ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 4, с. 129–138

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.93

ПОИСК ПОЧТИ ДУБЛИКАТОВ ИЗОБРАЖЕНИЙ РУКОПИСНЫХ ТЕКСТОВ ДЛЯ ВЫСОКОНАГРУЖЕННЫХ СЕРВИСОВ

© 2024 г. К. В. Варламова[a, b, *], М. С. Каприелова[a, b, c, **], И. О. Потяшин[a, b, *], Ю. В. Чехович[a, ****]**

[a]*Компания «Антиплагиат», Москва, Россия*
[b]*Московский физико-технический институт, Москва, Россия;*
[a]*ФИЦ ИУ РАН, Москва, Россия*
**e-mail: kvarlamova@ap-team.ru*
***e-mail: kaprielova@ap-team.ru*
****e-mail: potyashin@ap-team.ru*
*****e-mail: chehovich@ap-team.ru*

Поступила в редакцию 22.05.2024 г.
После доработки 27.05.2024 г.
Принята к публикации 15.07.2024 г.

Решение задачи поиска заимствований в рукописных текстах становится год от года более актуальным. Одним из видов заимствований является почти дублирование рукописной работы – съемка того же рукописного текста в других условиях или использование различных аугментаций. Существующие подходы к обнаружению почти дубликатов не приспособлены к работе с большими коллекциями, что существенно ограничивает их использование на практике. Представлен метод на основе машинного обучения, который позволяет производить обнаружение почти дубликатов изображений рукописных текстов среди больших коллекций потенциальных источников. Процесс включает в себя три основных этапа: перевод изображения в векторное представление, поиск кандидатов и последующий отбор источника дублирования среди кандидатов. Приведены результаты экспериментов по оценке качества и производительности разработанной системы: достигнуты 59 и 80% полноты и 5.5 и 4.8% доли ложноположительных срабатываний приближенных к реальным и синтетических данных соответственно, время работы метода составляет 5.5 с/запрос при размере коллекции около 10 тыс. изображений. Результаты показали, что созданный метод может быть использован для решения задач, требующих проверки рукописных документов по большому количеству потенциальных источников заимствований.

Ключевые слова: компьютерное зрение, поиск почти дубликатов, анализ рукописных документов, большие базы данных, русский рукописный текст.

DOI: 10.31857/S0002338824040085 **EDN:** UEFADS

HANDWRITTEN DOCUMENTS NEAR-DUPLICATE SEARCH FOR DATA INTENSIVE APPLICATIONS

K. Varlamova[a, b, *], M. Kaprielova[a, b, c, **], I. Potyashin[a, b, *], Yu. Chekhovich[a, ****]**

[a]*Antiplagiat Company, Moscow, Russian Federation*
[b]*Moscow Institute of Physics and Technology, Moscow, Russian Federation*
[c]*FRC CSC RAS, Moscow, Russian Federation*
**e-mail: kvarlamova@ap-team.ru*
***e-mail: kaprielova@ap-team.ru*
****e-mail: potyashin@ap-team.ru*
*****e-mail: chehovich@ap-team.ru*

The problem of cheating in handwritten academic essays has become more significant over last several years. One of the cheating cases is submitting the same paper, photographed in different environment (for

example, from another angle, in different light or in lower quality), or changed by means of automatic augmentation. The existing methods are not designed to work on large collections of handwritten documents. The proposed approach consists of three stages. The first stage is embedding generation, the second one is finding closest candidates in the collection of handwritten documents and the final one is similarity estimation between query image and each of candidates obtained at previous step. Our solution showed Recall@1 80% and 59% with FPR 4.8% and 5.5% on Synthetic and Real data respectively. The search latency is 5.5 seconds per query for the collection of 10 000 images. The results showed that the developed method is robust enough to work on large collections of handwritten documents.

Keywords: computer vision; near-duplicate detection; handwritten document analysis; large collections; Russian cursive.

Введение. Поиск заимствований в академических и учебных работах является актуальной задачей. Уже существуют системы, позволяющие обнаруживать много типов нарушений академической этики в текстах [1], такие, как переводные заимствования [2,3], парафраз, машинная генерация [4,5] и др. Однако проблеме поиска заимствований в рукописных текстах уделяется гораздо меньше внимания. С бурным развитием онлайн-образования и необходимостью повышения автоматизации проверок работ школьников проблема заимствований в рукописных работах становится все более актуальной [6,7]. Требуется повышение уровня автоматизации проверки работ, причем с возможностью ее осуществления для больших коллекций. В частности, необходимость поиска заимствований в рукописных текстах школьников обусловлена важностью обучения принципам работы с информацией и формированием правильных представлений о правовых и этических нормах использования материалов, находящихся в открытом доступе.

Заимствования в рукописных работах можно разделить на две крупные категории. К первой категории относится переписывание текста работы с возможным изменением его части. Второй категорией заимствований является визуальное изменение той же работы с помощью различных манипуляций, таких, как фотографирование под другим углом, освещением, изменение качества фотографии или применение различных аугментаций по отношению к одному и тому же изображению рукописного текста. Второй тип заимствования, рассматриваемый в текущей статье, назовем *почти дублированием* изображения рукописного текста. Представлен новый метод детекции почти дублированных изображений рукописных текстов, написанных на русском языке.

