УДК 004.93'1:912.43:81'373.21
DOI 10.52575/2712-7443-2025-49-1-128-145

Ранний этап в методологии автоматического распознавания топонимов на географических картах и в текстах

Дмитриев А.В.
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, г. Санкт-Петербург, вн. тер. г. муниципальный округ Академическое, ул. Политехническая, 29Б
avd84@list.ru

Аннотация. В статье исследуется эволюция методологических подходов к автоматическому распознаванию топонимов в картографических материалах и текстах с конца 1980-х до середины 2010-х годов. Анализируются ключевые исследования 2000-х годов, заложившие основу современных методов компьютерного зрения и машинного обучения в данной области. Особое внимание уделяется проблемам распознавания топонимов на исторических картах. Рассматриваются математические методы и алгоритмические решения, их эффективность и ограничения. Исследование демонстрирует методологическую преемственность в развитии данного направления и актуальность классических подходов в контексте современных технологий искусственного интеллекта.

Ключевые слова: исторические карты, топонимы, автоматическое распознавание, компьютерное зрение, машинное обучение

Для цитирования: Дмитриев А.В. 2025. Ранний этап в методологии автоматического распознавания топонимов на географических картах и в текстах. Региональные геосистемы, 49(1): 128–145. DOI: 10.52575/2712-7443-2025-49-1-128-145

The Early Stage in the Methodology of Automatic Toponym Recognition on Geographic Maps and in Texts

Alexander V. Dmitriev
Peter the Great St. Petersburg Polytechnic University
29B Polytechnicheskaya St, municipal district Academicheskoe,
St. Petersburg 195251, Russia
avd84@list.ru

Abstract. The article examines the evolution of methodological approaches to automatic toponym recognition in cartographic materials and texts from the late 1980s to the mid-2010s. It analyzes key research works of the 2000s that laid the foundation for modern computer vision and machine learning methods in this field. Special attention is paid to the challenges of toponym recognition on historical maps, including issues of homonymy, temporal ambiguity, and multilingualism. The study investigates mathematical methods and algorithmic solutions from various research groups, evaluating their effectiveness and limitations in handling complex cartographic materials. The works by Smith and Crane, Gelbukh and Levachkine, Pouderoux, and other scientists who made significant contributions to methodology development are analyzed in detail. Particular emphasis is placed on the three-phase approach to semi-automatic toponym recognition and the specific challenges encountered when processing historical maps. The research demonstrates methodological continuity in the development of this direction and highlights the enduring relevance of classical approaches in the context of modern artificial intelligence technologies, suggesting ways to integrate traditional methods with contemporary neural network architectures.

Keywords: historical maps, toponyms, automatic recognition, computer vision, machine learning

For citation: Dmitriev A.V. 2025. The Early Stage in the Methodology of Automatic Toponym Recognition on Geographic Maps and in Texts. Regional Geosystems, 49(1): 128–145 (in Russian). DOI: 10.52575/2712-7443-2025-49-1-128-145

Введение

Современные тенденции в развитии методологии автоматического распознавания и извлечения топонимов указывают на растущую важность интеграции различных источников данных и методов анализа, поэтому многие современные работы демонстрируют постепенный переход от простых методов компьютерного зрения к комплексным решениям, интегрирующим различные подходы и инструменты [Вицентий и др. 2020; Колесников и др. 2020; Вицентий, Шишаев, 2021].

Особенно перспективным представляется направление, связанное с анализом временных изменений топонимов и учетом исторического контекста, а также развитие методов обработки многословных топонимов и учет временного аспекта изменений географических названий. При этом остаются актуальными проблемы производительности алгоритмов, особенно при обработке больших массивов текстовой информации, в том числе современных и исторических карт [Milleville et al., 2020; Schlegel, 2021; Lenc et al. 2022; Olson et al. 2024]. Тема больших данных как инструмента в топонимических исследованиях также начинает выходить на первый план [Красовский, 2024].

Не теряют своей новизны практические решения, посвященные интеграции картографических методов в топонимические исследования [Гордова и др., 2021], что в сущности имеет давнюю традицию и описано во множестве работ (см. например: [Поспелов, 1971]).

Не менее важной является проблема геоинформационного обеспечения проектов сохранения культурного и природного наследия. В этом смысле становится многообещающим изучение топонимии культовых сооружений в историко-географической перспективе с опорой на картографические методы [Герцен и др., 2023].

Революционные достижения в области архитектур глубокого обучения, в частности появление моделей на основе механизма внимания и трансформерных архитектур [Peters et al., 2018], открыли новые перспективы в решении данной проблемы. Особый исследовательский интерес представляет развитие нейросетевых подходов к обработке естественного языка, продемонстрировавших значительный прогресс в задачах распознавания именованных сущностей [Wang et al., 2020; Zhou et al., 2023].

В данном исследовании анализируется ранний этап методологических подходов к автоматическому распознаванию топонимов в картографических материалах и текстах с конца 1980-х до середины 2010-х годов, с особым фокусом на ключевых исследованиях 2000-х годов, заложивших фундамент для современных методов компьютерного зрения и машинного обучения в данной области. С этой целью выявляются и систематизируются основные методологические подходы к автоматическому распознаванию топонимов в рассматриваемый период; анализируется специфика применения различных математических методов и алгоритмов в ключевых исследованиях; сопоставляется эффективность различных методологических решений на основе статистических показателей точности распознавания топонимических референций; определяются характерные проблемы и ограничения методов автоматического распознавания топонимов на исторических картах; исследуется вклад российских и европейских научных школ в развитие методологии автоматического распознавания топонимов.

Актуальность исследования обусловлена необходимостью систематизации и критического осмысления методологического опыта в области автоматического распознавания топонимов, накопленного за последние десятилетия. Особую значимость данная про-

блематика приобретает в контексте существующего разрыва между российскими и европейскими исследованиями, а также недостаточного освещения в русскоязычной научной литературе ключевых зарубежных работ в этой области. Фундаментальные проблемы автоматического распознавания топонимов, включая вопросы омонимии, исторической изменчивости названий и многоязычия, требуют комплексного анализа существующих методологических решений. Современное развитие технологий искусственного интеллекта и методов глубокого обучения создает потребность в переосмыслении классических подходов к распознаванию топонимов для их эффективной интеграции с новейшими алгоритмическими решениями. Все это определяет необходимость тщательного изучения данной проблематики как основы для совершенствования существующих и разработки новых методов в этой области.

Объекты и методы исследования

В работе использован комплекс взаимодополняющих методов научного исследования. Историко-генетический метод применяется для анализа эволюции подходов к распознаванию топонимов с конца 1980-х до середины 2010-х годов, что позволяет проследить развитие методологической базы в исторической перспективе. Сравнительный анализ используется при сопоставлении различных алгоритмических и методологических решений, их эффективности и ограничений. Системный анализ дает возможность рассмотреть комплексные подходы к распознаванию топонимов как целостных систем взаимосвязанных элементов. Метод «кейс-стади» применяется при детальном разборе конкретных примеров использования методов распознавания топонимов, что позволяет выявить специфические особенности их практического применения. Количественный анализ используется для оценки эффективности различных методов на основе статистических показателей точности распознавания и других метрик производительности.

Критический обзор магистральных работ. На протяжении 2000-х гг. процесс автоматической обработки топонимов представлял собой комплексную систему, работающую параллельно с двумя типами источников данных – текстовыми документами и картами. При обработке текстовых документов сначала выполняется распознавание именованных сущностей, включающее обнаружение потенциальных топонимов и их классификацию, после чего происходит разрешение топонимов путем устранения неоднозначности и связывания с базами данных. При работе с картами сначала осуществляется детекция топонимов через сегментацию изображения, обнаружение текстовых областей и группировку элементов, затем выполняется распознавание текста с применением *OCR* для печатного текста, *HTR* для рукописных надписей и учетом исторических шрифтов. Далее оба потока объединяются на этапе нормализации топонимов, где производится стандартизация написания, учет исторических вариантов и обработка разных языков. Завершающим этапом становится верификация и валидация, включающая проверку по газеттирам и пространственный анализ. Таким образом обеспечивается полный цикл обработки географических названий с их стандартизацией и проверкой достоверности полученных результатов (рис. 1).

