописные документы

1 Введение

Задача идентификации автора конкретного почерка ставилась с давних времён. Начиная с 5-6 веков в законодательстве Византии находило закрепление исследование почерка в судебных целях, а в России уже в XV веке при установлении подлинности документов использовалось сравнение рукописей. Ближе к настоящему времени, прикладных задач становилось всё больше. Работа с разного рода документами, от рецептов врачей до исторических скриптов, судебная аналитика и криминалистика - все эти области могут потребовать установить авторство или подлинность текста. Несмотря на всё большую цифровизацию нашей жизни, во всех упомянутых областях всё ещё используются рукописные тексты, а потому задача идентификации автора по его почерку по-прежнему сохраняется.

Распознавание почерка является одним из важнейших инструментов криминалистики, позволяющим идентифицировать автора документа или записи. Однако, существующие методы распознавания почерка требуют наличия текстовых образцов написания конкретного писателя, что затрудняет их применение в ряде случаев. В связи с этим, возникает необходимость в разработке методов распознавания почерка, которые не зависят от конкретного текста.

Один из таких методов - использование избыточных шаблонов, основанный на том, что каждый человек имеет свой индивидуальный почерк, который проявляется в определенных чертах написания букв и цифр. Эти черты могут быть выделены из большого числа образцов почерка и использованы для создания уникального шаблона писателя.

Актуальность проблемы независимого от текста распознавания почерка заключается в том, что это может помочь ускорить и улучшить процесс идентификации автора документов и записей. Например, при расследовании преступлений, где имеется только небольшой фрагмент написанного текста, использование избыточных шаблонов может значительно сократить время на поиск подходящих образцов почерка для сравнения. Кроме того, этот метод может быть полезен при работе с историческими доку-

ментами или при анализе письменных свидетельств в судебных процессах. Таким образом, разработка методов независимого от текста распознавания почерка является актуальной и важной задачей для криминалистики и других областей, где требуется идентификация авторства документов и записей.

Данная работа будет направлена на поиск признаков для создания такого уникального шаблона писателя.

2 Постановка задачи

Статья будет опираться на следующее предположение: идентификация автора по его почерку в большей степени зависит не от способа рисования или сегментации символов, когда извлеченные графемы могут нести некоторую семантическую информацию, а от физического способа создания линий или петель. Это означает, что масштаб рассматриваемых единиц письменности будет меньше. Иными словами, мы считаем, что писатель может использовать один и тот же жест рукой и, следовательно, один и тот же шаблон при написании разных символов, имеющих сходные базовые формы, как, например, одинаковые петли у разных букв (рис.1). Такие единицы письменности и назовём "штрихами".

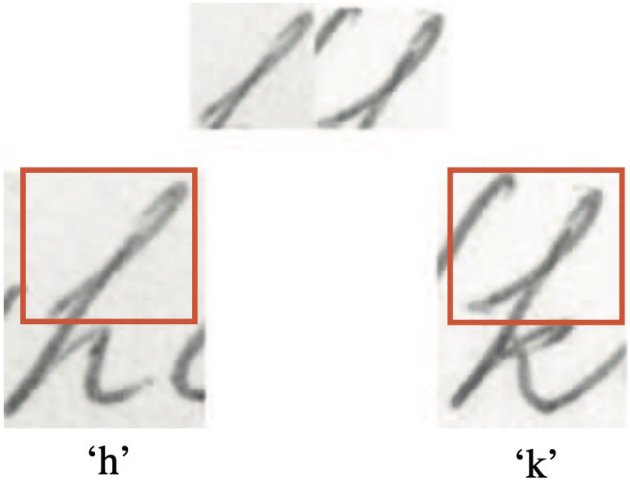


Рис. 1: Одинаковые петли у двух разных букв.

Основной задачей данной работы будет являться оценка возможности классификации почерков по шаблонности и уникальности штрихов для каждого автора.

3 Признак строк без опоры на графемы

Как утверждалось ранее, поставлена гипотеза, что с помощью штрихового представления можно отличать почерки разных людей. Опустимся ниже в иерархии представления рукописного текста и проанализируем межстрочные и межсимвольные расстояния.

Предлагается построить признак, отвечающий за плотность чёрных пикселей, просто просуммировав чёрные пиксели вдоль оси ординат (рис.2).