1. Краткий обзор литературы. Задачу поиска почти дублированных изображений рукописного текста можно назвать сложной и при этом критически важной для образовательной системы [6,7]. В настоящее время становится все более востребованно работать с большими объемами данных, что усложняет задачу, так как в литературе не было описано достаточно эффективных для крупных коллекций методов. В [8] представлен подход к задаче детекции заимствований изображений, специализирующийся на поиске по крупным коллекциям. Однако рукописные тексты составляют специфичный домен данных, поэтому требуется более детализированный по отношению к сравнению изображений подход.

Подход к сравнению двух рукописных документов представлен в [9]. Он основан на сопоставлении рамок, ограничивающих слова (bounding box), с помощию сверточной нейронной сети. В [8] описан метод, полагающийся на сегментацию слов с последующим анализом их длин. Ряд подходов, связанных с анализом рукописного текста, базируется на его распознавании [10,11]. В сфере распознавания рукописного текста были достигнуты значительные результаты [12,13]. Это делает потенциально возможным использование систем распознавания рукописного текста (optical character recognition (OCR)) совместно с существующими методами поиска заимствований в текстах [14]. Однако такой подход имеет существенный недостаток: для обучения модели распознавания рукописного текста требуется большой объем данных для разметки. Для кириллических языков такого количества открытых размеченных данных нет, что приводит к невысокому качеству моделей распознавания рукописного текста на русском языке. Для моделей OCR также могут быть критичны изменения цвета, качества, поворот изображения и т.д., поэтому разные варианты съемки одного и того же рукописного текста могут привести к различным выходам модели. В связи с этим, в рамках задачи поиска почти дубликатов пока нельзя назвать эффективными подходы, основанные на распознавании рукописного текста.

В статье ключевая задача – разработка метода детекции почти дублированных изображений рукописных текстов. Основными достоинствами предлагаемого метода является то, что он не требует большого количества размеченных данных и достигает высокой эффективности при поиске по большим коллекциям. Решение состоит из трех логических частей: генерация векторных представлений, быстрый поиск кандидатов на источник заимствования и отбор наиболее релевантных кандидатов путем подсчета уровня сходства между изображениями. Проведены эксперименты на двух коллекциях, состоящих из рукописных сочинений на русском языке.

2. Постановка задачи. Будем называть *почти дубликатами* такие изображения, которые полностью или почти копируют оригинал: используется та же фотография работы либо работа сфотографирована под другим углом, с другим фоном, тенью, фотография изменена по качеству, яркости и пр.

Задачу детекции почти дубликатов можно сформулировать следующим образом.
Существует два множества:

1) множество изображений-запросов Q,

2) множество изображений-источников S, из которого могли происходить почти дублирования, содержащиеся в Q; S также имеет нулевой элемент s_0.

Существует отображение $FindSourse : Q \to S$. Для всех запросов q^{orig} из Q, не являющихся почти дубликатами изображений из S, выполняется

$$FindSourse(q^{orig}) = s_0.$$

Множество таких q^{orig} обозначим как $Ker(FindSourse)$, его мощность – как $M = |Ker(FindSourse)|$. Введем модель поиска почти дубликатов $f(q)$, $q \in Q$, выходом которой является множество $\{s1, \ldots, s_K, s_k \in S, k = 1,K\}$, $K \in \mathbb{N}$ – заранее фиксированное число кандидатов на источник почти дублирования.

Основными метриками качества будем считать метрику полноты $Recall@K$ на выбранном числе кандидатов K и метрику доли ложноположительных срабатываний (false positive rate (FPR)):

$$Recall@K = \frac{100\%}{|Q|}\sum_{i=1}^{|Q|}\left|f(q_i)@K \cap \left\{FindSourse(q_i)\right\}\right|, q_i \in Q \tag{2.1}$$

$$FPR = \frac{100\%}{|M|}\sum_{i=1}^{M} f \left| \left(q_i^{orig}\right)@\cap S\backslash s_0 \right|, q_i^{orig} \in Ker(FindSourse). \tag{2.1}$$

Задачей является поиск модели $\hat{f}$, наилучшей в смысле полноты, при ограничении на долю ложноположительных срабатываний:

$$\begin{cases} \hat{f} = \arg\max\limits_{f \in F} Recall@1(f, FindSourse, Q, S), \\ FPR(\hat{f}, FindSourse, Q, S) < \alpha, \end{cases}$$

где α – установленный порог для ограничения ложных срабатываний, в данной работе рассматривался $\alpha = 7\%$; F – пространство моделей.

3. Предлагаемый подход. Решение задачи, описанной в разд. 2, разделено на три последовательные части: генерация векторных представлений для изображений; векторный поиск кандидатов-источников заимствования для изображения-запроса; выбор кандидатов с помощью подсчета уровня сходства между кандидатом и изображением-запросом.

