МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Сюй Минчуань

Анализ и распознавание иероглифических текстов на основе непрерывных морфологических моделей

Специальность 1.2.1 —
«Искусственный интеллект и машинное обучение»

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор технических наук, профессор
Местецкий Леонид Моисеевич

Москва — 2025

Оглавление

Стр.

Стр.

Введение

Оптическое распознавание символов – это технология, которая преобразует цифровые изображения рукописного или печатного текста в машинный текст для дальнейшего хранения и обработки. В последние годы были разработаны передовые методы работы с изображениями печатных текстов, но лишь небольшая часть исследований посвящена распознаванию текстов из рукописных архивов. В то же время, древние иероглифы в рукописных текстах имеют огромную исследовательскую ценность как гуманитарное наследие, фиксирующее историческую культуру, язык, развитие науки.

Поиск информации и навигация по запросам в огромном количестве древних документов всегда остаются одной из центральных и актуальных проблем в исследовании древних текстов. Задача поиска состоит в нахождении изображения отдельного заданного иероглифа, называемого запросом, в большом множестве изображений иероглифов, называемом файлом. Целью является локализация в файле всех иероглифов, совпадающих с запросом. В распознавании образов задача использования одного образца для распознавания в литературе часто называется *однократным обучением.*

Применение современных методов распознавания, использующих машинное обучение и глубокие нейронные сети, сталкивается с большими трудностями при подготовке обучающих данных по древнекитайским иероглифам:

- **Огромный словарь письменности**. Для обучения требуется очень большой объем данных, поскольку очень велико число символов древнекитайской письменности. В настоящее время общее число уникальных иероглифов, встречающихся в исторических источниках, превышает 6×10^4, включая устаревшие и редко используемые.
- **Сложность и вариативность документов**. Сложность структуры китайских иероглифов в сочетании с вариативностью, обусловленной индивидуальными стилями почерка, деградацией бумаги и неравномерностью штрихов создает существенные трудности для традиционных методов извлечения и сопоставления признаков. Сбор обучающей выборки, захватывающей все виды и стили документов практически невозможно.

- **Высокая трудоёмкость аннотирования**. Весьма велики трудозатраты на аннотирование символов даже для специалиста в области древних иероглифов. На распознавание символов и извлечение ключевой информации из нескольких страниц древних рукописей тратятся несколько часов или даже несколько дней.

Кроме того, как для обучения, так и для практического применения требуются современные ускорители, стоимость их использования относительно высока. Современные нейронные сети не дают возможности интерпретировать лежащий в основе распознавания процесс. Но понимание мотивов формируемого в сети решения очень важно для исследователей, поскольку это помогает лучше понять структуру символов, облегчает адаптацию к другим языкам и пути совершенствования алгоритма распознавания.

Таким образом, существует острая потребность в **интерпретируемых**, **вычислительно устойчивых** и **обучением-независимых** методах сравнения графических форм иероглифов для задач поиска и навигации в архивных коллекциях, которые могут сильно упростить работу исторических исследователей при минимальным требовании к ресурсам, и этот факт определяет актуальность данного исследования.

Целью диссертационной работы является разработка математического и программного обеспечения для решения задач поиска и навигации в массивных древнекитайских рукописях с использованием методов непрерывных морфологических моделей и методов определения меры сходства.

Объектом исследования диссертационной работы являются модели построения точечных признаковых описаниях и модели вычисления меры сходства между признаками изображений иероглифов. Предметом исследования диссертационной работы является разработка алгоритмов построения моделей генерации точечных признаков и оптимизационных моделей получения меры схожести, применимых к древнекитайским рукописным документам.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Формирование состава задач автоматизации работы с цифровыми архивами рукописных иероглифических документов.
2. Структурный анализ рукописных китайских иероглифов и получение признакового описания с использованием медиального представления

формы многоугольной фигуры, полученное на основе диаграммы Вороного.

3. Построение меры сходства рукописных изображений иероглифов и разработка алгоритма его вычисления.
4. Практическая реализация разработанных методов построения моделей для контекстного поиска в больших рукописных архивах по ключевым словам и проведение экспериментов для проверки корректности полученных результатов.

Диссертация соответствует специальности 1.2.1 «Искусственный интеллект и машинное обучение» в части направления разработки методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных, поскольку целью данной работы является исследование, разработка методов для анализа и обработки изображений древнекитайских рукописей и текстов естественного языка, содержащих в документах рукописей.

Научная новизна:

1. Предложены новые математические модели, позволяющие эффективно анализировать форму древного иероглифа на основе непрерывных морфологиях, и представляют признаковое описание изображения иероглифа в виде конечного множества критичных точек, характеризующих топологическое и геометрическое свойство иероглифа.
2. Предложены модели для сравнения иероглифических точечных признаков, полученных предложенными моделями построения точечных признаков, по принципу сопоставления соответствия пар критичных точек.
3. Разработана оптимизационная модель через задачу линейного программирования для вычисления количественной меры сходства между иероглифами.

Научная и практическая значимость. Научная значимость заключается в разработке методов построения признакового описания древнекитайского рукописного иероглифа в виде конечного набора критичных точек через медиальное представления формы бинарного изображения, также в разработке методов сравнения форм на основе оптимизационной модели линейного программирования, используя построенные точечные признаки иероглифов.

Предложенные подходы позволяют построить представление иероглифов прозрачным математическим аппаратом с вычислительно эффективной процедурой, в то же время семантично достаточно для последующего сравнения форм иероглифов по принципу сопоставления точек.

