

УДК 004.9

Поиск изображений в графических базах данных по их содержанию

Д. Ф. Ходарев, А. В. Боднар

Донецкий национальный технический университет, г. Донецк

кафедра программной инженерии им. Л.П. Фельдмана

E-mail: vip.khodarev@gmail.com

Аннотация

В статье рассмотрены важность и проблематика поиска изображения по их содержанию. Проведен анализ подходов, включая алгоритмы сопоставления ключевых точек, графовые методы и использование сверточных нейронных сетей. Выполнено сравнение основных известных принципов, а также достоинства и недостатки различных подходов поиска изображений. Представлены результаты кластеризации дубликатов, которые применяются для группировки похожих или идентичных элементов в наборе. Выполнен анализ применения кластеризации дубликатов в различных областях.

Введение

Исследования в области поиска изображений по содержанию стали чрезвычайно актуальны в последние десятилетия в связи с увеличением емкости накопителей информации, распространением цифровой фотографии, а также стремительного роста популярности сети интернет. В текущей ситуации, когда ежедневно в сети публикуются десятки, если не сотни миллионов изображений, становится всё сложнее найти то, что нужно. Раньше использовались различные текстовые метаданные, например, теги описания или ключевые слова. Этот способ потерял эффективность с ростом объема в базах данных из-за невозможности точного отслеживания присвоения им этих метаданных. С появлением более сложных алгоритмов обработки изображений и развитием методов глубокого обучения, системы поиска стали более точными и способны обрабатывать сложные запросы.

Существует множество разных реализаций, например, цветовые гистограммы [1], поиск по ключевым точкам (SIFT, SURF, ORB), использование векторных представлений

(Embeddings) и др [2]. Каждый из них нашел своё применение в поиске, обработке и анализе изображений. В завершение следует отметить, что поиск изображений по их содержанию представляет собой активно развивающуюся область исследований в области компьютерного зрения. Современные подходы, такие как использование сверточных нейронных сетей и алгоритмов сопоставления ключевых точек, открывают новые перспективы для повышения эффективности и точности поиска визуальных данных.

Поиск похожих изображений (CBIR)

Поиск похожих изображений, также известный как Content-Based Image Retrieval (CBIR), представляет собой метод, основанный на анализе содержания изображений для поиска и извлечения визуально схожих объектов или сцен. Этот подход учитывает не только метаданные или ключевые слова, но и собственные визуальные характеристики изображений. Процесс поиска можно наблюдать на рис.1.

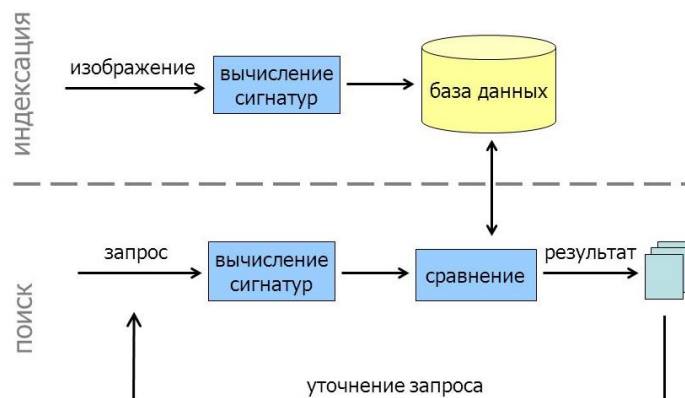


Рисунок 1 – Схема работы CBIR

На рисунке чётко видна последовательность занесения в БД новых данных, а также собственно поиск и сравнение с ней [3].

К основным компонентам этого метода, часто используемым вместе, относятся [4]:

- Извлечение признаков (Feature Extraction): Процесс извлечения важных визуальных характеристик изображений, которые могут включать цвета, текстуры, формы, и другие параметры. Эти признаки помогают создать уникальное представление для каждого изображения.

- Индексация (Indexing): Создание эффективной структуры данных для хранения и организации извлеченных признаков. Индексация ускоряет процесс поиска и обеспечивает более быстрый доступ к хранимым в базе данных изображениям.

- Метрики сходства (Similarity Metrics): Определение методов измерения сходства между изображениями на основе их извлеченных признаков. Эти метрики позволяют оценить, насколько два изображения похожи друг на друга.

- Методы поиска (Search Methods): Разработка алгоритмов для эффективного поиска изображений в базе данных на основе запроса пользователя. Эти методы могут включать в себя индексированный поиск, использование метрик сходства и оптимизированные алгоритмы поиска.