3.1. О п и с а н и е д а н н ы х. В работе используется три набора данных. Первый датасет – закрытый. Он состоит из 1 млн фотографий рукописных школьных сочинений, написанных на русском языке. Для взятых работ были искусственно сгенерированы почти дублирования с помощью наложений таких аугментаций, как изменения по цвету, углу, геометрических преобразований. Пример пары изображение – сгенерированный дубликат (рис. 1). Коллекцию, полученную из этого набора данных с помощью таких преобразований, будем называть *Synthetic*. Она использовалась для обучения моделей, входящих в итоговое решение.

Второй используемый датасет HWR200 [15] специализирован для задачи обнаружения заимствований из рукописных текстов, но обладает структурой, подходящей и для задачи поиска почти дубликатов. Каждое изображение коллекции – фотография или скан страницы школьного сочинения на русском языке. При этом каждая страница представлена в трех ви-

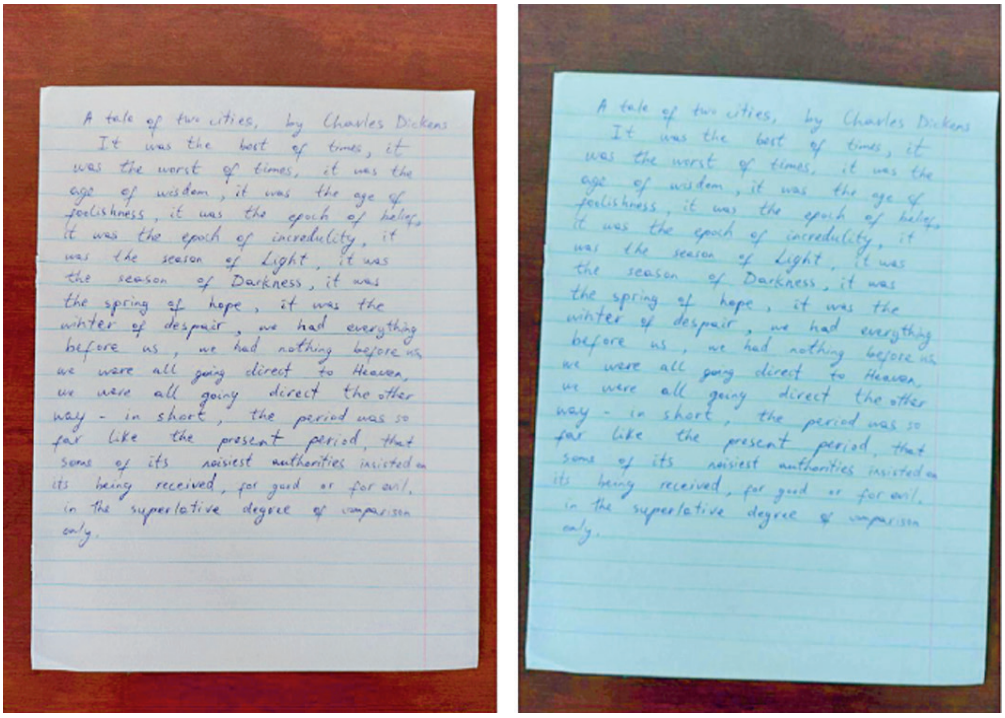

Рис. 1. Пример изображений из датасета с синтетическими почти дубликатами: оригинал и сгенерированный почти дубликат

дах: скан, светлая фотография и темная фотография. Светлая фотография сделана под хорошим освещением и с хорошим качеством. Темная фотография обладает меньшей яркостью, а также чаще содержит посторонние предметы – объекты, помимо самой страницы сочинения содержащиеся на фотографии. Пример элемента датасета можно увидеть на рис. 2. Таким образом, эти три вида изображений для одной и той же страницы являются друг для друга почти дубликатами. Коллекция содержит около 30 тыс. элементов по 10 тыс. изображений каждого типа. Она применялась для экспериментов, описанных ниже.