Исследования в области автоматического распознавания топонимов начались в конце 1980-х гг. Об этом свидетельствует ранняя работа Л. Флетчера и Р. Кастури, где был представлен один из первых алгоритмов для извлечения текста из бинарных изображений. Их алгоритм работал следующим образом: сначала анализировались связные компоненты в бинарном изображении, затем выбирались возможные символы на основе характеристик этих компонентов, после чего символы объединялись в слова с помощью преобразования Хафа (*Hough Transform*) [Fletcher, Kasturi, 1988].

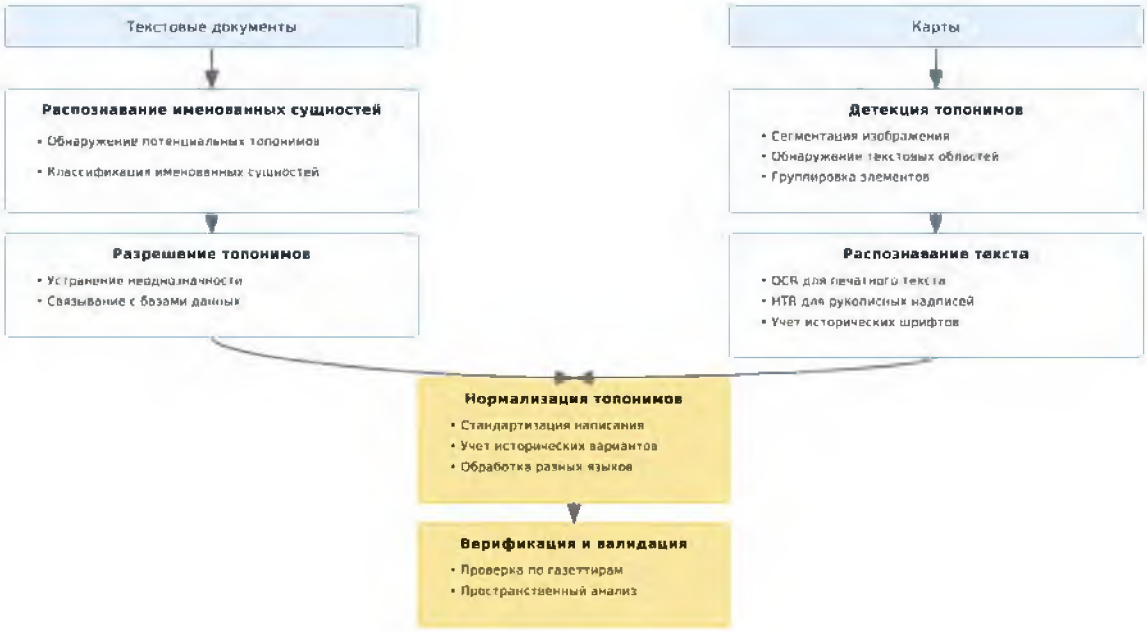

Рис. 1. Схема автоматической обработки топонимов из географических карт и текстов
Fig. 1. Diagram of automatic processing of toponyms from geographical maps and texts

По оценкам, этот метод давал хорошие результаты, так как мог работать с текстом разной ориентации и размера, и был одним из первых универсальных подходов к этой задаче. Однако у метода было важное ограничение – он не мог работать в ситуациях, когда текст пересекался с графическими элементами изображения. Это существенно ограничивало его применение для картографических материалов, где такие пересечения встречаются часто. Многим позже подход Л. Флетчера и Р. Кастури был усовершенствован в коллективной работе [Gllavata et al., 2003].

Это направление получило более активное развитие в 1990-х годах. Pierrot-Deseilligny et al. [1995] разработали метод реконструкции строк символов для картографических материалов. Их подход включал анализ связных компонентов и их распознавание в наборе возможных шрифтов. Реконструкция строк сводилась к задаче оптимизации графа. Требовалось предварительное знание шрифтов, используемых на картах. Главным ограничением было то, что не рассматривался вопрос сегментации цветной карты [Pierrot-Deseilligny et al., 1995].

В 1994 году была создана система *GIPSY* для автоматического географического индексирования текстовых, а именно картографических, документов. Эта система делала несколько интересных вещей: сопоставляла географические названия в тексте с пространственными координатами, пыталась понимать такие фразы, как *to the south of Lake Tahoe* – «к югу от озера Тахо», и представлять их в виде нечетких полигонов. Это была одна из первых систем, которая пыталась не просто найти название на карте, но и понять его пространственный контекст. Интересно, что методы *GIPSY* по работе с пространственными отношениями (*south of*, *near* и т. д.) оказались очень перспективными – этот подход продолжает развиваться и в современных геоинформационных системах [Woodruff, Plaunt, 1994].

До конца 1990-х гг. большинство работ по распознаванию текста фокусировалось на обычных печатных документах, а не на картографических материалах, которые представляют особую сложность из-за пересечения текста с другими элементами карты.

В 2001 году Д.Э. Смит и Г.Р. Крэйн, работая над проектом *Perseus* в Университете Тафтса, создали систему распознавания географических названий в исторических текстах, охватывающую колоссальный временной диапазон от античности до XIX века, исследование охватывало впечатляющий объем материала – более миллиона географических ссы-

лок в текстах. Исследователи проанализировали пять репрезентативных корпусов: древнегреческие и древнеримские тексты, коллекцию Боллеса по истории и топографии Лондона, а также две коллекции Библиотеки Конгресса о заселении Калифорнии и Верхнего Среднего Запада. Их подход, основанный на двухэтапной методологии, напоминал археологические раскопки: сначала происходила первичная идентификация и категоризация названий с использованием эвристических методов, сходных с системой *IBM Nominator*, а затем следовало более тонкое «просеивание» данных через сита локального контекста, документного окружения и общегеографических знаний [Smith, Crane, 2001].

Новую страницу в развитии методологии открыло исследование А. Гельбуха и С. Левачкина [Gelbukh, Levachkine, 2002]. Они сфокусировались на специфической проблематике распознавания текста на современных картографических материалах. Их работа выявила четыре фундаментальных вызова: отсутствие контекстной информации, зашумленность фона картографической нотацией, нестандартное расположение текста и необходимость установления связей между текстом и картографическими объектами. Авторская методология опиралась на комплексное использование различных источников данных: текстовой информации, анализа пространственного распределения букв, географической информации и нотационных данных. Исследователи использовали широкий спектр инструментов, включая статистические модели, географические базы данных и лингвистические фильтры [Gelbukh, Levachkine, 2002]. Эта же методология была продолжена в работах [Levachkine et al., 2002; Levachkine, 2003; Velázquez, Levachkine, 2003].

К 2003 году А. Гельбух и С. Левачкин углубили свое исследование, разработав более совершенную методологию, сочетающую технологии *OCR* со специализированными эвристиками для картографических данных. Особое внимание уделялось векторному представлению данных, что открывало новые возможности для решения семантических задач, включая анализ пространственных отношений и создание специализированных карт и эффективное масштабирование. Это исследование уже сфокусировалось на специфической проблеме распознавания текста на отсканированных картографических материалах [Gelbukh et al., 2003].

В это же время появляются и первые российские исследования. Одно из них посвящено разработке алгоритма программы, предназначенного для автоматического распознавания и выделения в текстах на русском языке именных групп с именами собственными, которые обозначают некоторые индивидуализированные объекты, а именно людей, учреждения и географические объекты. Исследование проводили ученые из ИППИ РАН в 2004–2005 гг.

В методологическом плане авторы опирались на контекстный анализ, отказавшись от статистического подхода из-за отсутствия достаточного размеченного корпуса русских текстов. Обработка текста происходила линейно и включала несколько последовательных этапов: токенизацию, морфологический анализ, работу со списками слов, контекстный анализ, разрешение неоднозначности и финальную разметку.

Для реализации этого подхода был разработан специальный инструментарий – программа *TagLite* (тэггер), использующая морфологический анализатор ИППИ РАН. Программа опиралась на систему из 16 списков слов и более 100 правил контекстного анализа, а также включала гессер для работы с неопознанными словами.

Для тестирования системы использовался корпус из 5748 слов, состоящий из новостей и статей на общественно-политические темы, содержащий 369 индивидуализированных именных групп: например, финансовый директор ЮКОСа Брюс Мизамор для категории людей, город Ногинск Костромской области для географических объектов и стадион им. Кирова для организаций. Кстати, в распознавании географических объектов (*GEOGR*) система показала наилучшие результаты с показателями *precision* 0,951, *recall* 0,957 и *f-value* 0,954, что существенно превышает показатели по другим категориям и демонстрирует особую эффективность разработанного метода для распознавания топонимов.