К преимуществам метода можно отнести:

- основан на визуальных характеристиках, что обеспечивает более точные результаты, особенно при многозначном индексировании;

- имеет возможность распознавания объектов, что выделяет его среди других для поиска визуально связанных данных;

- широкий спектр применения, включая медицинскую диагностику, организацию поиска изображений в графических базах данных, в безопасности и видеонаблюдении, а также дизайне и моде.

Несмотря на преимущества, есть и недостатки. Рассмотрим основные.

1) Субъективность восприятия. CBIR основан на визуальных характеристиках, которые могут варьироваться в зависимости от субъективного восприятия разных людей. Таким образом, различия в оценке цветов, текстур или форм могут сказаться на его эффективности [5].

2) Проблемы масштабирования. Некоторые методы извлечения признаков могут быть чувствительны к масштабированию изображений. Это означает, что изображения одного объекта с разными масштабами могут восприниматься как различные, что усложняет поиск.

3) Недостаточный учет разнообразия контента. Некоторые визуальные характеристики, например, форма объектов или текстуры, не всегда полностью передают смысл содержания изображения. В результате, CBIR может не учитывать некоторые аспекты контента, которые важны для пользователя.

4) Сложности в работе с большими базами данных. С увеличением размера базы данных возрастает сложность поиска и сопоставления изображений, требуя более эффективных методов индексации и поиска для поддержания производительности.

5) Неучет семантического контекста. CBIR часто игнорирует семантический контекст, связанный с изображением. Например, для изображения льва CBIR может выделить цвета и форму, но не обязательно понять, что это изображение льва.

6) Чувствительность к изменениям в освещении и угле обзора. Извлечение признаков, таких как цвет, может быть чувствительным к изменениям в освещении, что может затруднить сопоставление изображений при различных условиях освещения.

Поиск по ключевым точкам

Методы поиска по ключевым точкам, такие как SIFT, SURF и ORB, являются популярными в области компьютерного зрения [6]. В этих алгоритмах, точки выбираются так, чтобы они были устойчивы к изменениям масштаба, вращениям и изменениям в яркости.

Каждая ключевая точка описывается локальной окрестностью, называемой дескриптором. Дескрипторы содержат информацию о цвете, текстуре, и других визуальных характеристиках вокруг ключевой точки. Описания ключевых точек индексируются и хранятся в базе данных. Для эффективного поиска создаётся структура данных, позволяющая быстро находить сопоставления.

Когда поступает запрос, система извлекает ключевые точки и их дескрипторы, сравнивает их с данными, хранящимися в базе, и определяет наилучшие сопоставления. Процедура может включать в себя использование метрик сходства, таких как евклидово расстояние или косинусное сходство.

На следующем шаге происходит ранжирование этих сопоставлений, и пользователь получает на выходе набор изображений, наиболее близких к запросу.

Вкратце рассмотрим упомянутые методы и определим их преимущества и недостатки.

SIFT (Scale-Invariant Feature Transform)

Разработан Дэвидом Лоу в 1999 году, SIFT является одним из первых и наиболее

широко используемых методов. Он выделяет ключевые точки, устойчивые к изменениям масштаба и вращения.

Преимущества:

- инвариантность к масштабированию и вращению;
- робастность к изменениям в освещении и частичным сокрытиям.

Недостатки:

- требователен к вычислительным ресурсам;
- большой объем хранимых данных, что может сказаться на производительности.

SURF (Speeded-Up Robust Features)

Разработан в 2006 году Гербертом Бэйем и другими. SURF был создан для улучшения производительности по сравнению с SIFT, сохраняя его преимущества. Он использует быстрый алгоритм определения местоположения ключевых точек.

Преимущества:

- более быстрое вычисление по сравнению с SIFT;
- эффективен для реального времени.

Недостаток – менее устойчив к некоторым изменениям в изображении по сравнению с SIFT.

ORB (Oriented FAST and Rotated BRIEF)

Разработан в 2011 году Эдвардом Рублингом и др. ORB является более новым методом, созданным для обеспечения быстрого и эффективного поиска по ключевым точкам, поддерживая высокую производительность.

Преимущества:

- очень быстрый и эффективный;
- меньшее потребление памяти.

Недостаток - может быть менее точным в некоторых случаях по сравнению с SIFT и SURF.

Эти методы широко используются в задачах компьютерного зрения, включая поиск похожих изображений, реконструкцию 3D-сцен, распознавание объектов и другие приложения. Применяются в областях, таких как медицинская диагностика, робототехника, андройды, виртуальная реальность и многое другое. Выбор метода зависит от требований конкретной задачи, производительности и ресурсов, доступных для вычислений.