На исходном рукописном тексте можно распознать два разных почерка: первый - это первые пять строк, второй - все оставшиеся. Вытянутые столбцы на графике соответствуют строкам исходного текста, их количество совпадает.

Построим разбиение по оси ординат по точкам глобальных минимумов на отрезках между красными точками. Также добавим в разбиение две крайние точки соответствующие примерно 150 и 1850 ординатам. Каждая из полученных областей из разбиения будет содержать ровно одну красную точку, на которой будет достигаться максимум этого отрезка.

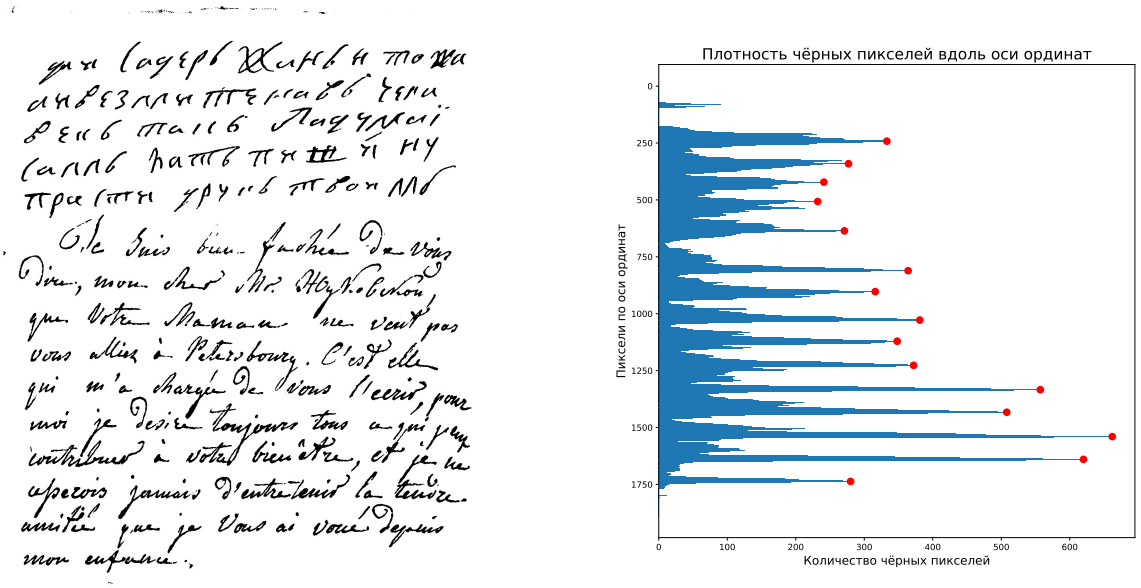


Рис. 2: Рукописный текст и его плотность.

Можно выделить два типа распределения таких областей: первые пять точек экстремума (они обозначены красными точками) с окрестностями и десять последующих. Первый тип характеризуется меньшими абсолютными значениями в точке максимума и большими абсолютными значениями в точках минимума относительно второго типа, хвосты у первого типа более тяжёлые. На основе приведённых замечаний можно однозначно установить, что на исследуемом изображении рукописного текста представлено два автора.

Описанный подход оказывается действенным, подходящим для применения одновременно с несколькими другими признаками для создания уникального шаблона автора в поставленной ранее задаче определения количества различных почерков. Но признак имеет существенное ограничение: его разумно использовать в случае, когда строки рукописного текста близки к горизонтальному расположению, что, очевидно, выполняется далеко не всегда.

4 Генерация признаков на основе штрихового представления

4.1 Сегментация по штрихам

Для построения штрихового представления рукописного почерка на изображении будет использована программа Л.М.Местецкого, принцип работы которой описан в источнике [3]. Пример работы программы отражён на рис.3.

Программа получает на вход изображение рукописного текста, бинаризует его и строит штриховое представление исходного текста, визуализируя, как на рис.3, и записывая в текстовый файл результат. Каждый штрих представляет собой набор нецелочисленных координат точек, из которых состоит штрих, и информацию о взаимном расположении этих точек, такую как тип штриха (цепь, кольцо или отрезок), его уровень (выступающий, базовый или свисающий) и номер строки, к которой штрих относится.