Очевидно, что использование линейного представления текста вместо структурного значительно ограничивало возможности анализа сложных языковых конструкций. Система опирается на заранее составленные списки слов и правила, что делает ее недостаточно гибкой для обработки новых, ранее не встречавшихся форм именных групп и требует постоянного ручного обновления баз данных. Относительно небольшой размер тестового корпуса (5748 слов) не позволяет делать надежные выводы о работоспособности системы на более разнообразных текстах. Отсутствие учета семантических связей между словами и более широкого контекста ограничивает точность распознавания в сложных случаях[1]. Система не учитывает вариативность написания имен собственных и их возможные искажения в реальных текстах. Кроме того, тестирование проводилось только на новостных и общественно-политических текстах, что не позволяет судить об эффективности системы при работе с текстами других жанров и стилей [Крейдлин, 2006].

Существенный вклад в развитие методологии распознавания топонимов в цветных топографических картах современного периода внесла работа Ж. Пудеро и др. Их четырехэтапный подход напоминал послойную реставрацию старинной картины: сначала производилась сегментация карты, затем анализ компонентов, далее анализ строк и, наконец, *OCR*-распознавание с постобработкой. В работе использовался алгоритм *GOCR* и морфологические операции для точного распознавания топонимов [Pouderoux et al., 2007].

Результаты и их обсуждение

Статистические показатели эффективности систем демонстрировали значительную вариативность в зависимости от типа обрабатываемых материалов. Система Perseus достигала наивысшей точности при работе с античными текстами (91–93 %), в то время как современные географические названия распознавались с меньшей точностью (74–86 %). Подход Ж. Пудеро и др. с показателем 89 % для картографических материалов занимал промежуточную позицию.

Математический аппарат и технические решения в рассматриваемых исследованиях существенно различались. Проект *Perseus* опирался на сложный математический аппарат, включающий вычисление центроидов на основе взвешенных карт упоминаний мест, расчет стандартных отклонений точек от центроида и оценку качества работы системы по формуле *F*-меры. Его техническая реализация использовала реляционные базы данных с информацией из *Getty Thesaurus of Geographic Names*. Ж. Пудеро и др. применяли преимущественно алгоритмические методы обработки изображений, включая *GOCR* и морфологические операции для анализа картографических материалов. Исследование А. Гельбуха и С. Левачкина [Gelbukh, Levachkine, 2002] базировалось на использовании статистических моделей и лингвистических фильтров (табл. 1).

Разберем пошагово, как работает математика во всех этих исследованиях. Сначала карта разбивается на сетку и делится на квадраты размером 1 × 1 градус (как шахматная доска). В каждом квадрате подсчитывается, сколько раз упоминаются разные места. Получается «тепловая карта» упоминаний мест. Затем вычисляется центроид. Это как поиск «центра тяжести» всех упоминаний: чем больше упоминаний места в квадрате, тем больший «вес» имеет этот квадрат. Центроид показывает, где сконцентрировано большинство упоминаемых мест. Далее вычисляется стандартное отклонение, то есть измеряется, насколько далеко разбросаны точки от центроида. Точки, которые находятся дальше двух стандартных отклонений, считаются выбросами и отбрасываются. После удаления выбросов центроид пересчитывается заново. Финальный этап вычислений – оценка качества через *F*-меру. *Precision* (*P*) показывает, сколько правильных из всех найденных названий, *Recall* (*R*) показывает, сколько нашли из всех реальных названий.

[1] Использование *Perl* для реализации также является устаревшим решением с точки зрения современной разработки систем обработки естественного языка.

Таблица 1
Table 1

Математический аппарат и технические решения в исследованиях 2000-х гг.
Mathematical apparatus and technical solutions in research of the 2000s.

Проект Perseus	1. Пространственную статистику с вычислением центроидов на сетке 1×1 градус 2. Формулу для расчета весовых коэффициентов точек на карте 3. Вычисление стандартного отклонения расстояний точек от центроида 4. *F*-меру для оценки качества распознавания: $F = ((\beta^2 + 1)RP) / (\beta^2 R + P)$, где: R – полнота (*recall*), P – точность (*precision*), β – весовой коэффициент между полнотой и точностью, обычно равный 1
Gelbukh и Levachkine	1. Вероятностную модель для оценки распознавания строк: $exp(-bd^2)$, где: 1.1. b – коэффициент масштаба карты и шрифтов 1.2. d – расстояние от надписи до объекта 2. Интегральную оценку для площадных объектов: $\iint_{S'} f(x,y)dxdy$, где: 2.1. $f(x,y)$ – минимальное расстояние от точки до букв надписи 2.2. S' – пересечение области S с границами карты 3. Биграммную/триграммную статистику для лингвистической верификации топонимов
Pouderoux et al.	1. Морфологические операции: 1.1. Закрытие (эрозия, затем дилатация) для удаления мелких линейных объектов 1.2. Анализ плотности связных компонент 2. Геометрические критерии фильтрации: 2.1. $max(height, width) < 0.5 \times min(\mu, \nu)$ 2.2. $max(height, width) > 2.0 \times max(\mu, \nu)$ где μ и ν – средние высота и ширина компонент

F-мера объединяет эти показатели в один, чтобы оценить общее качество работы.

Например, система анализирует текст про Москву. Она находит упоминания разных мест – Кремль, Арбат, Тверская улица и т. д. Все эти места концентрируются в центре Москвы (центроид), а упоминание, например, Владивостока будет отброшено как выброс, так как находится слишком далеко от основной группы мест. Такой подход позволяет понять основной географический фокус текста, отфильтровать случайные упоминания далеких мест, правильно идентифицировать, о каком именно месте идет речь, когда есть несколько мест с одинаковым названием.

В исследовании А. Гельбуха и С. Левачкина [Gelbukh, Levachkine, 2002] центральное место занимает вероятностная модель распознавания строк $(exp(-bd^2))$. Допустим, что вы нашли на карте название «Москва». Система должна понять, к какому объекту это название относится. Чем дальше надпись от объекта, тем меньше вероятность, что она к нему относится. Например, есть точка (город) и рядом надпись «Москва». D – это расстояние между точкой и надписью в миллиметрах, b – это коэффициент, учитывающий размер карты (чем крупнее масштаб, тем больше b). Если надпись прямо у точки (d маленькое), то $exp(-bd^2)$ близко к 1, если надпись далеко (d большое), то $exp(-bd^2)$ близко к 0.

Не менее важной в исследовании является интегральная оценка для площадных объектов. Например, мы видим надпись: «РОССИЯ» на карте, где буквы разбросаны по очерченной территории страны. Для каждой точки внутри страны (x,y) измеряется расстояние до ближайшей буквы. Эти расстояния суммируются по всей территории (это и есть интеграл). Чем лучше буквы распределены по территории, тем меньше будет сумма расстояний. Учитывается только та часть страны, которая видна на карте (S').

Третий компонент математических вычислений – это биграммная/триграммная статистика. Это способ проверить, похоже ли слово на настоящий топоним. Допустим, в русском языке часто встречаются сочетания «ск» (Минск, Омск), но сочетание «щщ» невоз-

можно. Система проверяет, какие пары/тройки букв возможны в географических названиях. Если встречается невозможное сочетание букв, вероятно это ошибка распознавания.

Таким образом, все эти методы работают вместе. Сначала проверяется, похоже ли слово на настоящий топоним, затем проверяется, насколько оно удалено от объекта, а для больших объектов проверяется, как буквы распределены по обозначенной территории.

Математический аппарат Ж. Пудеро и др. работает следующим образом. Например, на отсканированной карте есть названия городов и рек, линии дорог, границы областей, разные символы. Метод работает в два этапа. Сначала делается эрозия – все объекты как бы «съеживаются», тонкие линии исчезают; затем дилатация – оставшиеся объекты «расширяются» обратно. В результате тонкие линии (дороги, границы) пропадают, а буквы остаются.

При геометрической фильтрации система измеряет все оставшиеся объекты и проверяет их размеры: сначала вычисляет среднюю высоту (μ) и ширину (ν) всех объектов, потом отбрасывает слишком маленькие (меньше половины средних размеров) и слишком большие (больше чем в два раза превышают средние размеры). Например, средняя высота букв на карте 5 мм, а ширина 3 мм. Объект высотой 1 мм будет отброшен как слишком маленький, объект шириной 7 мм будет отброшен как слишком большой, а буква «А» высотой 4 мм и шириной 3 мм будет сохранена как подходящая. Это позволяет убрать с карты все, что не похоже на буквы по размеру, сохранить только те объекты, которые, вероятно, являются буквами, подготовить изображение для дальнейшего распознавания текста.