Свёрточные нейронные сети

Сверточные нейронные сети (СНС) представляют собой мощный класс алгоритмов машинного обучения, широко применяемых в обработке изображений и анализе визуальных данных [7]. Эти сети спроектированы для автоматического извлечения иерархических признаков из входных изображений.

Основная идея СНС заключается в использовании сверточных слоев для

локального анализа изображений. Свертки – это фильтры, которые применяются к небольшим областям входного изображения, что позволяет сети выделять локальные шаблоны и признаки. Постепенно увеличивается абстрактность признаков на более высоких уровнях иерархии.

Пулинговые слои применяются для уменьшения размерности данных, сохраняя важные признаки. Это способствует локализации и обобщению выделенных черт. После нескольких слоев свертки и пулинга, данные передаются полносвязным слоям для классификации или других задач обработки информации.

Сверточные нейронные сети обладают уникальной способностью к автоматическому обучению пространственных и иерархических представлений изображений. Это делает их особенно эффективными для решения задач распознавания образов, классификации изображений, обнаружения объектов и даже создания искусственных генеративных моделей. Их применение не ограничивается только областью компьютерного зрения, и сверточные нейронные сети успешно применяются в различных областях, включая медицинскую диагностику, автономные транспортные средства и анализ естественного языка.

Использование сверточных нейронных сетей (СНС) предоставляет ряд значительных преимуществ. Среди плюсов стоит выделить способность СНС автоматически извлекать иерархические признаки из входных данных, что делает их эффективными в задачах обработки изображений. СНС также демонстрируют хорошую способность к обучению и обобщению, что делает их подходящими для различных задач машинного обучения [7].

Однако, использование СНС сопряжено с некоторыми недостатками. Требовательность к вычислительным ресурсам и объему данных может создавать проблемы в реализации на устройствах с ограниченными возможностями. Также, интерпретация и визуализация внутренней работы СНС могут быть сложными, что усложняет понимание принимаемых ими решений. Кроме того, не всегда легко подобрать оптимальную архитектуру и настройки для конкретной задачи, что может потребовать значительных усилий в процессе настройки модели.

Проблема дубликатов

Среди задач поиска изображений на основе их содержимого есть проблема в наличии дубликатов. Допустим, мы хотим найти какую-то вещь, но из-за чрезмерной популярности данного изображения, мы получим лишь сотни или тысячи дубликатов,

один из которых отображен на рис. 2. Для решения данной проблемы предложена кластеризация дубликатов [8].

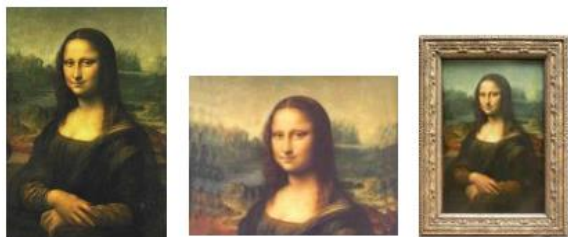


Рисунок 2 – Тумбнейлерные дубликаты

Процесс кластеризации дубликатов часто применяется для группировки похожих или идентичных элементов в наборе (рис.3). Это полезно в случаях, когда при работе с базой данных изображений или текстовыми документами требуется выявить, объединить или удалить эти самые дубликаты.

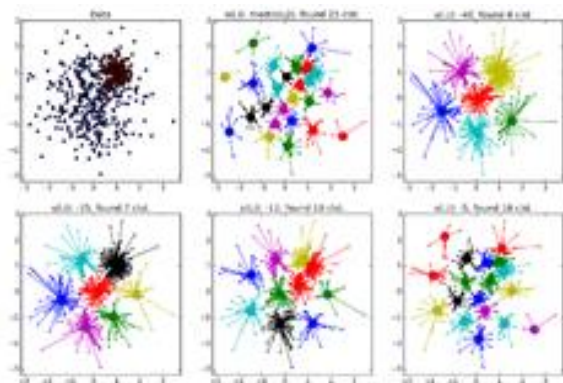


Рисунок 3 – Наглядный пример кластеризации

Рассмотрим несколько методов и подходов реализации этого процесса.

Хэш-функции.

Использование хэш-функций для быстрого обнаружения дубликатов. Если два элемента имеют одинаковый хэш, они могут считаться кандидатами в дубликаты, и затем дополнительные шаги могут быть предприняты для подтверждения этого.