Используя приведённую информацию о штрихах, постараемся построить полезные признаки, позволяющие оценить возможность классификации почерка по отдельным штрихам.

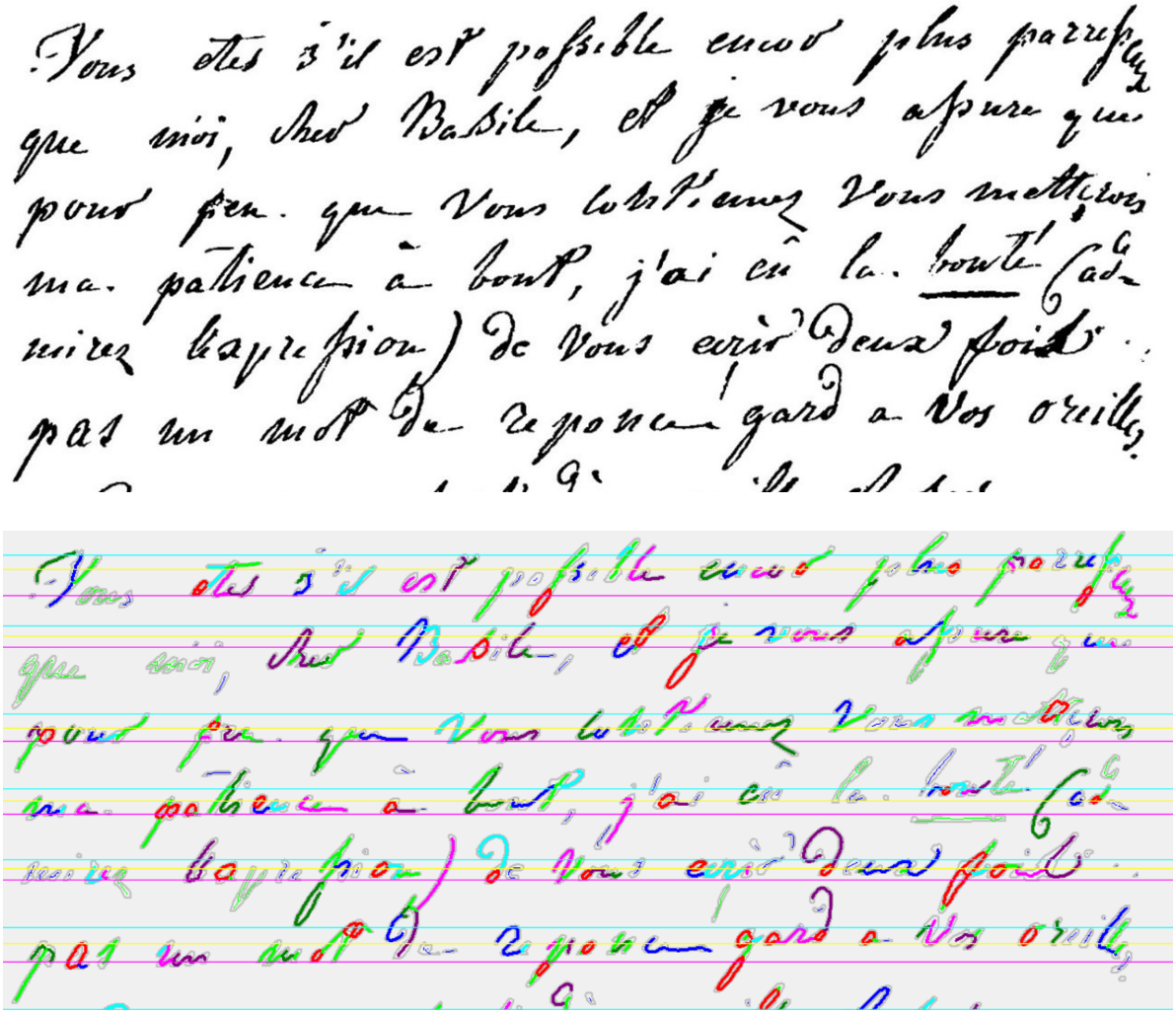


Рис. 3: Исходный рукописный текст и визуализированные штрихи.

4.2 Агрегация штрихов: наклон и кривизна

Базовыми признаками, рассматриваемыми в задаче распознавания почерка, являются наклон и кривизна текста. Эти признаки являются стабильными параметрами [4] при допущении, что рассматриваемый документ представляет собой естественный стиль письма писателя, а не является поддельным.