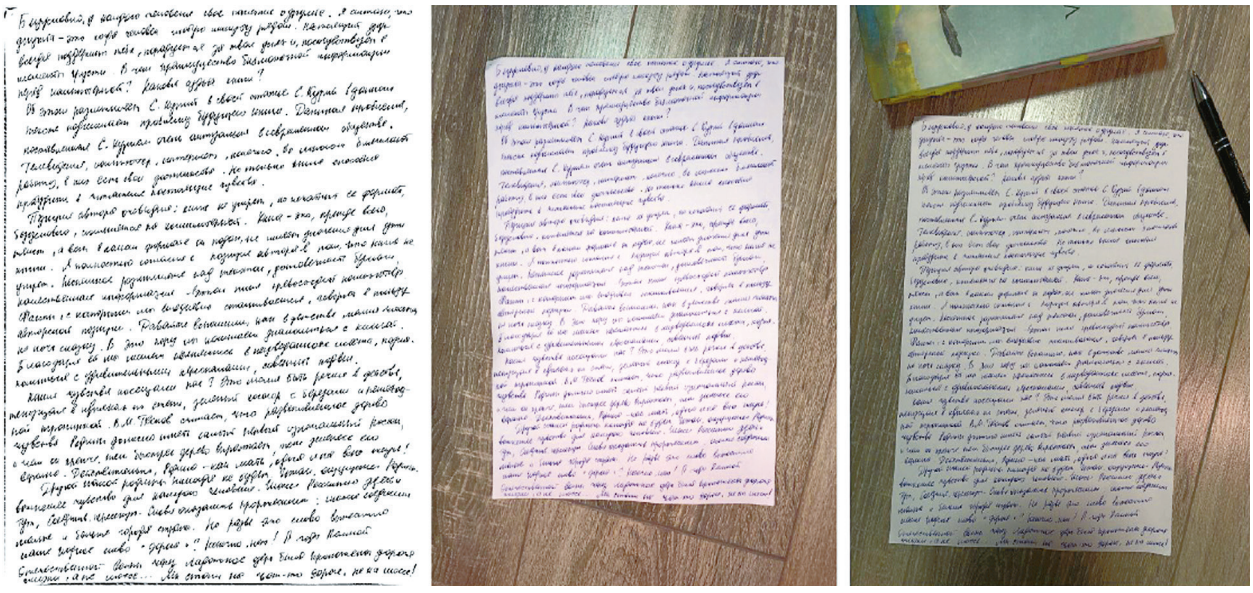

Рис. 2. Типы изображений из датасета HWR200 слева направо: скан, светлая фотография и темная фотография

Третий набор данных, полученный из открытых источников [16,17], и с помощью генерации почти дубликатов исходных изображений, также использовался для экспериментов.

3.2. Генерация векторных представлений. Для перевода пространства изображений с рукописными текстами в пространство векторных представлений применялась архитектура *Deep Ranking*, описанная в [18]. Авторы работы предложили архитектуру, целью которой было производить ранжирование не просто на основании классов, к которым принадлежат объекты, а на основании каких-либо характеристик объектов, которыми могут обладать экземпляры класса. Таким образом, в пространстве векторных представлений изображения, имеющие большее сходство, будут расположены ближе друг к другу.

Модель обучалась на *триплетах* – множествах из трех изображений $\{h,h_{pos},h_{neg}\}$, где h – входное изображение, h_{pos} – изображение из того же *положительного* класса, h_{neg} – изображение из другого *отрицательного* класса. Каждое изображение подавалось на вход сверточной нейронной сети для получения векторного представления. Затем векторные представления триплета подавались на вход нейросети, где в качестве функции ошибок использовалась функция *Triplet Loss*, рассмотренная в [19]:

$$L\left(e,e_{pos},e_{neg}\right)=\max\left\{0,dist\left(e,e_{pos}\right)-dist\left(e,e_{neg}\right)+g\right\}$$

где e,e_{pos},e_{neg} – векторные представления h,h_{pos},h_{neg} соответственно, *dist* – Евклидово расстояние, g – параметр сдвига (margin) между положительными и отрицательными классами. В нашем случае для обучения использовалось 100 тыс. случайно выбранных работ из датасета *Synthetic*, описанного в разд. 3.1, при этом для каждого изображения генерировалось пять почти дубликатов.

Таким образом, положительным классом считались сгенерированные почти дубликаты, отрицательным – другие изображения из взятых оригиналов. В качестве сверточной нейросети применялась ResNet-50 [20] с выбранной размерностью пространства векторных представлений 1024. Более детально процесс обучения описан в [18].

3.3. Векторный поиск. Имея обученную модель архитектуры *Deep Ranking*, можно попарно сравнивать изображения по L2-расстоянию между их векторными представлениями. Однако с увеличением коллекций до многотысячных и многомиллионных наборов изображений такой подход перестает быть реализуемым, так как имеет высокую вычислительную сложность. В связи с этим требуется найти подход, работающий более эффективно и способный осуществлять примерный поиск источников среди большого количества изображений.

Для приближенного решения задач поиска ближайших объектов существует широкий спектр методов, основанных на построении индекса, который дает возможность оптимизировать поиск и ускорить вычислительные процессы. Готовые библиотеки и фреймворки, такие, как Annoy [21], Pinecone [22] и Faiss [23], представляют собой библиотеки, специально разработанные для реализации методов поиска ближайших объектов. Они предлагают различные варианты построения индекса и алгоритмов поиска, позволяя настраивать параметры для достижения оптимальной производительности и точности результатов. Использование подобных фреймворков ускоряет процесс поиска ближайших объектов и повышает эффективность вычислительных операций при работе с большими объемами данных.