Практическая значимость состоит в детальной программной реализации всех предложенных методов, а также их приложений для решения различных прикладных задач, связанных с поиском информации в архивных документах. Апробация реализации разработанных методов проводилась на реальных данных отсканированных древнекитайских документов. По результатам экспериментального исследования, предложенные методы сравнимы по качеству с существующими методами распознавания иероглифических изображений, но имеют минимальное требование к ресурсам поскольку не нуждаются в тяжелой процедуре обучения и сбора бучающей выборки, требующего значительное человеческое усилие, и преимущество интерпретируемости для облегчения адаптации к другим языкам и задачам распознавания.

Методология и методы исследования. При получении основных результатов диссертационной работы использовались методы обработки и анализа изображений, методы вычислительной геометрии, теория графов, методы оптимизации. Работа носит экспериментально-теоретический характер. Разработка программного кода велась на языке Python с использованием библиотеки скелетизации, разработанной Л.М. Местецким. Эксперименты проводились на модельных данных и открытых базах изображений иероглифов.

Основные положения, выносимые на защиту:

1. Методы построения признакового описания для древнекитайского иероглифа в виде конечного набора критичных точек на основе медиального представления и непрерывных морфологий. Предложен метода выбора критичных точек с учетом максимальной кривизной.
2. Методы построения оптимизационной модели для сравнения точечных признаков иероглифа. Разработаны методы получения меры близости путем решения задачи о назначениях, которые могут быть реализованы эффективной процедурой вычисления.
3. Обоснование работоспособности предложенных методов путём реализации программного комплекса. Экспериментально доказано, что разработанные методы не уступают по качеству существующим современным методам с использованием нейронных сетей, при этом

обладают преимуществом отсутствия требования обучения на гиганской обучающей выборки, низкого запроса к вычислительным ресурсам и интерпретируемости алгоритма.

Все результаты, выносимые на защиту, получены автором самостоятельно под руководством научного руководители Л.М.Местецкого.

Достоверность полученных результатов обеспечивается проведенными экспериментами, корректным тестированием разработанных решений, публикациями в рецензируемых журналах и апробацией на российских и международных конференциях.

Апробация работы. Основные результаты работы докладывались на:

- Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2025» (Россия, Москва, 2025).
- 6th International workshop on Photogrammetric techniques for environmental and infrastructure monitoring, Biometry and Biomedicine (Россия, Москва, 2025).
- 3rd International Conference on Machine Intelligence and Digital Applications (Китай, Сиань, 2026).
- ??? Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2026» (Россия, Москва, 2026).

Личный вклад. Автор принимал активное участие в выполнении основного объема теоретических и экспериментальных исследований, а также в разработке реализации разработанных методов. Подготовка части материалов к публикации проводилась совместно с соавторами, причем вклад диссертанта был определяющим. Диссертационное исследование является самостоятельным и законченным трудом автора.

Публикации. Основные результаты по теме диссертации изложены в XX печатных изданиях, X из которых изданы в журналах, рекомендованных ВАК, X — в тезисах докладов.

Объем и структура работы. Диссертация состоит из введения, 4 глав, заключения и 3 приложений. Полный объём диссертации составляет 18 страниц, включая 0 рисунков и 0 таблиц. Список литературы содержит 0 наименований.

Глава 1. Обзор существующих методов и математических средств

1.1 Обзор методов распознавание иероглифов

Мы можем сделать **жирный текст** и *курсив*.

1.2 Основы математических средств

Глава 2. Методы построения точечного признакового описания иероглифических текстов

2.1 Генерация признаков по аналоги созвездий

2.2 Построение медианной оси и скелетного графа

2.3 Получение описания иероглифа конечным набором точек

2.3.1 Выбор критичных точек с максимальной кривизной

2.3.2 Метод нормализации

2.3.3 Методы борьбы с аномальными точками: смерживание близких по расстоянию вершин для снижения размерности задачи (merging_nodes), удаление мелких связанных компонент

Глава 3. Методв построения меры сходства изображений иероглифических текстов

3.1 Модель минимальной стоимости

3.1.1 Проверка соединения соответвующих вершин через ребра (connectivity checking)

3.1.2 Минимальная стоимость по ребрам

3.2 Модель максимального количества близких пар

3.3 Сравнение и обсуждение

Глава 4. Практические приложения контекстного поска

4.1 Методы ускорения алгоритма

4.1.1 Распараллеливание сравнений изображений

Некоторый текст.

4.1.2 Предварительный подсчет признаков изображений

4.2 Задача однократного распознавания

4.3 Поисковый запрос в рукописных документах

Заключение

Основные результаты работы заключаются в следующем.

1. На основе анализа ...
2. Численные исследования показали, что ...
3. Математическое моделирование показало ...
4. Для выполнения поставленных задач был создан ...

И какая-нибудь заключающая фраза.

Последний параграф может включать благодарности. В заключение автор выражает благодарность и большую признательность научному руководителю Иванову И. И. за поддержку, помощь, обсуждение результатов и научное руководство. Также автор благодарит Сидорова А. А. и Петрова Б. Б. за помощь в работе с образцами, Рабиновича В. В. за предоставленные образцы и обсуждение результатов, Занудятину Г. Г. и авторов шаблона *Russian-Phd-LaTeX-Dissertation-Template* за помощь в оформлении диссертации. Автор также благодарит много разных людей и всех, кто сделал настоящую работу автора возможной.

Словарь терминов

TeX : Система компьютерной вёрстки, разработанная американским профессором информатики Дональдом Кнутом

панграмма : Короткий текст, использующий все или почти все буквы алфавита

Список рисунков

Список таблиц

Приложение А

Примеры вставки листингов программного кода

Приложение Б

Очень длинное название второго приложения, в котором продемонстрирована работа с длинными таблицами

Б.1 Очередной подраздел приложения

Нужно больше подразделов приложения!

Б.2 И ещё один подраздел приложения

Нужно больше подразделов приложения!