Проанализируем применимость данных характеристик при оперировании штриховым представлением.

4.2.1 Признак-наклон

Имея набор пар координат точек, из которых состоит отдельный штрих, величину наклона можно определить как коэффициент incline в модели линейной регрессии с одной независимой переменной:

$$y = \text{inline } x + \text{bias} \tag{1}$$

Выражая из этого уравнения incline и подставляя данные точки штрихов, можно получить формулу для вычисления наклона, смещение интересовать нас не будет:

$$\text{incline} = \frac{\sum_{i=1}^{N}(x_i - \mathbb{E}x)(y_i - \mathbb{E}y)}{\sum_{i=1}^{N}(x_i - \mathbb{E}x)^2} \tag{2}$$

Репрезентативность такого подхода можно наблюдать на рис.4. Распределения величины наклона для строк текста одного (1-5 строки) и другого (6-15 строки) почерка заметно отличаются. Второй почерк имеет выраженный наклон по сравнению с первым, что и отражено на графике смещением центра распределения.

afterotsu-cropped.pdf

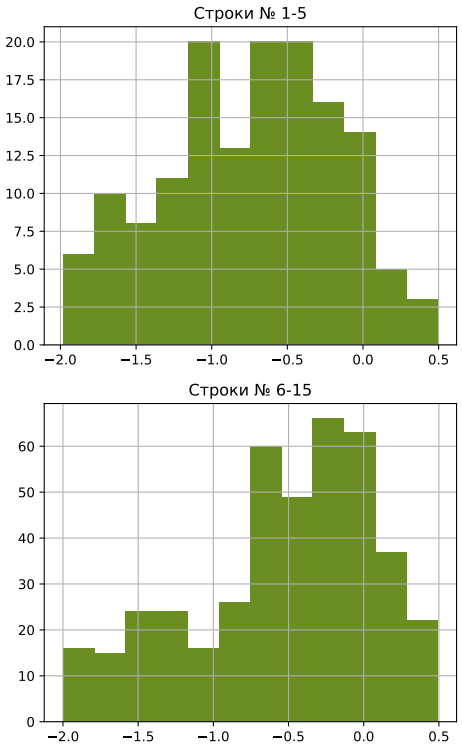


Рис. 4: Рукописный текст и его распределение величины наклона по строкам.

4.2.2 Признак-кривизна

Вторую фундаментальную характеристику письма, кривизну, можно определить следующим образом. Будем считать элементарную степень кривизны трёх последовательно взятых точек, как разность суммы расстояний между соседними точками и расстояния между крайними. Итоговую величину положим равной сумме всех таких элементарных степеней кривизны. Формульно это записывается так:

$$\text{curvature} = \sum_{i=2}^{N-1} (dist(p_{i-1}, p_i) + dist(p_i, p_{i+1}) - dist(p_{i-1}, p_{i+1})), \tag{3}$$

$$dist(p(x_1, y_1), p(x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{4}$$

Установим зависимость построенного признака-кривизны на том же изображении с двумя почерками (рис.5). Первый почерк можно охарактеризовать как прерывистый. Штрихи в нём в большей степени представляют линии, нарисованные между последовательными отрываниями руки от листа. При таком стиле писания особенно наглядно возникает определённый набор шаблонов писателя. Из элементов такого набора и состоит каждый рисуемый символ. Так, на графике рис.5 видно, что штрихи по величине кривизны разбиваются на три отделимых кластера по значению: 0-2, 2-5, 6-10. Второй почерк таким свойством не обладает, написанный им текст имеет межсимвольные переходы, более гладкую форму, в следствие чего и распределение получается более гладким.

Рассмотренные классические признаки оказались чувствительны к разным почеркам, поэтому можно утверждать о состоятельности данного штрихового представления.