Сравнение по признакам

Реализация сравнения по признакам для выявления сходства между элементами может включать в себя сравнение векторных представлений (например, векторов признаков изображений или текста) и применение метрик сходства.

Методы машинного обучения

Применяются для обучения модели, которая может классифицировать, являются ли пары элементов дубликатами. Может включать в себя использование пар признаков и классификацию на основе близости.

Блочная обработка

Разделение данных на блоки и сравнение элементов только внутри блоков, что может ускорить процесс и сделать его более масштабируемым.

Сверточные нейронные сети

Использование сверточных нейронных сетей для извлечения признаков и определения сходства между изображениями. Этот метод часто используется для кластеризации дубликатов изображений.

Методы на основе графов

Представление данных в виде графа и использование алгоритмов кластеризации для выявления подграфов, представляющих собой дубликаты.

В зависимости от ситуации, типа данных, точности, объема, скорости и вычислительных способностей, могут применяться различные методы. Поэтому выбор метода кластеризации дубликатов зависит от требований конкретной задачи.

Применение кластеризации дубликатов в различных областях

Кластеризация в различных задачах имеет ряд проблем, с которым сталкиваются разработчики и что влияет на эффективность и скорость работы. Среди основных проблем можно выделить следующие:

- выбор подходящего метода;
- определение числа кластеров (неправильно выбранное число может негативно повлиять на результаты);
- сложность алгоритмов (некоторые алгоритмы требуют высокие вычислительные ресурсы);
- чувствительность к выбросам;
- интерпретация результатов;
- необходимость в предварительной обработке данных и др.

В независимости от сложности, данные методы высоко востребованы в различных областях. Среди множества направлений применений можно выделить несколько наиболее важных в данное время.

Медицинская диагностика

В [9] дано обсуждение использования кластеризации дубликатов в медицинской области для обработки и анализа медицинских изображений, включающее в себя выявление дубликатов снимков, что особенно важно для точной диагностики и мониторинга состояния пациентов.

Промышленность и обработка изображений

Исследование использования методов кластеризации для обработки изображений в промышленности. Это может включать в себя поиск и управление дубликатами в графических

базах данных, применение в качестве инструмента для качественного контроля, анализа производственных процессов и т.д.

Архивация и управление данными

Рассмотрение применения кластеризации дубликатов в области архивации и управления данными. Может включать в себя оптимизацию хранения изображений, выявление и удаление дубликатов при архивировании больших объемов данных.

Интернет-платформы и социальные сети

Исследование роли кластеризации дубликатов в сфере социальных сетей и интернет-платформ. Здесь кластеризация может применяться для улучшения качества поиска изображений, а также для обнаружения и управления дубликатами, что особенно актуально в условиях постоянного потока новых данных.

Культурное наследие и цифровые архивы

Рассмотрение применения в сфере сохранения культурного наследия и создания цифровых архивов. Кластеризация дубликатов может помочь в эффективной организации и поддержке цифровых коллекций, облегчая поиск и управление контентом [10].

Образование и научные исследования

Обзор использования кластеризации дубликатов в образовательных и научных целях. Это может включать в себя работу с графическими базами данных для создания обучающих материалов, анализа результатов исследований и обеспечения эффективной обработки изображений в научных проектах.

Безопасность и анализ изображений

Рассмотрение применения кластеризации дубликатов в сфере безопасности и анализа изображений. Это может включать в себя выявление схожих изображений для обеспечения безопасности, а также их кластеризацию для выявления закономерностей и трендов.

Заключение

Выполненное исследование позволило выявить множество аспектов и вызовов в области обработки и анализа графической информации. Методы поиска похожих изображений (CBIR) и по ключевым точкам предоставляют эффективные инструменты для поиска и анализа визуальных данных. Однако, существуют проблемы, такие как сложность выбора оптимального метода, чувствительность к различным типам данных, вопросы интерпретации результатов, а также проблемы, связанные с дубликатами.

Анализ проблем дубликатов выделяет важность кластеризации для обработки и

группировки похожих или идентичных элементов в графических базах данных. Этот процесс включает в себя использование различных методов, таких как хэш-функции, сравнение по признакам, методы машинного обучения и сверточные нейронные сети. Кластеризация дубликатов играет ключевую роль в решении проблемы избыточности и обеспечивает эффективное управление данными в графических базах данных.

