УДК 004.422+004.9

ПОИСК СЛОВ В РУКОПИСНОМ ТЕКСТЕ НА ОСНОВЕ ШТРИХОВОЙ СЕГМЕНТАЦИИ

И. Д. Морозов[1] [0009-0004-1813-3474], Л. М. Местецкий[2] [0000-0001-6387-167X]

[1]*Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия*

[2]*НИУ Высшая школа экономики, г. Москва, Россия*

[1]morozov-ivan-2003@yandex.ru, [2]mestlm@mail.ru

Аннотация

Рукописные архивные документы составляют фундаментальную часть культурного наследия человечества, однако их анализ остается трудоемкой задачей для профессиональных исследователей-историков, филологов и лингвистов. В отличие от коммерческих приложений систем OCR (Optical Character Recognition, оптического распознавания символов), работа с историческими рукописями требует принципиально иного подхода из-за чрезвычайного многообразия почерков, наличия правок и деградации материалов.

Предложен метод поиска в рукописных текстах, основанный на штриховой сегментации. Вместо полного распознавания текста, часто недостижимого для исторических документов, метод позволяет эффективно отвечать на поисковые запросы исследователей. Ключевая идея заключается в декомпозиции текста на элементарные штрихи, формировании семантических векторных представлений с помощью контрастного обучения, последующей кластеризации и классификации для создания адаптивного словаря почерка.

Экспериментально показано, что поиск сравнением кортежей редуцированных последовательностей наиболее информативных штрихов по расстоянию Левенштейна обеспечивает достаточное качество для рассматриваемой задачи. Метод демонстрирует устойчивость к индивидуальным особенностям почерка и вариациям написания, что особенно важно для работы с авторскими архивами и историческими документами.

© И. Д. Морозов, Л. М. Местецкий, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Предложенный подход открывает новые возможности для ускорения научных исследований в гуманитарной сфере, позволяя сократить время поиска нужной информации с недель до минут, что качественно меняет возможности исследовательской работы с большими архивами рукописных документов.

Ключевые слова: *рукописный текст, поиск, штриховый анализ, сегментация, векторное представление, контрастное обучение, кластеризация.*

ВВЕДЕНИЕ

Автоматизированный поиск в рукописных документах – одна из ключевых задач для исторических архивов. Однако традиционные OCR-системы демонстрируют низкую эффективность из-за нестандартности почерков, деградации носителей и сложной структуры рукописного текста [1, 2].

Перечислим существующие подходы и отметим их ограничения: оптическое распознавание символов (OCR) требует точного определения границ и форм символов, что трудно достижимо в рукописях [3]; поиск по визуальному сходству оперирует целыми фрагментами изображений, обладает высокой вычислительной сложностью и плохо масштабируется [4].

В настоящей работе предложен штриховой подход, который преодолевает указанные ограничения. Его основная идея – это декомпозиция текста на элементарные графические единицы (штрихи) и последующий анализ их устойчивых комбинаций
[5–7]. Это позволяет перейти от распознавания символов к выявлению структурных паттернов, специфичных для почерка.

Основные этапы метода:

- Сегментация текста на элементарные штрихи и их предобработка.
- Создание семантических эмбеддингов штрихов с помощью контрастного обучения.
- Кластеризация полученных семантических эмбеддингов для выделения основных типов штрихов и формирования «словаря почерка».
- Классификация штрихов документа, в котором будет проводиться поиск.
- Поиск сравнением кортежей редуцированных последовательностей наиболее информативных штрихов по расстоянию Левенштейна.

ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ

Штрих s задается ломаной линией из k точек (k для каждого штриха разное) в декартовой системе координат:

$$s = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}, \quad (x_m, y_m) \in R^2.$$

Текстовый запрос $q_{\text{text}} = c_1 c_2 \dots c_L$ длины L – это последовательность из L символов алфавита Σ. Здесь Σ – алфавит символов, из которых составляются запросы (например, буквы русского алфавита).

Имеется множество D изображений из N рукописных документов (фрагмент приведен на рис. 1), где каждый документ D_i представлен как кортеж S_i из M_i элементарных штрихов: $S_i = \{s_{i1}, s_{i2}, \dots, s_{iM_i}\}$.

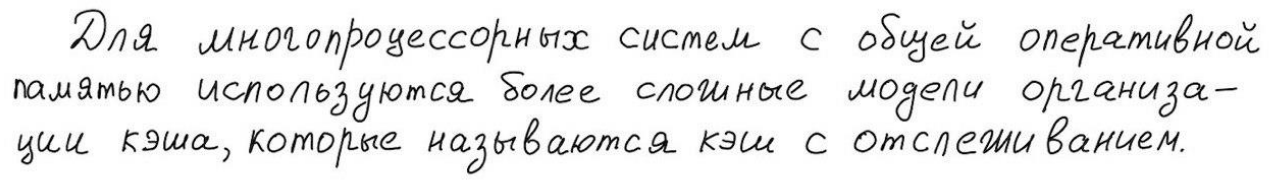


Рис. 1. Фрагмент изображения рукописного документа.