Следует отметить, что все эти методы были на тот момент развития технологий и данной проблематики основополагающими в исследованиях по автоматической обработке топонимов в текстах как таковых, не только картах. В монографии Й.Л. Ляйднера [Leidner, 2008] впервые систематически исследована комбинация лингвистических эвристик с экстралингвистическими знаниями (например, данными о населении). Особенно важным вкладом стало создание эталонного газеттира и размеченного корпуса для оценки алгоритмов, что заложило основу для последующих исследований в этой области. Использование модели максимальной энтропии для автоматической разметки также было прогрессивным решением для того времени. Работа Й.Л. Ляйднера определила базовые проблемы: различение между обычными словами и географическими названиями, множественность референтов (как в случае с различными Лондонами), необходимость точной пространственной привязки. В сущности, все эти вопросы точечно решались в работах Д.Э. Смита и Г.Р. Крэйна, А. Гельбуха и С. Левачкина, Ж. Пудеро и др.; Й.Л. Ляйднер же подвел некий промежуточный итог.

Каждое исследование демонстрировало свои уникальные методологические решения через конкретные примеры обработки топонимов. Все эти примеры иллюстрируют различные аспекты проблемы распознавания топонимов: омонимию, исторический контекст, пересечение с графикой, составные названия и масштабирование.

Так, А. Гельбух и С. Левачкин существенно продвинулись в решении проблемы топонимической омонимии на географических картах, что демонстрирует их анализ топонима «*Xalapa*». Хотя это название существует в штате Веракрус, система корректно определяет, что в контексте штата Оахака это вероятная ошибка распознавания, и верным вариантом является «*Jalapa*». В этой же работе рассматривается пример «*Moscow*» в координатах (57°N, 35°E), где система должна определить характер объекта (город или река) на основе вероятностного анализа географических координат. Проблема языковой специфики топонимов ярко проявляется в примере «*Río de Janeiro*», где система должна корректно различать географические маркеры (*río* – река) от составных частей самого топонима. Д.Э. Смит и Г.Р. Крэйн также обращают внимание на проблему конвенций именования в различных типах текстов, сравнивая структуру «*London, Ontario*» в новостных материалах с менее формализованными историческими документами.

Методология А. Гельбуха и С. Левачкина [Gelbukh, Levachkine, 2002] по комплексному использованию различных источников данных предвосхитила современные мультимодальные подходы в машинном обучении. Их идея интеграции текстовой, пространственной и нотационной информации соответствует современным тенденциям в обработке данных, хотя и с требованием значительных вычислительных ресурсов и наличия дополнительных баз данных.

Исследование Ж. Пудеро и др. фокусируется на проблемах распознавания топонимов в картографических материалах. Характерный пример – название "*Moncuquet*", ошибочно распознанное как "*Mqncuquet*" из-за пересечения с линией дороги. Аналогично, топоним "*Pechambert*" система неверно интерпретирует как "*Pe_ham_ert*" из-за шумов в изображении. Для решения этих проблем применяются морфологические операции и алгоритм *GOCR* с последующей контекстной фильтрацией. Работа также демонстрирует сложности распознавания составных названий на примере "*Pradal Haut*" и "*les Martres*", где требуется правильное объединение отдельно распознанных компонентов в единое составное название. Для этого применяется анализ пространственного расположения текстовых компонентов и их геометрических характеристик. Исследования выявили ряд фундаментальных проблем в области топонимического распознавания. Особенно острой оказалась проблема неоднозначности в различных географических регионах – в Северной и Центральной Америке 57,1 % названий имели несколько возможных локаций, тогда как в Европе этот показатель составлял всего 16,6 %. Временная неоднозначность также представляла существенную проблему, особенно при работе с историческими текстами.

Исследование Ж. Пудеро и др. заложило основы для современных систем компьютерного зрения в картографии. Однако их точность 89 % сегодня уже не может считаться удовлетворительной – современные системы на основе глубокого обучения достигают существенно более высоких показателей. Кроме того, их метод имел ограничения при работе с перекрывающимися текстовыми элементами.

Основным ограничением всех этих исследований, если смотреть из современной перспективы, является отсутствие методов глубокого обучения и нейронных сетей. Однако проанализированные примеры демонстрируют фундаментальную ценность разработанных подходов для понимания природы проблем топонимического распознавания. Их методологические принципы и алгоритмические решения заложили основу для развития новых технологий в области географической информатики, предоставляя богатый материал для анализа типичных проблем и способов их решения[2].

Усовершенствование методологии на материале исторических карт.

В 2014 году вышла коллективная статья Р. Саймона и др., которая в каком-то смысле подвела итог более чем 10-летнему изучению проблемы автоматического распознавания топонимов [Simon et al., 2014].

Работа представляет собой вершину развития классических методов компьютерного зрения, основанных на инженерном подходе и эвристических алгоритмах. При этом именно в эти годы начинается стремительный рост эффективности методов глубокого обучения. Этому способствовал ряд ключевых факторов:

1. Прорыв в области сверточных нейронных сетей, ознаменованный победой *AlexNet* в соревновании *ImageNet* в 2012 году и последующим развитием архитектур (*VGGNet, GoogLeNet* в 2014, *ResNet* в 2015).

2. Появление мощных графических процессоров и специализированных фреймворков для глубокого обучения.

3. Накопление больших массивов размеченных данных.

[2] Фундаментальная проблема неоднозначности топонимов, особенно в историческом контексте, до сих пор не имеет исчерпывающего решения.

4. Развитие методов переноса обучения (*transfer learning*).

Ценность исследования обусловлена не только тем, что оно символически маркирует границу между двумя эпохами в распознавании топонимов, но и тем, что оно посвящено проблеме автоматического анализа именно исторических карт, что затрагивалось в работах предшественников лишь вскользь [Smith, Crane, 2001] и не составляло предмет отдельного рассмотрения.

Исторические карты представляют собой сложные гибридные изображения, содержащие текстуры, линии и различные графические элементы, что существенно затрудняло процесс автоматического распознавания текста. Этот тип источников характеризуется значительной вариативностью в стилях написания и отображения географических названий. В отличие от современных карт, где существуют устоявшиеся стандарты картографической нотации, исторические карты отражают разнообразие каллиграфических традиций своего времени.

Существенную проблему создавала также физическая сохранность исторических карт. Выцветание чернил, повреждения бумаги, пятна и различные дефекты значительно усложняли процесс автоматического распознавания текста. Даже при качественном сканировании эти артефакты создавали дополнительные шумы, с которыми системы распознавания того времени справлялись плохо [Schlegel, 2021].

Другим серьезным вызовом являлась проблема исторической изменчивости топонимов. Один и тот же географический объект мог иметь различные написания не только в разные исторические периоды, но даже в работах разных картографов одного периода. Это создавало сложности при создании эталонных баз данных для обучения и верификации систем распознавания [Smith, Crane, 2001].

Исследователи сталкивались и с проблемой многоязычности исторических карт, где топонимы часто представлены в различных языковых традициях, с использованием разных алфавитов и систем транслитерации [Olson et al., 2024]. Ранние системы распознавания, ориентированные преимущественно на работу с одним языком и алфавитом, оказывались неэффективными при столкновении с такой языковой вариативностью.

Важным методологическим вкладом Р. Саймона и др. стала разработка трехфазного подхода к полуавтоматическому распознаванию топонимов, что, пожалуй, в хорошем смысле отличало данное исследование от своих предшественников (рис. 2).