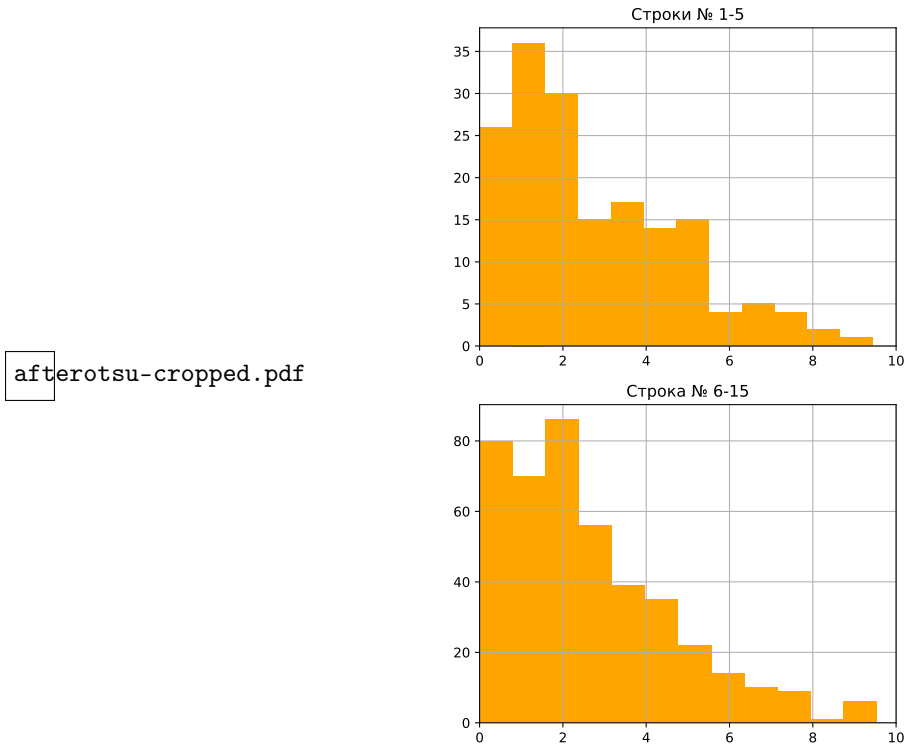


Рис. 5: Рукописный текст и его распределение величины кривизны по строкам.

4.3 Особенность почерка: кольцевые штрихи

Предлагается обратить внимание на возможность классификации штрихов по типу: цепь, кольцо или отрезок. Наиболее способным отразить стиль почерка писателя можно назвать кольцевой тип штриха, ведь помимо того, что данный тип можно охарактеризовать метрическими признаками: наклоном, кривизной или размером - каждому почерку свойственна своя частота встречаемости такого типа. Имеется в виду, что одни и те же буквы можно писать с петлёй и без (рис.1), и такие петли относятся к кольцевым штрихам.

В данной части работы построим признаки, характеризующие кольцевой тип штрихов.

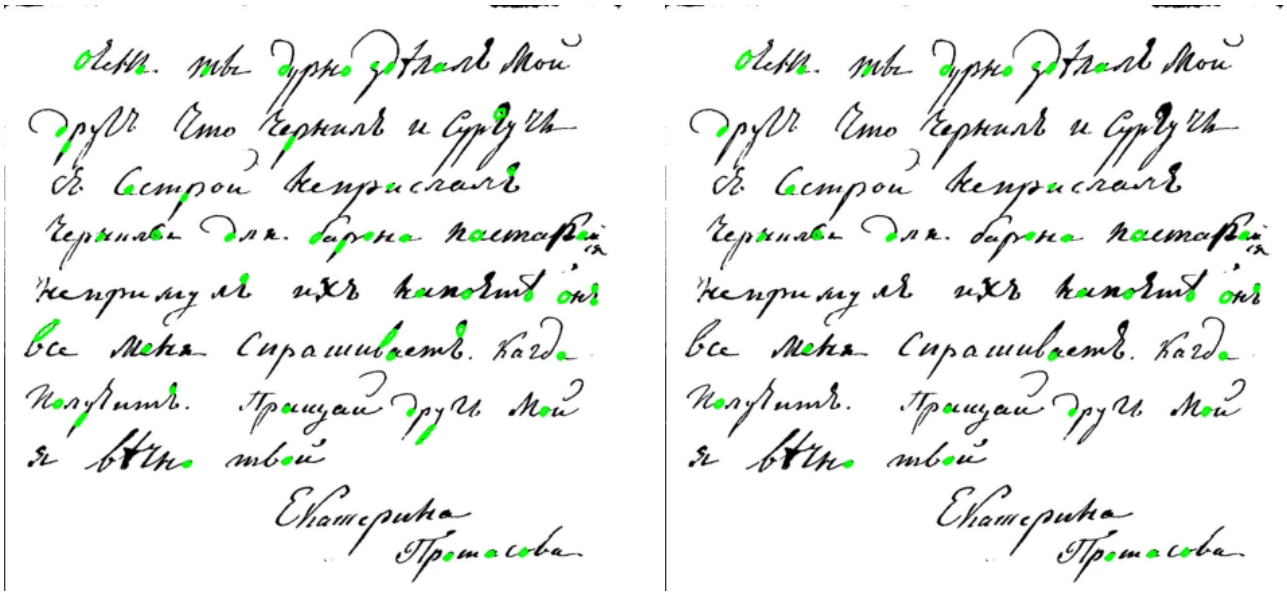
На рис.6 представлены выделенные кольцевые штрихи исследуемого представления. Можно заметить, что детектируются не все распознаваемые глазом кольцевые штрихи, но тем не менее значительная их часть, по которой возможно провести анализ, приводящий к результату.