Введем обозначения:

- $C = \cup_{i=1}^{N} S_i$ – все штрихи всех документов;
- $Q_l \subseteq C$ – множество штрихов, соответствующих отдельному символу $c_l \in \Sigma$;
- $P(C)$ – множество всех подмножеств C.

Тогда запрос q, получающийся из текстового запроса q_{text}, представляет собой последовательность $q = (Q_1, Q_2, \dots, Q_L)$.

Для текстового запроса q_{text} необходимо найти такое отображение

$$F: D \times \prod_{l=1}^{L} P(C) \to P(C),$$

что

$$\forall D_j \in D \quad F(D_j, q) = \{ s \in S_j \mid s \text{ входит в } Q_l \}.$$

Таким образом, по множеству изображений рукописных документов и кортежу штрихов, составляющих запрос, нужно найти кортежи штрихов в документах, относящиеся к поисковому запросу.

ПРИНЦИП ПОИСКА И АЛГОРИТМ КЛАСТЕРИЗАЦИИ

Сегментация

Первым шагом является преобразование исходного изображения документа в набор элементарных штрихов. Для алгоритма поиска неважен метод сегментации. Мы применили подход [6], который анализирует топологию скелета текста. Метод выделяет связные компоненты, соответствующие отдельным движениям пера, и восстанавливает порядок их написания – выделяются элементарные штрихи в виде упорядоченных ломаных линий. Метод основан на анализе топологии скелета текста и обеспечивает сохранение порядка написания штрихов, идентификацию связных компонент как цепочек или циклов. Традиционные методы OCR плохо работают с рукописным текстом из-за вариативности почерков. Разбиение на штрихи позволяет перейти от распознавания целых символов к анализу их составляющих.

Предобработка штрихов

Полученные штрихи – это ломаные линии с переменным числом точек и различным расстоянием между ними. Поэтому выполняем аппроксимацию кубическими сплайнами – каждый штрих заменяется гладкой параметрической кривой, проходящей через все его точки. Это решает три проблемы:

- устраняет ломаный характер линии – настоящие штрихи гладкие;
- дает возможность получения гладкой линии после аугментации штриха – изменения координат составляющих его точек;
- позволяет единообразно семплировать точки (например, 100 точек на штрих).

Семплированные точки преобразуются в бинарные изображения 64×64. Это обусловлено входом выбранной архитектуры нейросети – ResNet [8]. Таким образом, предобработка позволяет превратить штрихи из ломаных в реальные сглаженные изображения следа пера (рис. 2).

Рис. 2. Преобразование штриха: исходная ломаная, сплайн и растеризованное изображение.

Построение признакового описания штрихов

Для построения эмбеддингов штрихов использована модифицированная архитектура ResNet-18. Одноканальный входной слой адаптирован для черно-белых изображений штрихов вместо стандартных RGB. Каждый блок содержит две свертки 3×3 с пакетной нормализацией и остаточным соединением, что предотвращает затухание градиентов в глубоких слоях. Два полносвязных слоя преобразуют 512-мерные признаки в 128-мерные векторы, которые затем нормируются и проектируются на единичную сферу соответствующей головой. Это позволяет сравнивать штрихи через косинусное расстояние. Модель обучается отличать похожие штрихи от непохожих с помощью контрастной функцией потерь NT-Xent Loss (Normalized Temperature-Scaled Cross Entropy). Для каждого штриха s генерируются две аугментированные версии s_i и s_j. Пример аугментации с добавлением шума и масштабированием показан на рис. 3. Их эмбеддинги v_i и v_j должны стать ближе в векторном пространстве, а эмбеддинги других штрихов – отдалиться.

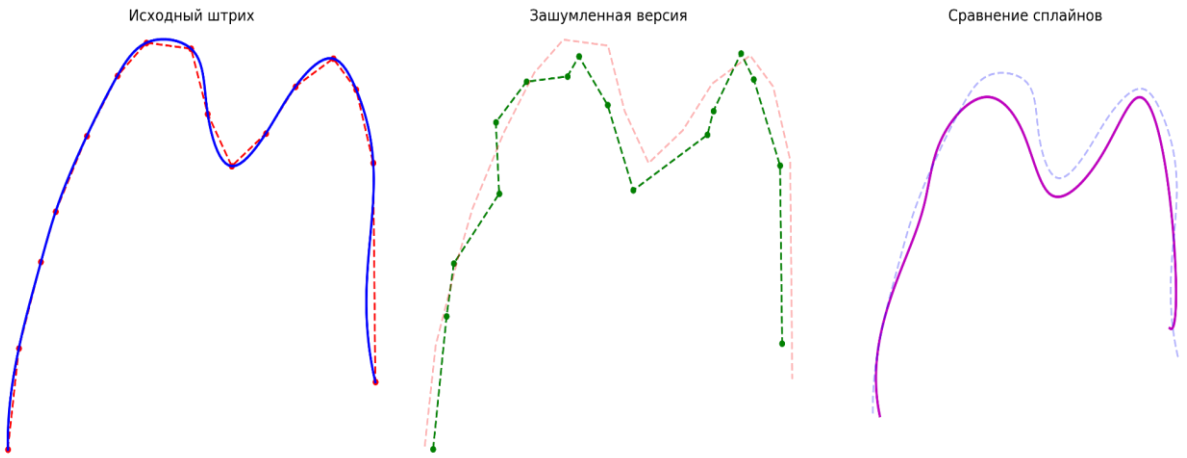


Рис. 3. Аугментация штриха: исходный штрих, зашумленные точки, сплайн аугментированного штриха.