Для решения задачи был выбран IVF индекс библиотеки Faiss [23], так как библиотека имеет открытый код. Коллекция изображений, среди которых планируется искать источники заимствований, помещается в индекс. Идея IVF индекса заключается в создании заданного количества N кластеров, на которые делится пространство векторов. Каждый вектор принадлежит одному из кластеров. Во время поиска вектор-запрос сначала сравнивается с центральными элементами кластеров, из них отбирается *nprobe* ближайших. Дальнейший поиск производится только по векторам этих ближайших кластеров. В нашей работе были выбраны значения $N = 65536$, *nprobe* = 128. Также использована *Product Quantization* (PQ32), описанная в [23], для уменьшения размерности индекса и более эффективного поиска. Таким образом, коллекция источников индексируется, и по индексу производится поиск для каждого запроса. Результатом поиска является заданное количество ближайших векторов – кандидатов на источники дублирования.

3.4. Оценка схожести изображений. Для отбора источников из кандидатов, полученных на предыдущем шаге, предлагается искать некую меру схожести между изображением-запросом и изображениями, соответствующими найденным кандидатам. Построение этой части модели оказалось наиболее трудоемким. Далее описаны несколько подходов к решению данной части задачи и подбор наиболее релевантного из них.

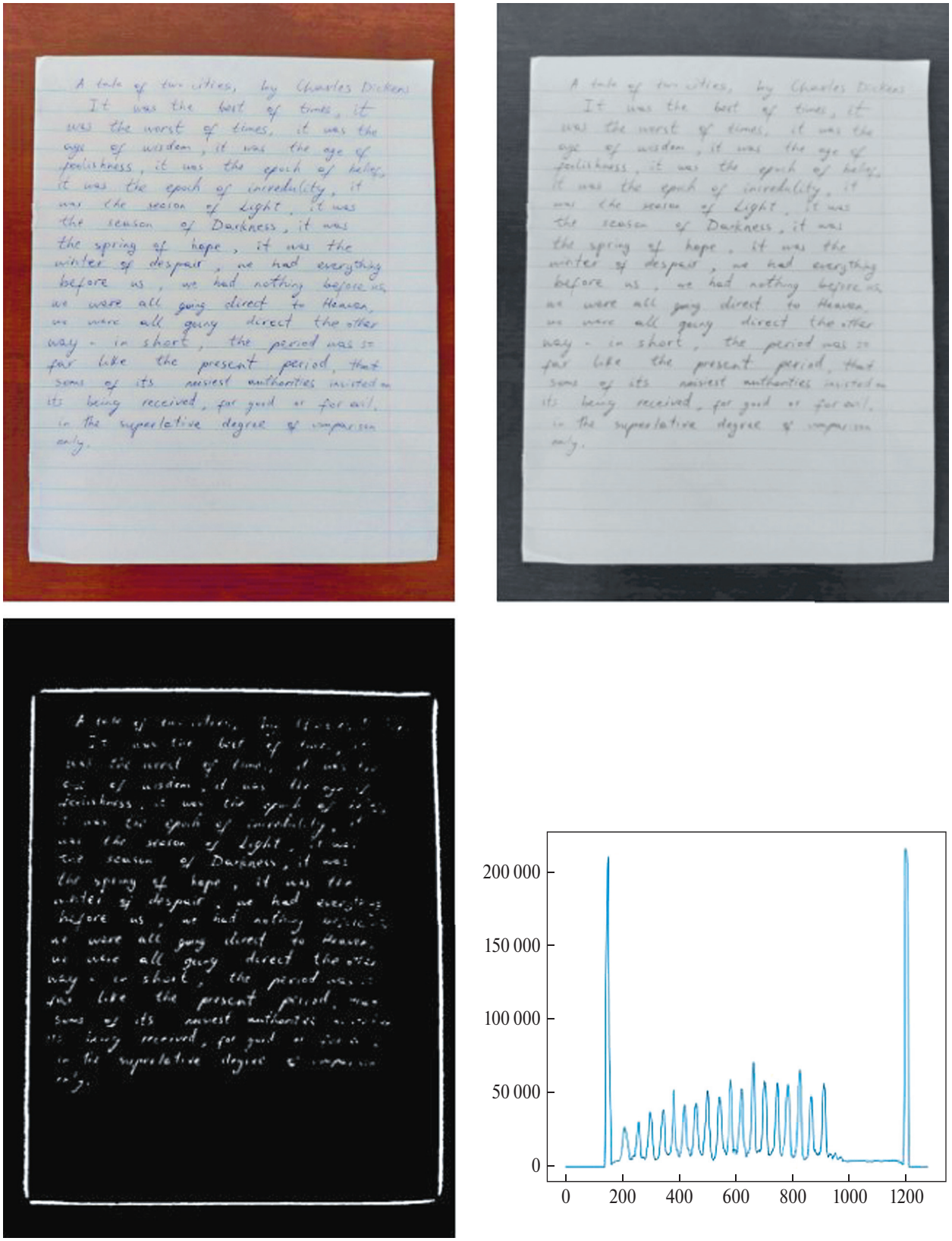

Рис. 3. Процесс извлечения сигнала из изображения: перевод в оттенки серого, применение адаптивного порога, выделение сигнала

Одним из распространенных подходов к поиску схожести между изображениями является использование Сиамских нейросетей [24]. Они позволяли достичь высоких результатов для случая, когда коллекция содержит большое количество изображений из разных доменов [25]. Однако в нашей задаче все изображения коллекции фактически принадлежат одному домену – фотографиям рукописных текстов. Согласно проведенному исследованию, применение Сиамских нейронных сетей не обеспечило приемлемого качества.