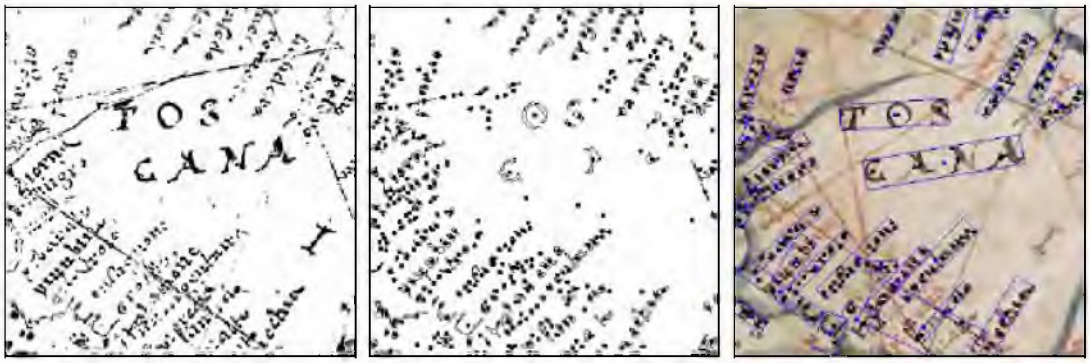

Рис. 2. Этапы обработки: сегментация фона и переднего плана (слева), обнаруженные контурные признаки (в центре), признаки, связанные в «группы признаков», обозначающие кандидаты в топонимы, наложенные поверх исходного изображения (справа) [Simon et al., 2014]
Fig. 2. Processing stages: background and foreground segmentation (left), detected contour features (center), features linked into "feature groups" denoting toponym candidates overlaid on the original image (right) [Simon et al., 2014]

Первая фаза ***Background-Foreground Segmentation (Сегментация фона и переднего плана)*** включала сегментацию фона и переднего плана с созданием черно-белой маски изображения. Этот этап оказался наиболее критичным для качества результатов, но и самым сложным из-за высокой вариативности типов карт.

Сначала проводился ручной этап, где либо определялись конкретные цветовые диапазоны, которые будут считаться фоном, либо применялся сильный медианный фильтр к карте. Медианная фильтрация была выбрана, потому что она удаляет тонкие структуры (такие как тонкие линии и текст) и при этом сохраняет общее распределение цвета карты.

После этого отфильтрованное изображение вычитывалось из оригинального, что дало базовую маску для сегментации фона. Затем проводилась очистка маски: удалялись линии (сетка, румбовые линии) с помощью преобразования Хафа; удалялись части с низким цветовым градиентом (потому что текст обычно имеет сильные края); применялись морфологические операции обработки изображения для уменьшения дефектов, появившихся из-за цветовой маскировки.

Вторая фаза ***Feature Detection (Обнаружение признаков)*** состояла в обнаружении характерных признаков с помощью алгоритмов определения контуров. Хотя этот метод был вычислительно эффективен, он сталкивался с проблемами в случаях, когда топонимы сливались с линейными элементами схожей яркости.

После создания маски-изображения, этот этап локализовал и характеризовал особенности на переднем плане изображения, в данном случае – связанные объекты. На этом этапе использовался алгоритм обнаружения контуров. Этот подход был выбран, потому что он вычислительно относительно легкий. На этом же этапе были введены правила для фильтрации недопустимых объектов на основе эвристик, касающихся покрываемой площади, ширины и соотношения сторон.

Третья фаза ***Feature Linking (Связывание признаков)*** заключалась в связывании обнаруженных признаков, которые, вероятно, относились к одному топониму, на основе эмпирических ограничений и эвристик относительно расстояния, площади, направления и абсолютных границ. Связывание особенностей не идентифицирует топонимы напрямую, а только отдельные связанные объекты на изображении. Топонимы обычно состоят из любого количества особенностей, поэтому этот этап постобрабатывает обнаруженные особенности, связывая их в группы, которые, вероятно, представляют один топоним.

Экспериментальные результаты существенно различались в зависимости от типа исторических карт. На карте Птолемея XV века, содержащей 41 топоним, точность локализации достигла 93 %: из 41 топонима 38 были локализованы корректно, 2 имели смещение рамки, 2 были ошибочно объединены, 1 был ошибочно разделен. Однако при работе с более сложным портоланом XVII века система обнаруживала только около 50 % видимых топонимов с точностью 31 %.

Авторы тестировали свой подход на трех типах карт разной сложности. Секция карты Британских островов эпохи Птолемея представляла наименее сложный случай благодаря однородному фону. На этой карте из 41 топонима 38 были локализованы корректно, хотя у двух топонимов наблюдалось небольшое смещение ограничивающей рамки. Из трех оставшихся топонимов два были ошибочно идентифицированы как один объединенный топоним, а последний топоним *ALVION INSVULA BRITANNICA* был ошибочно разделен на две группы признаков.

Лист австрийской топографической съемки середины XVIII века представлял более сложный случай из-за низкого контраста между топонимами и фоном. Фрагмент портолана XVII века являлся наиболее сложным примером с областями низкого цветового контраста, плохой читаемостью и проблемными структурами вроде пересекающихся линий и декоративных элементов. На этой карте визуально различалось 323 топонима для населенных пунктов, написанных мелким шрифтом, и 10 для регионов, обозначенных крупным шрифтом. Тест выдал 532 возможных обнаружения.

Среди характерных ошибок авторы показывают пример орнаментальных помех, где декоративные элементы и символы схожего размера и плотности с топонимами вызывают ложные срабатывания. В другом примере демонстрируется проблема наложения линий – пятый топоним слева имеет смещенные границы из-за береговой линии, а чуть правее ложное обнаружение вызвано другим сегментом береговой линии.

Проблема перекрестных помех между топонимами иллюстрируется случаем, где диагональная ограничивающая рамка пересекает другую рамку. Сложность работы с разделенными топонимами показана на примере *MUHR VORSTADTVON GRATZ*, разбитого на несколько строк. Отдельно выделяется проблема крупноформатных топонимов – авторы приводят пример топонима, идущего снизу вверх, который система не смогла корректно обработать.

Исследователи выявили ряд специфических проблем при работе с историческими картами: помехи от декоративных элементов, визуально похожих на текст; искажения при пересечении топонимов с картографическими элементами; перекрестные помехи между соседними названиями; разрывы топонимов на несколько строк; сложности с распознаванием крупноформатных и криволинейных надписей.

Эта работа легла в основу проекта Pelagios 3, целью которого стало создание поисковых индексов исторических топонимов и инструментов для научного аннотирования оцифрованных карт.

Вся методология работала итеративно: топонимы «выращивались» из отдельных признаков путем последовательного связывания (рис. 3).

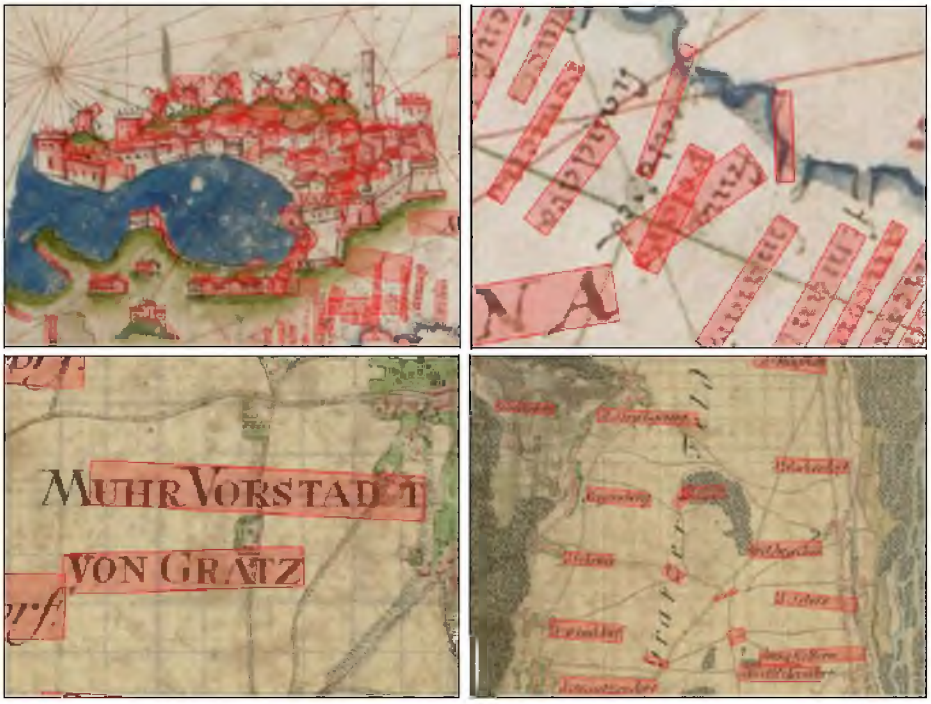

Рис. 3. Типичные ошибочные ситуации, встреченные в ходе экспериментов: помехи из-за символов и декоративных элементов (верхний левый); помехи из-за линейных элементов, сливающихся с топонимами, и «перекрестные помехи» между топонимами (верхний правый); разделенные топонимы (нижний левый), необнаруженный крупноформатный топоним (нижний правый) [Simon et al., 2014]

Fig. 3. Typical error situations encountered during experiments: interference from symbols and decorativem elements (upper left); interference from linear elements merging with toponyms and "cross-interference" between toponyms (upper right); separated toponyms (lower left), undetected large-format toponym (lower right) [Simon et al., 2014]

Как видно из общего «пайплайна», исследование Р. Саймона и др. сфокусировалось не просто на распознавании, а именно на детекции топонимов на картах, что послужило первым, но критически важным шагом в полном процессе обработки исторических картографических материалов (рис. 4).