Для батча из N штрихов с двумя аугментациями каждый (всего $2N$ эмбеддингов) функция потерь для i-го штриха вычисляется как

$$L_i = -\log \frac{\exp\left(\frac{v_i^\top v_j}{\tau}\right)}{\sum_{k=1}^{2N} 1_{k\neq i} \exp\left(\frac{v_i^\top v_k}{\tau}\right)},$$

где $v_i^\top v_j$ – косинусное сходство, τ (температура) – гиперпараметр, управляющий «резкостью» распределения (чем меньше τ, тем выше штраф за трудные негативные примеры), $1_{k\neq i}$ – индикатор, исключающий сравнение эмбеддинга с самим собой.

Матрица сходств $V^\top V$ строится для всех $2N$ векторов. Диагональные элементы (сравнение с собой) исключаются маской. Минимизация L сближает эмбеддинги аугментаций одного штриха и разводит разные штрихи. Если v_i и v_j – аугментации одного штриха, а v_k – другого, то

$$\exp\left(v_i^\top v_j\right) \gg \exp(v_i^\top v_k) \Longrightarrow L_i \to 0.$$

Сверточные сети эффективно ищут локальные геометрические паттерны, а остаточные связи позволяют обучать глубокие модели без переобучения. Это критично для работы с мелкими деталями штрихов. Принцип работы контрастного обучения проиллюстрирован на рис. 4 (двумерное пространство эмбеддин-

гов для наглядности): похожие штрихи (A и B) сближаются в эмбеддинг-пространстве, непохожий штрих (C) отдаляется от них, расстояние определяется косинусной мерой между векторами.

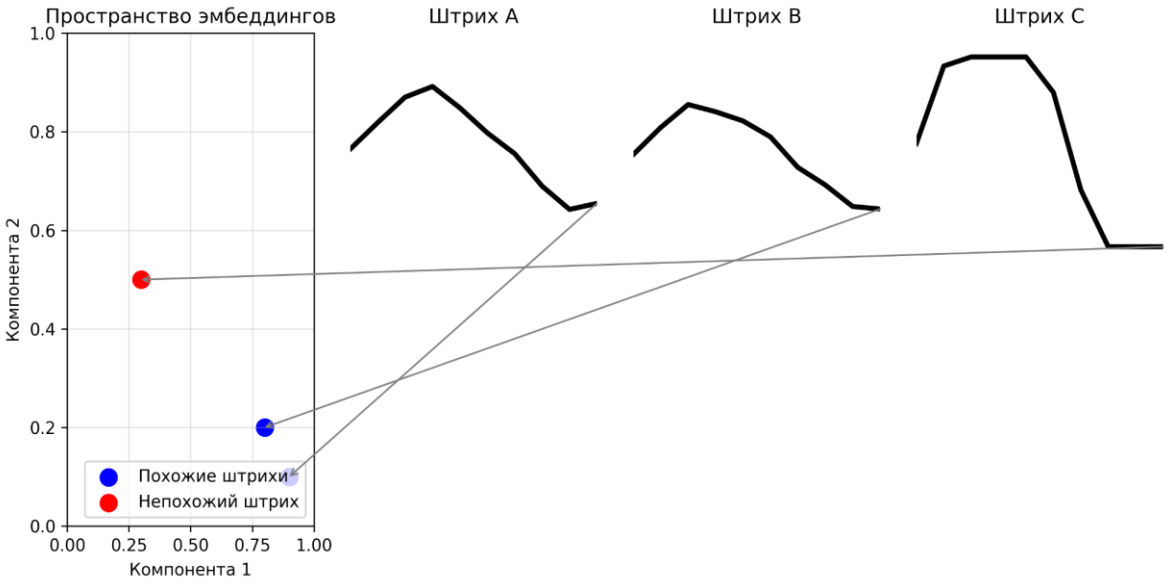


Рис. 4. Принцип контрастного обучения: похожие штрихи A и B; отличающийся штрих C.