Другим подходом к данной части задачи выступают методы, используемые для анализа временных рядов. Так как почти дубликатом работы служит изображение, где текст расположен на том же (или почти том же) регионе, то текст входного изображения работы можно определить как некоторый сигнал. В рамках такого рассмотрения задачи в нашей работе извлекался сигнал из изображения (рис. 3). Для получения сигнала сначала цветное изображение преобразовывалось в черно-белое, затем выделялся рукописный текст с помощью фильтрации по яркости. Затем матрица изображения суммировалась по строкам. Далее из сигнала извлекались максимумы как наиболее информативная часть. После этого подсчитывалась схожесть между сигналом работы запроса и работы-кандидата. Схожесть определялась как Евклидово расстояние между последовательностями. Для вычисления Евклидова расстояния был использован алгоритм динамической трансформации временной шкалы (dynamic time wrapping (DTW)). Одним из практических применений алгоритма DTW является онлайн проверка подписи [26]. В работе был использован алгоритм FastDTW [27] – аппроксимация DTW, имеющая линейную вычислительную сложность вместо квадратичной у полной DTW. Для классификации изображения как почти дубликат или оригинал найденная схожесть между запросом и кандидатом сравнивалась с подобранным порогом. При превышении порога изображение считалось почти дубликатом, а кандидат отбирался как источник почти дублирования. Распространенными методами решения задачи сравнения изображений являются методы извлечения опорных точек (keypoint-extraction) от подходов классического компьютерного зрения: SIFT [28], ORB [29], до нейросетевых [30, 31]. Различные алгоритмы извлечения ключевых точек (SIFT, ORB или нейронные сети) используются для выбора ключевых точек и вычисления их дескрипторов для каждого изображения-кандидата и изображения-запроса. Дескриптор – это объект, содержащий некоторую информацию о ключевой точке, на основе которой можно сделать вывод о сходстве или различии между двумя ключевыми точками. Подсчитывалось и нормировалось количество схожих ключевых точек. При превышении установленного порога запрос считался почти дубликатом, а кандидат отбирался как итоговый кандидат на оригинал.

Долгое время разница в результатах между классическими и глубинными методами keypoint-extraction была незначительна [32]. Большинство нейросетевых подходов к сопоставлению ключевых точек требовало двух отдельных архитектур. Одна нейронная сеть определяла ключевые точки, а затем другая сравнивала их. Однако с появлением в компьютерном зрении архитектур типа Трансформер [33] возникли решения, объединяющие эти два этапа. Авторы LoFTR [34] представили новую архитектуру, основанную на архитектуре Трансформера. Вдохновленные их работой, мы модифицировали архитектуру LoFTR [34] в соответствии с конкретными требованиями нашей задачи. Исследования, описанные ниже, показали, что наиболее подходящим решением является использование именно этой архитектуры.

4. Эксперименты. Для выбора модели в последней части метода был проведен эсперимент по сравнению нескольких подходов: LoFTR, SIFT, Signal+FastDTW (разд. 3.4). Для этого был создан специальный набор данных, состоящий из 1000 элементов коллекции *Synthetic* (разд. 3.1). Для каждого почти дубликата сравнение проводилось по 511 кандидатам, помимо истинного оригинала. Кандидаты были получены с помощью поиска, чтобы приблизить эксперимент к реальному использованию. По результатам (представлены в табл. 1) был выбран подход LoFTR ввиду явного преимущества в качестве.

Таблица 1. Качество итогового отбора кандидатов при использовании различных подходов

Метод	*Recall*@1, %
SIFT	30.9
Signal+FastDTW	78.8
LoFTR modification	98.4

Эксперимент по измерению качества всего метода проводился на двух коллекциях: с синтетическими и реальными данными. В качестве метрик применялись *Recall*@1 (2.1) и FPR (2.2). Минимизация доли ложноположительных стабатываний представляла особую важность как минимизация ложных обвинений в дублировании.

Для первой коллекции, *Synthetic*, использовались открытые данные IAM [16] и READ2016 [17]. Были взяты все train и test примеры (747 IAM train, 336 IAM test; 350 Read2016 train, 50 Read2016 test). На тестовых примерах из обоих датасетов были сгенерированы почти дубликаты по два на каждое изображение. Для оценки FPR взяли все validation примеры (116 IAM; 50 Read2016).

С целью проверки метода на данных, приближенных к реальным (*Real*), применялась коллекция, составленная из датасета HWR200 (разд. 3.1). В качестве источника заимствования и почти дубликата поочередно использовались *светлые* (*RealLight*) и *темные* (*RealDark*) изображения. Значения метрик качества вычислялись как среднее для этих двух экспериментов. Для оценки доли ложноположительных стабатываний применялась отложенная выборка из объектов того же датасета, для которых не индексировались почти дубликаты.