На рис.7 приведены примеры разных почерков. Обращая внимание на кольцевые штрихи, можно заметить, что у первого почерка рассматриваемые элементы ближе к форме кольца, отчётливо видна "дырка". У второго же почерка такие штрихи меньше и больше напоминают форму овала. Предлагается построить признаки, отражающие данные различия.

4.3.1 Признак-форма

Рассматриваемые в этом разделе штрихи имеют эллиптическую структуру, поэтому признак-форму можно определить как отношение большой полуоси к малой. Чем данная величина ближе к единице, тем штрих больше напоминает кольцо.

Длины полуосей эллипса определим следующим образом (5)-(7). Найдём приближённый центр, а затем из набора точек, входящих в штрих, найдём такие точки, расстояния до которых было бы наибольшим и наименьшим. Эти расстояние и будут определять большую и малую полуоси соответсвтенно. Как и раньше, будем считать евклидово расстояние между двумя точками (4).



Все Базовые

Рис. 6: Пример кольцевых штрихов.

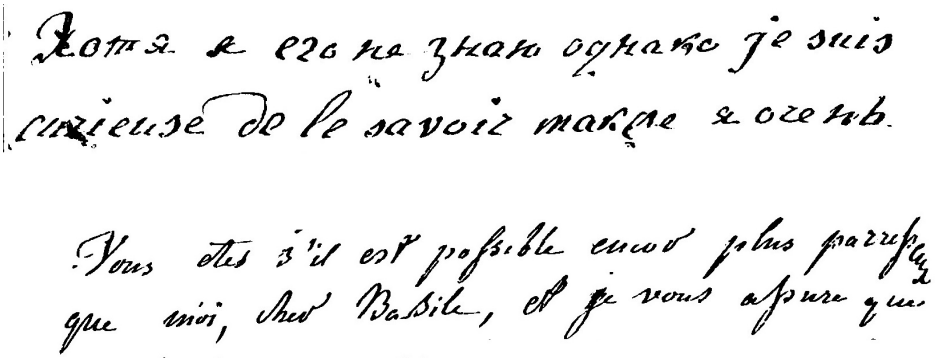


Рис. 7: Пример разных почерков.

$$x_c = \mathbb{E}x,\ y_c = \mathbb{E}y,\ p_c(x_c, y_c) = (x_c, y_c) \tag{5}$$

$$a = \max_{i \in 1,\dots,N} dist(p_i, p_c) \tag{6}$$

$$b = \min_{i \in 1,\dots,N} dist(p_i, p_c) \tag{7}$$

Тогда итоговая величина, отражающая форму штриха, будет определяться формулой (8).

$$\text{shape} = \frac{a}{b} \tag{8}$$

4.3.2 Признак-размер

Аналогично форме штриха определим величину его размера (9). Домножение на π осуществляется для интерпретируемости настоящей площади эллипса.

$$\text{area} = ab\pi \tag{9}$$

4.3.3 Признак-частота

Как видно будет далее, построенное к настоящему моменту признаковое пространство позволяет идентифицировать различные почерки, но некоторые стили писания сливаются. Поэтому остаётся необходимость в большем количестве признаков.