Одновременно с этим, необходимость решения проблем выбора метода кластеризации, обработки разнородных данных и поддержания сбалансированности весов признаков представляет перспективы для дальнейших исследований в области поиска изображений на основе их содержимого. Тенденции в развитии включают в себя интеграцию с технологиями глубокого обучения, семантическое понимание изображений и применение в различных сферах, от медицинской диагностики до обработки изображений в промышленности.

Постоянное внимание к проблемам и направлениям исследований в этой области позволит эффективно решать вызовы и создавать инновационные методы для поиска изображений на основе их содержимого.

Литература

1. Полный перечень методов CBIR [Интернет-ресурс]. - URL: https://en.wikipedia.org/wiki/List_of_CBIR_engines
2. Поиск изображений на основе содержания habr [Интернет-ресурс]. - URL: <https://habr.com/ru/articles/103385/>
3. Васильева, Н. Поиск изображений. Синтез различных методов поиска при формировании результатов / Н. Васильева, А. Дольник, И. Марков // Интернет-математика 2007: сб. работ участников конкурса науч. проектов по информ. Поиску. — Екатеринбург : Изд-во Урал. ун-та, 2007. — С. 46–55.
4. Karami, E. Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images [Electronic resource] / Ebrahim Karami, Siva Prasad, Mohamed Shehata. - URL: <https://arxiv.org/abs/1710.02726>
5. Десятников, И. Е. Поиск изображений по визуальному содержанию в графических базах данных и сети Интернет [Электронный ресурс] / И. Е. Десятников // Информационные технологии и вычислительные системы, 2013. - № 2. - С. 88-95. - URL: https://www.isa.ru/jitcs/images/documents/2013-02/88_95.pdf
6. Шозда, Н. С. Поиск изображений в больших БД с использованием коэффициента корреляции цветовых гистограмм [Электронный ресурс] / Н. С. Шозда, Е. А. Башков // Графикон, 2002. - URL: <http://www.graphicon.ru/2002/s89>.

7. Романов, А. А. Сверточные нейронные сети [Электронный ресурс] / А. А. Романов // Научные исследования, 2018. - № 1(21). – URL: <https://scientificresearch.ru/index.php/blizhajshij-nomer/01-00-00-fiziko-matematicheskie-nauki/266-svertochnye>

8. Ouhda, M. Using Image Segmentation in Content Based Image Retrieval Method [Электронный ресурс] / Mohamed Ouhda, Khalid El Asnaoui, Mohammed Ouanan, Brahim Aksasse // Advanced Information Technology, Services and Systems (AIT2S 2017). Conference Paper. - P. 179-195. – URL: https://www.researchgate.net/publication/321019137_Using_Image_Segmentation_in_Content_Based_Image_Retrieval_Method

9. Левчук, В. А. Компьютеризированная диагностика меланомы на базе поиска похожих дерматоскопических изображений в базе данных / В. А. Левчук, В. А. Ковалев, В. В. Баркалин, В. Э. Лозовский // Весті Національної академії наук України, 2016. – № 2. – С. 86-91.

10. Тюрин, А. Г. Кластерный анализ, методы и алгоритмы кластеризации / А. Г. Тюрин, И. О. Зуев // Вестник МГТУ МИРЭА. - М.: Изд-во МГТУ, 2014. - № 3.

Ходарев Д. Ф., Боднар А. В. Поиск изображений в графических базах данных на основе их содержания. В статье рассматривается важность и проблематика поиска изображения по их содержанию. Проведен анализ подходов, включая алгоритмы сопоставления ключевых точек, графовые методы и использование сверточных нейронных сетей. Проведено сравнение основных известных принципов, а также плюсы и минусы различных подходов поиска изображений. Представлены результаты кластеризации дубликатов, которые применяются для группировки похожих или идентичных элементов в наборе. Выполнен анализ применения кластеризации дубликатов в различных областях.

Ключевые слова: поиск изображений, CBIR, ключевые точки, проблема дубликатов, кластеризация, методы CBIR

Khodarev D. F., Bodnar A. V. Search for images in graphic databases using their contents. The article discusses the importance and problems of image search by their content. The analysis of approaches, including algorithms for matching key points, graph methods and the use of convolutional neural networks, is carried out. The main known principles are compared, as well as the pros and cons of various image search approaches. The results of clustering duplicates are presented, which are used to group similar or identical elements in a set. The analysis of the application of duplicate clustering in various fields is performed.

Keywords: image search, CBIR, key points, duplicate problem, clustering, CBIR methods

Статья поступила в редакцию 05.12.2023
Рекомендована к публикации профессором Мальчевой Р. В.