Результаты эксперимента показаны в табл. 2. Видно, что даже на данных из HWR200 уровень качества остается достаточно высоким для использования в высоконагруженных системах обнаружения заимствований при достижении низкого количества ложных срабатываний. При этом модель обучалась только на синтетических данных.

Таблица 2. Качество работы полного метода

Метрика, %	Коллекция	
	Synthetic	*Real*
Recall@1	80	59
FPR	4.8	5.5

Также был проведен эксперимент на производительность, которая особенно важна при использовании больших коллекций данных. Производительность измерялась на машине с процессором AMD Ryzen 9 3900XT 12-Core Processor с помощью 8 ядер и 64 Gb RAM. Для модели-модификации LoFTR применялся графический процессор NVIDIA GeForce RTX 3090. Для оценки производительности было подано 500 изображений-запросов из датасета HWR200. Время работы полного метода поиска почти дубликатов составило в среднем 5.5 с на один запрос при общем размере коллекции гипотетических источников около 10 тыс. изображений. Стоит отметить, что основные временные затраты относятся к работе заключительного модуля метода, который обрабатывает уже фиксированное количество кандидатов. Следовательно, при увеличении коллекции скорость работы этого модуля изменяться не будет. Таким образом, увеличение коллекции не приведет к значительному изменению скорости работы системы.

Заключение. Представлен метод обнаружения почти дубликатов изображений рукописного текста, способный обеспечивать высокую производительность при работе с большими коллекциями документов. Подход состоит из нескольких этапов. Первый этап — построение векторных представлений изображений. Второй этап, поиск кандидатов, включает в себя поиск объектов, который сужает число возможных источников почти дублирования. Последний этап – сравнение изображений, в результате которого остается небольшое количество кандидатов, с большой долей вероятности являющихся источником почти дублирования. Были проведены эксперименты на двух коллекциях изображений рукописных текстов. В первом наборе данных содержатся изображения рукописных текстов, дубликаты которых были созданы синтетически, а второй набор состоит из фотографий рукописных текстов, имеющих реальные дубликаты. Было достигнуто 59 и 80% полноты и 5.5 и 4.8% доли ложноположительных срабатываний для набора рукописных сочинений, приближенного к реальным данным, и синтетического набора соответственно. При этом время обработки составило в среднем 5.5 с на запрос. Полученные результаты свидетельствуют о том, что предложенный подход может быть использован в качестве решения для высоконагруженных систем обнаружения заимствований.

Дальнейшие исследования могут быть направлены на работу с изображениями рукописных текстов низкого качества, а также на обработку изображений рукописей на других языках, например, имеющих иероглифическую письменность.