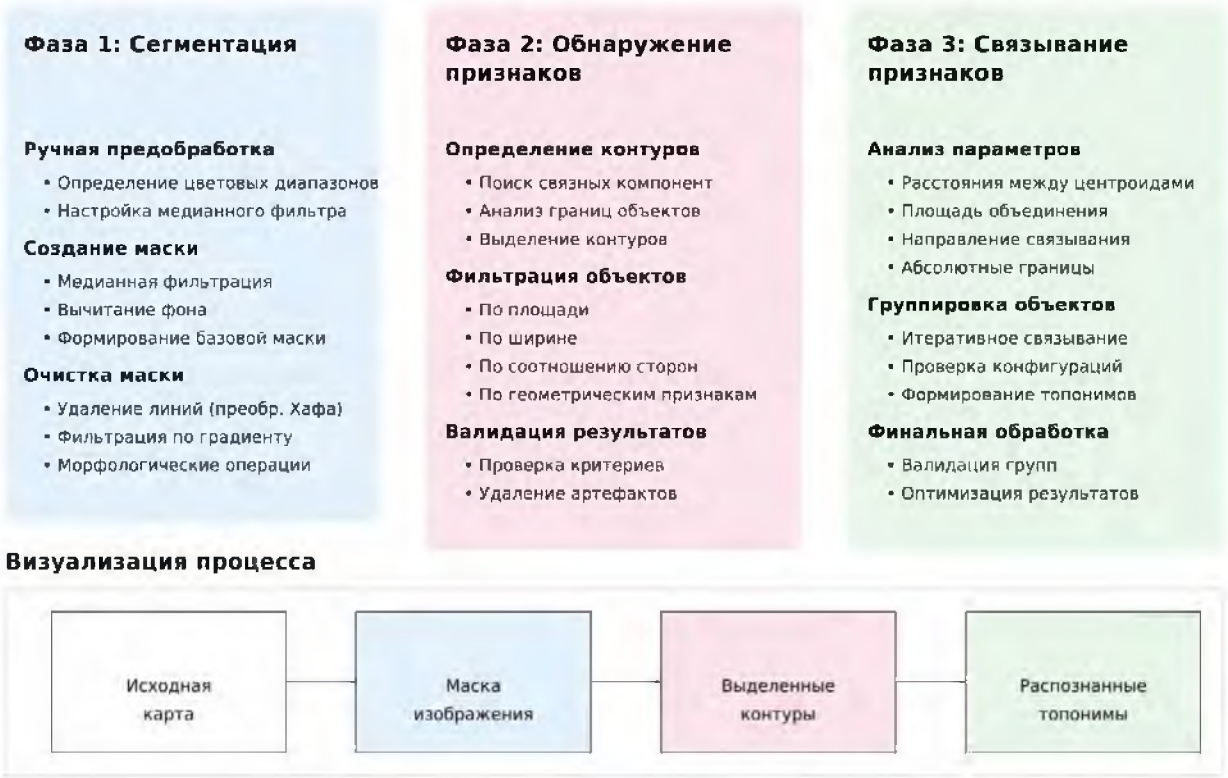

Рис. 4. Трехфазовый подход Р. Саймона и др. [Simon et al., 2014]
Fig. 4. Three-phase approach by R. Simon et al. [Simon et al., 2014]

Разумеется, в эпоху глубокого обучения методология Р. Саймона и др. была существенно усовершенствована. Например, фаза сегментации фона и переднего плана сегодня реализована с использованием современных трансформных архитектур, специализирующихся на сегментации изображений. Фаза обнаружения признаков уже улучшена применением детекторов текста на основе глубокого обучения, таких как *EAST* или *TextFuseNet*. Фаза связывания признаков оптимизирована с помощью графовых нейронных сетей, способных учитывать сложные пространственные отношения между элементами текста.

Исследование Р. Саймона и др. можно рассматривать как своеобразную кульминацию «классического периода» в развитии методов автоматического распознавания топонимов на картах. Трехфазный подход обобщил и систематизировал накопленный опыт предшественников, создав методологический мост между традиционными методами компьютерного зрения и современными подходами глубокого обучения. Работа предоставила важные технические детали о практических сложностях и подходах к автоматизированному распознаванию топонимов в исторической картографии этого периода. Их результаты помогают понять, почему полностью автоматизированные подходы оставались проблематичными, и почему полуавтоматические методы, сочетающие компьютерное обнаружение с человеческой верификацией, оказались более перспективным направлением развития. Особенно важно, что авторы не только разработали комплексную методологию, но и детально описали ограничения и проблемы своего подхода, что впоследствии помогло исследователям сфокусироваться на наиболее критических аспектах при разработке нейросетевых архитектур. Это делает работу Р. Саймона и др. особенно ценной для понимания эволюции методов распознавания топонимов и их постепенной трансформации в эпоху глубокого обучения.

Выводы. Бесспорно, эволюция методологических подходов к автоматическому распознаванию топонимов представляет собой многогранный процесс, где каждое исследование внесло уникальный вклад в развитие этого направления. От пионерских работ Л. Флетчера и Р. Кастури, предложивших базовые алгоритмы извлечения текста из бинарных изображений, до комплексных исследований 2000-х гг. мы наблюдаем последовательное усложнение и совершенствование методологического аппарата. Особенно впечатляет то, как исследователи 2000-х гг., несмотря на ограниченные технические возможности, смогли создать настолько глубокие и продуманные подходы к решению сложных задач распознавания исторических топонимов.

Особого внимания заслуживает методологическая преемственность в развитии данного направления. Некоторые из этих методов сохраняют актуальность даже в эпоху нейронных сетей, но часто упускаются из виду современными исследователями.

Работа Д.Э. Смита и Г.Р. Крэйна заложила фундаментальные принципы контекстного анализа топонимов, которые остаются актуальными и в современную эпоху глубокого обучения. Их новаторский подход к обработке исторических текстов, охватывающих колоссальный временной диапазон от античности до XIX в., предвосхитил современные методы работы с временными рядами и контекстно-зависимой обработкой текста. В контексте современных языковых моделей типа *BERT* и *GPT* их идеи о важности контекстного окружения и временной привязки топонимов приобретают новое звучание. Кроме того, яркий пример с этнонимом "*Germans*" в тексте о Галльской войне Цезаря, который система ошибочно связывает с современной Германией, показывает, что даже современные системы часто игнорируют историческую динамику топонимов.

Исследования А. Гельбуха и С. Левачкина внесли существенный вклад в развитие комплексного подхода к анализу картографических материалов. Их методология, основанная на интеграции различных источников данных, предвосхитила современные мультимодальные подходы в машинном обучении. Особенно ценным представляется их опыт работы с проблемой омонимии топонимов, что в современном контексте может быть усовершенствовано применением методов графового представления знаний и механизмов внимания.

Работа Ж. Пудеро и др. с цветными топографическими картами современного периода заслуживает особого внимания в контексте развития методов компьютерного зрения. Их четырехэтапный подход к обработке изображений, включающий сегментацию, анализ компонентов, анализ строк и *OCR*-распознавание, создал методологическую основу для современных *end-to-end* архитектур глубокого обучения.

Последующее исследование Р. Саймона и др. можно рассматривать как синтез предшествующих подходов, где трехфазная методология объединила лучшие практики предыдущих исследований. Их работа продемонстрировала эффективность полуавтоматического подхода, что особенно важно в контексте современных тенденций к созданию *human-in-the-loop* систем.