Кластеризация и формирование словаря

После преобразования всего архива документов в эмбеддинги штрихов (всего M штрихов) $V = \{v_i\}_{i=1}^{M} \subset R^{128}$ осуществляется группировка семантически близких элементов с помощью алгоритма DBSCAN [9]. Этот метод был выбран из-за его способность и обнаруживать кластеры произвольной формы и идентифицировать выбросы (шум), что соответствует природе рукописных данных, где один и тот же штрих может иметь вариации. Метод основан на критерии плотности распределения точек в пространстве признаков: кластер формируется как максимальное множество точек, где каждая точка имеет не менее minPts соседей в ϵ-окрестности. Формально это свойство задается условиями

$$N_\epsilon(v_i) = \left\{ v_j \in V \middle| \parallel v_i - v_j \parallel_2 \leq \epsilon \right\}, \quad |N_\epsilon(v_i)| \geq \mathrm{minPts}.$$

Для каждого кластера C_k вычисляется эталонный вектор $\mu_k = \frac{1}{|C_k|}\sum_{v \in C_k} v$, играющий роль «цифрового прототипа» каллиграфического элемента. Множество центроидов $M = \{\mu_k\}_{k=1}^{K}$ образует базовый словарь системы. Поиск по запросу q изначально сводится к решению задачи многокритериальной оптимизации в пространстве эталонов. Сначала каждый символ c_i преобразуется в эмбеддинг e_i (или их комбинацию) через нейросетевую модель, после чего осуществляется поиск ближайших центроидов:

$$\mu_i^* = \arg\min_{\mu_k \in M} \left(1 - \frac{e_i\, \mu_k}{\|e_i\| \|\mu_k\|}\right), \quad i = 1, \ldots,.$$

На рис. 5b показано, как разные реализации одного штриха (синие, зеленые, серые, коричневые точки) группируются вокруг общего центра, а рис. 5c демонстрирует механизм поиска через сопоставление с ближайшими эталонами.

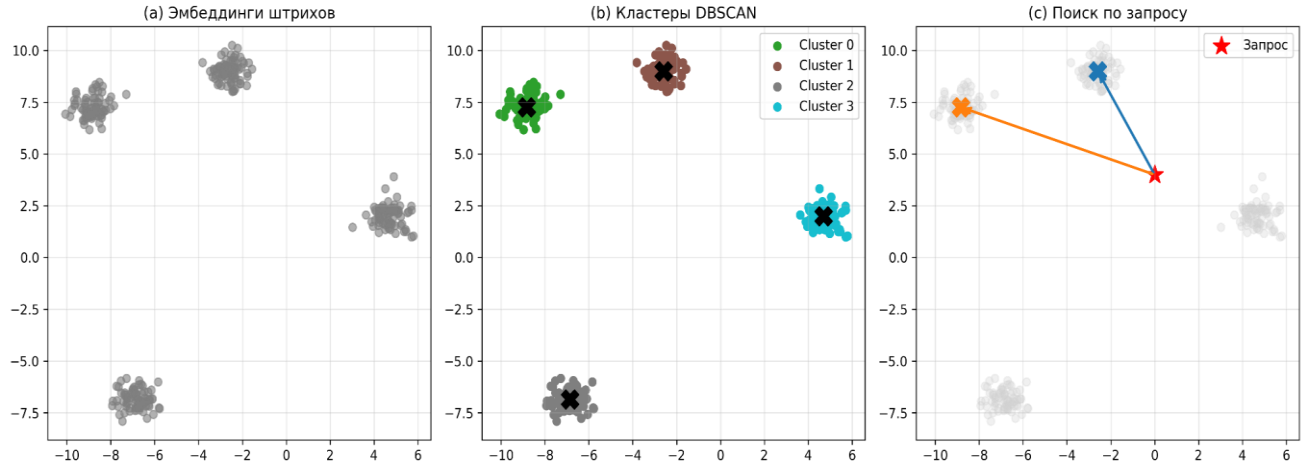


Рис. 5. Визуализация процесса кластеризации: (a) исходное распределение эмбеддингов; (b) выделенные кластеры с центрами (черные кресты); (c) поиск по запросу (красная звезда).

ЭКСПЕРИМЕНТЫ

Обучение модели и настройка кластеризации

Прежде чем перейти к поиску, необходимо убедиться в качестве векторных представлений и адекватности словаря.

Процесс обучения нейросетевой модели проводился на выборке из 10000 штрихов, полученных из рукописных конспектов. Каждый штрих предварительно

обрабатывался по схеме: сегментация → сплайновая аппроксимация → растеризация в изображение 64×64. Для аугментации применялись гауссово зашумление с $\sigma = 1$ и масштабирование в диапазоне $\alpha \in [0{,}9;\ 1{,}1]$.

Архитектура ResNet-18 обучалась с контрастной функцией потерь NT-Xent:

$$L = -\frac{1}{2N}\sum_{i=1}^{2N} \log \frac{\exp\left(\frac{v_i^{\top} v_j}{\tau}\right)}{\sum_{k \neq i} \exp\left(\frac{v_i^{\top} v_k}{\tau}\right)},$$

где температура $\tau = 0.5$ использована для калибровки шкалы сходств, а косинусная мера вычислялась между аугментированными парами. Оптимизация выполнялась алгоритмом Adam с параметром learning rate = $3 \cdot 10^{-4}$. Размер батча составлял 64 примера, каждый из которых содержал две аугментированные версии штриха. Обучение продолжалось 100 эпох с уменьшением потерь от 3.3576 до 3.0479, демонстрируя устойчивую сходимость модели (рис. 6). Характерные колебания потерь (например, локальный максимум 3,0865 на 50-й эпохе при общем тренде снижения) типичны для контрастных методов обучения. Общее время обучения составило 16 мин. на GPU NVIDIA A100.