СПИСОК ЛИТЕРАТУРЫ

1. *Bakhteev O., Ogaltsov A., Khazov A., Safin K., Kuznetsova R.* CrossLang: the System of Cross-lingual Plagiarism Detection // Workshop on Document Intelligence at NeurIPS. Vancouver, 2019.
2. *Avetisyan K., Gritsay G., Grabovoy. A.* Cross-Lingual Plagiarism Detection: Two Are Better Than One // Programming and Computer Software. 2023. V. 49. P. 346–354.
3. *Kuznetsova M., Bakhteev O., Chekhovich Y.* Methods of Cross-lingual Text Reuse Detection in Large Textual Collections // Informatika I Ee Primeneniya [Informatics and Its Applications]. 2021. V. 15. P. 30–41.
4. *Gritsay G., Grabovoy A., Kildyakov A., Chekhovich Y.* Artificially Generated Text Fragments Search in Academic Documents // Doklady Rossijskoj Akademii Nauk. Matematika, Informatika, Processy Upravlenia. 2023. V. 108. P. 308–317.
5. *Gritsay G., Grabovoy A., Chekhovich Y.* Automatic Detection of Machine Generated Texts: Need More Tokens // Ivannikov Memorial Workshop (IVMEM). Kazan, 2022. V. 108. P. 20–26.
6. *Ma H.J., Wan G., Lu E.Y.* Digital Cheating and Plagiarism in Schools // Theory Into Practice. 2008. V. 47. P. 197–203.
7. *Wrigley S.* Avoiding 'de-plagiarism': Exploring the Affordances of Handwriting in the Essay-writing Process // Active Learning in Higher Education. 2019. V. 20. P. 167–179.
8. *Bakhteev O., Kuznetsova R., Khazov A., Ogaltsov A., Safin K., Gorlenko T., Suvorova M., Ivahnenko A., Botov P. et. al.* Near-duplicate Handwritten Document Detection Without Text Recognition // Intern. Conf. on Computational Linguistics and Intellectual Technologies. Moscow, 2021. P. 47–57.
9. *Krishnan P., Jawahar C.V.* Matching Handwritten Document Images // Europ. Conf. on Computer Vision. Amsterdam, 2016. P. 766–782.
10. *Rowtula V., Bhargavan V., Kumar M., Jawahar C.V.* Scaling Handwritten Student Assessments with a Document Image Workflow System // IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Salt Lake City, 2018. P. 2307–2314.
11. *Pandey O., Gupta I., Mishra B.S.P.* A Robust Approach to Plagiarism Detection in Handwritten Documents // Intern. Sympos. on Visual Computing. San Diego, 2020. P. 682–693.
12. *Coquenet D., Chatelain C., Paquet T.* End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network // ArXiv 2021. ArXiv Preprint ArXiv:2012.03868.
13. *Voigtlaender P., Doetsch P., Ney H. H*andwriting Recognition With Large Multidimensional Long Short-term Memory Recurrent Neural Networks // 15th Intern. Conf. on Frontiers in Handwriting Recognition (ICFHR). Shenzhen, 2016. P. 228–233.
14. *Khritankov A., Botov P., Surovenko N., Tsarkov S., Viuchnov D., Chekhovich Y.* Discovering Text Reuse in Large Collections of Documents: A Study of Theses in History Sciences // Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conf. (AINLISMW FRUCT). St. Petersburg, 2015. P. 26–32.
15. *Potyashin I., Kaprielova M., Chekhovich Y., Kildyakov A., Seil T., Finogeev E., Grabovoy A.* HWR200: New Open Access Dataset of Handwritten Texts Images in Russian // Intern. Conf. on Computational Linguistics and Intellectual Technologies. Moscow, 2023.
16. *Grieggs S., Shen B., Rauch G., Li P., Ma J., Chiang D., Price B., Scheirer W.J.* Measuring Human Perception to Improve Handwritten Document Transcription // ArXiv 2019. ArXiv Preprint ArXiv:1904.03734 .
17. *Toselli A., Romero V., Villegas M., Vidal E., Sanchez J.* HTR Dataset // Intern. Conf. on Frontiers in Handwriting Recognition (ICFHR). Shenzhen, 2016. P. 630635.
18. *Wang J., Song Y., Leung T., Rosenberg C., Wang J., Philbin J., Chen B., Wu Y.* Learning Fine-grained Image Similarity With Deep Ranking // IEEE Conf. on Computer Vision and Pattern Recognition. Columbus, 2014. P. 1386–1393.
19. *Balntas V., Riba E., Ponsa D., Mikolajczyk K.* Learning Local Feature Descriptors With Triplets and Shallow Convolutional Neural Networks // The British Machine Vision Conference (BMVC). 2016. V. 1. №2. P. 3.
20. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016.
21. Annoy // https://github.com/spotify/annoy.
22. Pinecone // https://github.com/pinecone-io.
23. *Johnson J., Douze M., J'egou H.* Billion-scale Similarity Search With GPUs // IEEE Transactions on Big Data. 2019. V. 7. P. 535–547.
24. *Melekhov I., Kannala J., Rahtu E.* Siamese Network Features for Image Matching // 23rd Intern. Conf. on Pattern Recognition, ICPR. Cancun, 2016. P. 378–383.
25. *Bakhteev O., Chekhovich Y., Finogeev E., Gorlenko T., Kaprielova M., Kildyakov A., Ogaltsov A.* Image Reuse Detection in Large-scale Document Scientific Collection // ENAI Conf., Concurrent Sessions 12. Porto, 2022. P. 107.
26. *Patil B. V., Patil P. R.* An Efficient DTW Algorithm for Online Signature Verification // Intern. Conf. On Advances in Communication and Computing Technology (ICACCT). Painpat, 2018. P. 1–5.
27. *Salvador S., Chan P.* Toward Accurate Dynamic Time Warping in Linear Time and Space // Intellectual Data Analysis. 2007. V. 11. P. 561–580.
28. *Lowe D.G.* Distinctive Image Features from Scale-invariant Keypoints // Intern. J. of Computer Vision. 2004. V. 60. P. 91–110.

29. *Rublee E., Rabaud V., Konolige K., Bradski G.* ORB: An Efficient Alternative to SIFT or SURF // Intern. Conf. on Computer Vision. Barcelona, 2011. P. 25642571.
30. *DeTone D., Malisiewicz T., Rabinovich A.* Superpoint: Self-supervised Interest Point Detection and Description // IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Salt Lake City. 2018, P. 224–236.
31. *Barroso-Laguna A., Riba E., Ponsa D., Mikolajczyk K.* Key. net: Keypoint Detection by Handcrafted and Learned cnn Filters // IEEE/CVF Intern. Conf. on Computer Vision. Seoul, 2019. P. 5836–5844.
32. *Mishkin D.* Local Features: from Paper to Practice // Computer Vision and Pattern Recognition (CVPR) Workshops. Seattle, 2020.
33. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Polosukhin I. et. al.* Attention Is All You Need // ArXiv 2017. ArXiv Preprint ArXiv:1706.03762.
34. *Sun J., Shen Z., Wang Y., Bao H., Zhou X.* LoFTR: Detector-Free Local Feature Matching With Transformers // ArXiv 2021. ArXiv Preprint ArXiv:2104.00680. P. 8922–8931.