В перспективе развития современного инструментария особенно важным представляется симбиоз различных методологических подходов. Контекстный анализ Д.Э. Смита и Г.Р. Крэйна может быть усовершенствован применением современных языковых моделей. Комплексный подход А. Гельбуха и С. Левачкина к интеграции данных может быть реализован с использованием трансформных архитектур, способных одновременно обрабатывать текстовую и визуальную информацию. Методология сегментации изображений Ж. Пудеро и др. может быть улучшена применением современных сегментационных моделей, а итеративный подход Р. Саймона и др. может быть оптимизирован с помощью рекуррентных нейронных сетей. В частности, использование медианной фильтрации и преобразования Хафа для сегментации изображений может показаться архаичным на фоне современных сегментационных моделей типа *U-Net* или *Mask R-CNN*, однако именно эти базовые методы заложили концептуальный фундамент для развития более совершенных алгоритмов. Современные архитектуры глубоких нейронных сетей во многом повторяют логику этого многоступенчатого процесса, но реализуют ее на более высоком уровне абстракции.

Отдельного внимания заслуживает подход к анализу ошибок распознавания. Все исследователи уделяли особое внимание систематизации и классификации типичных ошибок: орнаментальные помехи, наложение линий, перекрестные помехи между топонимами, разрывы топонимов, проблемы с крупноформатными топонимами. Такая детальная каталогизация проблемных случаев создает ценную основу для улучшения современных систем распознавания.

Наконец, стоит отметить принципиальную позицию исследователей относительно полуавтоматического подхода. Их опыт показывает, что в некоторых случаях более эффективным может быть сочетание автоматических методов с человеческой верификацией, особенно при работе с историческими материалами, где контекст и нюансы интерпретации играют критическую роль.

Практическая значимость этих исследований проявляется в создании методологического фундамента для современных систем геоинформационного анализа. Разработанные подходы к решению проблем омонимии, временной неоднозначности, многоязычия и пространственных отношений остаются актуальными и в контексте современных технологий. Более того, накопленный опыт работы с различными типами картографических материалов и текстовых источников создает бесценную базу для проектирования и обучения современных систем искусственного интеллекта.

В свою очередь, хотелось бы обратить внимание на то, что в текстах проанализированных работ не хватает некоторых разъяснений. Например, в статьях не приводится полной детальной статистики по каждому типу ошибок для каждого метода, при том, что описание тестовых наборов в статьях присутствует. Что касается математического аппарата, то конкретные диапазоны значений также не приводятся. Кроме того, в статьях нет полноценного анализа вычислительной сложности алгоритмов. В статье Р. Саймона и др. вообще говорится, что алгоритм обнаружения контуров «был выбран, потому что он вычислительно относительно легкий» [Simon et al., 2014].

Заключение

Перспективы исследований в области автоматического распознавания топонимов лежат в плоскости интеграции традиционных подходов с новейшими достижениями в сфере компьютерного зрения и обработки естественного языка. Особенно многообещающей представляется разработка мультимодальных систем, способных одновременно анализировать визуальные, текстовые и пространственные данные, учитывая при этом временную динамику изменения географических названий. Такие системы могли бы не только распознавать топонимы на картах разных эпох, но и автоматически отслеживать эволюцию названий, их вариативность в разных языках и культурных традициях.

Отдельного внимания заслуживает возможность создания универсальных инструментов для работы с историческими картографическими материалами, способных адаптироваться к различным стилям оформления, системам нотации и степени сохранности источников. Это откроет новые горизонты для масштабной цифровизации картографического наследия и создания детальных пространственно-временных моделей развития топонимических систем. Такой подход позволит не только автоматизировать процесс обработки исторических карт, но и получить новые данные о закономерностях эволюции географических названий в различных регионах мира.

Список литературы

Вицентий А.В., Диковицкий В.В., Шишаев М.Г. 2020. Технология извлечения и визуализации пространственных данных, полученных при анализе текстов. Труды Кольского научного центра РАН, 11(8–11): 115–119. https://doi.org/10.37614/2307-5252.2020.8.11.012

Вицентий А.В., Шишаев М.Г. 2021. Технология извлечения геоатрибутированных сущностей для визуального представления пространственной связности объектов на основе автоматизированной генерации картосхем. Труды Кольского научного центра РАН, 12(5): 35–49. https://doi.org/10.37614/2307-5252.2021.5.12.003

Герцен А.А., Герцен О.А., Гордова Ю.Ю., Костовска С.К., Костовска Ст.К., Хропов А.Г. 2023. Аспекты картографии и топонимии культовых сооружений в историко-географической перспективе. ИнтерКарто. ИнтерГИС, 29: 180–203. https://doi.org/10.35595/2414-9179-2023-2-29-180-203

Гордова Ю.Ю., Герцен О.А., Герцен А.А., Костовска С.К. 2021. Применение картографических методов в топонимике (история вопроса и современные исследования). ИнтерКарто. ИнтерГИС, 27(4): 520–536. https://doi.org/10.35595/2414-9179-2021-4-27-520-536

Колесников А.А., Кикин П.М., Нико Д., Комиссарова Е.В. 2020. Системы обработки естественного языка для извлечения данных и картографирования на основе неструктурированных блоков текста. ИнтерКарто. ИнтерГИС, 26(1): 375–384. https://doi.org/10.35595/2414-9179-2020-1-26-375-384

Красовский А.П. 2024. Большие данные как инструмент исследований в топонимике и истории межевания. ИнтерКарто. ИнтерГИС, 30(1): 321–341. https://doi.org/10.35595/2414-9179-2024-1-30-321-341

Крейдлин Л.Г. 2006. Программа выделения русских индивидуализированных именных групп TagLite. В кн.: Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог, Москва, 01–04 июня 2016. М., Российский государственный гуманитарный университет: 292–297.

Поспелов Е.М. 1971. Топонимика и картография. М, Мысль, 256 с.

Fletcher L.A., Kasturi R. 1988. A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE transactions on pattern analysis and machine intelligence, 10(6): 910–918. https://doi.org/10.1109/34.9112

Gelbukh A., Levachkine S. 2002. Error Detection and Correction in Toponym Recognition in Cartographic Maps. IASTED International Conference Geopro-2002: 1–7

Gelbukh A., Levachkine S., Han S.Y. 2003. Resolving Ambiguities in Toponym Recognition in Cartographic Maps. In: Graphics Recognition. Recent Advances and Perspectives. GREC 2003. Lecture Notes in Computer Science. Ed. by Lladós J., Kwon Y.B. Springer, Berlin, Heidelberg: 75–86.

Gllavata J., Ewerth R., Freisleben B. 2003. A Robust Algorithm for Text Detection in Images. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, Rome, Italy: 611–616.

Leidner J.L. 2008. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. USA, Florida, Universal-Publishers, 261 p.

Lenc L., Martínek J., Baloun J., Prantl M., Král P. 2022. Historical Map Toponym Extraction for Efficient Information Retrieval. In: Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science. Springer: 171–183. https://doi.org/10.1007/978-3-031-06555-2_12

Levachkine S. 2003. Raster to Vector Conversion of Color Cartographic Maps. In: Graphics Recognition. Recent Advances and Perspectives. GREC 2003. Lecture Notes in Computer Science. Berlin, Heidelberg, Springer: 50–62. https://doi.org/10.1007/978-3-540-25977-0_5.

Levachkine S., Vel´azquez A., Alexandrov V., Kharinov M. 2002. Semantic Analysis and Recognition of Raster-Scanned Color Cartographic Images. In: Graphics Recognition Algorithms and Applications. GREC 2001. Lecture Notes in Computer Science. Berlin, Heidelberg, Springer-Verlag: 178–189.

Milleville K., Verstockt S., Van de Weghe N. 2020. Improving Toponym Recognition Accuracy of Historical Topographic Maps. Automatic Vectorisation of Historical Maps, Proceedings of the International Workshop on Automatic Vectorisation of Historical Maps, Budapest, Hungary, 13: 63–72.

Olson R., Kim J., Chiang Y.Y. 2024. Automatic Search of Multiword Place Names on Historical Maps. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data, 9–12. https://doi.org/10.1145/3681769.3698577

Peters M., Neumann M., Iyyer M., Gardner M., Clark Chr., Lee K, Zettlemoyer L. 2018. Deep Contextualized Word Representations. In: Human Language Technologies. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, New Orleans, Association for Computational Linguistics, Vol. 1: 2227–2237.