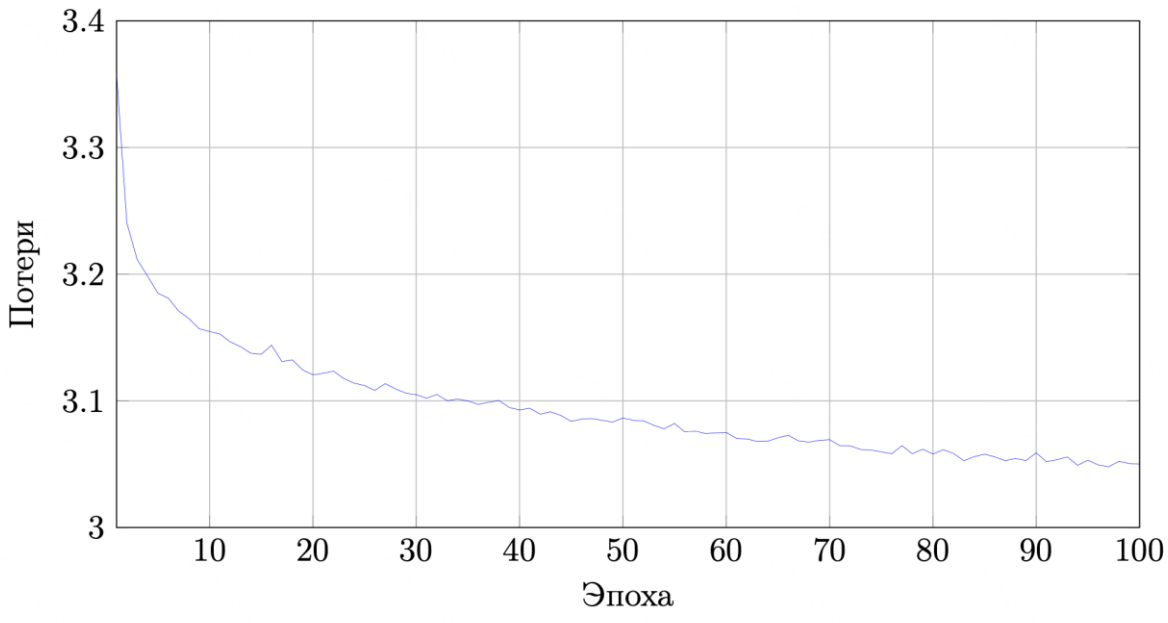


Рис. 6. Динамика функции потерь на обучении.

Для формирования базового словаря элементарных штрихов был проведен анализ зависимости числа кластеров K от параметра ϵ в алгоритме DBSCAN. График $K(\epsilon)$ (рис. 7) демонстрирует три следующих характерных режима.

Область малых значений $\boldsymbol{\epsilon} \in [0.7;\ 0.78)$. Алгоритм выделяет крупные кластеры, объединяя семантически различные штрихи. Например, при $\epsilon = 0.7$ все штрихи группируются всего в два кластера, что явно недостаточно для описания почерка.

Оптимальный диапазон $\boldsymbol{\epsilon} \in [0.78;\ 0.94]$. Рост $K(\epsilon)$ отражает обнаружение устойчивых структур. Каждый новый кластер соответствует уникальному типу штриха, удовлетворяющему условию:

$$|N_\epsilon(v)| \geq 6 \quad \text{для} \quad v \in C,$$

где $N_\epsilon(v)$ – соседи точки v в ϵ-окрестности. При $\epsilon \in [0.9;\ 0.94]$ достигается максимум $K = 10$, что соответствует набору элементарных компонент почерка.

Область больших значений $\boldsymbol{\epsilon} > 0.94$. Основная причина снижения K при $\epsilon > 0.94$ – это нарушение условия минимальной плотности. Согласно алгоритму DBSCAN кластером считается группа точек, где каждая точка имеет менее чем minPts соседей в ϵ-окрестности и все точки кластера достижимы через цепочку ϵ-соседей.

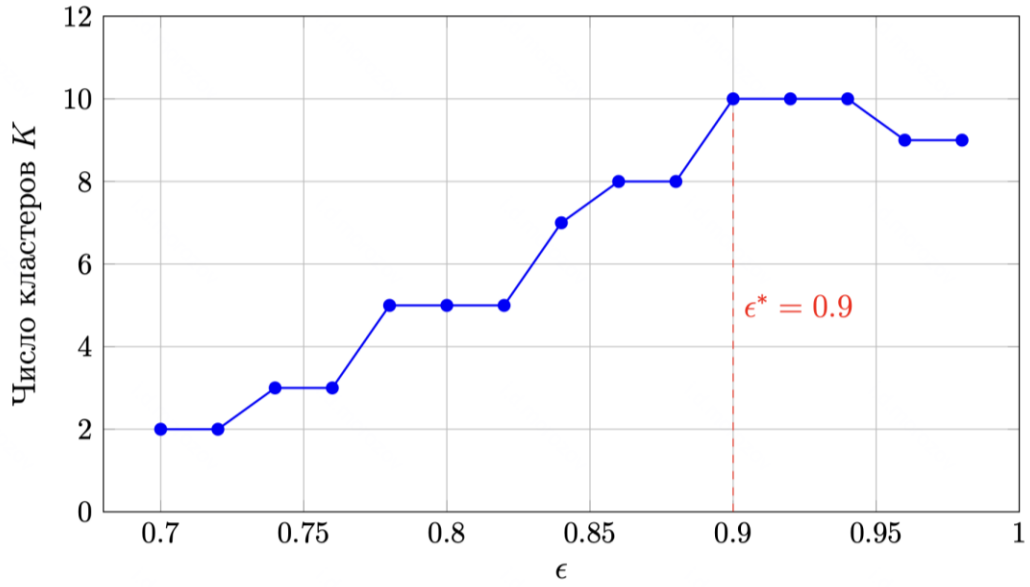


Рис. 7. График зависимости числа кластеров от ϵ. Пунктирной линией обозначено оптимальное значение параметра.

При увеличении ϵ происходит потеря связности – группы распадаются на подкластеры с менее чем minPts элементами. Например, рассмотрим кластер $C = \{v_1, \ldots, v_5\}$ из 5 точек при $\epsilon = 0.96$. Несмотря на выполнение условия связности,

$$\exists v_i \in C: |N_{0.96}(v_i)| = 4 < 6.$$

Согласно правилам DBSCAN весь кластер C классифицируется как шум.

Оптимальное значение $\epsilon^* = 0.92$ выбрано как точка, где достигается насыщение – дальнейшее увеличение ϵ не дает новых кластеров ($K(\epsilon) = 10$ при $\epsilon \geq 0.92$); сохраняется устойчивость: не менее 85% штрихов остаются в тех же кластерах при $\epsilon \pm 0.02$.

Сформированный словарь (см. рис. 8) содержит все необходимые элементы для декомпозиции рукописных символов. Таким образом, предложенная методика позволяет автоматически выделять структурные элементы почерка, обеспечивая баланс между детализацией и устойчивостью представления.

Рис. 8. Базовый словарь из 10 элементарных штрихов.

Модифицированный поиск по кортежам штрихов

Сформированный словарь позволяет искать слова, сопоставляя последовательности штрихов. Однако полная последовательность может быть избыточной и чувствительной к шуму.

В проведенном эксперименте была исследована возможность поиска рукописных слов с использованием редуцированных последовательностей штрихов. Основная гипотеза заключается в том, что для каждого текстового запроса существует оптимальная подпоследовательность штрихов, позволяющая обеспечить поиск при снижении влияния шума менее информативных штрихов. Штрихи, составляющие эти подпоследовательности, назовем главными. Обозначим эти подмножества размера k через $T_{\text{core}}^{(k)}$.

Для каждого типа штриха t_i (после отнесения к кластеру) вычислим дискриминативную силу на основе TF-IDF:

$$DS(t_i) = TF(t_i, q) \times IDF(t_i),$$

где

- N – общее количество штрихов в запросе,
- n_i – количество вхождений штриха t_i в запрос,
- $TF(t_i, q) = n_i/N$ – относительная частота штриха t_i в запросе q,
- $|D|$ – общее количество строк во множестве документов,
- $|\{d \in D: t_i \in d\}|$ – количество строк, содержащих штрих t_i,
- $IDF(t_i) = \log \frac{|D|}{|\{d \in D: t_i \in d\}|}$.

Для запроса q и документа d построим редуцированные последовательности штрихов:

$$q_{\text{core}}^{(k)} = \left\{t \in q: t \in T_{\text{core}}^{(k)}\right\}, \quad d_{\text{core}}^{(k)} = \left\{t \in d: t \in T_{\text{core}}^{(k)}\right\}.$$

Решение о наличии слова в документе примем на основе порогового значения нормализованного расстояния Левенштейна:

$$\text{match} = \left\{1, \text{если } \frac{\text{Lev}(q_{\text{core}}, d_{\text{core}})}{\max(|q_{\text{core}}|, |d_{\text{core}}|)} \leq \theta;\ 0 - \text{ иначе}\right\},$$

где θ – пороговое значение. Размер окна выберем, исходя из длины оригинальной (нередуцированной) последовательности штрихов.

Для эксперимента были выбраны слова различной длины и частоты встречаемости в тестовом документе объемом 15267 штрихов. Процедура эксперимента включала последовательное добавление штрихов в поисковый шаблон в порядке убывания дискриминативной силы и вычисление показателей качества для каждого k.

Эксперимент показал, что для всех тестовых слов существует оптимальное подмножество штрихов, обеспечивающее показательное по полноте (Recall) качество поиска при редукции признакового пространства (см. табл. 1).

Табл. 1. Оптимальное k и оценки качества.

Слово	Оптимальное k (количество штрихов)	Точность (Precision)	Полнота (Recall)
«конъюнкция»	6	0.56	0.75
«доказательство»	7	0.6	0.67
«следовательно»	7	0.67	0.83
«теорема»	5	0.45	0.8

Наблюдались характерный рост эффективности с добавлением наиболее информативных штрихов и последующее насыщение после достижения оптимального подмножества, что свидетельствует о наличии небольшого набора высокоинформативных признаков.

Предложенный метод продемонстрировал несколько ключевых преимуществ. Вычислительная эффективность достигается за счет редукции признакового пространства, что особенно важно при работе с крупными архивами рукописных документов. Адаптивность метода позволяет автоматически определять оптимальное подмножество штрихов для каждого слова на основе объективных метрик, без необходимости ручной настройки. Устойчивость к вариациям почерка обеспечивается за счет использования главных штрихов, которые, как пра-

вило, остаются стабильными при различных стилях написания. Снижается чувствительность к «лишним штрихам», вносящим шум. Это свойство особенно ценно при работе с историческими документами, где часто встречаются индивидуальные особенности почерка.

Практическая значимость метода состоит в возможности создания эффективных систем поиска для крупных архивов рукописных текстов. Сокращение вычислительной сложности позволяет масштабировать систему для работы с коллекциями, содержащими миллионы штрихов, что открывает новые возможности для цифровой археографии и исторических исследований.

Следует отметить, что метод имеет определенные ограничения. Качество поиска зависит от точности сегментации текста на элементарные штрихи, которая может оказаться низкой для документов плохого качества или со сложной структурой. Кроме того, для новых слов требуется предварительный анализ дискриминативной силы штрихов, что добавляет этап обучения в рабочий процесс.

ЗАКЛЮЧЕНИЕ

Предложен и экспериментально обоснован штриховой подход к поиску в рукописных документах. Ключевым достижением является создание полноценной системы, которая:

- автоматически сегментирует текст на элементарные штрихи;
- строит их семантические эмбеддинги с помощью контрастного обучения;
- формирует адаптивный словарь типичных штрихов методами кластеризации без учителя;
- реализует механизм поиска, основанный на классификации штрихов и анализе редуцированных последовательностей классов штрихов.

Показана возможность редукции признакового пространства (штрихов) для достижения приемлемой точности поиска, что открывает пути к созданию более быстрых и эффективных алгоритмов. Универсальность алгоритма (независимость от языка и типа документа) расширяет область его применения.

Перспективы дальнейших исследований включают разработку адаптивных алгоритмов автоматического определения оптимального размера подмножества штрихов, интеграцию контекстной информации для улучшения точности поиска,

а также применение методов глубокого обучения для более точной оценки дискриминативной силы штрихов.

Благодарность

Работа поддержана грантом РНФ №22-68-00066 «Культурное наследие России: интеллектуальный анализ и тематическое моделирование корпуса рукописных текстов».

СПИСОК ЛИТЕРАТУРЫ

1. *Zhang X.-Y., Sun Z., Jin L., Ni H. & Lyons T. J.* Learning Spatial–Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. Vol. 40, no. 8. P. 1903–1917.
https://doi.org/10.1109/tpami.2017.2732978

2. *Rahal N., Vögtlin L., Ingold R.* Historical Document Image Analysis Using Controlled Data for Pre-training // International Journal on Document Analysis and Recognition (IJDAR). 2023. Vol. 26, no. 3. P. 241–254.
https://doi.org/10.1007/s10032-023-00437-8

3. *Puigcerver J.* Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2017. Vol. 1. P. 67–72.
https://doi.org/10.1109/ICDAR.2017.20

4. *Rath T. M., Manmatha R.* Word Spotting for Historical Documents // International Journal on Document Analysis and Recognition (IJDAR). 2007. Vol. 9, no. 2–4. P. 139–152.
https://doi.org/10.1007/s10032-006-0027-8

5. *Местецкий Л.М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. М.: ФИЗМАТЛИТ. 2009. 231 с.

6. *Mestetskiy L.M.* Stroke Segmentation of Handwritten Text Based on Medial Representation // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. 2024. Vol. 34, no. 4. P. 1185-1191.
https://doi.org/10.1134/S1054661824701256

7. *Dias C. da S., Britto Jr. A. de S., Barddal J. P., Heutte L., Koerich A. L.* Pattern Spotting and Image Retrieval in Historical Documents using Deep Hashing. 2022. arXiv:2208.02397

8. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778. https://doi.org/10.1109/CVPR.2016.90

9. *Ester M., Kriegel H.-P., Sander J., Xu X.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // 2nd International Conference on Knowledge Discovery and Data Mining (KDD). 1996. P. 226–231.

WORD SEARCH IN HANDWRITTEN TEXT BASED ON STROKE SEGMENTATION

I. D. Morozov[1] [0009-0004-1813-3474], L. M. Mestetskiy[2] [0000-0001-6387-167X]

[1, 2]*Lomonosov Moscow State University, Moscow, Russia*

[2]*Higher School of Economics, Moscow, Russia*

[1]morozov-ivan-2003@yandex.ru, [2]mestlm@mail.ru

Abstract

Handwritten archival documents form a fundamental part of humanity's cultural heritage. However, their analysis remains a labor-intensive task for professional researchers, such as historians, philologists, and linguists. Unlike commercial OCR applications, working with historical manuscripts requires a fundamentally different approach due to the extreme diversity of handwriting, the presence of corrections, and material degradation.

This paper proposes a method for searching within handwritten texts based on stroke segmentation. Instead of performing full text recognition, which is often unattainable for historical documents, this method allows for efficiently answering researcher search queries. The key idea involves decomposing the text into elementary strokes, forming semantic vector representations using contrastive learning, followed by clustering and classification to create an adaptive handwriting dictionary.

It is experimentally shown that search by comparing tuples of reduced sequences of the most informative strokes using the Levenshtein distance provides sufficient quality for the task at hand. The method demonstrates resilience to individual handwriting characteristics and writing variations, which is particularly important for working with authors' archives and historical documents.

The proposed approach opens up new possibilities for accelerating scientific research in the humanities, reducing the time required to find relevant information from weeks to minutes, thereby qualitatively transforming research capabilities when working with large archives of handwritten documents.

Keywords: *handwritten text, search, stroke analysis, segmentation, vector representation, contrastive learning, clustering.*

REFERENCES

1. *Zhang X.-Y., Sun Z., Jin L., Ni H. & Lyons T. J.* Learning Spatial–Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. Vol. 40, no. 8. P. 1903–1917.
https://doi.org/10.1109/tpami.2017.2732978

2. *Rahal N., Vögtlin L., Ingold R.* Historical Document Image Analysis Using Controlled Data for Pre-training // International Journal on Document Analysis and Recognition (IJDAR). 2023. Vol. 26, no. 3. P. 241–254.
https://doi.org/10.1007/s10032-023-00437-8

3. *Puigcerver J.* Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2017. Vol. 1. P. 67–72.
https://doi.org/10.1109/ICDAR.2017.20

4. Rath T. M., Manmatha R. Word Spotting for Historical Documents // International Journal on Document Analysis and Recognition (IJDAR). 2007. Vol. 9, no. 2–4. P. 139–152. https://doi.org/10.1007/s10032-006-0027-8

5. *Mestetskii L.M.* Continuous Morphology of Binary Images: Figures, Skeletons, Circulars. M.: FIZMATLIT, 2009. 231 p.

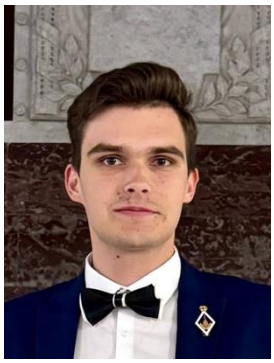
6. *Mestetskiy L.M.* Stroke Segmentation of Handwritten Text Based on Medial Representation // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. 2024. Vol. 34, no. 4. P. 1185-1191. https://doi.org/10.1134/S1054661824701256

7. *Dias C. da S., Britto Jr. A. de S., Barddal J. P., Heutte L., Koerich A. L.* Pattern Spotting and Image Retrieval in Historical Documents using Deep Hashing. 2022. arXiv:2208.02397

8. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778. https://doi.org/10.1109/CVPR.2016.90

9. *Ester M., Kriegel H.-P., Sander J., Xu X.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // 2nd International Conference on Knowledge Discovery and Data Mining (KDD). 1996. P. 226–231.

СВЕДЕНИЯ ОБ АВТОРАХ



МОРОЗОВ Иван Дмитриевич – магистр кафедры Математические методы прогнозирования факультета вычислительной математики и кибернетики МГУ имени М. В. Ломоносова. Область научных интересов: машинное обучение, распознавание рукописных текстов, математика.

Ivan Dmitrievich MOROZOV – Master's student at the Department Mathematical Forecasting Methods of the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University. Research interests: machine learning, handwriting recognition, mathematics.

email: morozov-ivan-2003@yandex.ru

ORCID: 0009-0004-1813-3474

МЕСТЕЦКИЙ Леонид Моисеевич – доктор технических наук, академик РАЕН, профессор кафедры математических методов прогнозирования МГУ. Научные интересы: вычислительная геометрия, анализ и распознавание изображений.

Leonid Moiseevich MESTETSKIY – Doctor of Engineering Sciences, Academician of the Russian Academy of Natural Sciences, Professor at the Department of Mathematical Forecasting Methods at Moscow State University. His research interests include computational geometry and image analysis and recognition.

email: mestlm@mail.ru

ORCID: 0000-0001-6387-167X

Материал поступил в редакцию 2 ноября 2025 года