Так же как писатель может сделать петлю в том месте, где не сделает её другой автор, он может делать отличительно выступающие за базовую линию штрихи в разных местах. Поэтому предлагается добавить признак, отвечающий за частоту встречаемости выступающих штрихов по отношению к другим для каждой строки (10), где M - число строк, K_i - количество штрихов в i-ой строке.

$$\text{ledge} = \frac{1}{M}\sum_{i=1}^{M}\frac{\sum_{j=1}^{K_i}[stroke_j = \text{' '}]}{K_i} \tag{10}$$

5 Вычислительный эксперимент

Оценим возможность классифицировать почерк по отдельным штрихам на основе построенных признаков. Эксперимент будет заключаться в классификации четырёх почерков: 5 изображений первого почерка, 3 - второго, 5 - третьего, 3 - четвёртого, - преимущественно опираясь на особенности кольцевых штрихов. Примеры двух типов почерков представлены на рис.7, со всеми рассматриваемыми типами можно ознакомиться в приложении.

Будем оценивать наиболее часто встречающиеся кольцевые штрихи: оставим для рассмотрения только базовые штрихи с не слишком вытянутой формой (shape < 10) и выбросим 15% самых больших. Полученные значения признаков для рассматриваемых почерков приведены в Таблице 1.

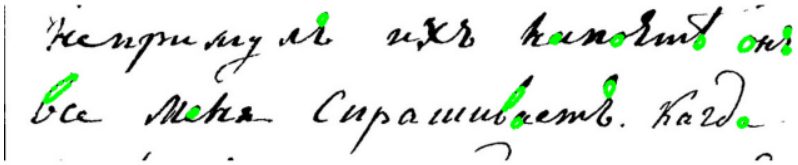
Тип почерка	shape	area	ledge
1	5.73	81	0.108
1	7.30	75	0.085
1	6.15	83	0.050
1	4.11	105	0.133
1	5.64	77	0.107
2	-	220	-
2	-	178	-
2	-	188	-
3	4.42	93	0.141
3	3.83	82	0.121
3	4.41	86	0.124
3	4.58	140	0.13
3	5.20	81	0.135
4	6.82	132	0.041
4	6.32	102	0.022
4	6.33	107	0.154

Таблица 1: Значения признаков для разных почерков.

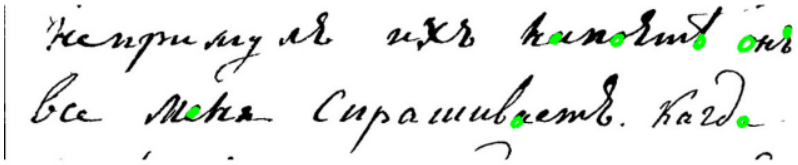
Анализируя полученные результаты, однозначно можно опредлить второй тип почерка по размеру кольцевых штрихов, эти значения самые большие среди рассматриваемых типов почерков. Также можно заметить, что по форме этих штрихов можно отличить, за исключением одного объекта, 3-ий тип почерка от 4-ого, по размеру - 1-ый от 4-ого и по частоте выступающих штрихов - 1-ый от 3-его. Что в итоге позволяет определить тип каждого отдельного почерка.

Рассмотрим поближе четвёртый объект из таблицы (1 тип почерка), который классифицируется неправильно (рис. 8). У данного фрагмента почерка кольцевых штрихов немного, в частности, слишком маленькая выборка базовых штрихов. Вследствие чего значениями признаков данный фрагмент классифицируется как почерк 3-его типа.

Из эксперимента можно сделать вывод, что построенные признаки являются чувствительными к оценке возможности классифицировать почерк на выделенных штрихах. В частности, специфика кольцевых штрихов была установлена как особенность каждого писателя, позволяющая различать разных авторов. Однако такой подход является переобученным и далеко не все признаки показывают значения, по которым можно примерно отличать каждый почерк от других. Требуется строить более сложные конструкции, учитывающие разные аспекты письма. Кроме того, обучение должно происходить на большом объёме изображений почерков, когда величины каждого признака почерка будут объективно отражать исследуемые характеристики. Вместе с тем входные изображения должны представлять собой хотя бы несколько строк текста, иначе алгоритм будет работать некорректно.



Кольцевые штрихи



Базовые штрихи

Рис. 8: Изображение, на котором происходит ошибка.