Pierrot-Deseilligny M., Men H.L., Stamon G. 1995. Characters String Recognition on Maps, a Method for High Level Reconstruction. Montreal, QC, Canada, Proceedings of ICDAR: 249–252.

Pouderoux J., Gonzato J.-C., Pereira A., Guitton P. 2007. Toponym Recognition in Scanned Color Topographic Maps. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, 1: 531–535.

Schlegel I. 2021. Automated Extraction of Labels from Large-Scale Historical Maps. AGILE: GIScience Series, 2: 12. https://doi.org/10.5194/agile-giss-2-12-2021

Simon R., Pilgerstorfer P., Isaksen L., Barker E. 2014. Towards Semi-Automatic Annotation of Toponyms on old maps. e-Perimetron, 9(3): 105–128.

Smith D.A., Crane G.R. 2001. Disambiguating Geographic Names in A Historical Digital Library. In: Research and Advanced Technology for Digital Libraries. ECDL 2001. Lecture Notes in Computer Science, Springer, Berlin: 127–136.

Vel´azquez A., Levachkine S. 2003. Text/graphics separation and recognition in raster-scanned color cartographic maps. Proceedings: Automated geographic indexing of text documents. Journal of the American Society for Information Science, 45(9): 645–655.

Zhou B., Zou L., Hu Y., Qiang Y., Goldberg D. 2023. TopoBERT: a Plug and Play Toponym Recognition Module Harnessing Fine-Tuned BERT. International Journal of Digital Earth, 16(1): 3045–3064.

References

Vitsentiy A.V., Dikovitskiy V.V., Shishayev M.G. 2020. The Technology of Extraction and Visualization of Spatial Data Obtained by Texts Analysis. Kola Science Centre Publisher, 11(8–11): 115–119 (in Russian). https://doi.org/10.37614/2307-5252.2020.8.11.012

Vitsentiy A.V., Shishayev M.G. 2021. The Geoattributed Entity Extraction Technology for Visual Representation of Objects Spatial Relations Based on Automated Schematic Map Generation. Kola Science Centre Publisher, 12(5): 35–49 (in Russian). https://doi.org/10.37614/2307-5252.2021.5.12.003

Herzen A.A., Herzen O.A., Gordova Yu.Yu., Kostovska S.K., Kostovska St.K., Khropov A.G. 2023. Aspects of Cartography and Toponymy of Religious Buildings in the Historical-Geographical Perspective. InterKarto. InterGIS, 29: 180–203 (in Russian). https://doi.org/10.35595/2414-9179-2023-2-29-180-203

Gordova Yu.Yu., Herzen O.A., Herzen A.A., Kostovska S.K. 2021. Usage of Cartographic Methods in Place-Name Study (History of the Problem and Actual Research). InterKarto. InterGIS, 27(4): 520–536 (in Russian). https://doi.org/10.35595/2414-9179-2021-4-27-520-536

Kolesnikov A.A., Kikin P.M., Niko D., Komissarova E.V. 2020. Natural Language Processing Systems for Data Extraction and Mapping on the Basis of Unstructured Text Blocks. InterKarto. InterGIS, 26(1): 375–384 (in Russian). https://doi.org/10.35595/2414-9179-2020-1-26-375-384

Krassowski A.P. 2024. Big data as a research tool in toponymy and the history of land surveying. InterCarto. InterGIS, 30(1): 321–341 (in Russian). https://doi.org/10.35595/2414-9179-2024-1-30-321-341

Kreydlin L.G. 2006. Program for Extraction of Russian Individualized Name Groups TagLite. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue, Moscow, 1–4 June 2016. Moscow, Pabl. Rossiyskiy gosudarstvennyy gumanitarnyy universitet: 292–297 (in Russian).

Pospelov E.M. 1971. Toponimika i kartografiya [Toponymy and Cartography]. Moscow, Pabl. Mysl, 256 p.

Fletcher L.A., Kasturi R. 1988. A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE transactions on pattern analysis and machine intelligence, 10(6): 910–918. https://doi.org/10.1109/34.9112

Gelbukh A., Levachkine S. 2002. Error Detection and Correction in Toponym Recognition in Cartographic Maps. IASTED International Conference Geopro-2002: 1–7

Gelbukh A., Levachkine S., Han S.Y. 2003. Resolving Ambiguities in Toponym Recognition in Cartographic Maps. In: Graphics Recognition. Recent Advances and Perspectives. GREC 2003. Lecture Notes in Computer Science. Ed. by Lladós J., Kwon Y.B. Springer, Berlin, Heidelberg: 75–86.

Gllavata J., Ewerth R., Freisleben B. 2003. A Robust Algorithm for Text Detection in Images. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, Rome, Italy: 611–616.

Leidner J.L. 2008. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. USA, Florida, Universal-Publishers, 261 p.

Lenc L., Martínek J., Baloun J., Prantl M., Král P. 2022. Historical Map Toponym Extraction for Efficient Information Retrieval. In: Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science. Springer: 171–183. https://doi.org/10.1007/978-3-031-06555-2_12

Levachkine S. 2003. Raster to Vector Conversion of Color Cartographic Maps. In: Graphics Recognition. Recent Advances and Perspectives. GREC 2003. Lecture Notes in Computer Science. Berlin, Heidelberg, Springer: 50–62. https://doi.org/10.1007/978-3-540-25977-0_5.

Levachkine S., Vel´azquez A., Alexandrov V., Kharinov M. 2002. Semantic Analysis and Recognition of Raster-Scanned Color Cartographic Images. In: Graphics Recognition Algorithms and Applications. GREC 2001. Lecture Notes in Computer Science. Berlin, Heidelberg, Springer-Verlag: 178–189.

Milleville K., Verstockt S., Van de Weghe N. 2020. Improving Toponym Recognition Accuracy of Historical Topographic Maps. Automatic Vectorisation of Historical Maps, Proceedings of the International Workshop on Automatic Vectorisation of Historical Maps, Budapest, Hungary, 13: 63–72.

Olson R., Kim J., Chiang Y.Y. 2024. Automatic Search of Multiword Place Names on Historical Maps. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data, 9–12. https://doi.org/10.1145/3681769.3698577

Peters M., Neumann M., Iyyer M., Gardner M., Clark Chr., Lee K, Zettlemoyer L. 2018. Deep Contextualized Word Representations. In: Human Language Technologies. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, New Orleans, Association for Computational Linguistics, Vol. 1: 2227–2237.

Pierrot-Deseilligny M., Men H.L., Stamon G. 1995. Characters String Recognition on Maps, a Method for High Level Reconstruction. Montreal, QC, Canada, Proceedings of ICDAR: 249–252.

Pouderoux J., Gonzato J.-C., Pereira A., Guitton P. 2007. Toponym Recognition in Scanned Color Topographic Maps. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, 1: 531–535.

Schlegel I. 2021. Automated Extraction of Labels from Large-Scale Historical Maps. AGILE: GIScience Series, 2: 12. https://doi.org/10.5194/agile-giss-2-12-2021

Simon R., Pilgerstorfer P., Isaksen L., Barker E. 2014. Towards Semi-Automatic Annotation of Toponyms on old maps. e-Perimetron, 9(3): 105–128.

Smith D.A., Crane G.R. 2001. Disambiguating Geographic Names in A Historical Digital Library. In: Research and Advanced Technology for Digital Libraries. ECDL 2001. Lecture Notes in Computer Science, Springer, Berlin: 127–136.

Vel´azquez A., Levachkine S. 2003. Text/graphics separation and recognition in raster-scanned color cartographic maps. Proceedings: Automated geographic indexing of text documents. Journal of the American Society for Information Science, 45(9): 645–655.

Zhou B., Zou L., Hu Y., Qiang Y., Goldberg D. 2023. TopoBERT: a Plug and Play Toponym Recognition Module Harnessing Fine-Tuned BERT. International Journal of Digital Earth, 16(1): 3045–3064.

Поступила в редакцию 16.12.2024;
поступила после рецензирования 15.01.2025;
принята к публикации 11.02.2025

Received December 16, 2024;
Revised January 15, 2025;
Accepted February 11, 2025

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Дмитриев Александр Владиславович, кандидат филологических наук, доцент, докторант, доцент гуманитарного института, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

INFORMATION ABOUT THE AUTHOR

Alexander V. Dmitriev, Candidate of Philological Sciences, Associate Professor, Doctoral Candidate, Associate Professor of the Humanities Institute, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia