Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Сибирский государственный аэрокосмический университет
имени академика М. Ф. Решетнева»

На правах рукописи

Проскурин Александр Викторович

МЕТОДЫ И АЛГОРИТМЫ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ИЗОБРАЖЕНИЙ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

Специальность 05.13.17 – Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель –
доктор технических наук
Фаворская М. Н.

Красноярск – 2017

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность работы. В последние десятилетия широкое распространение устройств со встроенными видеокамерами привело к экспоненциальному росту количества изображений в сети интернет, что вызвало необходимость их эффективного поиска. Существующие методы поиска изображений можно разделить на три типа: поиск по текстовым аннотациям, анализ изображений по визуальному содержанию и методы на основе автоматического аннотирования. В поисковых методах первого типа изображениям вручную присваиваются субъективные текстовые описания, а поиск осуществляется как в текстовых документах. Методы поиска изображений по содержанию, требующие изображение-запрос, выполняют поиск на основе анализа и сравнения низкоуровневых признаков изображения, таких как цвет или текстура. Однако при этом часто наблюдается проблема семантического разрыва – отсутствия связи между низкоуровневыми признаками изображения и его интерпретацией человеком. Основной идеей методов автоматического аннотирования изображений (ААИ) является формирование семантической модели из обучающей выборки изображений большого объема. С помощью семантической модели автоматически определяются ключевые слова для новых изображений. Таким образом, методы автоматического аннотирования предполагают поиск по ключевым словам, полученным на основе анализа содержания изображений, и используют преимущества первых двух подходов.

Наиболее активные исследования в области автоматического аннотирования изображений проводятся в таких университетах, как: University of California (США), Massachusetts Institute of Technology (США), University of Central Florida (США), Pennsylvania State University (США), University of Florence (Италия), International Institute of Information Technology (Индия). Среди отечественных учреждений, занимающихся данной тематикой, можно отметить Томский политехнический университет (Томск),

Южный федеральный университет (Таганрог). Большой вклад в развитие методов автоматического аннотирования изображений внесли P. Duygulu, A. Makadia, Y. Verma, L. Ballan, S.L. Feng, M. Guillaumin, V. Lavrenko, A.C. Мельниченко, А.А. Друки и другие.

Однако до сих пор существует ряд проблем, связанных с автоматическим аннотированием изображений. Разработанные экспериментальные системы с большой долей достоверности определяют только 2–3 ключевых слова, при этом для формирования семантической модели необходимы большие вычислительные затраты, а добавление новых ключевых слов требует повторного обучения поисковой системы.

Целью диссертационной работы является повышение эффективности автоматического аннотирования изображений в информационно-поисковых системах.

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

1. Провести анализ методов и алгоритмов автоматического аннотирования изображений, кластеризации данных, описания изображений с помощью низкоуровневых признаков.
2. Разработать алгоритм быстрого параллельного вычисления набора локальных дескрипторов для описания изображения.
3. Разработать алгоритм восстановления пропущенных ключевых слов в аннотациях обучающих изображений.
4. Разработать метод кластеризации изображений в однородные текстово-визуальные группы с помощью самоорганизующейся нейронной сети.
5. Создать алгоритм автоматического аннотирования изображений на основе однородных текстово-визуальных групп.
6. Разработать программное обеспечение, реализующее алгоритмы вычисления дескрипторов, восстановления пропущенных ключевых слов,

формирования однородных текстово-визуальных групп и автоматического аннотирования изображений.

7. Провести экспериментальные исследования эффективности разработанных алгоритмов на тестовых наборах изображений.

Область исследования. Работа выполнена в соответствии с пунктами 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» и 7 «Разработка методов распознавания образов, фильтрации, распознавания и синтеза изображений, решающих правил. Моделирование формирования эмпирического знания» паспорта специальностей ВАК (технические науки, специальность 05.13.17 – Теоретические основы информатики).

Методы исследования. Для решения поставленных в работе задач использовались методы теории цифровой обработки изображений, теории обработки информации, методы теории распознавания образов и анализа данных, методы объектно-ориентированного программирования.

Новые научные результаты, выносимые на защиту:

1. Впервые разработан метод автоматического аннотирования изображений, основанный на разделении обучающего набора изображений на однородные текстово-визуальные группы. Метод отличается тем, что аннотирование нового изображения осуществляется с помощью обучающих изображений небольшого количества визуально похожих групп, что обеспечивает повышение точности и полноты аннотирования изображений.

2. Разработан новый метод двухэтапной кластеризации изображений с помощью модифицированной самоорганизующейся нейронной сети на основе текстовых и визуальных дескрипторов. Метод позволяет формировать однородные текстово-визуальные группы, которые представляют собой контекст для аннотирования новых изображений, и уточнять их в течение жизненного цикла системы.

3. Предложен новый метод расширения аннотаций обучающих изображений, позволяющий восстановить ключевые слова, пропущенные при составлении обучающих выборок. Метод отличается автоматическим определением количества пропущенных ключевых слов и позволяет повысить точность аннотирования изображений.

4. Разработан алгоритм быстрого извлечения набора локальных дескрипторов, описывающих все части изображения, позволяющий существенно ускорить процесс аннотирования и формировать более полный глобальный визуальный дескриптор изображения.

Практическая значимость. Предложенные в диссертационной работе методы и алгоритмы предназначены для практического применения в программном обеспечении информационно-поисковых систем интернета, а также могут использоваться для анализа и аннотирования изображений, полученных с помощью мобильных платформ. В рамках диссертационного исследования разработано экспериментальное программное обеспечение для автоматического аннотирования изображений.

Реализация результатов работы. Материалы диссертационного исследования переданы для дальнейшего использования в ООО «НПП «Бевард», о чем получен акт от 12.08.2015. Получен акт о внедрении результатов диссертационного исследования в учебный процесс кафедры информатики и вычислительной техники Института информатики и телекоммуникаций от 15.02.2017. Получены свидетельства о регистрации программ для ЭВМ: программа «Система автоматического формирования визуальных слов (ForVW)» (№2015611845 от 6.02.2015), программа «Система автоматического аннотирования изображений (AIA)» (№2016611307 от 29.01.2016).

Апробация работы. Основные положения и результаты диссертации докладывались и обсуждались на XVI, XVIII, XIX международных научных конференциях «Решетневские чтения» (Красноярск, 2012, 2014, 2015 гг.), всероссийской научной конференции студентов, аспирантов и молодых

ученых «Наука. Технологии. Инновации» (Новосибирск, 2013 г.), международной научно-практической конференции «Электронные средства и системы управления» (Томск, 2013 г.), 16-й, 17-й, 18-й международных конференциях и выставках «Цифровая обработка сигналов и ее применение» (Москва, 2014, 2015, 2016 гг.), международной научной конференции «Региональные проблемы дистанционного зондирования Земли» (Красноярск, 2014 г.), 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (Сингапур, 2015 г.).

Публикации. По результатам диссертационного исследования опубликовано 20 печатных работ, из которых 4 изданы в журналах, рекомендованных ВАК, 2 в журналах и книгах, индексируемых в Scopus, 12 в материалах докладов, 2 свидетельства, зарегистрированных в Российском реестре программ для ЭВМ.

Структура работы. Работа состоит из введения, трех глав, заключения, списка литературы и четырех приложений. Основной текст диссертации содержит 129 страниц, изложение иллюстрируется 28 рисунками и 15 таблицами. Библиографический список включает 108 наименований.

ГЛАВА 1. АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ И АЛГОРИТМОВ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ИЗОБРАЖЕНИЙ

В главе представлен обзор существующих методов автоматического аннотирования изображений, кластеризации данных и описания изображений с помощью низкоуровневых признаков, приведена их классификация. Также рассмотрен ряд программных систем, реализующих автоматическое аннотирование изображений.

1.1 Анализ существующих методов автоматического аннотирования изображений

Существующие методы ААИ можно разделить на две категории, аннотирующие изображения с помощью одного и нескольких ключевых слов соответственно. Классификация методов ААИ по категориям приведена в таблице 1.1.

Таблица 1.1

Классификация методов автоматического аннотирования изображений

Категории	Подходы	Методы
Аннотирование одним ключевым словом	Классификационный	– На основе неотрицательного матричного разложения – На основе метода опорных векторов – На основе многовариантного обучения
Аннотирование несколькими ключевыми словами	Генеративный	– Модель совместной встречаемости – Модель машинного перевода – На основе моделей релевантности
	Поисковый	– Joint Equal Contribution (JEC) – Tag Propagation (TagProp) – 2-Pass K-Nearest Neighbor (2PKNN)

Рассмотрим подробнее основные методы ААИ, сгруппированные в три подхода: классификационный, генеративный и поисковый.

1.1.1 Классификационные методы

Методы классификационного подхода рассматривают процесс аннотирования изображений как проблему категоризации изображений. Для этого ключевые слова представляются в виде независимых классов, на примерах которых обучается классификатор. При аннотировании нового изображения классификатор определяет класс, к которому оно относится, и присваивает соответствующее ключевое слово. Несколько ключевых слов могут быть получены из предположения, что изображение принадлежит нескольким классам. Рассмотрим подробнее некоторые методы данного подхода.

Методы на основе неотрицательного матричного разложения

Неотрицательное матричное разложение (NMF, Non-negative Matrix Factorization) [64] является одним из методов разложения матриц, благодаря ограничению на неотрицательность получивший распространение для обработки данных (таких как текстовые документы и изображения) на основе анализа их частей [48, 98, 104]. В работе [48] метод NMF использовался для классификации изображений. Авторы работы создали коллекцию, состоящую из плиток (квадратных фрагментов) изображений, и разделили ее на 10 классов. Из этой коллекции случайным образом выбиралось по 1000 плиток для формирования обучающей и тестовой выборок. При обучении метод NMF формировал подпространства для каждого класса, на которых в дальнейшем обучался классификатор. При классификации тестовое изображение вначале отображалось в каждое из 10 созданных подпространств, после чего выбирался класс, получивший наибольшее отклик классификатора.

В дальнейшей работе данные авторы [47] сравнивали несколько различных метрик в пространствах, полученных с помощью метода NMF. В своих экспериментах по классификации объектов они обнаружили, что в случае, когда объекты частично перекрывают друг друга, метод NMF с

косинусной метрикой показывает наилучшие результаты. Однако в работе [67] было показано, что базис, полученный с помощью метода NMF, не подходит для непосредственного распознавания объектов с помощью методов ближайшего соседа. Они предложили проводить ортонормализацию базиса перед дальнейшим анализом, вследствие чего повышалась точность распознавания объектов.

Подходы на основе метода опорных векторов

Метод (машина) опорных векторов (SVM, Support Vector Machine) является одним из наиболее популярных методов для классификации данных [99]. Основная идея линейного метода опорных векторов заключается в том, что множество признаков, принадлежащее двум классам, можно разделить оптимальной гиперплоскостью. Оптимальная гиперплоскость формирует компактные множества из наибольшего количества признаков одного и того же класса, при этом максимизируются расстояния от обоих классов до гиперплоскости.

В работе [30] авторы одними из первых применили метод опорных векторов для классификации изображений. Для описания изображений использовались цветовые гистограммы, а метод опорных векторов, изначально разработанный для классификации двух классов, обучался по принципу «один против всех» для классификации семи классов.

Также был предложен метод для классификации областей изображений с помощью ансамбля SVM-классификаторов [44]. В данном методе на первом этапе изображение разбивается сетками на прямоугольные плитки с использованием кратных 8 пикселам масштабов. Из каждой плитки извлекается 90-мерный вектор признаков, после чего полученное 90-мерное пространство разнородных признаков (значения и диапазоны одних признаков существенно отличается от других) разбивается на 9 однородных подмножеств. На втором этапе «слабый» SVM-классификатор обучается для каждого однородного подмножества признаков. В результате обучения выбираются наиболее эффективные классификаторы, а также

соответствующие подмножества признаков и размеры плиток. На последнем этапе выбранные «слабые» SVM-классификаторы объединяются с использованием метода бустинга (boosting), формируя ансамбль классификаторов (рис. 1.1). Отметим, что данный метод способен аннотировать изображения на уровне объектов.

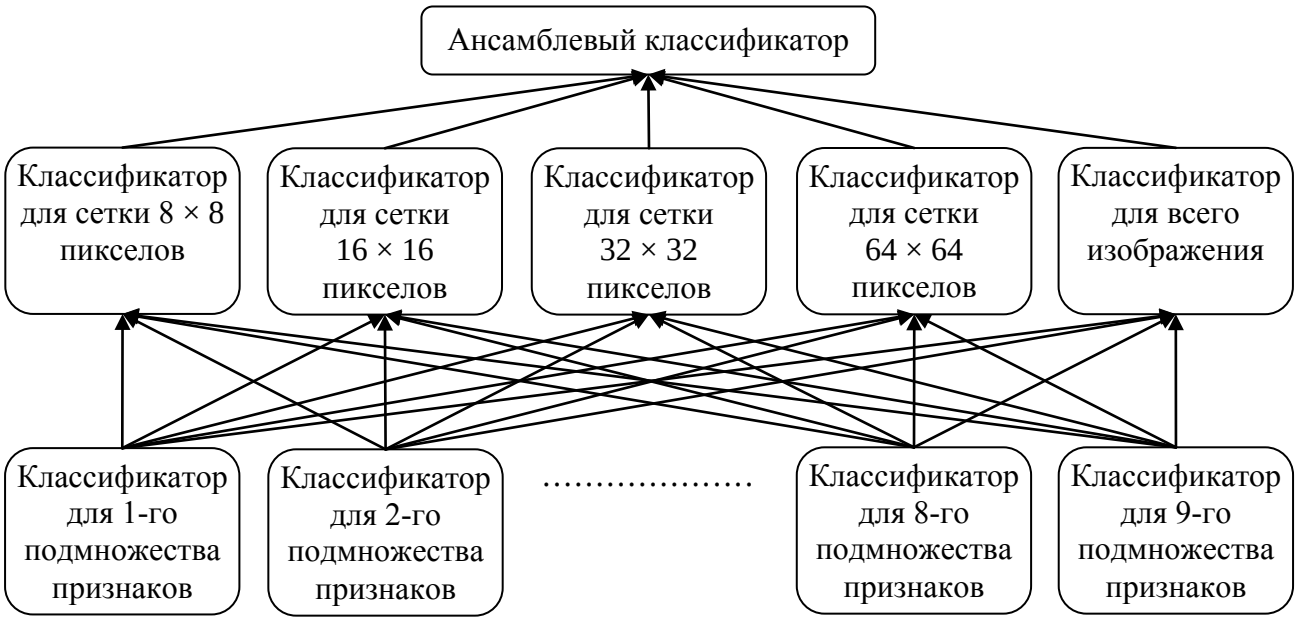

Рисунок 1.1. Подход к классификации областей изображения на основе разделения пространства признаков и иерархии SVM-классификаторов

Также был предложен метод, в котором для автоматического аннотирования изображений комбинируются два набора SVM-классификаторов [87]. Один набор классификаторов обучается на признаках областей изображений, полученных с помощью метода многовариантного обучения (MIL, Multiple Instance Learning) [76], а другой набор использует глобальные признаки изображений для обучения. Результаты работы обоих наборов классификаторов объединяются для аннотирования новых изображений.

Подходы на основе многовариантного обучения

Многовариантное обучение является разновидностью бинарного метода обучения с учителем [76]. Данный метод вместо обучения на наборе элементов, каждый из которых помечен как положительный или

отрицательный, получает набор положительных и отрицательных пакетов (bags). Каждый пакет содержит несколько элементов. Он помечается как отрицательный, если все его элементы отрицательные, и как положительный, если хотя бы один элемент пакета является положительным (рис. 1.2). Цель метода MIL заключается в обучении принципу, с помощью которого можно правильно помечать отдельные элементы.

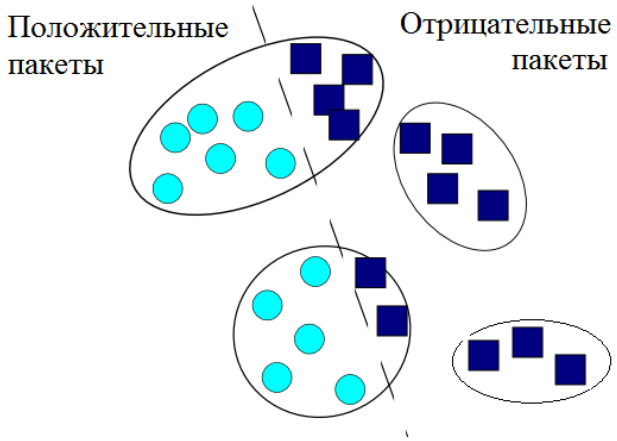

Рисунок 1.2. Пример пакетов, помеченных как положительные и отрицательные

Для решения данной проблемы был предложен подход, называемый Diverse Density (DD). Основная идея подхода заключается в вычислении для каждого элемента DD-значения, являющегося мерой того, сколько различных положительных пакетов имеют элементы вблизи данного элемента и как далеко от данного элемента расположены элементы отрицательных пакетов.

В некоторых работах изображение рассматривается как пакет элементов, каждый из которых представляет собой вектор признаков области изображения. По отношению к определенному ключевому слову изображения, проаннотированные этим ключевым словом, помечаются как положительные, в то время как другие помечаются как отрицательные. В работе [32] для классификации изображений предложен метод DD-SVM, объединяющий метод Diverse Density с классификатором SVM. На первом этапе данного метода изображения сегментируются на области, после чего из каждой извлекается 9-мерный вектор признаков, использующийся в качестве

элемента пакета. На следующем этапе определяются набор элементов-прототипов, используя DD-функцию. Каждый элемент-прототип является представителем класса элементов, которые с большей вероятностью появятся в пакетах с одной меткой. С использованием элементов-прототипов в качестве осей создается новое пространство, в которое отображаются обучающие пакеты (изображения). При этом координата пакета на конкретной оси равна расстоянию между соответствующим прототипом и ближайшим к нему элементом пакета. На последнем этапе SVM-классификатор обучается на основе расположения пакетов в созданном пространстве. В работе [28] использовался аналогичный подход, однако вместо DD-функции для выбора прототипов и классификации элементов был адаптирован метод разреженных опорных векторов (Sparse Support Vector Machine). Согласно полученным результатам данный подход более эффективен.

В работе [105] был предложен модифицированный метод DD, с помощью которого определялись области-образцы, соответствующие конкретным ключевым словам. При аннотировании нового изображения оно разделяется на области, каждому из которых ставится в соответствие ближайшая область-образец и ассоциированное с ней ключевое слово. Таким образом, осуществляется аннотирование на уровне объектов.

В целом, можно отметить, что методы классификационного подхода позволяют быстро и с достаточно большой точностью определить изображения или их области в ряд заранее известных категорий. Однако для этого требуется сбалансированная обучающая выборка (количество примеров для каждой категории должно быть сопоставимо), создание которой чаще всего осуществляется вручную. Также в работе [33] показано, что при увеличении количества категорий (ключевых слов) точность классификации значительно снижается. Кроме того, классификационные методы имеют низкую масштабируемость: каждый раз при добавлении

новых категорий или обучающих изображений необходимо обучать систему классификации заново, что требует значительных вычислительных затрат.

1.1.2 Генеративные методы

Основная идея генеративных (статистических) методов заключается в оценке вероятностей совместной встречаемости ключевых слов и низкоуровневых признаков изображений на основе набора обучающих изображений. Рассмотрим ряд наиболее популярных генеративных методов, предложенных для автоматического аннотирования изображений.

Модель совместной встречаемости является одной из первых попыток автоматического аннотирования изображений [79]. В данном методе вначале все изображения обучающей выборки разделяются на прямоугольные плитки одинакового размера. При этом каждая плитка наследует весь набор ключевых слов изображения, к которому она принадлежит. На следующем этапе из каждой плитки извлекается вектор цветовых и текстурных признаков. Все полученные векторы группируются в несколько кластеров методом k-средних, после чего оценивается вероятность принадлежности ключевого слова k_i кластеру c_j с помощью подсчета совместной встречаемости ключевого слова и плиток изображений в кластере:

Модель совместной встречаемости

$$P\left(k_i \mid c_j\right) = \frac{m_{i,j}}{\sum_{n=1}^{N} m_{n,j}}, \tag{1.1}$$

где $P(k_i \mid c_j)$ – вероятность принадлежности ключевого слова k_i кластеру c_j; $m_{i,j}$ – количество включений ключевого слова k_i в кластер c_j; N – общее количество ключевых слов.

Полученные вероятности используются для аннотирования новых изображений. Для этого новое изображение A также разделяется на прямоугольные плитки, из которых извлекают векторы низкоуровневых признаков. Для каждого вычисленного вектора определяется ближайший кластер, после чего вычисляется вероятность принадлежности каждого ключевого слова аннотируемому изображению:

$$P(k_i \mid A) = \frac{1}{B_A} \cdot \sum_{c_j : \mathbf{x}_b \in c_j}^{B_A} P(k_i \mid c_j), \tag{1.2}$$

где $P(k_i \mid A)$ – средняя вероятность принадлежности ключевого слова k_i изображению A; B_A – количество плиток в изображении A; $\mathbf{x}_b$ – вектор низкоуровневых признаков b-й плитки изображения A.

В качестве описания изображения выбираются ключевые слова, имеющие наибольшую среднюю вероятность по всем плиткам аннотируемого изображения. Также в работах [13, 19] было предложено использовать для группирования векторов признаков модифицированную сеть ESOINN [93], что позволило повысить точность аннотирования за счет создания более точных кластеров.

Модель машинного перевода

В работе [35] для автоматического аннотирования изображений предложено применять модель машинного перевода, ранее использовавшуюся в задачах автоматического перевода естественных языков. В этой модели изображение рассматривается как набор областей, а процесс присоединения ключевых слов к областям изображения как аналог перевода из одной формы представления (например, слово на французском языке) в другую форму (слово на английском языке).

В предложенном методе изображения сегментируются на области, после чего из областей, размер которых больше определенного порога,

извлекаются векторы признаков. Все полученные векторы группируются в кластеры визуально похожих областей изображений (авторы называют их «каплями»). В дальнейшем будем называть подобные кластеры визуальными словами, а их совокупность – словарем визуальных слов (рис. 1.3).

а б в

Рисунок 1.3. Пример формирования словаря визуальных слов: а) набор исходных изображений; б) сегментированные изображения; в) словарь визуальных слов

На следующем этапе проводится оценка вероятностей перевода между визуальными и ключевыми словами, с помощью которых строится таблица перевода. Таким образом, каждому визуальному слову ставится в соответствие одно ключевое слово. При аннотировании нового изображения оно сегментируется на области, для каждой из которых определяется ближайшее визуальное слово. Используя таблицу перевода, областям изображения присваиваются ключевые слова, ассоциированные с выбранными визуальными словами. Данный метод показал хорошие результаты аннотирования на выборке ландшафтных изображений, в которых природные объекты (скалы, облака и др.) часто состоят из одной текстуры, либо одна текстура занимает значительную их часть [7, 11].

Методы на основе моделей релевантности

В работе [60] был предложен метод Cross-Media Relevance Model (CMRM), в котором также используется вышеописанный процесс формирования визуальных слов, а каждое изображение I из обучающей выборки TS, представляется в виде набора визуальных и ключевых слов:

$I = \{vw_1, ..., vw_m; k_1, ..., k_n\}$, где m и n – количество визуальных и ключевых слов изображения. Однако, в отличие от модели машинного перевода, предполагающей наличие однозначного соответствия между визуальными и ключевыми словами, в методе CMRM лишь предполагается, что набор ключевых слов связан с набором визуальных слов. В этом случае для аннотирования нового изображения необходимо оценить вероятности наблюдения наборов визуальных и ключевых слов в этом изображении.

Пусть $P(k_i|A)$ и $P(vw_m|A)$ обозначают базовые распределения вероятностей всевозможных визуальных и ключевых слов, которые могут встречаться в не аннотированном изображении A. Если изображение A представить в виде набора визуальных слов, т. е. $A = \{vw_1, ..., vw_m\}$, то вероятность принадлежности ключевого слова k_i изображению A можно аппроксимировать следующим образом:

$$P(k_i \mid A) \approx P(k_i \mid vw_1, \ldots, vw_m). \tag{1.3}$$

Если предположить, что распределения вероятностей ключевых и визуальных слов являются независимыми, то вычисление условной вероятности $P(k_i \mid vw_1, \ldots, vw_m)$ эквивалентно вычисление совместной вероятности $P(k_i, vw_1, ..., vw_m)$ и рассчитывается следующим образом:

$$P(k_i, vw_1, ..., vw_m) = \sum_{I \in TS} P(I) P(k_i \mid I) \prod_{j=1}^{m} P(vw_j \mid I). \tag{1.4}$$

Априорные вероятности $P(I)$ выбираются одинаковыми для всех обучающих изображений, в то время как для вероятностей $P(k_i \mid I)$ и $P(vw_j \mid I)$ используются сглаженные оценки максимального правдоподобия.

В работе [63] авторы предположили, что процесс квантования непрерывных признаков изображений в дискретные визуальные слова в методах машинного перевода и CMRM приводит к потере полезной

информации. Для решения этой проблемы был предложен метод Continuous-space Relevance Model (CRM), в котором вероятность $P(vw_j \mid I)$ из метода CMRM заменена на вероятность $P(\mathbf{x}_b \mid I)$, где $\mathbf{x}_b$ – вектор низкоуровневых признаков b-й области изображения. Вероятность $P(\mathbf{x} \mid I)$ является непараметрической оценкой плотности вероятности и вычисляется следующим образом:

$$P(\mathbf{x} \mid I) = \frac{1}{m} \cdot \sum_{b=1}^{B_I} \frac{1}{\sqrt{2\pi|\Sigma|}} \cdot e^{(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}-\mathbf{x}_b)}, \quad (1.5)$$

где B_I – количество векторов признаков в изображении I; $\sum$ – ковариационная матрица для управления степенью сглаживания, $\sum = \beta \cdot E$, где E – единичная матрица, а β – значение, выбранное эмпирически на основе валидационной выборки.

В работе [40] метод CRM был модифицирован таким образом, что вероятность принадлежности ключевого слова обучающему изображению $P(k_i \mid I)$ моделируется в виде множественного распределения Бернулли, а не полиномиального распределения. Также было показано, что разделение изображения на прямоугольные плитки вместо использования методов автоматической сегментации приводит к повышению точности и полноты аннотирования. Данная модификация получила название Multiple Bernoulli Relevance Model (MBRM).

В работе [81] предложен гибридный классификационно-генеративный метод Support Vector Machine and Discrete Multiple Bernoulli Relevance Model (SVM-DMBRM). Классификационная часть данного метода заключается в обучении по одному бинарному классификатору SVM для каждого ключевого слова. На этапе аннотирования новое изображение подается на вход каждого классификатора, после чего полученные отклики нормализуются таким образом, чтобы представлять вероятность присвоения

соответствующего ключевого слова изображению. В качестве генеративной части метода SVM-DMBRM используется модификация MBRM, в которой изображение представляется в виде набора визуальных слов для ускорения вычислений. Вероятности присвоения ключевых слов изображению, полученные отдельно для классификационной и генеративной частей метода, линейно комбинируются, формируя ранжированный список ключевых слов.

Таким образом, генеративные методы рассматривают задачу ААИ как статистическую проблему, оценивая совместную вероятность изображений (областей изображений) и ключевых слов и ранжируя ключевые слова в соответствии с полученными значениями вероятностей. При этом большая часть вычислений выполняется на этапе обучения, в связи с чем аннотирование нового изображений осуществляется быстрее, чем в классификационных методах, а обучение системы – медленнее. Также, аналогично методам классификационного подхода, генеративные методы имеют низкую масштабируемость, требуя повторного обучения системы при каждом добавлении новых ключевых слов и обучающих изображений.

1.1.3 Поисковые методы

Поисковые методы автоматического аннотирования изображений основаны на предположении, что визуально похожие изображения должны аннотироваться одинаковыми ключевыми словами. Для нового изображения определяется набор визуально похожих изображений, уже имеющих текстовое описание, после чего аннотация формируется на основе значений схожести между изображениями. Рассмотрим подробнее основные поисковые методы.

Метод Joint Equal Contribution

В работе [72] впервые предложено рассматривать автоматическое аннотирование нового изображения как проблему поиска визуально похожих обучающих изображений (ближайших соседей). Для этого каждое

изображение описывается с помощью 7 глобальных цветовых и текстурных признаков, нормализованных таким образом, чтобы значения расстояний между парами признаков любых двух изображений находились в диапазоне [0; 1]. При сравнении двух изображений сначала отдельно вычисляются расстояния для каждого типа признаков, после чего полученные значения объединяются с равными весами (JEC, Joint Equal Contribution).

При аннотировании нового изображения для него определяется 5 ближайших обучающих изображений, отсортированных по увеличению расстояния. Обозначим их как I_i, где i = 1, …, 5 (I_1 – наиболее похожее изображение), а количество ассоциированных с ними ключевых слов как $|I_i|$. Для того, чтобы проаннотировать новое изображение A с помощью NK = 5 ключевых слов, используется следующий алгоритм:

1. Ранжировать ключевые слова изображения I_1 по убыванию их частоты встречаемости в обучающей выборке.

2. Выбрать первые NK ключевых слов изображения I_1 в качестве аннотации изображения A. Если $|I_1| < NK$, то перейти к шагу 3.

3. Ранжировать все ключевые слова изображений I_2, …, I_5 на основе двух факторов: совместной встречаемости в обучающей выборке с ключевыми словами, выбранными на шаге 2, и локальной частоты встречаемости в изображениях I_2, …, I_5.

4. Выбрать NK - $|I_1|$ ключевых слов с наивысшим рангом в качестве аннотации изображения A.

Метод Tag Propagation

В работе [46] предложен метод Tag Propagation, в котором на этапе обучения вычисляются веса значимости отдельных типов низкоуровневых признаков изображений. Пусть $y_{i,n} \in \{-1, +1\}$ обозначает наличие / отсутствие ключевого слова k_n в описании обучающего изображения I_i. В этом случае вероятность принадлежности ключевого слова k_n изображению I_i обозначается как $P(y_{i,n} = +1)$ и оценивается с помощью взвешенного

суммирования значений принадлежности ключевого слова k_n соседним изображениям I_j:

$$P\left(y_{i,n}=+1\right)=\sum_{I_j\in TS\left(I_i\right)}\pi_{i,j}\cdot P\left(y_{i,n}=+1\mid I_j\right), \tag{1.6}$$

$$P\left(y_{i,n}=+1\mid I_j\right)=\begin{cases}1-\varepsilon & \text{если } y_{j,n}=+1\\ \varepsilon & \text{иначе}\end{cases}, \tag{1.7}$$

где $TS(I_i)$ – набор обучающих изображений, являющихся ближайшими соседями изображения I_i; $\pi_{i,j}$ – весовой коэффициент изображения I_j при предсказании аннотации изображения I_i; ε – коэффициент, равный 10^{-5}.

Авторами предложено два способа определения весов $\pi_{i,j}$, из которых наилучший результат показал способ на основе расстояния между изображениями (TagProp-ML):

$$\pi_{i,j}=\frac{\exp\left(-d_\theta\left(I_i,I_j\right)\right)}{\sum_{I_l\in TS\left(I_i\right)}\exp\left(-d_\theta\left(I_i,I_l\right)\right)}, \tag{1.8}$$

$$d_\theta\left(I_i,I_j\right)=\boldsymbol{\theta}^{\mathrm{T}}\mathbf{d}_{i,j}, \tag{1.9}$$

где $\boldsymbol{\theta}$ – вектор параметров, оптимизируемый на этапе обучения; $\mathbf{d}_{i,j}$ – вектор значений расстояний между изображениями I_i и I_j, полученный для разных типов признаков.

Для оценки параметров $\boldsymbol{\theta}$ авторы максимизируют логарифмическую функцию правдоподобия корректного предсказания ключевых слов обучающих изображений. Также в работе предложена логистическая дискриминантная модель, использующая сигмоидальную функцию (TagProp-σML). Данная модель позволяет увеличивать вероятность для редких ключевых слов и снижать для частых, что положительно сказывается на

полноте аннотирования, однако снижает возможность масштабирования системы.

Метод 2-Pass K-Nearest-Neighbor

В работе [100] был предложен двухпроходный метод K ближайших соседей (2PKNN, 2-Pass K-Nearest Neighbor), реализующий альтернативный способ решения проблемы огромной разницы в частотах встречаемости разных ключевых слов. В этом методе проблема аннотирования сформулирована как задача поиска апостериорных вероятностей принадлежности ключевых слов k_n новому изображению A:

$$P(k_n \mid A) = \frac{P(A \mid k_n)P(k_n)}{P(A)}, \tag{1.10}$$

где $P(A \mid k_n)$ – вероятность принадлежности изображения A семантической группе обучающих изображений, имеющих в описании ключевое слово k_n; $P(k_n)$ – априорная вероятность ключевого слова k_n; $P(A)$ – априорная вероятность нового изображения A.

Для ее решения обучающий набор TS, состоящий из пар изображений I_i и соответствующих им наборов ключевых слов K_i, разделяется на семантические группы $TS_n \subseteq TS$, где $n \in \{1, \ldots, N\}$ (N – общее количество ключевых слов), в которых все изображения имеют в описании ключевое слово k_n. Так как изображения обычно проаннотированы несколькими ключевыми словами, то они могут принадлежать нескольким семантическим группам.

При аннотировании нового изображения A на первом проходе метода из каждой семантической группы выбирается по 2 наиболее похожих обучающих изображения, которые затем объединяются в набор $TS_A \subseteq TS$. Полученный набор включает наиболее информативные обучающие изображения для предсказания принадлежности ключевых слов новому

изображению A. На втором проходе метода осуществляется оценка вероятностей $P(A \mid k_n)$ с помощью изображений набора TS_A:

$$P(A \mid k_n) = \sum_{(I_i, K_i) \in TS_A} \exp(-d(A, I_i)) \cdot \delta(k_n \in K_i), \quad (1.11)$$

где $d(A,I_i)$ – расстояние между изображениями A и I_i; $\delta(k_n \in K_i)$ обозначает наличие / отсутствие ключевого слова k_n в описании изображения I_i (принимает значения 1 и 0 соответственно).

Поскольку в наборе TS_A ключевые слова имеют сопоставимые частоты встречаемости, то априорные вероятности $P(A)$ выбраны одинаковыми для всех ключевых слов. Полученные значения вероятности подставляются в выражение (1.10), после чего в качестве аннотации нового изображения A используется 5 ключевых слов с наибольшими значениями вероятностей $P(k_n \mid A)$. Приведенные авторами работы результаты экспериментов продемонстрировали повышение точности аннотирования в сравнении с методом TagProp. При этом метод 2PKNN требует значительно меньших вычислительных затрат, поскольку разные типы признаков используются с равными весами.

Таким образом, в поисковых методах процесс аннотирования нового изображения можно разделить на два этапа: поиск небольшого количества визуально похожих обучающих изображений и выбор их ключевых слов в качестве аннотации нового изображения. Благодаря этому большинство поисковых методов ААИ не требует списка заранее известных ключевых слов, а также способно продолжать работу при добавлении новых обучающих изображений без повторного обучения системы. Также в качестве обучающих выборок поисковые методы могут напрямую использовать данные специализированных веб-сайтов, в которых размещаются фотоматериалы с пользовательскими ключевыми словами (например, Flickr [43] и Instagram [56]).

1.1.4 Сравнение методов автоматического аннотирования изображений

Обычно для сравнения различных методов ААИ используются тестовые базы изображений, например [72]:

– IAPR TC-12 [53] содержит 19627 фотографий различных сцен, включающих людей, животных, города, ландшафты, а также другие аспекты современной жизни. Изначально содержание каждого изображения базы описано несколькими предложениями, из которых авторы статьи выбирали существительные с помощью программы TreeTagger [96]. Эти существительные в дальнейшем использовались в качестве ключевых слов. Таким образом, был получен словарь из 291 ключевого слова;

– ESP Game [37] содержит 20770 изображений, включающих как фотографии, так и изображения с искусственной графикой (анимационные картинки, логотипы и т.п.). Эти изображения получены с помощью игры ESP, предложенной в работе [101]. В этой игре два игрока независимо друг от друга присваивают одному и тому же изображению ключевое слово. Если слова совпадают, то оно принимается в качестве аннотации изображения. Всего в базе использовано 269 ключевых слов.

Авторы работы [72] разделили изображения баз на обучающие и тестовые выборки с соотношениями 90 % и 10 % соответственно. В дальнейшем такие выборки использовались другими исследователями при публикации результатов работы методов ААИ. При этом оценка эффективности заключается в вычислении средней точности и полноты аннотирования, вычисленных для ключевых слов, а также подсчете количества использованных при аннотировании ключевых слов (N+). В таблице 1.2 приведены опубликованные результаты аннотирования некоторых из рассмотренных методов.

Как видно из таблицы 1.2, лучшие результаты по точности и количеству использованных при аннотировании ключевых слов показывает метод SVM-DMBRM, в то время как метод TagProp-σML демонстрирует

лучшие показатели полноты для указанных выше тестовых баз. Оба метода имеют низкую степень масштабируемости, требуя повторного обучения системы при добавлении новых ключевых слов, а демонстрируемые лучшие результаты аннотирования являются недостаточными, что свидетельствует о необходимости дальнейшего развития методов ААИ.

Таблица 1.2

Сравнительные оценки эффективности методов ААИ

Метод	IAPR TC-12			ESP Game		
	Точность, %	Полнота, %	N+	Точность, %	Полнота, %	N+
MBRM [40]	24	23	223	18	19	209
JEC [72]	28	29	250	22	25	224
TagProp-ML [46]	48	25	227	49	20	213
TagProp-σML [46]	46	**35**	266	39	**27**	239
2PKNN [100]	49	32	274	51	23	245
SVM-DMBRM [81]	**56**	29	**283**	**55**	25	**259**

1.2 Анализ методов кластеризации данных

В рассмотренном ранее методе автоматического аннотирования изображений 2PKNN набор обучающих изображений предложено разделять на семантические группы, каждая из которых используется для аннотирования нового изображения. Подобное разделение выборки позволило повысить точность аннотирования, что свидетельствует о необходимости предварительного структурирования обучающего набора. Одним из возможных способов структурирования изображений является их кластеризация по текстовому описанию и визуальным признакам. Рассмотрим подробнее некоторые из существующих методов кластеризации.

1.2.1 Иерархические методы

Иерархические методы структурируют выборку векторов в виде системы вложенных разбиений, формируя дерево кластеров, корнем

которого является вся выборка, а листьями – отдельные векторы [23]. Выделяют два основных типа методов иерархической кластеризации [103]:

1. Нисходящие алгоритмы, работающие по принципу «сверху вниз». Вначале все векторы помещаются в один кластер, который затем разделяется на все более мелкие кластеры.

2. Восходящие алгоритмы, работающие по принципу «снизу вверх». В начале работы каждый вектор перемещается в отдельный кластер, после чего кластеры объединяются во все более крупные до тех пор, пока все векторы выборки не будут содержаться в одном кластере. Данный тип методов является наиболее распространенным.

В обоих типах алгоритмов при принятии решения о слиянии (разделении) кластеров используются следующие способы вычисления расстояний между кластерами [82]:

1. Метод одиночной связи (расстояние ближайшего соседа). Расстояние между двумя кластерами определяется как расстояние между двумя наиболее близкими векторами этих кластеров.

2. Метод полной связи (расстояние дальнего соседа). В этом методе расстояние между кластерами определяется наибольшим расстоянием между любыми двумя векторами этих кластеров.

3. Метод средней связи. Расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами векторов этих кластеров.

4. Центроидный метод. В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами масс.

Иерархические методы кластеризации представляют выборку векторов в виде древовидной системы кластеров, что в контексте задачи ААИ позволяет использовать выборку обучающих изображений на нескольких уровнях детализации. Однако иерархические методы имеют высокую вычислительную сложность, из-за чего не применяются для выборок, размер которых больше 10 000 изображений.

1.2.2 Методы квадратичной ошибки

Методы этой категории разделяют кластеризуемые векторы данных на группы, используя в качестве критерия оптимальности минимизацию среднеквадратической ошибки разбиения:

$$e = \sum_{j=1}^{NC} \sum_{\mathbf{x}_i \in C_j} \left(\mathbf{x}_i - \mathbf{c}_j\right)^2 \qquad (1.16)$$

где NC – количество кластеров; C_j – j-ый кластер; $\mathbf{c}_j$ – центр масс j-го кластера (вектор со средними значениями характеристик для данного кластера), $\mathbf{x}_i$ – кластеризуемый вектор.

Самым распространенным алгоритмом этой категории является метод k-средних [69]. Этот алгоритм разбивает векторы на заранее заданное количество кластеров, расположенных как можно дальше друг от друга. Работа алгоритма состоит из следующих шагов:

1. Выбрать случайным образом k векторов в качестве начальных центров масс кластеров.

2. Ассоциировать каждый вектор с кластером ближайшего центра масс.

3. Пересчитать центры масс кластеров согласно их текущему составу.

4. Если критерий остановки алгоритма не удовлетворен, обнулить состав кластеров и перейти к шагу 2. В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки.

Метод k-средних показывает хорошие результаты в случаях, когда кластеры представляют собой значительно разделенные между собой компактные группы, однако при этом алгоритм чувствителен к шуму и обособленным векторам, а результат работы алгоритма зависит от выбора исходных центров масс. Для решения этих проблем был предложен ряд

модифицированных методов, таких как алгоритм *k*-средних++ [25], реализующий эвристики для лучшего выбора начальных значений центров кластеров, и *k*-медоидов [61], в котором для представления центра кластера используется не центр масс, а наиболее близкий к нему кластеризуемый вектор, что позволяет снизить влияние шума. К недостаткам этих методов можно отнести необходимость задавать количество кластеров для разбиения.

1.2.3 Инкрементальные методы

Все вышерассмотренные методы кластеризации предполагают неизменность множества кластеризуемых векторов в течение работы алгоритмов, в связи с чем при изменении выборки необходимо осуществлять повторную кластеризацию. Для решения этой проблемы разработаны инкрементальные (онлайн) методы, в которых кластеризуемые векторы подаются на вход алгоритмов по одному, а общее количество векторов заранее неизвестно. Одним из таких методов является расширенная самоорганизующаяся инкрементальная нейронная сеть (ESOINN, Enhanced Self-Organizing Incremental Neural Network) [9, 93].

Сеть ESOINN начинает работу с выбора случайным образом двух входных векторов в качестве начальных узлов сети. Когда на вход сети подан вектор, сеть находит два ближайших узла (победитель и второй победитель).

Используя пороговые критерии подобия (максимальное расстояние между двумя связанными узлами), сеть определяет, относиться ли входной вектор к тому же самому кластеру, что и победитель или второй победитель. Если расстояние между входным вектором и победителем или вторым победителем будет больше соответствующих им порогов подобия, входной вектор вставляется в сеть как первый узел нового класса. В случае если входной вектор определен как принадлежащий к кластеру победителя или второго победителя, создается новая связь между победителем и вторым

победителем с «возрастом» равным 0, а «возраст» всех связей, соединенных с победителем, увеличивается на 1.

На следующем этапе обновляется плотность узла-победителя, определяемая с помощью среднего расстояния между победителем и связанными с ним узлами. После этого счетчик количества побед узла-победителя увеличивается на 1, а векторы весов победителя и его узлов-соседей смещаются в сторону входного вектора. На следующем шаге удаляются все связи, «возраст» которых превышает заранее предустановленное значение age_{max}.

Если период обучения сети еще не закончен, то на вход сети подается новый входной вектор. В противном случае удаляются связи в перекрывающихся областях. На последнем этапе удаляются узлы, имеющие двух или меньше топологических соседей и низкую плотность.

Таким образом, сеть ESOINN автоматически определяет количество кластеров, а также кластеризует данные инкрементально, что позволяет использовать ее совместно с поисковыми методами ААИ, не требующими повторного обучения при добавлении новых данных.

1.3 Анализ низкоуровневых признаков изображений

В рассмотренных методах аннотирования и кластеризации изображений ключевую роль играют визуальные признаки, извлекаемые из изображений и используемые для их компактного представления. Выбор значимых признаков может значительно повысить эффективность методов аннотирования. В общем случае используемые признаки можно разделить на признаки цвета, текстуры и формы [8, 88]. Также в последнее время получили большое распространение локальные дескрипторы, описывающие окрестности особенных точек, и методы их кодирования в глобальные дескрипторы. Рассмотрим данные категории подробнее.

1.3.1 Цветовые признаки

Цветовые признаки являются одними из наиболее важных признаков, применяемых для поиска изображений [5]. Среди предложенных признаков цвета наибольшее распространение получили цветовые моменты [36, 42], цветовые гистограммы [57, 92], цветовой когерентный вектор [83], цветовая коррелограмма [52] и другие. Также стандарт MPEG-7 устанавливает такие цветовые признаки, как масштабируемый цветовой дескриптор, дескриптор цветовой структуры и дескриптор доминирующего цвета [73]. При вычислении цветовых признаков также могут использоваться различные цветовые пространства: RGB, LUV, HSV, HMMD и другие [3, 70, 89].

Наиболее простыми и популярными признаками являются цветовые моменты, описывающие статистическое распределение отдельных цветовых каналов [90]. При этом чаще всего используются такие цветовые моменты, как среднее значение, дисперсия и асимметрия. В работе [91] авторы рассматривают распределения отдельных цветовых каналов, как компоненты трехмерного распределения. Для этого вводится пять фиксированных областей: центральная область в виде эллипса и четыре боковые области. Для каждой области вычисляется среднее значение по каждому из цветовых каналов, а также рассчитываются попарные ковариации распределений каналов.

Цветовая гистограмма является еще одним из простых способов описания характеристик цвета изображения. Для ее построения все допустимые значения цветового пространства квантуются в несколько подмножеств (столбцов), для каждого из которых подсчитывается количество пикселов, значения которых соответствуют этим подмножествам. Сформированные таким образом цветовые гистограммы устойчивы к смещениям и поворотам объектов на изображениях, однако не содержат информацию о пространственной конфигурации элементов сцены.

Цветовой когерентный вектор расширяет стандартную цветовую гистограмму за счет включения информации о пространственном распределении цветов. Для этого все столбцы гистограммы разделяются на когерентную и некогерентную части. В первой части столбца подсчитывается количество пикселов, относящихся к крупным пространственно связанным областям изображения с соответствующим цветом. Во вторую часть столбца включены изолированные пикселы и пикселы из небольших областей изображения. В общем случае цветовой когерентный вектор демонстрирует улучшение результатов в сравнение с обычной гистограммой, однако при этом размерность дескриптора удваивается.

Цветовая коррелограмма является цветовой версией матрицы смежности уровней яркости, в которой описывается пространственное распределение любых пар цветов на изображении. Цветовая коррелограмма может рассматриваться в качестве трехмерной гистограммы, первые два измерения которой соответствуют допустимым значениям цветов пикселов изображения, а третье – пространственному расстоянию между пикселами. В этом случае значение коррелограммы с индексом (i, j, d) содержит количество пар пикселов со значениями цветов i и j, находящимися на расстоянии d пикселов друг от друга. Благодаря описанию как уровней цвета, так и структуры распределения цветов на изображении, цветовая коррелограмма показывает лучшие результаты в сравнение с обычной гистограммой и цветовым когерентным вектором. Однако при этом значительно возрастает количество необходимых вычислений, а также размерность формируемого дескриптора.

В стандарте MPEG-7 масштабируемый цветовой дескриптор представляет собой цветовую гистограмму, вычисленную в цветовом пространстве HSV. Чтобы обеспечить масштабируемость с точки зрения количества столбцов, используемых для описания изображения, соседние элементы гистограммы кодируются с помощью преобразования Хаара. В

этом случае суммирование элементов эквивалентно вычислению гистограммы с вдвое меньшим количеством столбцов. Подобная организация гистограммы позволяет при поиске похожих изображений по первым столбцам гистограммы выбрать набор изображений-кандидатов, среди которых проводится дальнейшая селекция с помощью остальной части гистограммы. Это ускоряет поиск изображений, однако все проблемы обычной гистограммы, связанные с отсутствием пространственной информации, остаются.

Дескриптор цветовой структуры является гистограммой, вычисляемой в цветовом пространстве HMMD. Для ее создания по изображению методом скользящего окна передвигается структурный элемент (например, квадрат). Для каждого столбца гистограммы подсчитывается количество раз, когда в структурный элемент входил хотя бы один пиксел с соответствующим цветом. Эффективность дескриптора во многом зависит от выбора размера и формы структурного элемента. Также для его вычисления требуются большие вычислительные затраты, чем для масштабируемого цветового дескриптора.

Дескриптор доминирующего цвета также является разновидностью цветовой гистограммы, в которой вместо фиксированного разделения цветового пространства на столбцы используется адаптивный выбор цветов. Для этого цвета каждого изображения кластеризуются в 1–8 кластеров. Затем каждый кластер описывается с помощью нескольких параметров: цвета центра масс кластера; отношения количества пикселов кластера к общему количеству пикселов изображения; дисперсии цветов пикселов кластера. Окончательный дескриптор представляет собой небольшой набор векторов, описывающих доминирующие цвета изображения, а также параметр, характеризующий пространственную когерентность пикселов изображения. Также в работе [2] была предложена модификация дескриптора, в которой вместо дисперсии цветов пикселов кластера и пространственной когерентности пикселов используются пространственные координаты центра

масс пикселов кластера. Таким образом, данный тип дескрипторов является более точным и компактным цветовым признаком, чем другие гистограммы. Однако при вычислении расстояния между двумя изображениями необходимо сравнивать все пары векторов двух дескрипторов.

1.3.2 Текстурные признаки

Еще одной значимой характеристикой изображения является текстура. Текстура представляет собой пространственную организацию элементов в пределах некоторого участка изображения и является одним из ключевых компонентов в восприятии человеком изображения [1]. В отличие от цвета, являющегося свойством пиксела, текстура является свойством области и воспринимается как перепад уровней интенсивности. В связи с этим текстурные признаки вычисляются на областях изображений в оттенках серого. Методы извлечения текстурных признаков, получившие наибольшее распространение, можно разделить на два подхода: статистический и спектральный. Рассмотрим данные подходы подробнее.

Статистический подход

Статистический подход основан на предположении, что можно выявить статистические характеристики образца текстуры и считать их соответствующими любым другим образцам данной текстуры. Наиболее распространенными статистическими методами являются моменты [3, 70], признаки Харалика [49, 50] и признаки Тамуры [94, 95].

Статистические моменты являются одним из простейших способов вычисления текстурных признаков, определяемых по гистограмме яркости всего изображения или его области. При этом чаще всего используют центральные моменты порядка k, а также энтропию, характеризующую величину хаотичности текстуры, и энергию, вычисляемую как сумма квадратов элементов гистограммы. Данные оценки эффективны для описания текстур с невыраженной пространственной регулярностью.

Для того чтобы учесть информацию о взаимном расположении пикселов изображения, в работе [49] было предложено вычислять матрицы смежности уровней яркости. Размер матриц равен количеству уровней яркости, а значение элемента с индексами (*i, j*) обозначает количество пар пикселов со значениями яркости *i* и *j*, находящихся на расстоянии *d* пикселов друг от друга. В большинстве случаев параметр *d* выбирают равным 1, а сами матрицы рассчитывают для горизонтального, вертикального и диагональных направлений. После нормализации матрицы используются для вычисления признаков Харалика, среди которых наиболее информативными являются [6, 11]: момент обратной разности, коэффициент корреляции, энтропия, а также информационная мера корреляции и однородность элементов матриц смежности.

Признаки Тамуры – зернистость, контрастность и направленность – являются еще одними из известных текстурных признаков, выделенных с учетом особенностей зрительного восприятия человека [94]. Зернистость описывает размер структурных элементов текстуры, определяя на каком максимальном масштабе существует текстура. Контраст соответствует тому, насколько отчетливо воспринимаются структурные элементы текстуры. Направленность описывает степень ориентированности структурных элементов текстуры. Используя эти признаки, строится изображение Тамуры, в котором три значения пиксела являются оценками зернистости, контрастности и направленности, вычисленными в окрестности соответствующего пиксела исходного изображения. Сформированное изображение Тамуры в дальнейшем описывается с помощью цветовых признаков.

Спектральный подход

Идея спектрального подхода заключается в представлении изображения в виде двумерного сигнала и его последующем разложении на составляющие части. Для этого чаще всего используется вейвлет-анализ, с помощью которого осуществляется разложение сигнала по базисным

функциям (вейвлетам). При этом базисные функции строятся на основе порождающего вейвлета с использованием операций сдвига и масштабирования. Для исходного изображения строится первая проекция сигнала (свертка с первой базисной функцией), далее вычисляется разность полученного и исходного сигнала и строится вторая проекция этой разности (свертка со второй базисной функцией) и так далее. При вычислениях каждая последующая базисная функция является сдвигом предыдущей, растянутой в $2n$ раз (параметр n характеризует масштаб). Такие базисные функции обычно называют фильтрами. Наиболее распространенными и используемыми являются фильтры Габора [108] и фильтры анализа независимых компонент (ICA-фильтры) [27].

Фильтр Габора представляет собой свертку преобразований Фурье гармонической функции и гауссиана. В большинстве случаев используются сечения фильтра по нескольким заранее заданным масштабам и направлениям. Каждое из полученных таким образом сечений фильтра сворачивается с исходным изображением, в результате чего формируется матрицы отклика. На их основе вычисляются следующие признаки: фазовая информация, амплитуда, локальная энергия, ориентация.

ICA-фильтры строятся с помощью анализа независимых компонент по обучающему набору изображений. Данные фильтры являются локальными и подобны фильтрам Габора, однако в отличие от них носят естественный характер и отражают основные направления текстуры изображений, по которым они строились. Также проводились исследования [41], показывающие, что способ построения ICA-фильтров имитирует процесс зрительного обучения человека.

1.3.3 Признаки формы

Признаки формы являются наиболее сложно формализуемыми характеристиками для формирования дескрипторов изображений. Люди

воспринимают признаки формы на высоком уровне абстракции, не вдаваясь в математические подробности описания форм объектов на изображении. Также признаки формы чувствительны к качеству выделения контуров и сегментации изображений, в связи с чем в задачах поиска и аннотирования изображений используются только простейшие из них [1, 4]:

– площадь – количество пикселов объекта;

– периметр – количество пикселов, расположенных на внешнем контуре объекта;

– координаты центра масс пикселов объекта;

– компактность – отношение квадрата периметра объекта к его площади;

– ориентация оси наименьшей инерции. Эта ось проходит через центр масс объекта таким образом, чтобы сумма расстояний от оси до точек внешнего контура была минимальна;

– удлиненность, вычисляемая как соотношение сторон минимального охватывающего объект прямоугольника;

– цельность представляет собой меру выпуклости или вогнутости формы объекта и вычисляется как отношение площади объекта к площади охватывающего выпуклого многоугольника;

– отношение площади объекта к площади содержащихся в нем отверстий.

Существующие более сложные методы описания формы объектов используются в узкоспециализированных областях, таких как поиск торговых знаков [51, 65] и классификация объектов [68, 77].

1.3.4 Локальные дескрипторы

В последнее десятилетие в задачах поиска большое распространение получили локальные дескрипторы – векторы признаков, описывающие не все изображение в целом, а некоторые его значимые части (окрестности

особенных точек). В этом случае сходство двух изображений измеряется в количестве локальных дескрипторов, совпавших на обоих изображениях. Методы поиска по локальным дескрипторам применяются, в основном, для решения задач поиска нечетких дубликатов и заданного фрагмента на изображениях коллекции. На сегодняшний день наибольшую известность и распространение получили локальные дескрипторы SIFT [71], SURF [26], а также их модификации.

Дескриптор Scale Invariant Feature Transform

Метод SIFT формирует локальные признаки, инвариантные относительно масштаба и поворота, а также устойчивые к шуму. Данный алгоритм можно разделить на две части: определение точек интереса (особенных точек) и построение дескрипторов окрестностей данных точек. Для определения точек интереса для исходного изображения строится пирамида Гаусса. Изображения пирамиды затем приводятся к одному размеру, после чего попарно вычисляются разности между изображениями соседних масштабов (DoG, Difference-of-Gaussian). В качестве точек интереса выбираются пикселы, значение яркости которых является локальным экстремумом по сравнению со значениями яркостей соседних пикселов данного и двух соседних масштабов.

На следующем этапе для каждой такой точки интереса вычисляется локальный дескриптор, характеризующий направления и величины градиентов изображения соответствующего масштаба в области размером 16 × 16 пикселов, разбитой на 4 × 4 блока. Для обеспечения инвариантности относительно изменения ориентации производится преобразование локальных координат. Также величины градиентов взвешиваются с помощью гауссовой весовой функции для акцентирования центра окрестности точки интереса. Затем в каждом блоке направления градиентов распределяются по восьми ориентациям и их величины сохраняются в 8-мерные векторы. Полученные для каждого блока векторы затем объединяются в 128-мерный локальный дескриптор.

Главным минусом SIFT дескрипторов является их высокая размерность, а также большие вычислительные затраты, необходимые для их формирования. Для решения первой проблемы в работе [62] был предложен метод PCA-SIFT, в котором размерность дескриптора уменьшается до 36 элементов с помощью анализа главных компонент. Это повышает скорость сравнения двух изображений при некотором снижении точности. Также это приводит к дополнительным вычислительным затратам при формировании локальных дескрипторов.

Дескриптор Speeded Up Robust Features

Еще одним широко распространенным локальным дескриптором является SURF, требующий в несколько раз меньших вычислительных затрат и обеспечивающий примерно одинаковые результаты по сравнению с дескриптором SIFT [58]. В данном методе для ускорения вычислений используются интегральные изображения [34]. Значение в каждой точке интегрального изображения равно сумме значений соответствующего пиксела исходного изображения и всех пикселов выше и левее его. С помощью интегральных изображений прямоугольные фильтры, состоящие из нескольких прямоугольных областей, вычисляются за ограниченный интервал времени. Также как и SIFT, метод SURF можно разделить на этапы поиска точек интереса и их описания локальными дескрипторами.

Определение особых точек на изображении выполняется с помощью матрицы Гессе (детектор «быстрый Гессиан»):

$$\mathbf{H}(\mathbf{p},\sigma)=\begin{bmatrix} L_{xx}(\mathbf{p},\sigma) & L_{xy}(\mathbf{p},\sigma) \\ L_{xy}(\mathbf{p},\sigma) & L_{yy}(\mathbf{p},\sigma) \end{bmatrix}, \tag{1.12}$$

где $\mathbf{p}$ – точка в изображении I, σ – масштаб фильтра, $L_{xx}(\mathbf{p}, \sigma)$, $L_{xy}(\mathbf{p}, \sigma)$, и $L_{yy}(\mathbf{p}, \sigma)$ – свертки части изображения I в точке $\mathbf{p}$ с аппроксимациями вторых производных Гауссиана $g(\sigma)$.

Определитель матрицы Гессе (Гессиан) обладает инвариантностью относительно вращения, однако чувствителен к изменению масштаба. В связи с этим гессианы вычисляются для нескольких масштабов изображения. В качестве точек интереса выбираются локальные максимумы Гессианов, соответствующие локальным максимумам изменения градиента яркости (пятна, углы и края линий и т. п.).

На следующем этапе вокруг каждой найденной точки интереса выбирается квадратный регион с размером сторон 20*s*, где *s* – масштаб, на котором найдена точка интереса. Полученный регион интереса разбивается на 4×4 квадратных блока. В каждом блоке для 5×5 равномерно распределенных точек вычисляются отклики вейвлета Хаара по горизонтальному и вертикальному направлениям. При этом полученные значения взвешиваются с помощью фильтра Гаусса, центрированного на точке интереса, для подавления шумов. На следующем шаге для каждого блока формируется 4-мерный вектор из сумм значений и сумм абсолютных значений откликов Хаара. Полученные векторы объединяются, формируя 64-мерный локальный дескриптор.

В связи с использованием фильтра Гаусса при формировании SURF-дескриптора происходит размытие краев и деталей изображения, что снижает точность описания. Для решения этой проблемы в работе [24] предложен дескриптор Gauge SURF (G-SURF), в котором отсутствует этап взвешивания элементов дескриптора, а вместо откликов Хаара используются производные по калибровочным координатам:

$$L_{ww}(\mathbf{p},\sigma) = \frac{L_x^2(\mathbf{p},\sigma)L_{xx}(\mathbf{p},\sigma) + 2L_x(\mathbf{p},\sigma)L_{xy}(\mathbf{p},\sigma)L_y(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)L_{yy}(\mathbf{p},\sigma)}{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)}, \qquad (1.13)$$

$$L_{vv}(\mathbf{p},\sigma) = \frac{L_x^2(\mathbf{p},\sigma)L_{yy}(\mathbf{p},\sigma) - 2L_x(\mathbf{p},\sigma)L_{xy}(\mathbf{p},\sigma)L_y(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)L_{xx}(\mathbf{p},\sigma)}{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)}, \qquad (1.14)$$

где $L_x(\mathbf{p}, \sigma)$ и $L_y(\mathbf{p}, \sigma)$ – свертки части изображения I в точке $\mathbf{p}$ с аппроксимациями первых производных Гауссиана $g(\sigma)$.

Дескриптор G-SURF демонстрирует повышение точности сравнения некоторых типов изображений, однако требует дополнительных вычислительных затрат.

1.3.5 Кодирование локальных дескрипторов

Локальные дескрипторы демонстрируют более высокие результаты при поиске изображений, чем рассмотренные ранее глобальные признаки. Однако при измерении сходства двух изображений требуется сравнение большого количества локальных дескрипторов, что затрудняет их использование в задачах классификации и аннотирования изображений. Для решения этой проблемы предложено несколько методов кодирования набора локальных дескрипторов, извлеченных из изображения, в один глобальный дескриптор. В общем случае они состоят из четырех этапов [22].

На первом этапе определяются точки интереса на изображении, окрестности которых будут описаны с помощью локальных дескрипторов. Помимо рассмотренных выше детекторов на основе DoG и матриц Гессиана в работе [66] было предложено использовать в качестве точек интереса узлы регулярной сетки, наложенной на изображение. Также было показано, что применение узлов регулярной сетки существенно эффективнее при категоризации изображений по типу сцены по сравнению с другими детекторами. В работе [12] было предложено дополнительно сегментировать изображение и отсеивать узлы, находящиеся вблизи границ областей. Это позволяет повысить эффективность, но требует дополнительных вычислений.

На втором этапе окрестности выбранных точек интереса описываются с помощью локальных дескрипторов. Для этого могут использовать дескрипторы SIFT и SURF, а также векторы из цветовых и текстурных признаков [14].

На третьем этапе из обучающей выборки изображений случайным образом выбирается большой набор локальных дескрипторов. Данный набор кластеризуется и центры масс используются в качестве визуальных слов. В большинстве случаев для кластеризации используется алгоритм *k*-средних [69], являющийся одним из наиболее простых способов кластеризации, обеспечивающий заданное количество кластеров. Также в работе [39] было предложено формировать визуальные слова с помощью нейронной сети ESOINN [9, 93]. Нейронная сеть автоматически определяет количество кластеров, что позволяет формировать более точные визуальные слова, однако изменяющийся размер словаря визуальных слов не всегда удобен. В работе [10] было предложено дополнительно уточнять визуальные слова, полученные в результате кластеризации, с помощью контекстной информации о принадлежности дескрипторов областям изображений. Это позволяет избежать ситуации, когда части изображения, являющиеся различными в семантическом плане, относятся к одному и тому же визуальному слову из-за схожести их визуальных признаков, однако требует дополнительных затрат для выделения областей изображений.

На четвертом этапе набор локальных дескрипторов одного изображения кодируется с помощью сформированного словаря. Наиболее простым способом является метод Bag-of-Visual-Words (BoVW) [97]. В этом методе для каждого локального дескриптора определяется ближайшее визуальное слово, после чего подсчитывается, сколько раз то или иное визуальное слово встретилось в изображении. Таким образом, глобальный дескриптор представляет собой гистограмму визуальных слов. Чтобы снизить потери информации, связанные с заменой локальных дескрипторов на дискретные визуальные слова, в работах [102] и [107] были предложены два метода. Эти методы определяют для каждого локального дескриптора несколько визуальных слов и отличаются способом выбора ближайших визуальных слов и их весов при формировании гистограммы. Также эксперименты показали, что увеличение количества визуальных слов

повышает точность поиска изображений, в связи с чем в больших база изображений используется до миллиона визуальных слов [85]. Это приводит к значительным вычислительным затратам при сопоставлении локальных дескрипторов и визуальных слов.

Альтернативным подходом к кодированию является метод агрегирования локальных дескрипторов (VLAD, Vector of Locally Aggregated Descriptors) [59]. Суть метода заключается в описании отличия изображения от некоторого усредненного варианта. Это достигается путем определения для каждого локального дескриптора $\mathbf{x}_b$ ближайшего визуального слова $NN(\mathbf{x}_b)$ и накопления разницы $\mathbf{v}_s$ между s-ыми визуальными словами и ближайшими к ним локальными дескрипторами:

$$\mathbf{v}_s = \sum_{\mathbf{x}_b : NN(\mathbf{x}_b) = \mathbf{vw}_s} \mathbf{vw}_s - \mathbf{x}_b. \tag{1.15}$$

Глобальный дескриптор представляет собой объединение всех вычисленных векторов. Данный метод показывает лучшие результаты по сравнению с BoVW, используя на несколько порядков меньшее количество визуальных слов. Также в работе [22] предлагается формировать отдельно две части глобального дескриптора, описывающие все изображение и только наиболее значимые объекты. Для этого локальные дескрипторы взвешиваются с помощью карт заметности [106]. В этом случае локальные дескрипторы, находящиеся на наиболее выделяющихся объектах, получат большие веса, чем локальные дескрипторы фона. Для формирования части глобального дескриптора, описывающей объекты, веса локальных дескрипторов используются при вычислении разницы с визуальными словами. Данный метод позволяет повысить точность категоризации, но требует дополнительных вычислительных затрат для формирования карт значимости, а также удваивает длину глобального дескриптора.

1.4 Анализ существующего программного обеспечения

На сегодняшний день существует несколько программных систем для автоматического аннотирования изображений. Рассмотрим некоторые из них.

Исследовательский программный продукт «PiXiT» [84], разработанный компанией PEPITe на основе работы [75], предназначен для классификации изображений. Для того чтобы воспользоваться им, необходимо написать разработчикам письмо с указанием наименования вашей организации и кратким описанием проблемы, с которой собираетесь работать. Также необходимо дополнительно сообщить о результатах проведенных исследований. Программа предполагает наличие обучающих изображений, из которых случайным образом извлекается большое количество блоков. С использованием этих блоков обучается ансамбль деревьев решений, с помощью которого новым изображениям автоматически присваивается по одной метке категорий (рис. 1.4). Помимо настроек по умолчанию, пользователю предоставляется возможность настроить параметры разбиения изображений на блоки, извлечения признаков и обучения классификатора.

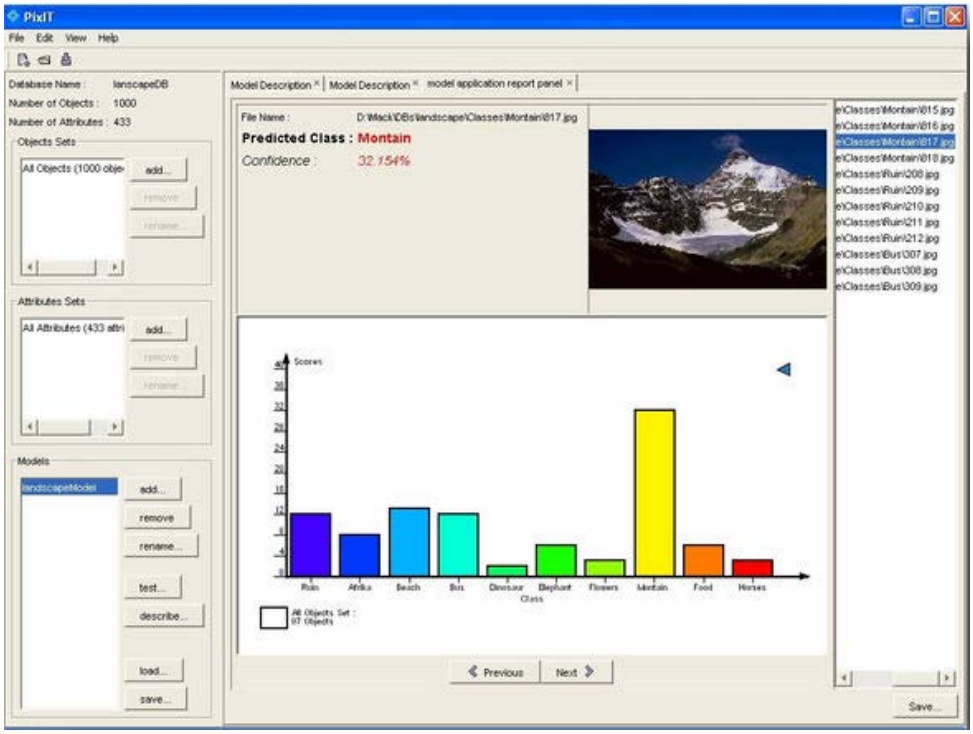

Рисунок 1.4. Интерфейс программы «PiXiT»

Программа «ImageTagger» [54], разработанная компанией Attrasoft, является коммерческим продуктом, позволяющим аннотировать изображения несколькими ключевыми словами. Для использования программы необходимо предварительное обучение, требующее от 1000 до 10 000 тренировочных изображений для каждого ключевого слова. При этом размер словаря ключевых слов должен быть небольшим (в пределах 10 – 20 слов). Интерфейс программы с примером аннотирования изображения представлен на рисунке 1.5.

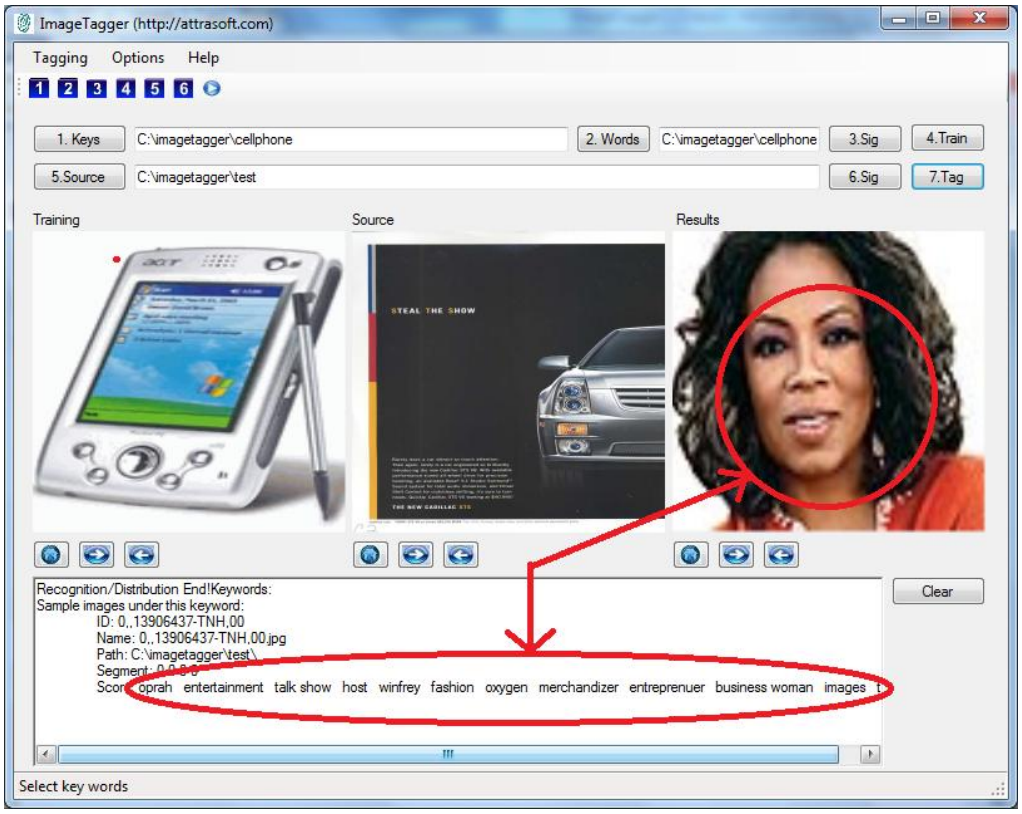

Рисунок 1.5. Интерфейс программы «ImageTagger»

Исследовательский веб-сервис «MUFIN Image Annotation» [80], разработанный лабораторией DISA (Laboratory of Data Intensive Systems and Applications), реализует поисковый метод аннотирования. Для загружаемого изображения определяется набор визуально похожих изображений среди 20 миллионов изображений фотобанка Profimedia [86]. После этого ключевые слова изображений набора кластеризуются по семантическому значению и взвешиваются на основании визуального сходства изображений и

семантической схожести ключевых слов. В качестве аннотации выбирается 20 ключевых слов с наибольшими весами (рис. 1.6).

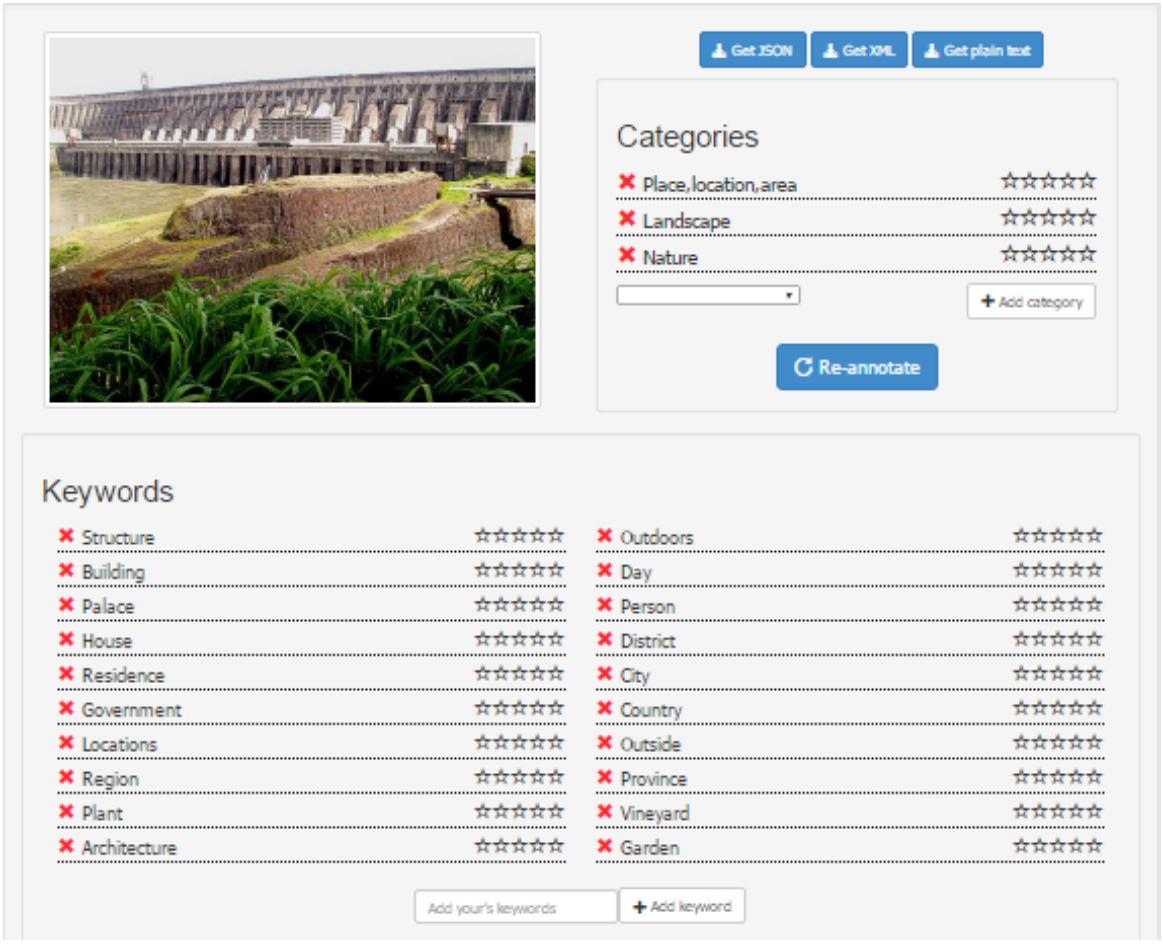

Рисунок 1.6. Пример аннотирования изображения в веб-приложении «MUFIN Image Annotation»

Коммерческий веб-сервис «Imagga» [55] включает API, предоставляющий доступ к автоматическому аннотированию изображений. В ответ на пользовательский запрос, включающий идентификационную информацию и изображение, API предоставляет список, состоящий из пар ключевое слово – уровень доверия (рис. 1.7). Разработчиками системы рекомендуется фильтровать полученный список по уровню доверия в 30 %, что позволяет сократить количество нерелевантных ключевых слов, однако нередко оставляет в аннотации только 2-3 слова.

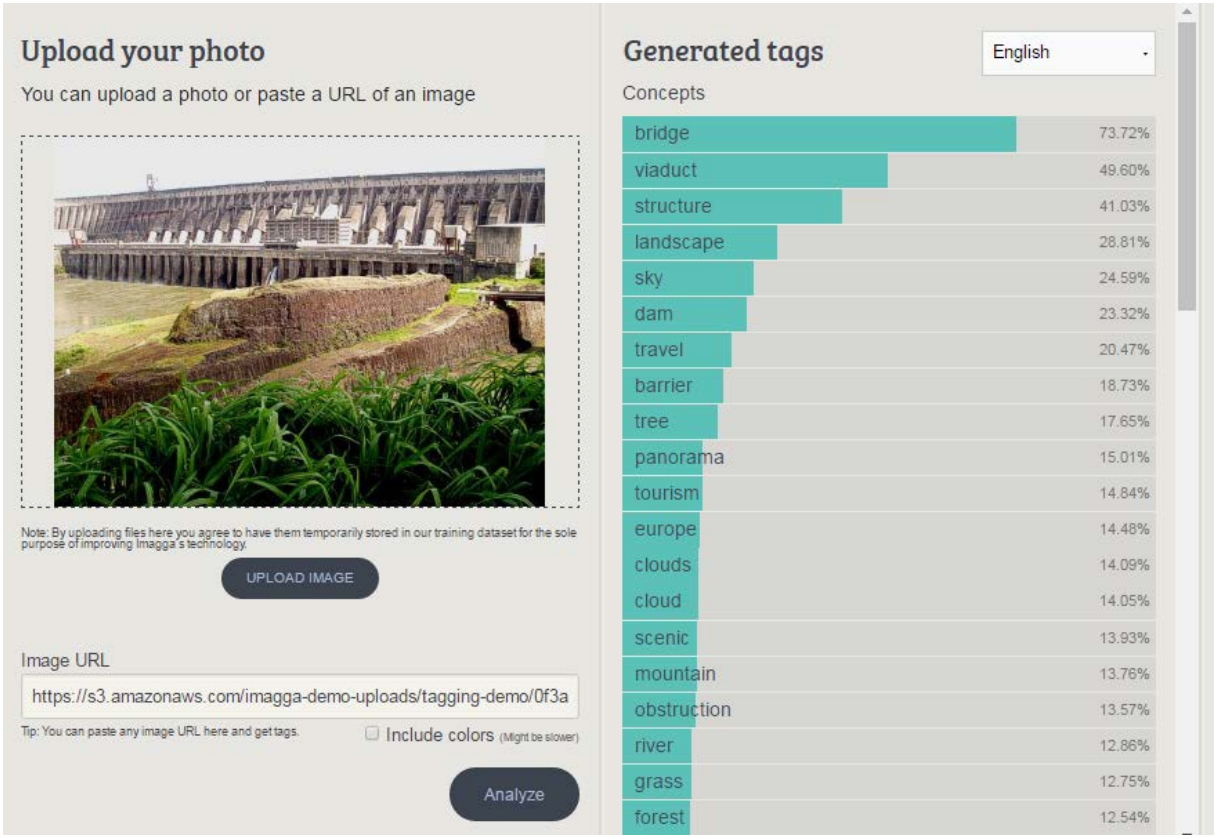

Рисунок 1.7. Пример аннотирования изображения в веб-сервисе «Imagga»

Бесплатный веб-сервис «Google Photos» [45], разработанный и поддерживаемый компанией Google, предназначен для загрузки, обработки и хранения пользовательских фотографий. Отличительной чертой сервиса является использование нейронной сети, позволяющей распознавать часть наиболее заметных объектов, благодаря чему пользователям предоставляется возможность поиска фотографий по ключевым словам. Однако следует отметить, что система не отображает присвоенные ключевые слова. Это затрудняет поиск фотографий, поскольку используемые ключевые слова не всегда очевидны. Также система не всегда распознает некоторые объекты на почти похожих изображениях (рис. 1.8).

Таким образом, несмотря на наличие коммерческих программных продуктов, существующие системы ААИ имеют достаточно низкие показатели точности и полноты аннотирования, в связи с чем требуется дальнейшее развитие методов автоматического аннотирования изображений.

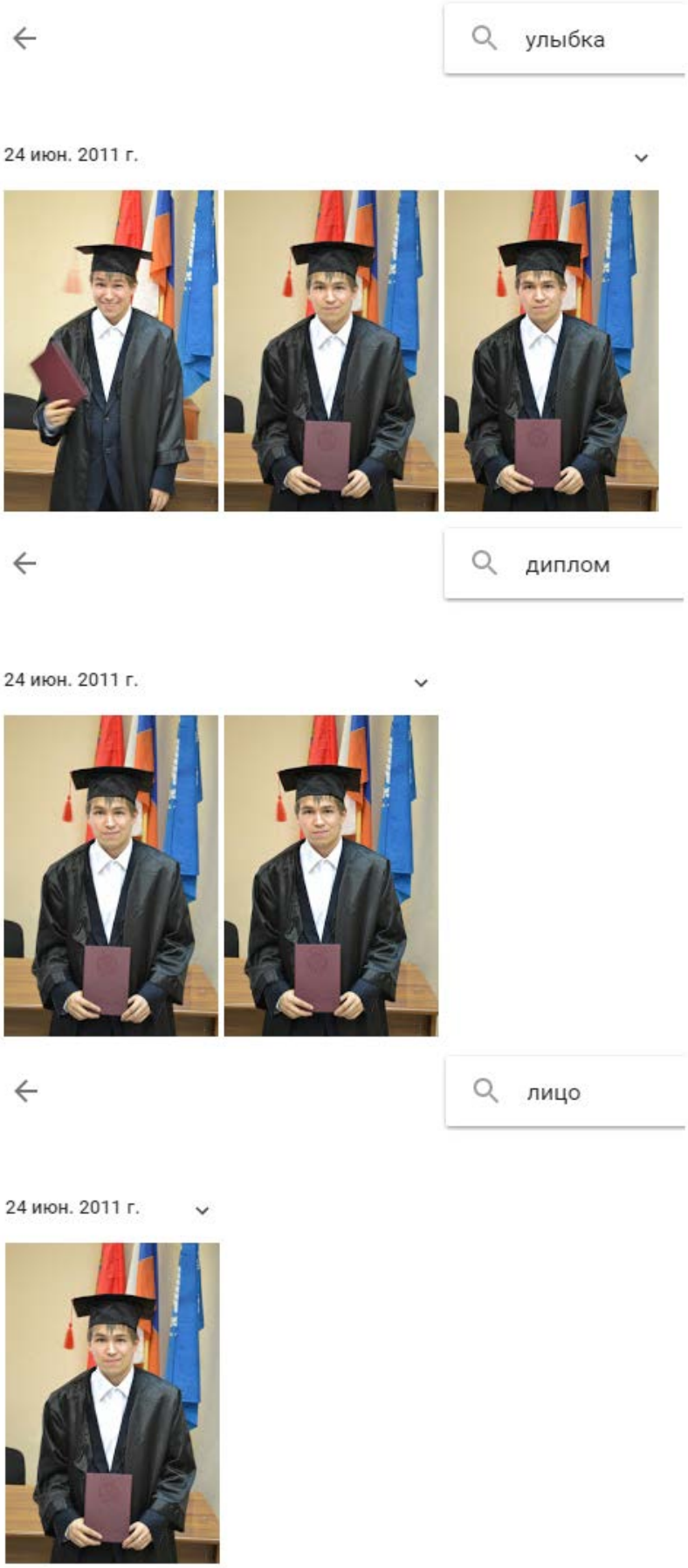

Рисунок 1.8. Примеры поиска фотографий в веб-сервисе «Google Photos»

1.5 Выводы по главе

В первой главе представлен обзор существующих методов автоматического аннотирования изображений, кластеризации данных и описания изображений с помощью низкоуровневых признаков, приведена их классификация. Также рассмотрен ряд программных систем, реализующих автоматическое аннотирование изображений.

Существующие методы ААИ можно разделить на три подхода: классификационный, генеративный и поисковый. Классификационные методы представляют ключевые слова в виде независимых классов, на примерах которых обучается классификатор. Это позволяет быстро присвоить изображениям или их областям метки категорий, однако для достижения достаточной точности необходима сбалансированная обучающая выборка. Также увеличение количества категорий (ключевых слов) приводит к значительному снижению точности классификации.

Методы генеративного подхода на основе набора обучающих изображений оценивают вероятности совместной встречаемости ключевых слов и низкоуровневых признаков областей изображений. Полученные значения вероятностей используются для определения вероятностей принадлежности ключевых слов новому изображению. Поскольку большая часть необходимых вычислений выполняется на этапе обучения, то аннотирование нового изображений осуществляется быстрее, чем в классификационных методах, а обучение системы – медленнее.

Как генеративные, так и классификационные методы имеют низкую масштабируемость, требуя повторного обучения системы при каждом добавлении новых ключевых слов и обучающих изображений. Методы поискового подхода лишены этого недостатка, поскольку аннотирование нового изображения заключается в поиске небольшого количества визуально похожих обучающих изображений и выборе их ключевых слов в качестве аннотации. При этом метод 2PKNN, предварительно разделяющий

обучающую выборку на семантические группы, продемонстрировал повышение точности аннотирования, что свидетельствует о необходимости предварительного структурирования обучающего набора.

Одним из возможных способов структурирования изображений является их кластеризация по текстовому описанию и визуальным признакам. Среди существующих методов кластеризации можно выделить иерархические и инкрементальные методы, алгоритмы квадратичной ошибки. Иерархические методы разделяют кластеризуемую выборку в систему вложенных разбиений, формирующих дерево кластеров. Это позволяет использовать выборку обучающих изображений на нескольких уровнях детализации, однако высокая вычислительная сложность затрудняет использование иерархических методов в поисковых системах.

Методы квадратичной ошибки кластеризуют выборку векторов в сферические группы, расположенные как можно дальше друг от друга. Наиболее распространенным методом этой категории является алгоритм *k*-средних и его модификации. Эти методы обладают простотой реализации и показывают хорошие результаты кластеризации, однако требуют заранее задавать количество кластеров.

В отличие от методов первых двух категорий, предполагающих неизменность множества кластеризуемых векторов в течение работы алгоритмов, инкрементальные методы способны кластеризовать заранее неизвестное количество векторов. Одним из таких методов является сеть ESOINN, автоматически определяющая количество кластеров, обрабатывая кластеризуемые векторы по одному. Благодаря способности кластеризовать новые данные без повторного обучения, сеть ESOINN может использоваться совместно с поисковыми методами ААИ.

Большую роль при кластеризации и аннотировании изображений играют низкоуровневые признаки, используемые для компактного представления изображений. В большинстве случаев используют признаки цвета, среди которых наиболее распространенны различные модификации

цветовых гистограмм, и текстурные признаки, такие как признаки Тамуры и Харалика, фильтры Габора и ICA-фильтры. Также в последнее время получили большое распространение локальные дескрипторы SIFT и SURF, описывающие окрестности особенных точек, и методы их кодирования в глобальные дескрипторы.

Проведенный обзор существующих систем аннотирования показал, что, несмотря на наличие коммерческих программных продуктов, точность и полнота аннотирования предлагаемых решений остается на достаточно низком уровне. Также представлено сравнение существующих методов ААИ, проведенное на двух современных базах изображений, показавшее, что лучшие результаты не превышают 56 % для точности и 35 % для полноты аннотирования. Таким образом, требуется дальнейшее развитие методов автоматического аннотирования изображений.

ГЛАВА 2. АВТОМАТИЧЕСКОЕ АННОТИРОВАНИЕ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ОДНОРОДНЫХ ТЕКСТОВО-ВИЗУАЛЬНЫХ ГРУПП

Анализ существующих подходов и методов ААИ показал целесообразность аннотирования новых изображений на основе обучающих изображений, наиболее похожих визуально, а также необходимость предварительного структурирования обучающего набора изображений. Сложность состоит в том, что в обучающем наборе каждое ключевое слово относится ко всему изображению, а не отдельным объектам, кроме того в аннотациях могут отсутствовать некоторые релевантные ключевые слова. Ограничения и условия, предъявляемые к обучающему набору и аннотируемым изображениям, представлены в таблице 2.1.

Таблица 2.1

Ограничения и условия, предъявляемые к обучающему набору и аннотируемым изображениям

Критерий	**Ограничение**
Тип изображений	Фотографии
Размер изображений	Не менее 480 × 360 пикселов
Размер аннотируемых объектов	Не менее 5% площади изображения
Качество изображений	Равномерно освещенные, контрастные, без размытия и смазов, уровень шума – не более 2-3 дБ
Частота встречаемости ключевых слов	Не менее 0,005

Предположим, что обучающий набор TS состоит из изображений и соответствующих им текстовых описаний. Пусть $J = \{I_1, \ldots, I_M\}$ – коллекция изображений, а $K = \{k_1, \ldots, k_N\}$ – словарь, состоящий из N ключевых слов, тогда обучающий набор $TS = \{(I_1, K_1), \ldots, (I_M, K_M)\}$, где $K_m \subseteq K$. Также предположим, что обучающий набор разделен на несколько непересекающихся однородных текстово-визуальных групп (ОТВ-групп) $H = \{H_1, \ldots, H_L\}$, а выбор ключевых слов в процессе аннотирования тестового изображения A зависит от ассоциации изображения с той или иной группой.

Обозначим вероятность принадлежности изображения A ОТВ-группе H_l как $P(A|H_l)$. Также введем вероятность $P_l(A|k_n)$ для оценки принадлежности изображения A семантической группе, сформированной из изображений ОТВ-группы H_l, имеющих в описании ключевое слово k_n. В этом случае аннотирование изображения моделируется как проблема поиска апостериорных вероятностей:

$$P(k_n \mid A) = \frac{\sum_{H_l \in H} \left[P(H_l) P(A \mid H_l) P_l(A \mid k_n) P_l(k_n) \right]}{P(A)}, \tag{2.1}$$

где $P(H_l)$ – априорная вероятность ОТВ-группы H_l; $P_l(k_n)$ – априорная вероятность ключевого слова k_n внутри ОТВ-группы H_l; $P(A)$ – априорная вероятность тестового изображения A.

Поскольку априорная вероятность $P(A)$ является константой, то будем анализировать числитель:

$$P(k_n \mid A) \approx \sum_{H_l \in H} \left[P(H_l) P(A \mid H_l) P_l(A \mid k_n) P_l(k_n) \right]. \tag{2.2}$$

Используя полученные значения $P(k_n|A)$ ключевые слова ранжируются по убыванию. В качестве аннотации используется Nkw ключевых слов с наибольшей вероятностью. Таким образом, для аннотирования тестового изображения A необходимо оценить вероятности $P(H_l)$, $P(A|H_l)$, $P_l(A|k_n)$ и $P_l(k_n)$. Предлагаемый для решения этой задачи алгоритм автоматического аннотирования изображений на основе однородных текстово-визуальных групп можно разделить на три этапа обучения и этап аннотирования:

I этап. Вычисление глобального визуального дескриптора:

– быстрое вычисление набора локальных дескрипторов;

– вычисление цветовых локальных дескрипторов;

– кодирование набора локальных дескрипторов.

II этап. Создание текстового дескриптора:

– формирование текстового дескриптора;

– восстановление пропущенных ключевых слов обучающих изображений.

III этап. Формирование однородных текстово-визуальных групп:

– первичное разделение изображений на основе совместной встречаемости ключевых слов;

– кластеризация изображений с использованием текстово-визуальных дескрипторов.

IV этап. Автоматическое аннотирование изображений.

Далее рассмотрим более подробно алгоритмическую реализацию каждого из представленных этапов.

2.1 Вычисление глобального визуального дескриптора

На первом этапе автоматического аннотирования изображений для каждого изображения I из коллекции обучающих изображений $J = \{I_1, \ldots, I_M\}$, а также набора аннотируемых изображений, вычисляется глобальный визуальный дескриптор $\mathbf{V} = \{V_1, \ldots, V_Z\}$. Для этого из изображения извлекается набор локальных дескрипторов, который кодируется с помощью словаря визуальных слов в один глобальный дескриптор. При этом вычисление локальных дескрипторов с использованием регулярной сетки эффективнее по сравнению с другими детекторами, а увеличение пересечения областей, на которых вычисляются дескрипторы, повышает точность аннотирования. Однако это приводит к значительному увеличению вычислительных затрат. Для решения этой проблемы предложен метод быстрого извлечения набора локальных дескрипторов, описывающих все части изображения, а также алгоритм для их кодирования в глобальный визуальный дескриптор.

2.1.1 Быстрое вычисление набора локальных дескрипторов

Предлагаемый алгоритм быстрого вычисления набора локальных дескрипторов (Fast Dense Speeded-Up Features – FD-SUF) [18] основан на локальном дескрипторе SURF (п. 1.3.4) и состоит из двух этапов: вычисления матрицы частей дескрипторов **M**, в которой каждый элемент $\mathbf{M}_{rx,ry}$ является вектором из 4 чисел, и построения с ее помощью набора локальных дескрипторов (рис. 2.1).

На первом этапе исходное изображение I (рис. 2.2, а) переводится из цветового пространства RGB в оттенки серого (рис. 2.2, б) для чего используется компонента Y из цветовой схемы YUV:

$$Y = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B. \tag{2.3}$$

Полученное изображение разделяется сеткой на блоки $I_{rx,ry}$ размером $5\sigma_{sc} \times 5\sigma_{sc}$ пикселов, где σ_{sc} – масштаб (для изображений размером 480 × 360 равен 1) (рис. 2.2, в). Масштаб подбирается пропорционально размеру изображений таким образом, чтобы размеры матриц **M**, вычисленных для любых двух изображений, были сопоставимы.

На следующем шаге в каждом блоке для 5 × 5 равномерно распределенных точек вычисляются первые производные:

$$L_x(p,\sigma_{sc}) = I(p) * \frac{\partial}{\partial x} g(\sigma_{sc}), \tag{2.4}$$

$$L_y(p,\sigma_{sc}) = I(p) * \frac{\partial}{\partial y} g(\sigma_{sc}), \tag{2.5}$$

где p – точка в изображении I с координатами (x, y); $L_x(p, \sigma_{sc})$ и $L_y(p, \sigma_{sc})$ – свертки части изображения I с центром в точке p с первыми производными Гауссиана $g(\sigma_{sc})$ по осям OX и OY соответственно.

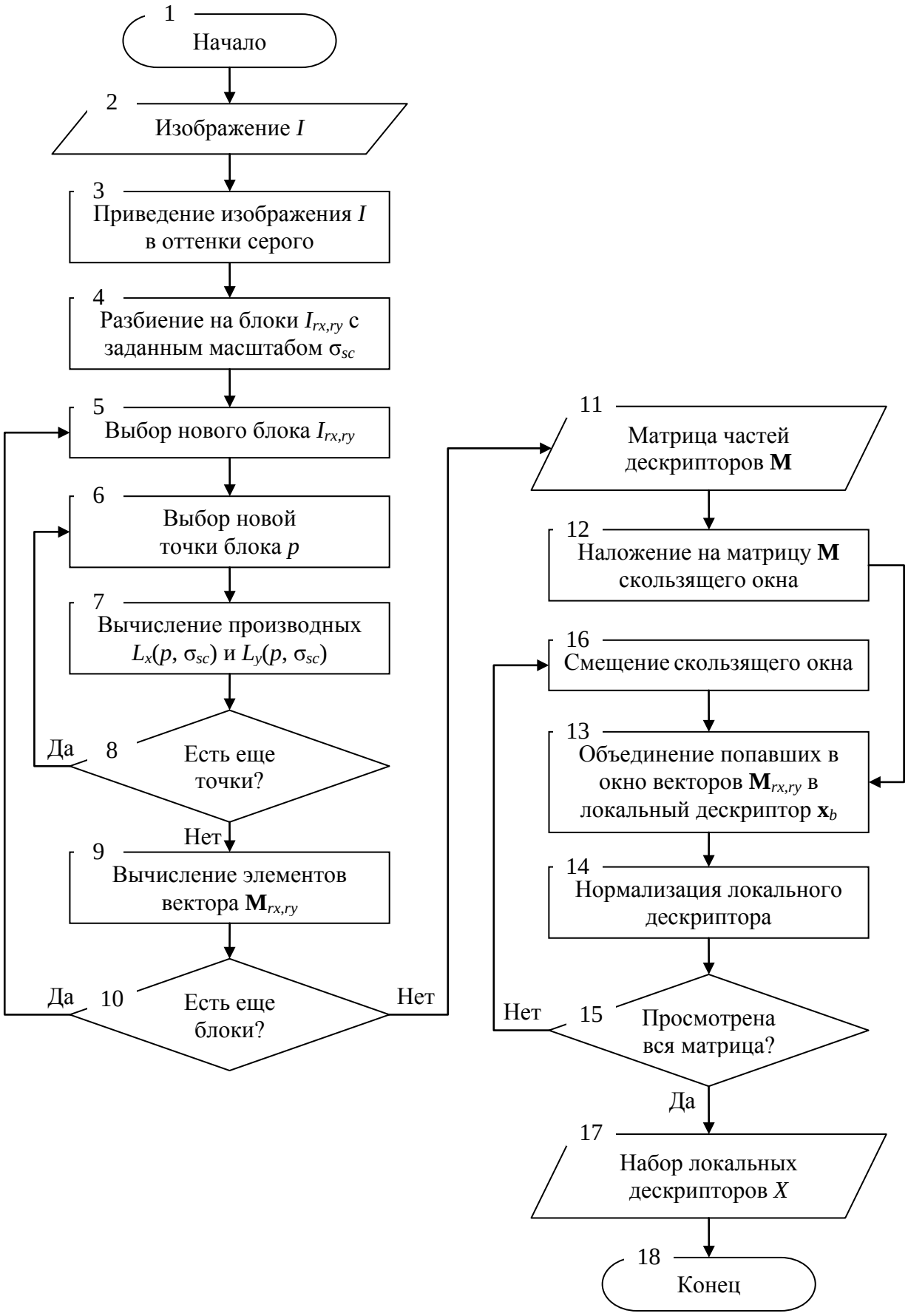

Рисунок 2.1. Блок-схема алгоритма быстрого вычисления набора локальных дескрипторов

Используя накопленные значения первых производных, каждый элемент матрицы частей дескрипторов формируется следующим образом (рис. 2.2, г):

$$\mathbf{M}_{rx,ry} = \left(\sum_{p \in I_{rx,ry}} L_x(p, \sigma_{sc}), \quad \sum_{p \in I_{rx,ry}} L_y(p, \sigma_{sc}), \quad \sum_{p \in I_{rx,ry}} \left| L_x(p, \sigma_{sc}) \right|, \quad \sum_{p \in I_{rx,ry}} \left| L_y(p, \sigma_{sc}) \right| \right). \quad (2.6)$$

На втором этапе алгоритма формируется набор локальных дескрипторов $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_B\}$, где $\mathbf{x}_b \in \mathbf{R}^D$. Для этого на матрицу $\mathbf{M}$ накладывается скользящее окно размером 4×4 элемента (рис. 2.2, д). В дескрипторе SURF каждый элемент векторов $\mathbf{M}_{rx,ry}$, попавших в скользящее окно, дополнительно взвешивается с помощью фильтра Гаусса, однако, как показали исследования, это приводит к размытию краев и деталей изображения, что снижает точность описания. В связи с этим, в данном исследовании предложено формировать локальный дескриптор объединением всех 16 векторов $\mathbf{M}_{rx,ry}$ скользящего окна без взвешивания элементов (рис. 2.2, е). Полученный дескриптор нормализуется с помощью l_2-нормы, после чего скользящее окно смещается. Изменяя шаг смещения скользящего окна можно регулировать количество извлекаемых локальных дескрипторов.

В предложенном алгоритме FD-SUF основные вычислительные затраты требуются на первом этапе, включающем два базовых типа циклов – внешний цикл по изображению размером $(w/5\sigma_{sc}) \times (h/5\sigma_{sc})$ и вложенные циклы по блокам, где w и h – ширина и высота изображения. Таким образом, повысить эффективность вычислений можно с помощью распараллеливания внешнего цикла [15]. Для этого изображение разделяется на полосы, обработка которых распределяется между ядрами процессора вычислительной системы. Для реализации таких параллельных алгоритмов широкое распространение получил стандарт OpenMP для распараллеливания

программ на языках Си, Си++ и Фортран. Помимо него в последнее время развивается расширение языка Си++, известное как Intel Cilk Plus.

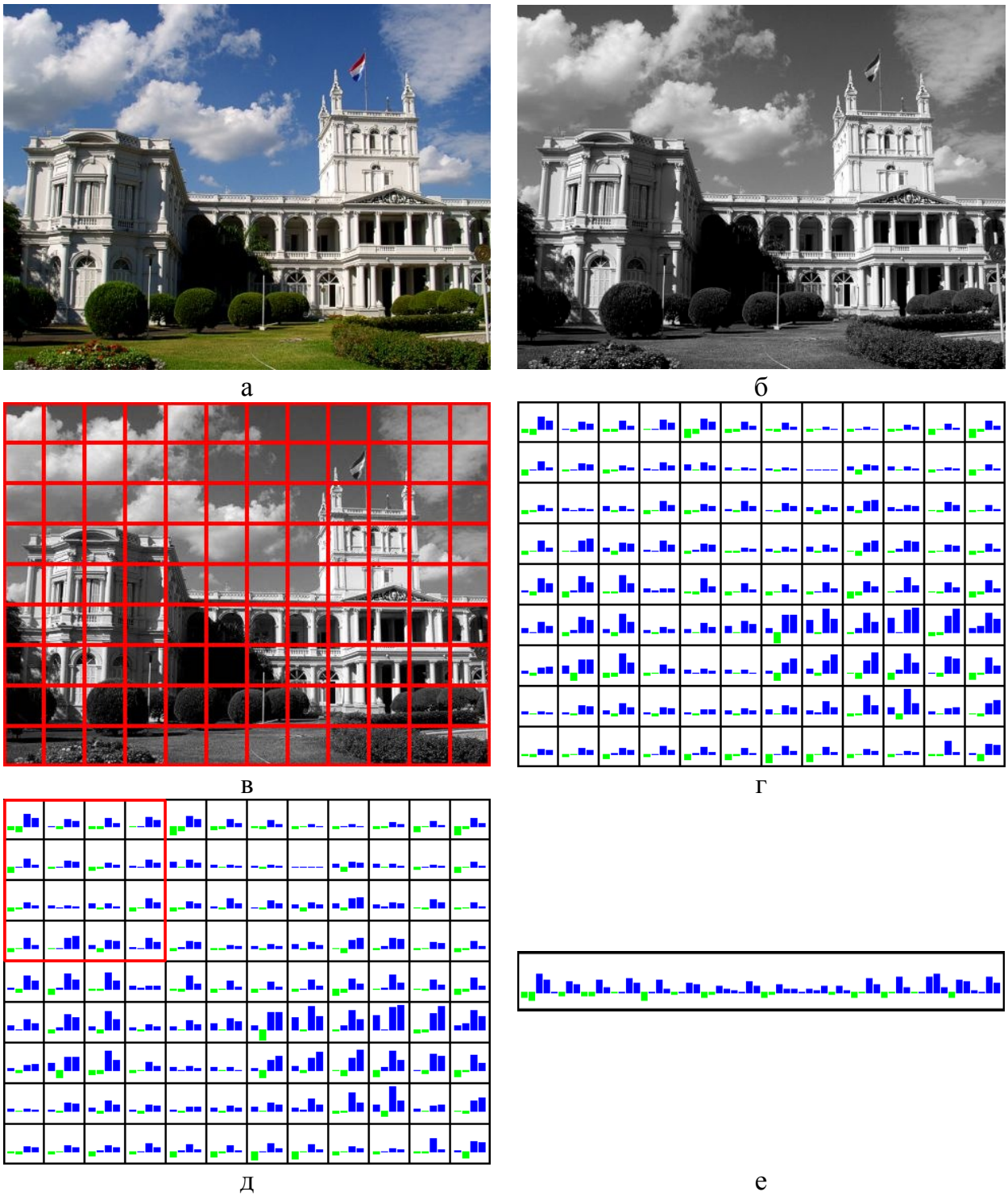

Рисунок 2.2. Иллюстрация вычисления локального дескриптора для изображения 58 из базы IAPR TC-12: а) исходное изображение; б) приведение изображения в оттенки серого; в) разбиение на блоки с масштабом $\sigma_{sc} = 8$; г) визуализация матрицы частей дескрипторов (синим цветом обозначены положительные значения, зеленым – отрицательные); д) наложение скользящего окна на матрицу частей дескрипторов; е) визуализация сформированного локального дескриптора

2.1.2 Вычисление цветовых локальных дескрипторов

Базовый алгоритм FD-SUF вычисляется только на изображениях в оттенках серого, в связи с чем подвержен сильному влиянию условий освещенности, а также не учитывает цветовую информацию. Для решения этой проблемы предложено формировать цветовые локальные дескрипторы [17]: дескрипторы FD-SUF вычисляются для каждой компоненты цветового пространства, после чего они объединяются (рис. 2.3).

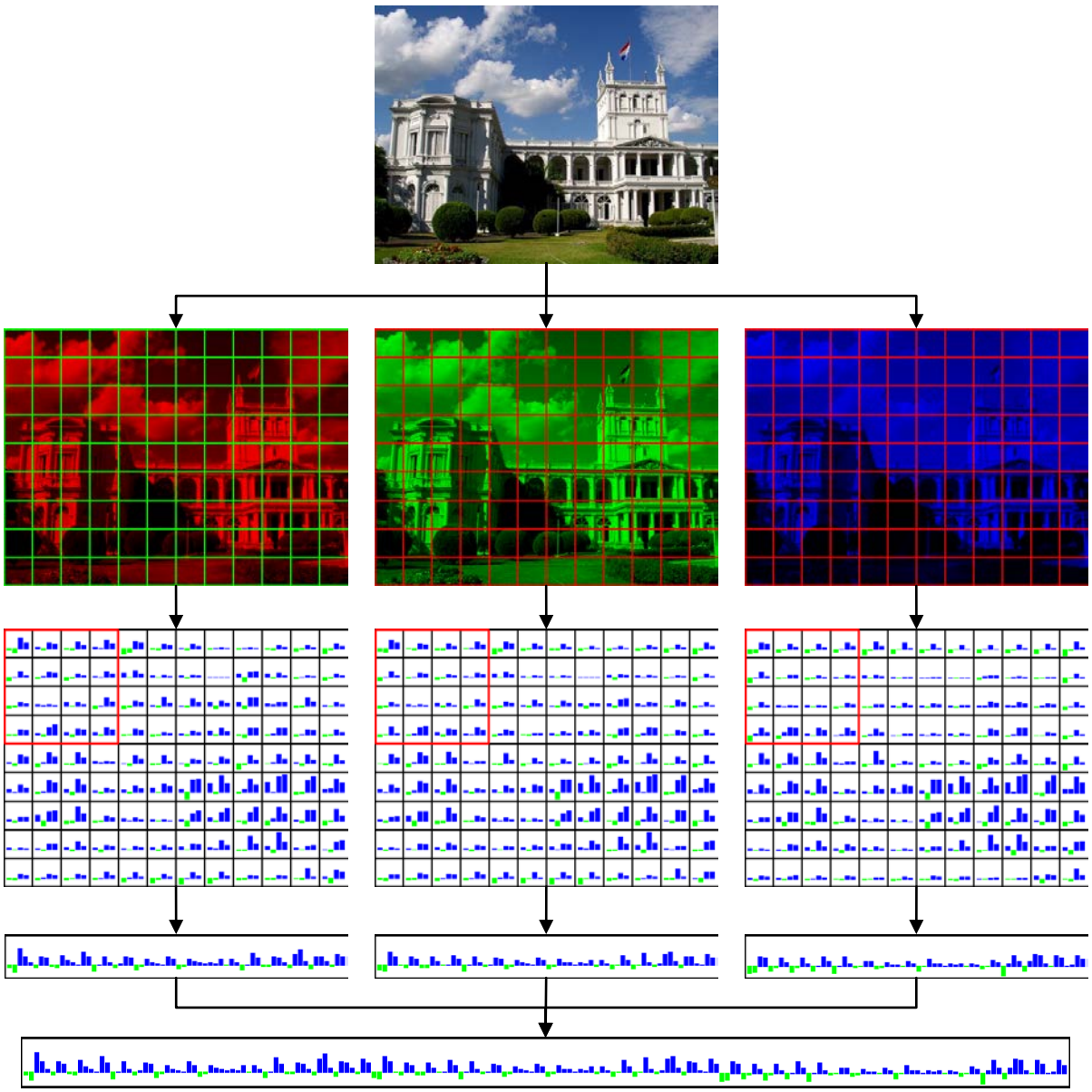

Рисунок 2.3. Иллюстрация вычисления цветового локального дескриптора для изображения 58 из базы IAPR TC-12 в цветовом пространстве RGB

В исследовании были использованы такие распространенные цветовые пространства, как HSV (Hue, Saturation, Value) и LUV (Lightness, Uniform chromaticity scale, Valence), а также пространства, обладающие некоторой степенью инвариантности к изменению интенсивности света [38]:

– rg-пространство, являющееся нормализацией цветовой модели RGB. В этом пространстве хроматические компоненты r и g описывают цветовую информацию, а компонент b является избыточным, поскольку $r + g + b = 1$:

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \dfrac{R}{R+G+B} \\ \dfrac{G}{R+G+B} \\ \dfrac{B}{R+G+B} \end{pmatrix} \tag{2.7}$$

– Цветовое пространство Opponent, включающее два цветовых канала O_1, O_2 и компоненту интенсивности O_3:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \dfrac{R-G}{\sqrt{2}} \\ \dfrac{R+G-2B}{\sqrt{6}} \\ \dfrac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{2.8}$$

– Цветовая модель HSI (Hue, Saturation, Intensity):

$$\begin{pmatrix} h \\ s \\ i \end{pmatrix} = \begin{pmatrix} \operatorname{actan}\left(O_1/O_2\right) \\ \sqrt{O_1^2 + O_2^2} \\ O_3 \end{pmatrix} \tag{2.9}$$

В этой модели компонент *h* (оттенок) обладает нестабильностью вблизи серого цвета. Однако определенность оттенка обратно пропорциональна насыщенности (компонент *s*), в связи с чем для большей устойчивости компоненты *h* и *s* заменяются на их произведение.

– Нормализация значений пикселов в RGB-пространстве (nRGB):

$$\begin{pmatrix} nR \\ nG \\ nB \end{pmatrix} = \begin{pmatrix} \dfrac{R-\mu_R}{\sigma_R} \\ \dfrac{G-\mu_G}{\sigma_G} \\ \dfrac{B-\mu_B}{\sigma_B} \end{pmatrix} \qquad (2.10)$$

где μ_i и σ_i – среднее значение и среднеквадратичное отклонение в *i*-м канале, вычисляемые по выбранной области (блок или все изображение).

После вычисления расширенных локальных дескрипторов, их размерность сокращаются до 64 элементов с помощью метода главных компонент.

2.1.3 Кодирование набора локальных дескрипторов

Извлеченный из изображения набор локальных дескрипторов кодируется в глобальный визуальный дескриптор с помощью алгоритма, представленного на рисунке 2.4 б. Данный алгоритм требует на этапе обучения с помощью коллекции обучающих изображений *J* создать словарь визуальных слов $VW = \{\mathbf{vw}_1, \ldots, \mathbf{vw}_S\}$, где $\mathbf{vw}_s \in \mathbf{R}^D$.

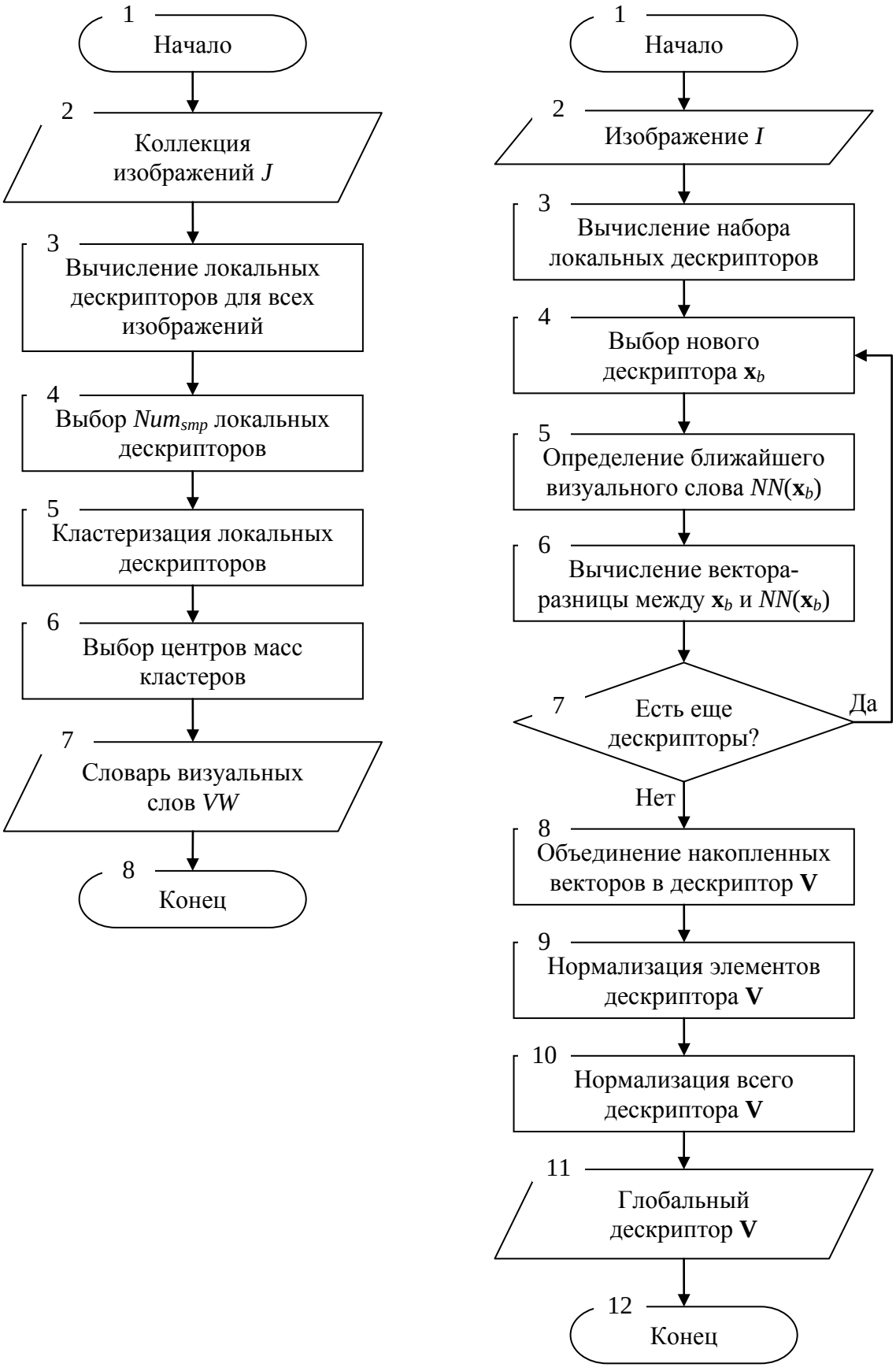

Рисунок 2.4. Блок-схемы алгоритмов: а) формирования словаря визуальных слов; б) кодирования набора локальных дескрипторов

Для этого из обучающих изображений случайным образом выбирается Num_{smp} локальных дескрипторов (в исследовании используется Num_{smp} = 200 000), которые затем кластеризуются с помощью алгоритма *k*-средних (п. 1.2.2). Количество кластеров обычно устанавливается в пределах от 16 до 256, а в качестве визуального слова $\mathbf{vw}_s$ выбирается центр масс *s*-го кластера.

Используя полученный словарь, локальные дескрипторы преобразуются в один глобальный вектор $\mathbf{V} \in \mathbf{R}^{S \times D}$ с помощью метода VLAD [59]. В этом методе для каждого локального дескриптора $\mathbf{x}_b$ определяется ближайшее визуальное слово $NN(\mathbf{x}_b)$, после чего для каждого *s*-го визуального слова накапливается разница $\mathbf{v}_s$ между ним и ассоциированными с ним локальными признаками. При этом вклад каждого локального дескриптора уравнивается:

$$\mathbf{v}_s = \sum_{\mathbf{x}_b : NN(\mathbf{x}_b) = \mathbf{vw}_s} \frac{\mathbf{vw}_s - \mathbf{x}_b}{\left\| \mathbf{vw}_s - \mathbf{x}_b \right\|}. \qquad (2.11)$$

После вычислений всех $\mathbf{v}_s$ они объединяются, формируя глобальный визуальный дескриптор $\mathbf{V}$, который нормализуется в два этапа. Вначале модифицируется каждый элемент дескриптора с помощью выражения (2.12), затем весь дескриптор нормализуются с помощью l_2-нормы:

$$V_e^{`} = \mathrm{sign}(V_e) \cdot |V_e|^{\gamma}, \qquad (2.12)$$

где γ – коэффициент нормализации, изменяющийся в пределах (0, 1].

В случае использования нескольких цветовых пространств, глобальные дескрипторы формируется отдельно для каждого пространства, после чего объединяются. Для снижения дальнейших вычислительных затрат, размерность расширенного дескриптора сокращается с помощью метода главных компонент до *Z* элементов.

При сравнении двух изображений с помощью их глобальных визуальных дескрипторов используется евклидово расстояние:

$$D_V\left(\mathbf{V}_i, \mathbf{V}_j\right) = \sqrt{\sum_{z=1}^{Z}\left(v_z^i - v_z^j\right)^2}. \qquad (2.13)$$

Таким образом, чем меньше значение $D_V(\mathbf{V}_i,\mathbf{V}_j)$, тем больше визуальное сходство между изображениями I_i и I_j.

2.2 Создание текстового дескриптора

На втором этапе обучения системы ААИ для каждого изображения I_m, принадлежащего обучающему набору $TS = \{(I_1, K_1), \ldots, (I_M, K_M)\}$, необходимо создать текстовый дескриптор $\mathbf{T}_m = \{t_1, \ldots, t_N\}$, длина которого равна размеру словаря ключевых слов. Для этого используется как текстовое описание изображения K_m, так и частота встречаемости каждого ключевого слова в обучающей выборке. Однако в связи с тем, что аннотации в обучающих выборках часто формируются вручную, то некоторые изображения могут быть проаннотированы не всеми релевантными ключевыми словами. Для решения этой проблемы предложен метод восстановления ключевых слов, определяющий количество пропущенных слов.

2.2.1 Формирование текстового дескриптора

Поскольку некоторые ключевые слова являются «общими» (встречаются в описании большого количества изображений разных категорий), то они являются менее полезными для описания изображений. В связи с этим, при формировании дескриптора они получают меньшее числовое значение, чем «характерные» ключевые слова (ключевые слова,

встречающиеся в описании небольшого количества изображений) (рис. 2.5). Для этого элементы текстового дескриптора вычисляются с помощью статистической меры TF-IDF [74]:

$$t_n^m = \frac{\delta(k_n \in K_m)}{|K_m|} \cdot \log\left(\frac{M}{F(k_n)}\right), \tag{2.14}$$

где $\delta(k_n \in K_m)$ обозначает наличие / отсутствие ключевого слова k_n в описании изображения I_m (принимает значения 1 и 0 соответственно); $|K_m|$ – количество ключевых слов в описании изображения I_m; M – размер обучающей выборки; $F(k_n)$ – частота встречаемости ключевого слова k_n в обучающей выборке.

man
mountain
peak
pole
snow
valley

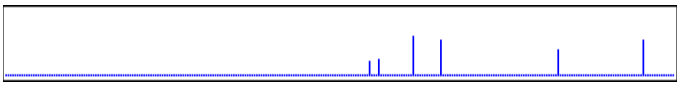

Рисунок 2.5. Иллюстрация текстового дескриптора, сформированного для изображения 2731 из базы IAPR TC-12

При сравнении двух изображений с помощью их текстовых дескрипторов используется косинусная метрика:

$$D_T(\mathbf{T}_i, \mathbf{T}_j) = \frac{\sum_{n=1}^{N} t_n^i \cdot t_n^j}{\sqrt{\sum_{n=1}^{N} \left(t_n^i\right)^2} \cdot \sqrt{\sum_{n=1}^{N} \left(t_n^j\right)^2}}. \tag{2.15}$$

Таким образом, чем ближе значение $D_T(\mathbf{T}_i,\mathbf{T}_j)$ к 1, тем больше сходство текстового описания изображений I_i и I_j.

2.2.2 Восстановление пропущенных ключевых слов

В полученных текстовых дескрипторах некоторые релевантные ключевые слова могут отсутствовать. Связано это с тем, что при составлении аннотаций обучающих изображений вручную часть «очевидных» ключевых слов часто пропускается. Например, на рисунке 2.6 можно заметить, что ключевое слово *landscape* подходит для всех четырех изображений, но встречается в аннотации только одного из них.

bush, coast, grey, sea, sky
39895

hill, house, mountain, sky, tree, village
39961

bay, gravel, house, landscape, meadow, road, shrub, sky
39896

bay, bush, coast, dirt, house, meadow, road, sea, sky
39897

Рисунок 2.6. Пример изображений из базы IAPR TC-12 и их аннотаций

Для восстановления таких пропущенных ключевых слов предлагается алгоритм [20], являющийся модификацией метода ААИ 2PKNN (п. 1.1.3). Блок-схема алгоритма приведена на рисунке 2.7.

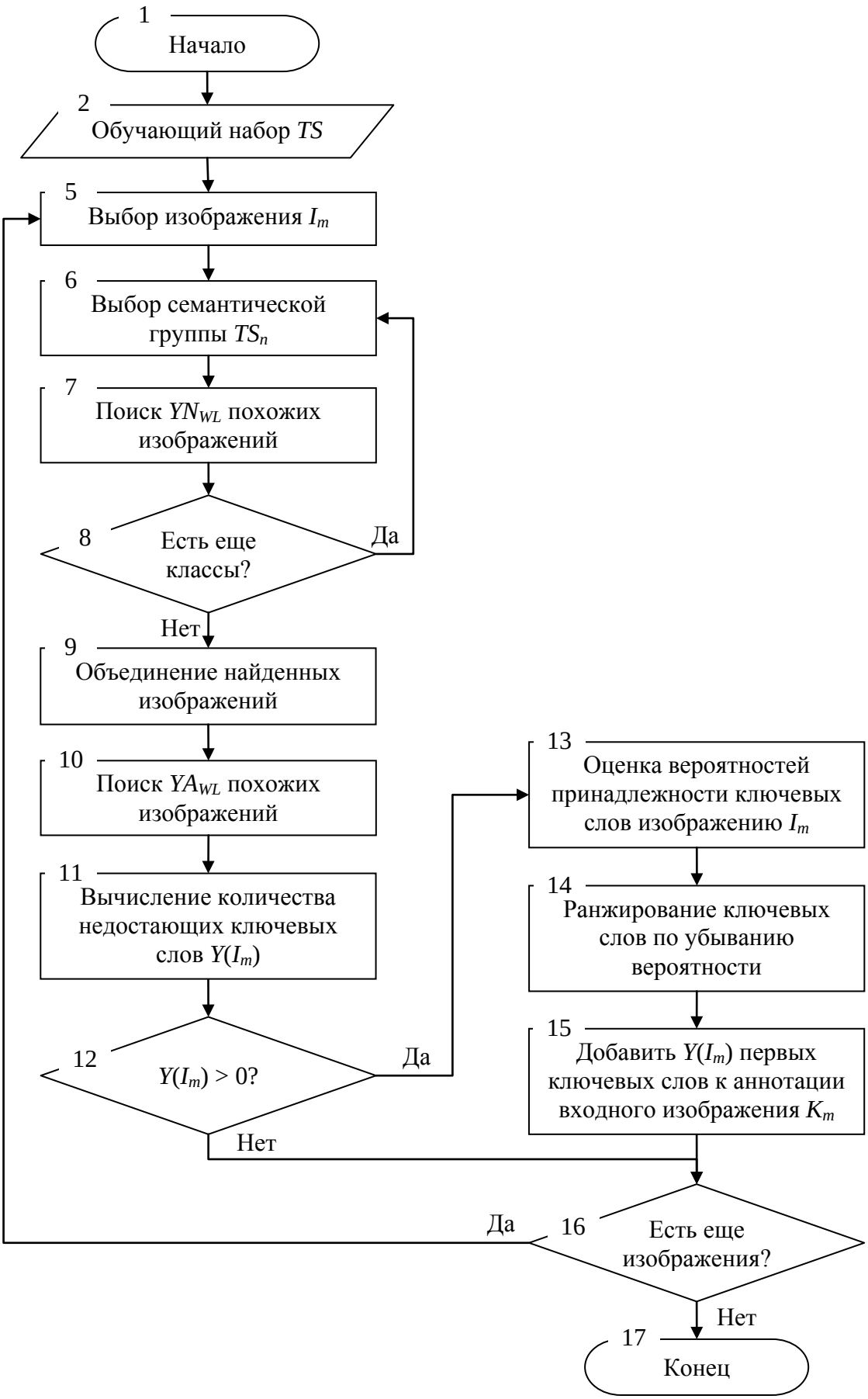

Рисунок 2.7. Блок-схема алгоритма восстановления пропущенных ключевых слов обучающих изображений

Пусть TS – набор обучающих изображений, в котором каждое изображение I_m имеет текстовый и визуальный дескрипторы. На первом этапе обучающий набор разделяется на несколько групп $TS_n \subseteq TS$, $n \in \{1, \ldots, N\}$, каждая из которых содержит все изображения с ключевым словом k_n. Поскольку изображения группы TS_n включают одно общее ключевое слово, будем называть такой набор семантической группой. Так как изображение обычно проаннотировано несколькими ключевыми словами, то оно может принадлежать нескольким семантическим группам.

Следующим этапом при восстановлении ключевых слов изображения I_m, из каждого семантической группы TS_n с помощью уравнения (2.13) выбирается YN_{WL} изображений с минимальным расстоянием (при этом само изображение I_m из наборов исключается). Таким образом, полученные семантические наборы $TS_n(I_m)$ содержат изображения, наиболее информативные в оценке принадлежности ключевого слова k_n изображению I_m. Объединив все семантические наборы в один $TS(I_m)$, определяется количество пропущенных ключевых слов. Для этого вычисляется разность $Y(I_m)$ между количеством ключевых слов изображения I_m и средним количеством ключевых слов в аннотациях YA_{WL} наиболее похожих изображений из набора $TS(I_m)$. Если $Y(I_m)$ положительно, то выполняется оценка вероятности принадлежности каждого ключевого слова k_n изображению I_m. При этом помимо визуального дескриптора используется текстовый, позволяющий исключить из оценки вероятности изображения, не имеющие общих ключевых слов с изображением I_m:

$$P(k_n \mid I_m) = \sum_{(I_i, K_i) \in TS(I_m)} \mathrm{D}_T(\mathbf{T}_m, \mathbf{T}_i) \cdot \exp(-\mathrm{D}_V(\mathbf{V}_m, \mathbf{V}_i)) \cdot \delta(k_n \in K_i). \qquad (2.16)$$

Аннотация изображения I_m пополняется $Y(I_m)$ ключевыми словами с наибольшими значениями вероятности $P(k_n|I_m)$ (рис. 2.8). В случае, если $Y(I_m)$ отрицательно, аннотация остается без изменений.

stage, view, pant, **spectator**
646

house, road, slope, gravel, **tree**
1214

palm, tree, front, **sky**
1384

building, street, entrance, dome, **night**
40396

Рисунок 2.8. Пример изображений из базы IAPR TC-12 с восстановленными ключевыми словами (выделены полужирным шрифтом)

2.3 Формирование однородных текстово-визуальных групп

После вычисления для каждого изображения I_m визуального $\mathbf{V}_m$ и текстового $\mathbf{T}_m$ дескрипторов, они объединяются в один текстово-визуальный $\mathbf{VT}_m$. С его помощью обучающий набор изображений предлагается разделять на однородные текстово-визуальные группы [20]. Идея заключается в том, что обучающие изображения одной ОТВ-группы формируют контекст для аннотируемого изображения, иными словами, если изображение отнесено к какой-либо группе, то оно аннотируется из ограниченного набора ключевых слов этой группы. Также предполагается, что тестовое изображение может принадлежать нескольким ОТВ-группам, однако их количество

ограничивается визуальным сходством. Это позволяет отсеять заведомо нерелевантные ключевые слова, не потеряв релевантных, а также снизить количество обучающих изображений, участвующих в аннотировании. Для этого каждая из ОТВ-групп должна соответствовать двум условиям:

1. Все изображения группы в своих аннотациях имеют общие «характерные» ключевые слова.

2. Изображения группы обладают существенным визуальным сходством.

Эта задача решается в два этапа:

1. Проводится первичное разделение изображений на группы на основе совместной встречаемости ключевых слов в описаниях изображений.

2. Изображения кластеризуются в автоматически определяемое количество ОТВ-групп, используя текстово-визуальные дескрипторы.

Далее рассмотрим более подробно каждый из представленных этапов.

2.3.1 Первичное разделение обучающих изображений

Для первичного разделения изображений необходимо построить взвешенный орграф $G = (K, E)$, где вершины являются ключевыми словами из словаря K. В этом случае дуга $e_{i,j}$ соединяет ключевые слова k_i и k_j, если одно или больше обучающих изображений одновременно проаннотировано ключевыми словами k_i и k_j. Вес этой дуги $w_{i,j}$ определяется с помощью следующей формулы:

$$w_{i,j} = \frac{Nimg(k_i, k_j)}{Nimg(k_i)}, \tag{2.17}$$

где $Nimg(k_i)$ – количество обучающих изображений проаннотированных ключевым словом k_i; $Nimg(k_i, k_j)$ – количество обучающих изображений, имеющих в описании ключевые слова k_i и k_j одновременно.

Полученный орграф разделяется на группы PH_f, $f \in \{1, \ldots, F\}$, с помощью алгоритма Louvain [29], использующего для проверки качества разделения понятие модульности, измеряемой как плотность связей внутри групп в сравнении с плотностью между группами. Таким образом, ключевые слова, часто встречающиеся совместно или имеющие похожие семантические значения, с большой вероятностью попадут в одну группу.

На следующем шаге для каждого обучающего изображения I_m определяется, к какой группе оно относится. Для этого суммируются соответствующие элементы текстового дескриптора:

$$PH_f(I_m) = \sum_{n:k_n \in PH_f} t_n^m. \tag{2.18}$$

Изображение присоединяется к той группе, для которой значение $PH_f(I_m)$ наибольшее. Подобное разделение обучающей выборки позволяет с минимальными затратами получить инициализацию ОТВ-групп, а также обеспечивает стабильный результат кластеризации.

2.3.2 Кластеризация обучающих изображений

Для последующей кластеризации обучающей выборки *TS* в ОТВ-группы используются текстово-визуальные дескрипторы изображений. При этом сходство между двумя изображениями I_i и I_j вычисляется по следующей формуле:

$$D_{VT}(\mathbf{VT}_i, \mathbf{VT}_j) = \alpha \cdot D_T(\mathbf{T}_i, \mathbf{T}_j) + (1-\alpha) \cdot \exp(-D_V(\mathbf{V}_i, \mathbf{V}_j)), \tag{2.19}$$

где α – эмпирический коэффициент, изменяющийся в пределах [0; 1].

Чем больше значение $D_{VT}(\mathbf{VT}_i, \mathbf{VT}_j)$, тем более похожи изображения I_i и I_j. Дескрипторы кластеризуются, используя модификацию расширенной самоорганизующейся инкрементальной нейронной сети (ESOINN) (п. 1.2.3), единственный слой которой постепенно подстраивается под структуру входных данных, определяя количество кластеров и их топологию. Модифицированный алгоритм ESOINN представлен на рисунке 2.9. Он включает следующие шаги [20]:

1. Структура нейронной сети инициализируется с помощью первичного разделения обучающей выборки (блок 3). Для этого из каждой сформированной группы PH_f случайным образом выбирается по два дескриптора, с помощью которых создаются узлы сети $r_i \in R$. Узлы, созданные из дескрипторов, принадлежавших одной группе, соединяются связями.

2. На вход сети подается новый текстово-визуальный дескриптор $\mathbf{VT}_m$ (блок 4).

3. Для входного дескриптора с помощью формулы (2.19) определяются два ближайших узла сети (победитель и второй победитель) (блок 5). Если сходство между входным дескриптором и победителем или вторым победителем меньше соответствующих порогов подобия, то входной дескриптор вставляется в сеть как первый узел нового класса, а алгоритм переходит к шагу 2 для получения нового дескриптора (блоки 6-7).

Поскольку распределение входных данных заранее неизвестно, то порог подобия s_i обновляется для каждого узла в отдельности, используя следующую формулу:

$$s_i = \min_{j \in R_i} \mathrm{D}_{VT}\left(\mathbf{W}_i, \mathbf{W}_j\right), \tag{2.20}$$

где R_i – набор узлов , соединенных с узлом r_i; $\mathbf{W}_i$ – вектор весов узла r_i.

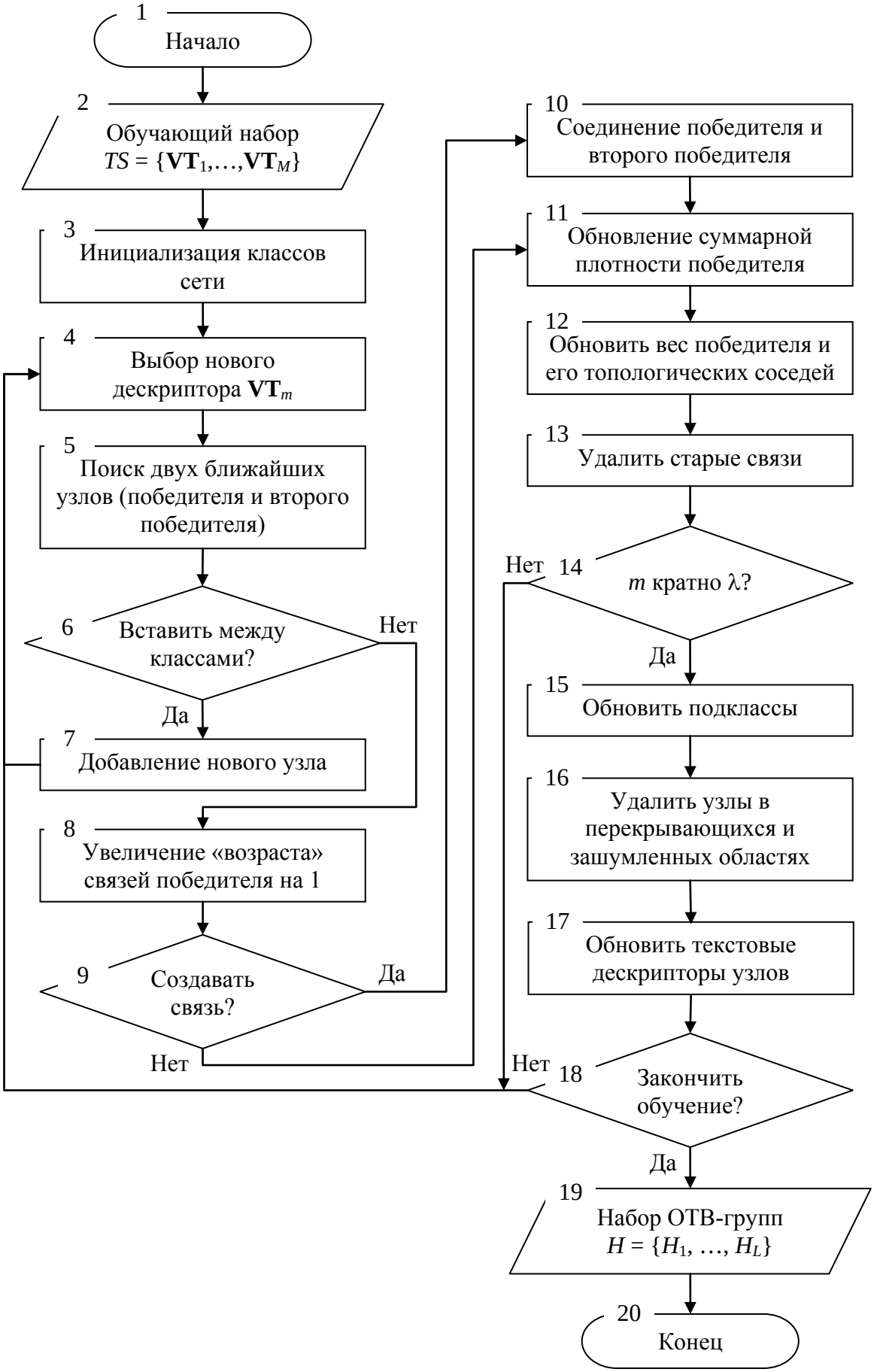

Рисунок 2.9. Блок-схема кластеризации обучающей выборки в ОТВ-группы

В случае если узел не имеет соседей, то порог подобия вычисляется с помощью всех узлов сети:

$$s_i = \max_{j \in R \setminus \{i\}} \mathrm{D}_{VT}\left(\mathbf{W}_i, \mathbf{W}_j\right). \tag{2.21}$$

4. «Возраст» (числовой коэффициент, при создании новой связи равный 0) всех связей победителя увеличивается на 1 (блок 8), после чего решается вопрос о необходимости создания новой связи между победителем и вторым победителем (блоки 9-10).

5. Обновляется суммарная плотность победителя (блок 11). Плотность узла p_{win} в m-й цикл работы сети вычисляется с помощью среднего расстояния $\overline{d}_{win}$ от узла до его соседей:

$$p_{win,m} = \frac{1}{\left(1 + \overline{d}_{win,m}\right)^2}. \tag{2.22}$$

Если среднее расстояние от узла до его соседей большое, то количество узлов в этой области небольшое и плотность будет низкой, и наоборот. В течение одной итерации вычисляется плотность только для победителя. Суммарная плотность узла h_i определяется следующим образом:

$$h_{win} = \frac{1}{q} \cdot \sum_{dp=1}^{Q} p_{win,dp}, \tag{2.23}$$

где q – количество периодов, в которые плотность узла r_{win} больше 0; Q – количество прошедших циклов обучения сети.

6. Счетчик количества побед U_{win} узла-победителя r_{win} увеличивается на 1, а вектора весов победителя и его узлов-соседей обновляются с помощью входного дескриптора следующим образом (блок 12):

$$\Delta \mathbf{W}_{win} = \frac{1}{U_{win}} \cdot \left(\mathbf{VT} - \mathbf{W}_{win} \right), \tag{2.24}$$

$$\Delta \mathbf{W}_j = \frac{1}{100 \cdot U_{win}} \cdot \left(\mathbf{VT} - \mathbf{W}_j \right), \; j \in R_{win}. \tag{2.25}$$

7. Удаляются все связи, «возраст» которых превышает заранее установленное значение age_{max} (блок 13).

8. Если период обучения сети закончен (количество входных дескрипторов кратно периоду сети λ) (блок 14), то существующие кластеры разбиваются на подклассы с целью обнаружения перекрывающихся областей (блок 15), после чего из нейронной сети удаляются узлы, являющиеся шумами (блок 16). Такими считаются узлы r_i, имеющие двух или меньше топологических соседей и удовлетворяющих условию следующего вида:

$$h_i < b_o \cdot \sum_{j=1}^{R} \frac{h_j}{|R|}, \tag{2.26}$$

где $b_{o,}$ o = {1, 2, 3} – эмпирические коэффициенты, используемые при удалении узлов с двумя топологическими соседями, одним соседом и не имеющих соседей соответственно.

9. Обновляются текстовые дескрипторы узлов сети. Для каждого узла вычисляются центры масс текстовых частей последних $\lambda / 10$ входных дескрипторов, ассоциированных с этим узлом.

10. Если процесс кластеризации закончен (на вход сети поданы все дескрипторы) (блок 17), то полученные узлы классифицируются по принадлежности к тому или иному кластеру, используя понятие пути между двумя узлами (узлы r_i и r_j связаны путем, если между ними существует непрерывная цепочка связей).

11. Если ESOINN продолжает работу, то переходим к шагу 2 для получения нового входного дескриптора.

После окончательного формирования структуры нейронной сети необходимо ассоциировать обучающие изображения с полученными кластерами, являющимися базисом ОТВ-групп.

coast, formation, group, photo, rock, sea
7848

coast, group, picture, rock, sea, tourist
8329

bridge, coast, rock, sea, tourist
10898

coast, sea, team, tourist
10995

bridge, coast, hill, rock, sea, tourist
12962

beach, blanket, coast, rock, sea, tourist
13672

coast, rock, sea, sky, tourist
14173

coast, formation, group, people, rock, sea
21827

bridge, coast, people, rock, sea, sky
23272

Рисунок 2.10. Пример изображений однородной текстово-визуальной группы, сформированной из базы IAPR TC-12

Вначале для каждого изображения определяется ближайший узел сети, используя только текстовый дескриптор, после чего этот же процесс

повторяется с использованием только визуального дескриптора. В случае, когда изображение по текстовому и визуальному дескрипторам ассоциировано с разными кластерами, изображение считается шумовым и исключается из обучающей выборки. Это необходимый шаг, поскольку при выполнении алгоритма аннотирования ассоциация тестового изображения *A* происходит только с помощью визуальных дескрипторов. Пример изображений сформированной ОТВ-группы представлен на рисунке 2.10.

Следует отметить, что предложенный алгоритм позволяет уточнять ОТВ-группы с помощью новых обучающих изображений в течение жизненного цикла системы ААИ, позволяя таким образом избежать полного переобучения системы и лишних вычислительных затрат.

2.4 Автоматическое аннотирование изображений

Используя сформированные ОТВ-группы, можно проаннотировать тестовое изображение *A*. Для этого необходимо вычислить априорные вероятности $P(H_l)$ и $P_l(k_n)$, а также оценить вероятности $P(A|H_l)$, $P_l(A|k_n)$. Подставив их в выражение (2.2), производится оценка вероятностей принадлежности ключевых слов изображению *A*.

В предлагаемом алгоритме ААИ вначале вычисляется априорная вероятность ОТВ-группы $P(H_l)$ как отношение количества обучающих изображений в ОТВ-группе к общему количеству обучающих изображений [20]:

$$P(H_l) = \frac{|H_l|}{\sum_{i=1}^{L} |H_i|}. \qquad (2.27)$$

Поскольку базы изображений часто несбалансированны (частота встречаемости разных ключевых слов сильно разнится), то при оценке

априорной вероятности $P_l(k_n)$ используется взвешивание, позволяющее увеличить вероятность для редких ключевых слов и уменьшить для частых:

$$P_l(k_n) = \exp\left(-\frac{Nimg(k_n)}{M}\right). \quad (2.28)$$

Стоит отметить, что апостериорные вероятности могут быть вычислены на этапе обучения. В этом случае при аннотировании используются ранее сохраненные значения. Процесс оценки вероятностей $P(A|H_l)$ с помощью визуального дескриптора тестового изображения V_A включает следующие шаги:

1. С помощью уравнения (2.13) выбрать из всей обучающей выборки YN_H = 200 наиболее похожих изображений, каждое из которых ассоциировано с определенной ОТВ-группой.

2. Для каждой ОТВ-группы, представленной в полученном наборе изображений, оценить условную вероятность $P(A|H_l)$ с помощью формулы:

$$P(A \mid H_l) = \exp(-\mathrm{D}_V(V_A, HV_l)), \quad (2.29)$$

где HV_l – визуальный дескриптор ближайшего изображения ОТВ-группы H_l.

3. Для всех оставшихся ОТВ-групп принять условную вероятность $P(A|H_l)$ равной 0.

4. Условные вероятности $P(A|H_l)$ нормализовать таким образом, чтобы их сумма равнялась 1.

На следующем шаге аннотирования используются только ОТВ-группы, для которых условная вероятность $P(A|H_l) > 0$. В каждой такой группе для тестового изображения A с помощью уравнения (2.13) выбирается YN_{AN} = 5 изображений с минимальным расстоянием, формирующих набор $H_l(A)$. Используя его, вероятность $P_l(A|k_n)$ оценивается следующим образом:

$$P_l(A \mid k_n) = \sum_{(V_i, K_i) \in H_l(A)} \exp\left(-\mathrm{D}_V(V_A, V_i)\right) \cdot \delta(k_n \in K_i). \quad (2.30)$$

Подставив полученные оценки условных вероятностей в уравнение (2.2) и отсортировав результаты по убыванию, формируется ранжированный список ключевых слов. В качестве аннотации нового изображения выбирается *Nkw* первых ключевых слов. Значение *Nkw* выбирается равным среднему количеству ключевых слов в описании обучающих изображений, выбранных из ОТВ-группы с наибольшей условной вероятностью $P(A|H_l)$.

Следует отметить, что эффективность предложенного метода повышается, если вместе с тестовым изображением будут предоставлены некоторые ключевые слова, полученные от пользователей или с помощью узкоспециализированных классификаторов. В этом случае при оценке вероятностей используется текстово-визуальный дескриптор.

2.5 Выводы по главе

Во второй главе приведена математическая модель автоматического аннотирования изображений на основе обучающего набора изображений, разделенного на однородные текстово-визуальные группы, а также предложен алгоритм для реализации данной модели, отличающийся тем, что аннотирование нового изображения осуществляется с помощью обучающих изображений небольшого количества визуально похожих групп. Данный алгоритм включает три этапа обучения и этап аннотирования.

На первом этапе для всех обучающих, а также аннотируемых изображений формируется глобальный визуальный дескриптор. Для этого из изображения извлекается набор локальных дескрипторов, который кодируется с помощью словаря визуальных слов. Поскольку данный этап является наиболее вычислительно затратным, то предложен быстрый метод извлечения набора локальных дескрипторов, позволяющий избежать

повторных вычислений при наложении областей расчета дескрипторов и, таким образом, существенно ускоряющий процесс аннотирования. Также рассмотрен процесс вычисления цветовых локальных дескрипторов, использование которых позволяет повысить точность аннотирования, и приведены алгоритмы формирования словаря визуальных слов и кодирования набора локальных дескрипторов в глобальный визуальный дескриптор.

На втором этапе для всех обучающих изображений на основе текстового описания формируется текстовый дескриптор с помощью статистической меры TF-IDF. При этом изображения могут быть проаннотированы не всеми релевантными ключевыми словами, поскольку обучающие выборки формируются на основе аннотаций, предоставленных человеком, в связи с чем часть «очевидных» ключевых слов может отсутствовать. Для восстановления таких ключевых слов предложен метод, отличающийся автоматическим определением количества пропущенных ключевых слов и использующий визуальные и текстовые дескрипторы. В главе также представлен алгоритм, реализующий данный метод.

На третьем этапе обучающий набор изображений разделяется на однородные текстово-визуальные группы. Для этого разработан новый метод кластеризации изображений, включающий предварительный этап разделения набора на основе совместной встречаемости ключевых слов в аннотациях изображений, а также этап кластеризации изображений с помощью модифицированной самоорганизующейся нейронной сети на основе текстовых и визуальных дескрипторов. Метод позволяет формировать однородные текстово-визуальные группы, которые представляют собой контекст для аннотирования новых изображений и уточняются в течение жизненного цикла системы.

Используя полученные ОТВ-группы, новое изображение аннотируется с помощью вычисления априорных вероятностей ОТВ-групп и ключевых слов, а также оценки вероятности нахождения изображения в каждой ОТВ-

группе и оценки распределения визуального дескриптора изображения для каждого ключевого слова в ОТВ-группах. На основе полученных значений, производится оценка вероятности принадлежности ключевых слов аннотируемому изображению, после чего выбираются ключевые слова с наибольшими значениями вероятности.

ГЛАВА 3. ПОСТРОЕНИЕ ЭКСПЕРИМЕНТАЛЬНОЙ СИСТЕМЫ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ИЗОБРАЖЕНИЙ И ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

В данной главе приводится практическая апробация разработанного метода автоматического аннотирования изображений. Рассматривается структурная схема экспериментального программного обеспечения и функциональное назначение его модулей. Приводится методика и результаты тестирования отдельно для модуля вычисления глобальных визуальных дескрипторов и всей системы автоматического аннотирования. Анализируются экспериментальные результаты, полученные при помощи разработанной системы, путем сравнения с результатами, представленными в научных статьях ученых, занимающихся проблемой ААИ.

3.1 Структурная схема и описание модулей системы автоматического аннотирования изображений

Разработанные методы и алгоритмы легли в основу экспериментальной системы автоматического аннотирования изображений. Система представляет собой модульное приложение, программные модули которого могут быть использованы как в совокупности, так и по отдельности для решения более узких задач, например, кластеризации данных с возможностью итеративного уточнения кластеров, вычисления визуальных дескрипторов для категоризации изображений и распознавания образов. Наименования разработанных модулей и их функциональные характеристики приведены в таблице 3.1, а структурная схема изображена на рисунке 3.1.

Таблица 3.1

Разработанные программные модули и их назначение

Название модуля	**Функциональная характеристика**
1. Модуль преобразований изображения	Осуществляет преобразования изображений в различные цветовые пространства и вычисление интегральных изображений
2. Модуль вычисления визуальных дескрипторов	Осуществляет вычисление набора локальных дескрипторов, формирование словаря визуальных слов и кодирование глобального визуального дескриптора
3. Модуль формирования текстового дескриптора	Осуществляет вычисление частот встречаемости ключевых слов и формирование текстового дескриптора
4. Модуль восстановления ключевых слов	Модуль формирует семантические группы и расширяет аннотации обучающих изображений путем восстановления пропущенных ключевых слов
5. Модуль формирования ОТВ-групп	Осуществляет первичное разделение обучающих изображений и кластеризацию текстово-визуальных дескрипторов
6. Ядро системы	Реализует автоматическое аннотирование изображений, используя другие модули системы
7. Пользовательский интерфейс	Осуществляет взаимодействие пользователя с ядром системы

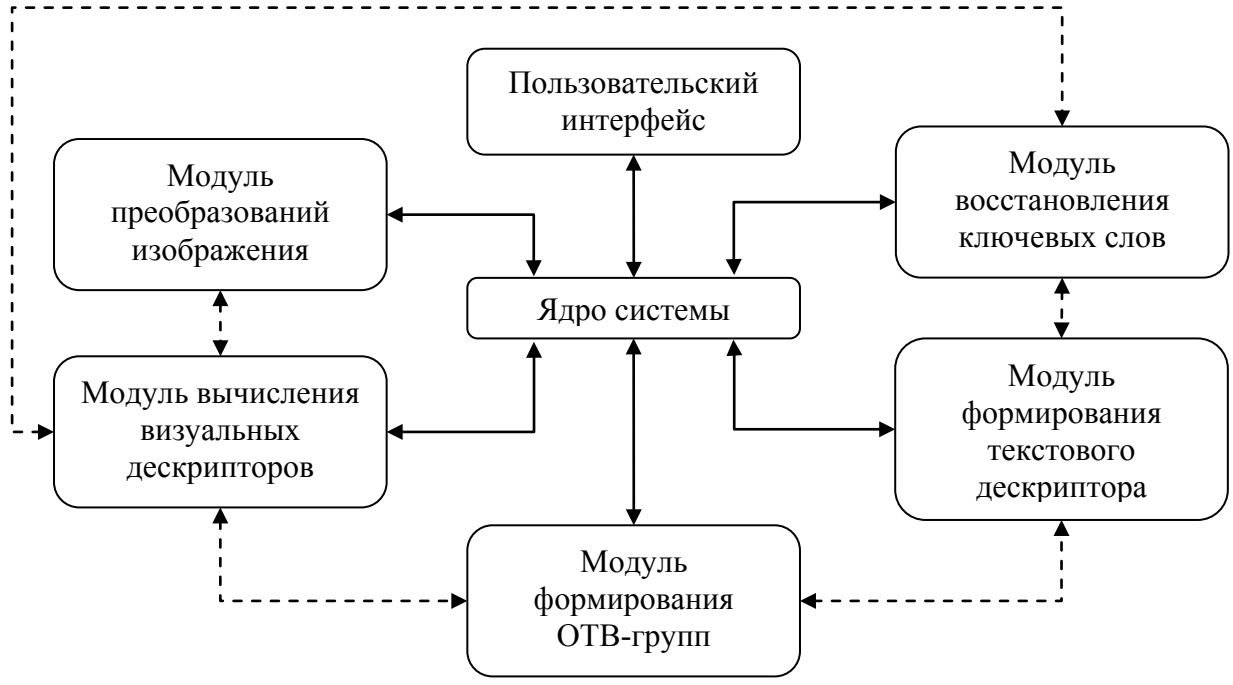

Рисунок 3.1. Структурная схема экспериментального программного комплекса

Далее приведем более подробное описание функционального назначения разработанных модулей системы.

Модуль преобразований изображения

Модуль обеспечивает преобразование входных изображений в необходимые цветовые пространства и каналы (RGB, nRGB, rg, Opponent, HSI, LUV, HSV, YUV). В каждом канале допустимое изменение значений приводится к диапазону [0; 1], после чего вычисляются интегральные изображения, используемые при вычислении локальных дескрипторов.

Модуль вычисления визуальных дескрипторов

Упрощенная блок-схема алгоритмов модуля вычисления визуальных дескрипторов представлена на рисунке 3.2.

На этапе обучения все изображения обучающей коллекции преобразуются в заданное цветовое пространство (блоки 2–3). После этого из всей коллекции случайным образом выбирается 200 000 локальных дескрипторов, которые используются для нахождения главных компонент, а также формирования словаря визуальных слов (блоки 4–6). Затем из обучающей коллекции случайным образом выбирается 2048 изображений, для которых с помощью визуальных слов вычисляются глобальные дескрипторы (блок 7). Полученные дескрипторы используются для нахождения главных компонент (блок 8). В случае, когда задано несколько цветовых пространств, блоки 2–6 выполняются для каждого цветового пространства отдельно.

Полученные визуальные слова и матрицы перевода в пространства главных компонент используются для вычислений глобальных визуальных дескрипторов для всей коллекции обучающих изображений, а также аннотируемых изображений. Для этого каждое изображение также переводится в заданное цветовое пространство (блок 13), после чего из изображения извлекается набор локальных дескрипторов, размерность каждого из которых сокращается (блоки 14–15). Используя полученные локальные дескрипторы, формируется глобальный дескриптор, размерность которого также сокращается (блоки 16–17). В случае, когда задано несколько

цветовых пространств, блоки 13–15 выполняются для каждого цветового пространства отдельно.

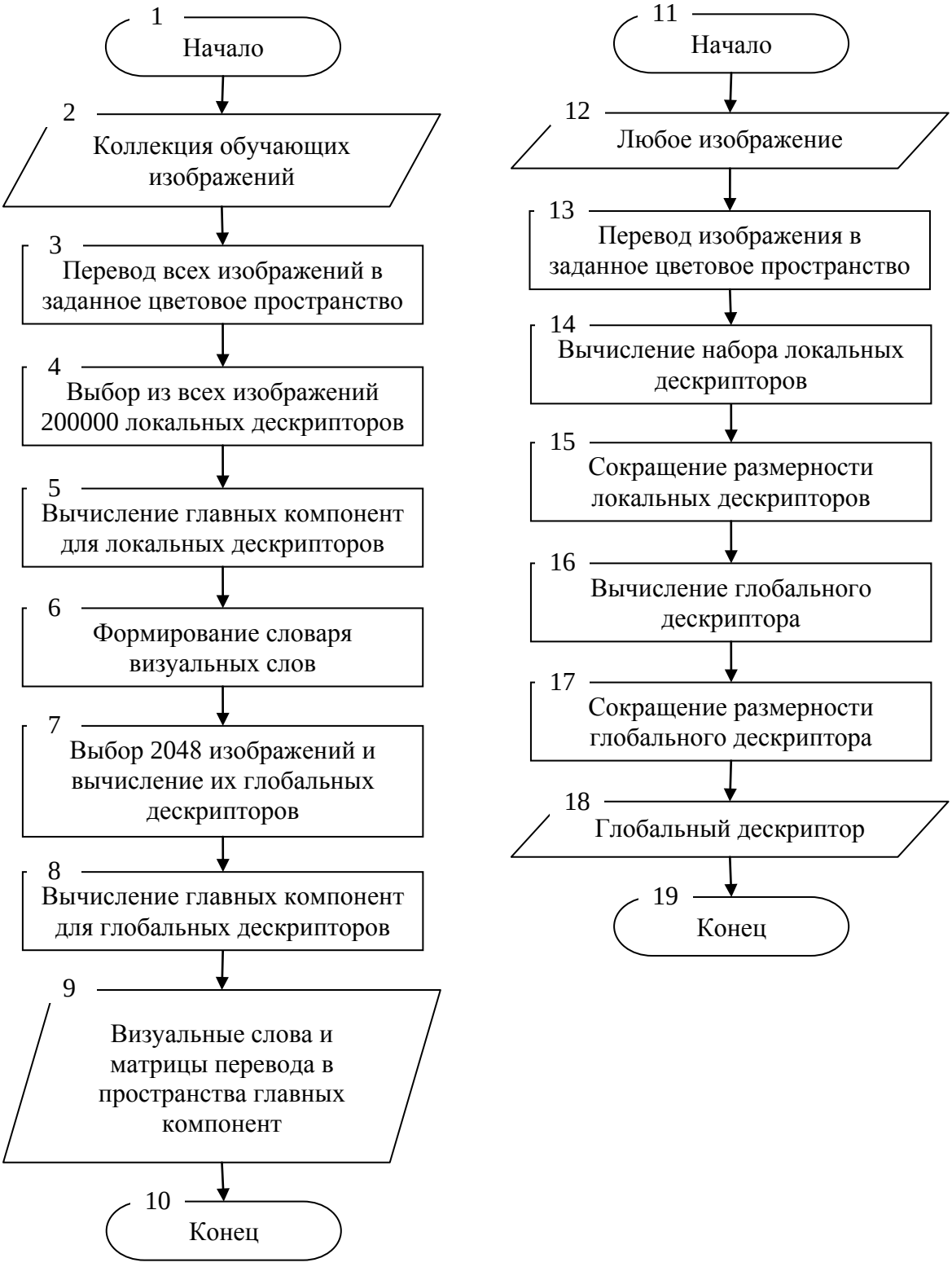

Рисунок 3.2. Упрощенная блок-схема алгоритмов модуля вычисления визуальных дескрипторов

Функции модуля вычисления визуальных дескрипторов реализованы в программном продукте «Система автоматического формирования визуальных слов (ForVW)», которая позволяет осуществлять вычисление набора локальных дескрипторов, формирование словаря визуальных слов и кодирование глобального визуального дескриптора (Приложение 1) [16]. Возможности настройки модуля приведены на рисунке 3.3.

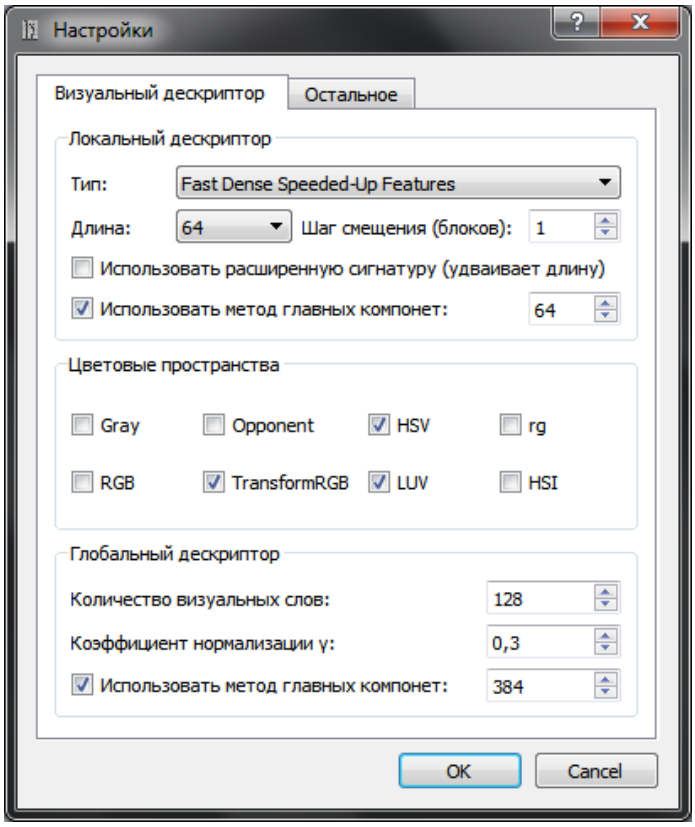

Рисунок 3.3. Экранная форма настройки модуля вычисления визуальных дескрипторов

Модуль формирования текстового дескриптора

В модуле реализованы функции для вычисления частот встречаемости ключевых слов в коллекции обучающих изображений, а также формирования текстовых дескрипторов с помощью статистической меры TF-IDF.

Модуль восстановления ключевых слов

В модуле реализованы функции для восстановления ключевых слов обучающих изображений. Исходными данными модуля является обучающий набор, в котором каждое изображение уже имеет визуальный и текстовый

дескрипторы. Все изображения распределяются по семантическим группам на основе их текстового описания, после чего изображения выбираются последовательно. Для каждого выбранного изображения (будем называть его исходным) в каждой семантической группе определяется по 2 визуально похожих обучающих изображения. Эти изображения объединяются в набор, с помощью которого оценивается количество пропущенных ключевых слов. В случае если полученное значение недостающих ключевых слов положительно, то набор используется для вычисления вероятностей принадлежности всех ключевых слов исходному изображению. После этого аннотация исходного изображения пополняется ключевыми словами с наибольшими вероятностями, и алгоритм выбирает новое изображение.

Модуль формирования ОТВ-групп

В этом модуле реализован разработанный двухэтапный алгоритм формирования ОТВ-групп. Упрощенная блок-схема алгоритма работы модуля представлена на рисунке 3.4.

Исходными данными модуля является набор изображений, описанных визуальными и текстовыми дескрипторами (блок 2). Для каждой пары ключевых слов вычисляются относительные значения совместной встречаемости, на основе которых строится орграф (блок 3). Этот орграф разделяется таким образом, чтобы ключевые слова, часто встречающиеся совместно или имеющие похожие семантические значения, попали в одну группу (блок 4). С полученными группами ключевых слов ассоциируются обучающие изображения, формируя первичные группы изображений (блок 5). Из каждой первичной группы случайным образом выбираются по 2 изображения, используемых для инициализации модифицированной нейронной сети ESOINN (блок 6). Последовательно выбирая текстово-визуальные дескрипторы изображений, сеть определяет базис ОТВ-групп, с которым ассоциируются обучающие изображения (блоки 7–8).

Рисунок 3.4. Упрощенная блок-схема алгоритма модуля формирования ОТВ-групп

Ядро системы

Модуль ядра системы отвечает за взаимодействие других модулей программного продукта между собой, а также реализует разработанный алгоритм автоматическое аннотирование изображений. Упрощенная блок-схема алгоритмов модуля представлена на рисунке 3.5.

Рисунок 3.5. Упрощенная блок-схема алгоритмов модуля ядра системы

На этапе обучения для всех изображений обучающей коллекции вычисляются глобальные визуальные дескрипторы, а также формируются текстовые дескрипторы на основе текстовых описаний (блоки 2–4). Поскольку некоторые релевантные ключевые слова могут отсутствовать, то дополнительно проводится восстановление ключевых слов обучающих изображений, после чего текстовые дескрипторы обновляются (блок 5). С

использованием полученных визуальных и текстовых дескрипторов формируются однородные текстово-визуальные группы (блок 6), для которых вычисляются априорные вероятности (блок 7).

При аннотировании нового изображения для него также вычисляется визуальный дескриптор (блоки 12–13), с помощью которого оцениваются вероятности принадлежности изображения ОТВ-группам и семантическим группам ключевых слов (блоки 14–15). Используя полученные значения условных вероятностей, а также вычисленные на этапе обучения априорные вероятности, осуществляется оценка принадлежности ключевых слов аннотируемому изображению. В качестве аннотации выбираются ключевые слова с наибольшими значениями вероятности (блок 16).

Функции модулей ядра системы и формирования текстовых дескрипторов и ОТВ-групп реализованы в программном продукте «Система автоматического аннотирования изображений (AIA)» (Приложение 2) [21]. Возможности настройки модулей приведены на рисунке 3.6.

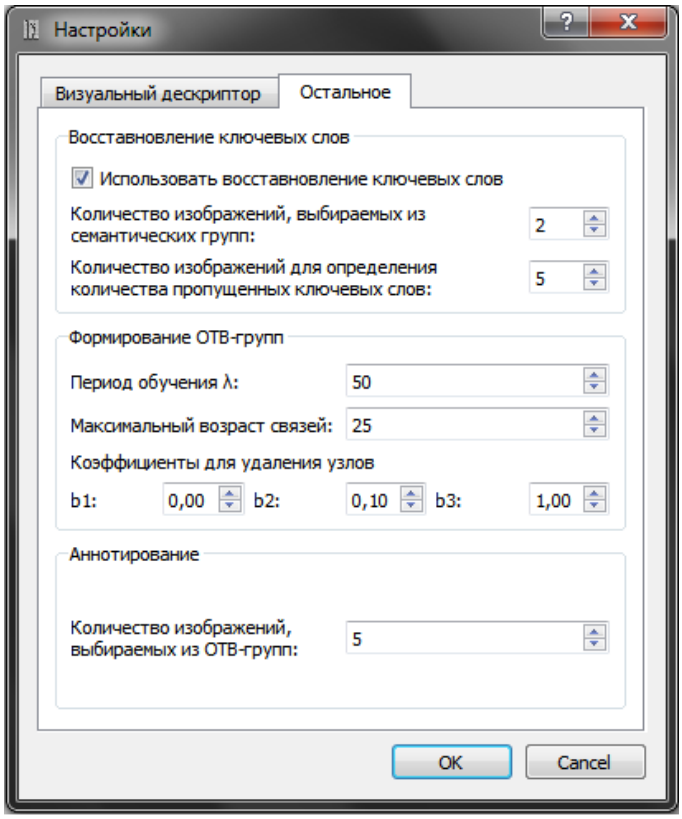

Рисунок 3.6. Экранная форма настройки модулей восстановления ключевых слов, формирования ОТВ-групп и ядра системы

Пользовательский интерфейс

Модуль пользовательского интерфейса позволяет загружать в программу обучающие и тестовые базы изображений, а также осуществлять настройку параметров реализованных алгоритмов и оценивать полученные результаты. Структура таблиц баз изображений описана в таблице 3.2.

Таблица 3.2

Структура таблиц базы данных аннотированных изображений

Наименование поля	Тип данных	Описание
Таблица *images* содержит обучающие, тестовые или проаннотированные изображения		
ID	Счетчик	Ключевое поле. Однозначно идентифицирует изображение
Name	Текстовый	Название файла изображения
Таблица *keywords* содержит список ключевых слов		
ID	Счетчик	Ключевое поле. Однозначно идентифицирует ключевое слово
Keyword	Текстовый	Название ключевого слова
Таблица *linking* содержит связи между изображениями и ключевыми словами		
ID	Счетчик	Ключевое поле. Однозначно идентифицирует связь ключевого слова с изображением
Image_ID	Числовой	Идентификатор изображения
Keyword_ID	Числовой	Идентификатор ключевого слова

3.2 Результаты экспериментальных исследований вычисления визуальных дескрипторов

Для проверки предложенных методов и алгоритмов вычисления визуальных дескрипторов использовался набор изображений OT8 [78], включающий 8 категорий сцен (рис. 3.7). OT8 состоит из 2688 изображений, размер каждого изображения 256 × 256 пикселов. Тестирование заключалось в исследовании влияния размера словаря визуальных слов и значения коэффициента нормализации γ на точность категоризации изображений. Дополнительно проведены исследования для определения оптимальной длины глобального дескриптора Z, выбора наиболее информативных цветовых пространств и количества потоков при параллельном вычислении локальных дескрипторов. Для вычисления точности категоризации отдельной категории *ctg* использовалась следующая формула:

$$accuracy_{ctg} = \frac{CC_{ctg}}{CT_{ctg}}, \qquad (3.1)$$

где CC_{ctg} – количество тестовых изображений, правильно отнесенных к категории *ctg*; CT_{ctg} – количество тестовых изображений категории *ctg*.

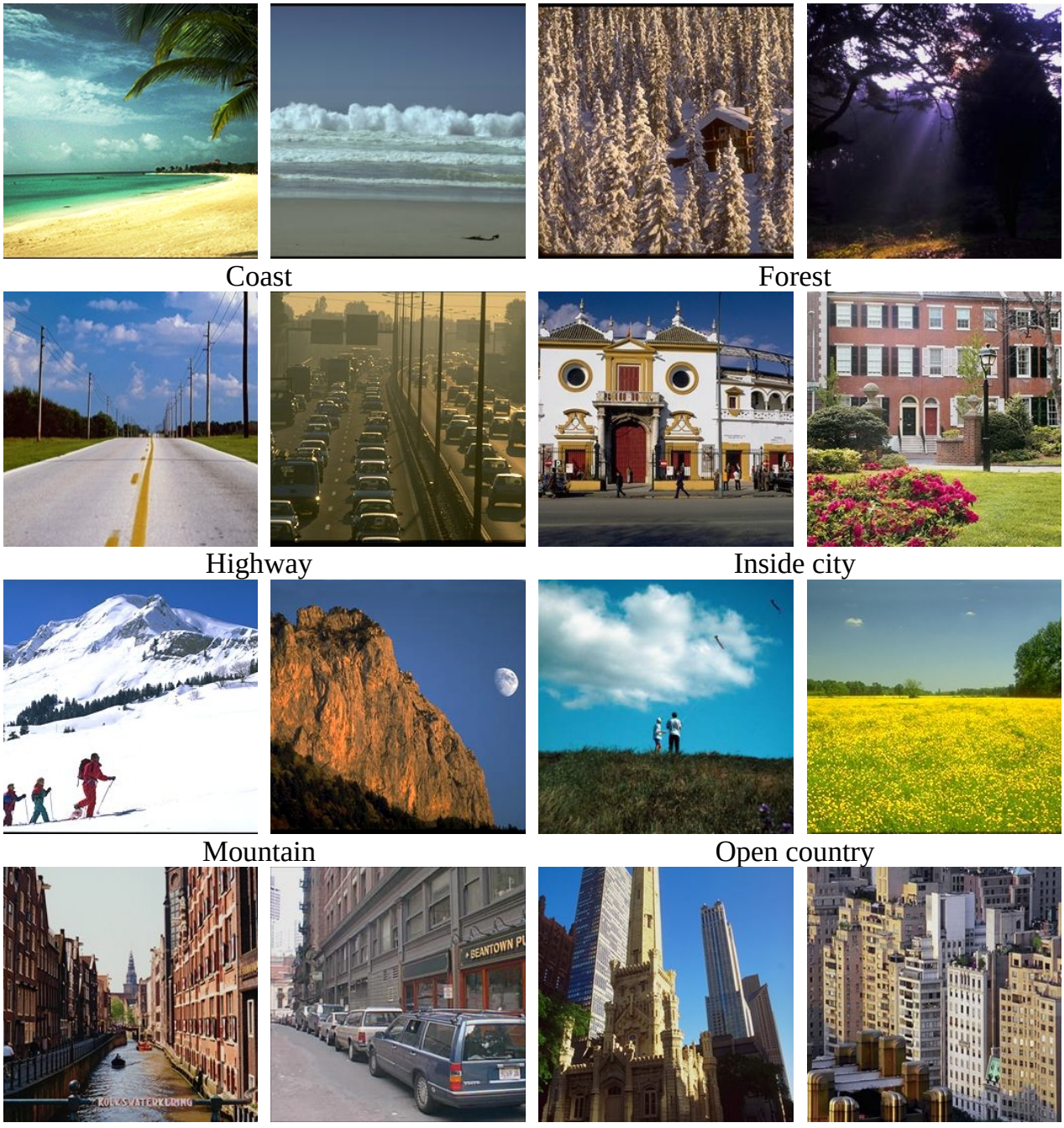

Рисунок 3.7. Примеры изображений каждой категории набора OT8

В проведенных исследованиях в качестве классификатора использовалась машина опорных векторов, для обучения которой из каждой категории случайным образом выбиралось по 100 изображений, а остальные использовались для тестирования. Также ряд параметров оставался неизменным для всех экспериментов:

- При вычислении FD-SUF и других локальных дескрипторов на точках интереса, полученных с помощью регулярной сетки, масштаб $\sigma_{sc} = 1$, а сдвиг точек интереса составлял 5 пикселов (1 блок).
- Для формирования словаря визуальных слов любой длины из обучающей выборки случайным образом выбиралось 200 000 локальных дескрипторов.
- Для экспериментов использовался компьютер с процессором Intel Core i5-2430M 2,4 ГГц и оперативной памятью Kingston 1333 МГц DDR3 8 Гб. Все расчеты повторялись 5 раз, после чего результаты усреднялись.

3.2.1 Сравнение с существующими локальными дескрипторами

В первом эксперименте проводилось сравнение локальных дескрипторов FD-SUF, SURF и G-SURF. Для последних двух вычисления осуществлялись на точках интереса, полученных с помощью регулярной сетки и детектора «быстрый Гессиан». Размер словаря визуальных слов S выбран равным 128. При формировании глобального дескриптора коэффициент $\gamma = 1$ (без нормализации элементов), длина дескриптора также не сокращалась. Результаты сравнения приведены в таблице 3.3. Вычисления производились с использованием одного процессорного ядра.

Как видно из приведенных данных, вычисление локальных дескрипторов на точках интереса, полученных с помощью регулярной сетки, существенно повышает точность категоризации. При этом точность категоризации для дескрипторов SURF и G-SURF различается для разных категорий, но в среднем одинакова. Также предложенный алгоритм

вычисления набора локальных дескрипторов позволяет повысить точность категоризации изображений в среднем на 2 % за счет исключения взвешивания элементов дескриптора с помощью фильтра Гаусса. При этом вычисления осуществляются в 3,3 и 5,6 раз быстрее, чем дескрипторы SURF и G-SURF, вычисляющих каждый локальный дескриптор отдельно.

Таблица 3.3

Основные показатели категоризации изображений при использовании различных локальных дескрипторов

	SURF (Гессиан)	G-SURF (Гессиан)	SURF (Сетка)	G-SURF (Сетка)	FD-SUF
Точность категоризации, %					
Coast	69,36	58,92	84,62	78,85	**85,51**
Forest	86,11	93,67	**94,30**	94,01	92,84
Highway	67,08	75,13	76,25	80,42	**82,08**
Inside city	78,52	80,85	90,06	87,02	**92,15**
Mountain	77,86	68,42	81,39	80,78	**82,36**
Open country	49,03	50,39	64,73	**70,43**	69,03
Street	69,96	70,23	82,12	80,38	**86,11**
Tall building	76,04	50,83	**87,50**	86,98	87,11
Средняя	71,75	68,55	82,62	82,36	**84,65**
Количество вычисленных дескрипторов					
Среднее	207	207	2304	2304	2304
Время вычисления дескрипторов, мс					
Среднее	12,35	15,78	57,26	96,35	17,18

3.2.2 Исследование параметров алгоритма формирования глобальных дескрипторов

Во втором эксперименте определялся оптимальный размер словаря визуальных слов при использовании локальных дескрипторов FD-SUF. Так же, как и в предыдущем эксперименте, коэффициент нормализации γ установлен равным 1. Результаты вычислений для отдельных категорий приведены в таблице 3.4, а график для средней точности категоризации представлен на рисунке 3.8.

Результаты эксперимента показали, что при увеличении размера словаря визуальных слов выше 128 элементов, точность категоризации изображений остается практически без изменений. Это связано с тем, что при

большом количестве визуальных слов некоторые из них могут располагаться достаточно близко в пространстве признаков, что приводит к фактическому дублированию отдельных участков глобального дескриптора без внесения новой информации. В связи с этим, во всех дальнейших экспериментах используется 128 визуальных слов.

Таблица 3.4

Точность категоризации изображений (%) в зависимости от размера словаря визуальных слов

Категории	Размер словаря визуальных слов						
	8	16	32	64	128	192	256
Coast	79,36	82,44	84,23	84,87	85,51	86,54	85,90
Forest	92,40	92,98	94,88	93,86	92,84	92,54	93,27
Highway	73,75	78,96	84,38	81,46	82,08	82,71	82,50
Inside city	89,10	87,34	89,42	93,11	92,15	92,63	91,99
Mountain	75,06	77,25	78,35	79,44	82,36	83,46	84,79
Open country	57,31	65,91	70,32	68,17	69,03	67,63	66,88
Street	71,88	75,35	80,56	84,20	86,11	86,98	87,67
Tall building	75,00	83,46	85,81	85,94	87,11	86,33	85,81
Средняя точность	76,73	80,46	83,49	83,88	84,65	84,85	84,85

Рисунок 3.8. График зависимости средней точности категоризации изображений от размера словаря визуальных слов ($\gamma = 1$)

В третьем эксперименте исследовалась зависимость средней точности категоризации от коэффициента нормализации γ при фиксированном размере словаря визуальных слов. Результаты проведенных вычислений для

отдельных категорий приведены в таблице 3.5, а на рисунке 3.9 представлен график для средней точности категоризации.

Таблица 3.5

Точность категоризации изображений (%) в зависимости от коэффициента нормализации γ

Категории	Коэффициент нормализации γ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Coast	90,38	90,51	91,15	90,26	89,23	89,87	89,36	89,23	88,59	85,51
Forest	93,86	93,27	93,71	94,15	94,30	93,86	94,15	94,30	92,98	92,84
Highway	84,79	84,58	87,08	86,04	86,88	86,04	85,00	82,50	83,13	82,08
Inside city	93,43	93,91	94,55	94,71	94,55	94,23	95,03	94,07	93,27	92,15
Mountain	90,63	92,09	90,75	91,00	91,24	89,66	89,29	87,71	84,67	82,36
Open country	75,48	76,56	77,85	77,31	76,67	76,99	75,91	75,05	70,97	69,03
Street	88,54	89,24	90,28	89,58	89,93	88,89	88,54	88,37	85,94	86,11
Tall building	90,10	90,36	90,89	91,41	90,36	90,23	90,10	88,67	87,76	87,11
Средняя точность	88,40	88,82	89,53	89,31	89,14	88,72	88,42	87,49	85,91	84,65

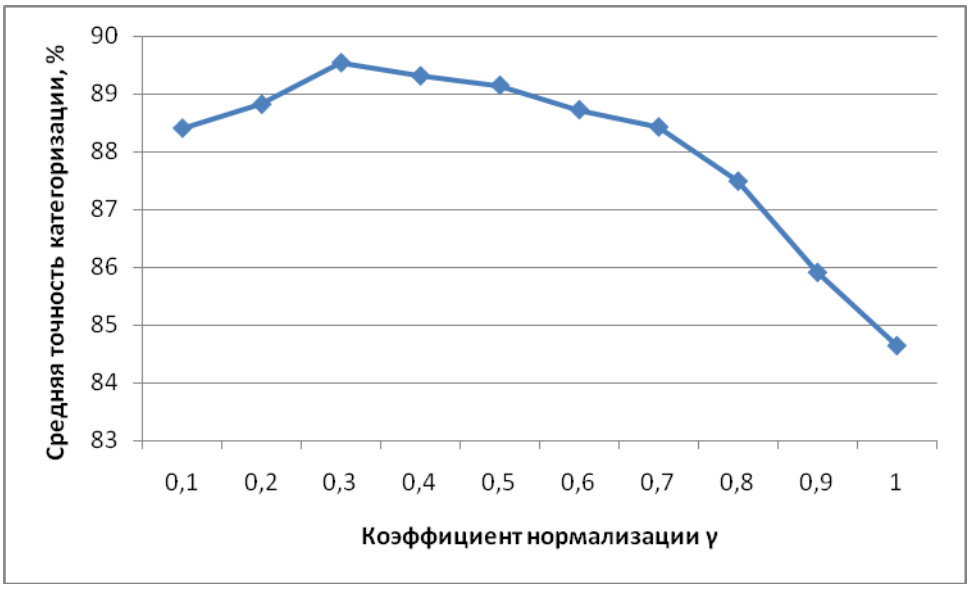

Рисунок 3.9. График зависимости средней точности категоризации изображений от коэффициента нормализации γ (S = 128)

Как видно из приведенных данных, любое значение коэффициента нормализации γ, отличное от 1, приводит к повышению точности категоризации, поскольку это позволяет снизить влияние локальных дескрипторов, вычисленных на области с изображением какой-либо структуры (например, рябь на воде), в случае, когда эта область занимает значительную часть изображения. При этом наилучший результат

достигается при $\gamma = 0{,}3$. Данное значение будет использоваться в дальнейших экспериментах.

Поскольку сформированные глобальные дескрипторы обладают большой размерностью (для 128 визуальных слов длина глобального дескриптора равна 8192), то для быстрых сравнений изображений большое значение имеет сокращение размерности с помощью метода главных компонент. В четвертом эксперименте определялась оптимальная размерность глобального дескриптора, позволяющая сократить вычислительные затраты с минимальными потерями. Результаты проведенных вычислений для отдельных категорий приведены в таблице 3.6, а на рисунке 3.10 представлен график для средней точности категоризации.

Таблица 3.6

Точность категоризации изображений (%) в зависимости от длины глобального дескриптора

Категории	Длина глобального дескриптора									
	32	64	128	192	256	384	512	768	1024	8192
Coast	83,59	86,54	87,05	89,10	88,33	90,00	90,26	89,74	88,97	91,15
Forest	93,27	93,86	94,44	94,30	94,30	94,30	94,44	94,01	93,86	93,71
Highway	80,21	81,67	83,75	85,00	84,38	84,79	84,58	85,83	84,38	87,08
Inside city	92,15	92,63	93,43	92,95	94,55	93,27	93,43	94,07	94,23	94,55
Mountain	87,71	89,29	90,02	89,66	90,27	91,73	90,88	91,48	91,61	90,75
Open country	71,18	73,76	74,09	75,16	74,62	75,81	75,48	75,48	76,45	77,85
Street	84,55	88,02	88,37	88,37	88,89	89,06	89,76	89,58	88,89	90,28
Tall building	88,02	88,67	88,93	90,10	89,19	90,23	89,71	89,45	90,49	90,89
Средняя точность	85,09	86,81	87,51	88,08	88,07	88,65	88,57	88,71	88,61	89,53

Как видно из приведенного на рисунке 3.10 графика, наименьшие потери достигаются при сокращении длины дескриптора до 384 элементов. В этом случае дескриптор уменьшается в 21,3 раза, а величина потери точности равна 0,88%.

Рисунок 3.10. График зависимости средней точности категоризации изображений от длины глобального дескриптора (S = 128, γ = 0,3)

3.2.3 Исследование цветовых локальных дескрипторов

Также были проведены дополнительные эксперименты для оценки точности определения отдельных категорий изображений с помощью цветовых локальных дескрипторов. Поскольку длина полученных цветовых дескрипторов изменяется в зависимости от количества используемых компонент цветового пространства, то для сравнения все глобальные дескрипторы сокращались до 384 элементов. Результаты проведенных вычислений представлены в таблице 3.7.

Результаты исследования показали, что наилучшие результаты в отдельных категориях достигаются с использованием трех дескрипторов, вычисленных в цветовых пространствах nRGB, HSV и LUV. Комбинация этих трех дескрипторов позволяет повысить среднюю точность категоризации на 2,1 % и 3 % по сравнению с дескрипторами, вычисленными только для nRGB и оттенков серого (Y) соответственно.

Таблица 3.7

Сравнение точности категоризации изображений с помощью разных цветовых дескрипторов (%)

Категории	Цветовое пространство (компонент)							
	Y	rg	Opponent	HSI	nRGB	HSV	LUV	nRGB + HSV + LUV
Coast	90,00	77,56	84,36	82,18	**90,13**	89,23	86,54	91,38
Forest	94,30	91,81	93,13	93,27	93,86	**95,61**	94,30	96,61
Highway	84,79	81,67	85,21	83,54	**85,63**	80,63	85,00	87,25
Inside city	93,27	87,34	89,26	85,42	**96,15**	90,38	88,94	94,27
Mountain	91,73	78,59	84,31	84,06	**94,16**	85,04	88,32	93,70
Open country	75,81	68,28	72,58	72,15	**81,29**	74,19	76,45	82,61
Street	89,06	81,25	89,24	88,02	85,42	88,54	**90,63**	94,23
Tall building	90,23	89,19	89,58	87,37	89,84	87,89	**91,41**	93,58
Средняя точность категоризации	88,65	81,96	85,96	84,50	89,56	86,44	87,70	91,70

3.2.4 Многопоточное вычисление локальных дескрипторов

Для оценки алгоритма распараллеливания вычисления набора локальных дескрипторов, была разработана тестовая система в среде Microsoft Visual C++ 2010, в которой были реализованы неоптимизированный вариант алгоритма, а также алгоритмы с использование OpenMP и Intel Cilk Plus. Распараллеливание в алгоритмах реализовано на основе линейного разделения изображения по высоте. В качестве тестовой базы выступали 6 наборов содержащих по 10 изображений со следующими разрешениями: 1920 × 1080, 2560 × 1600, 2800 × 2100, 3840 × 2160, 3646 × 2735, 4096 × 3072. Непосредственное тестирование выполнялось на персональных и серверных компьютерах разной конфигурации.

Для определения относительного коэффициента ускорения обработки изображения учитывалось время выполнения неоптимизированного алгоритма, принятое за 1. Коэффициент ускорения рассчитывался для каждого набора тестовых данных при использовании 2, 3 и 4-х потоков обработки. Значения коэффициентов ускорения распараллеленных алгоритмов для каждой технологии представлены в таблице 3.8.

Таблица 3.8

Значения коэффициентов ускорения средств распараллеливания

Размер изображения	Количество потоков								
	OpenMP			Intel Cilk Plus			Разница между OpenMP и Intel Cilk Plus, %		
	2	3	4	2	3	4	2	3	4
2 Мб	1,75	2,29	2,68	1,93	2,51	2,92	10,27	9,79	9,04
4 Мб	1,74	2,30	2,64	1,96	2,58	2,99	12,90	12,11	13,38
6 Мб	1,73	2,27	2,65	2,00	2,61	3,01	15,40	14,77	13,34
8 Мб	1,74	2,29	2,69	1,94	2,58	2,99	11,05	12,35	11,18
10 Мб	1,72	2,28	2,66	1,99	2,61	3,02	15,92	14,58	13,58
12 Мб	1,73	2,27	2,66	1,98	2,62	3,05	14,17	15,35	14,58
среднее	1,74	2,28	2,67	1,97	2,58	3,00	13,29	13,16	12,52

В результате исследования было выяснено, что наибольший прирост производительности достигается при использовании двух потоков вычислений, в то время как дальнейшее увеличение числа потоков приводит к снижению эффективности параллельной обработки изображений. Также эксперименты показали, что ускорение обработки с применением Intel Cilk Plus в среде Microsoft Visual C++ 2010 в среднем выше на 9–14%. Это связано с тем, что в среде Intel Cilk Plus реализована более эффективная балансировка нагрузки.

3.3 Результаты экспериментальных исследований автоматического аннотирования изображений

Для проверки предложенного метода автоматического аннотирования изображений, использовалась база изображений IAPR TC-12 [53]. Эта база содержит 19627 изображений размером 480×360 пикселов, каждое из которых описано несколькими ключевыми словами. Некоторые характеристики базы представлены в таблице 3.9.

Эксперименты проводились с целью исследования влияния параметров модифицированной сети ESOINN и количества выбираемых изображений при восстановлении ключевых слов обучающих изображений и оценке

условных вероятностей принадлежности новых изображений ОТВ-группам и семантическим группам на качество аннотирования этих изображений.

Таблица 3.9

Статистические данные базы изображений IAPR TC-12

Параметр		Значение
Количество изображений	обучающих	17665
	тестовых	1962
Количество изображений на ключевое слово	минимальное	44
	медианное	153
	среднее	347,7
	максимальное	4999
Количество ключевых слов	всего	291
	с частотой меньше среднего	217 (74,6 %)
Количество ключевых слов на изображение	минимальное	1
	медианное	5
	среднее	5,7
	максимальное	23

Для оценки качества ААИ вычислялись средняя точность (precision), средняя полнота (recall), а также совместная оценка этих двух показателей с помощью F_β-меры:

$$precision = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{AA(k_n)}, \tag{3.2}$$

$$recall = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{GT(k_n)}, \tag{3.3}$$

$$F_\beta = \frac{(\beta^2+1)\cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{3.4}$$

где $AA(k_n)$ – количество изображений, автоматически аннотированных ключевым словом k_n; $CA(k_n)$ – количество изображений, правильно аннотированных ключевым словом k_n; $GT(k_n)$ – количество изображений, содержащих в тестовой аннотации ключевое слово k_n; β – коэффициент, при значениях [0; 1) дающий больший вес точности, а при $\beta > 1$ – полноте аннотирования.

При тестировании приоритет отдавался точности аннотирования, в связи с чем использовалась $F_{0,5}$-мера ($\beta = 0,5$). Также в качестве дополнительной информации подсчитывалось количество ключевых слов, использованных при автоматическом аннотировании (N+).

Во всех экспериментах при вычислении визуальных дескрипторов использовались значения параметров, полученные на этапе исследования вычисления визуальных дескрипторов ($S = 128$, $\gamma = 0,3$, $Z = 384$), а при формировании ОТВ-групп текстовый дескриптор изображения имеет больший вес ($\alpha = 0,75$). Все расчеты повторялись 5 раз, после чего результаты усреднялись.

3.3.1 Исследование параметров алгоритмов формирования ОТВ-групп и автоматического аннотирования изображений

В первом эксперименте исследовался разработанный алгоритм ААИ без этапа восстановления ключевых слов обучающих изображений при различных параметрах. Исследовалось 5 наборов параметров сети ESOINN, при использовании которых средний размер формируемых ОТВ-групп постепенно увеличивается, начиная с первого:

- Set1: $\lambda = 50$, $age_{max} = 25$, $b_1 = 0,0$, $b_2 = 0,1$, $b_3 = 1,0$.
- Set2: $\lambda = 50$, $age_{max} = 50$, $b_1 = 0,0$, $b_2 = 0,1$, $b_3 = 1,0$.
- Set3: $\lambda = 100$, $age_{max} = 50$, $b_1 = 0,0$, $b_2 = 0,01$, $b_3 = 1,0$.
- Set4: $\lambda = 100$, $age_{max} = 100$, $b_1 = 0,0$, $b_2 = 0,005$, $b_3 = 1,0$.
- Set5: $\lambda = 200$, $age_{max} = 100$, $b_1 = 0,0$, $b_2 = 0,001$, $b_3 = 1,0$.

Также исследовано 4 значения количества изображений, выбираемых для определения вероятностей принадлежности нового изображения ОТВ-группам $YN_H = \{50, 100, 200, 300\}$, и 5 значений количества изображений, выбираемых из каждой ОТВ-группы $YN_{AN} = \{1, 3, 5, 7, 10\}$. Результаты вычислений для всех комбинаций параметров приведены в таблице 3.10.

Таблица 3.10

Результаты аннотирования изображений базы IAPR TC-12 при различных параметрах алгоритма ААИ (без этапа восстановления ключевых слов)

ESOINN	YN_H	YN_{AN}	Точность	Полнота	$F_{0,5}$	N+
Set1	50	1	50,24	37,80	47,14	278
		3	54,13	36,23	49,26	271
		5	55,89	35,53	50,14	270
		7	55,14	35,15	49,51	266
		10	55,01	35,00	49,37	265
	100	1	52,95	36,66	48,63	276
		3	57,46	35,00	50,92	270
		5	57,40	34,04	50,47	266
		7	57,32	33,75	50,30	264
		10	57,15	33,68	50,16	263
	200	1	55,21	35,33	49,62	273
		3	59,11	33,70	51,36	266
		5	59,98	33,02	**51,56**	265
		7	59,92	32,59	51,31	262
		10	59,63	32,38	51,04	260
	300	1	55,76	34,66	49,71	270
		3	59,17	33,00	51,07	264
		5	59,91	32,43	51,23	262
		7	60,14	31,76	51,02	260
		10	59,10	31,30	50,19	257
Set2	50	1	49,24	37,72	46,40	277
		3	54,06	36,15	49,18	271
		5	55,76	35,21	49,93	269
		7	56,15	34,62	49,94	266
		10	56,45	34,59	50,11	264
	100	1	51,57	36,70	47,71	275
		3	56,74	34,51	50,27	269
		5	58,17	33,86	50,87	268
		7	58,19	33,38	50,66	265
		10	58,11	33,14	50,50	263
	200	1	53,44	35,68	48,60	273
		3	57,88	33,40	50,48	265
		5	60,08	32,75	51,48	265
		7	59,67	32,06	50,90	262
		10	59,01	31,66	50,31	259
	300	1	54,46	35,34	49,14	273
		3	58,96	33,14	51,01	265
		5	**60,24**	32,36	51,39	265
		7	**60,24**	31,64	51,02	261
		10	59,38	31,14	50,26	257

Таблица 3.10. Продолжение

ESOINN	YN_H	YN_{AN}	Точность	Полнота	$F_{0,5}$	N+
Set3	50	1	47,59	37,80	45,25	280
		3	53,98	35,88	49,03	273
		5	55,86	34,95	49,89	270
		7	56,60	34,49	50,17	268
		10	56,50	34,12	49,95	265
	100	1	50,10	37,11	46,82	277
		3	55,81	34,49	49,67	269
		5	58,00	33,70	50,69	267
		7	58,64	33,07	50,79	266
		10	57,92	32,79	50,22	262
	200	1	51,22	36,36	47,35	276
		3	57,28	34,05	50,40	269
		5	58,75	33,10	50,86	265
		7	58,61	32,27	50,38	262
		10	59,18	32,02	50,60	261
	300	1	51,98	35,84	47,69	275
		3	58,25	33,73	50,86	270
		5	60,09	32,79	51,51	266
		7	60,01	31,94	51,04	263
		10	59,31	31,31	50,31	259
Set4	50	1	47,03	37,82	44,85	281
		3	52,45	35,45	47,86	269
		5	55,06	34,63	49,25	267
		7	56,75	34,04	50,07	268
		10	56,39	33,75	49,72	263
	100	1	48,73	37,42	45,95	280
		3	54,68	34,88	49,10	270
		5	58,09	34,00	50,88	269
		7	57,76	33,33	50,38	265
		10	58,08	32,91	50,37	263
	200	1	50,19	36,85	46,80	279
		3	56,25	34,18	49,82	269
		5	58,14	33,15	50,52	265
		7	59,00	32,64	50,79	264
		10	59,27	32,14	50,71	262
	300	1	49,80	36,52	46,42	278
		3	56,93	33,80	50,08	269
		5	58,64	32,86	50,68	265
		7	59,67	32,36	51,05	265
		10	59,54	31,62	50,60	262

Таблица 3.10. Продолжение

ESOINN	YN_H	YN_{AN}	Точность	Полнота	$F_{0,5}$	N+
Set5	50	1	46,01	**38,12**	44,18	**283**
		3	52,61	36,02	48,17	272
		5	54,31	34,71	48,80	267
		7	55,11	34,15	49,08	265
		10	56,19	33,78	49,60	265
	100	1	47,42	37,50	45,04	281
		3	53,55	34,92	48,39	269
		5	56,39	33,85	49,76	265
		7	57,34	33,26	50,09	265
		10	57,67	33,01	50,17	263
	200	1	48,47	36,97	45,63	279
		3	55,21	34,23	49,18	269
		5	56,95	33,34	49,88	265
		7	57,83	32,69	50,12	263
		10	57,61	32,27	49,79	261
	300	1	48,77	36,80	45,79	280
		3	55,93	34,07	49,57	270
		5	58,99	33,30	51,10	269
		7	59,34	32,70	51,03	266
		10	58,47	31,96	50,15	262

Проанализировав результат проведенных исследований можно сделать вывод, что увеличение размера ОТВ-групп приводит к снижению точности с одновременным увеличением полноты аннотирования, поскольку в этом случае снижается однородность: в группах встречается большее количество ключевых слов с меньшим количеством примеров. Увеличение значений параметров YN_H и YN_{AN} наоборот, приводит к увеличению количества обучающих изображений, используемых при аннотировании нового, и, как следствие, увеличению частоты встречаемости некоторых ключевых слов, что повышает точность и снижает полноту аннотирования. Наибольшее значение $F_{0,5}$ = 51,56 достигается при использовании набора Set1 и значениях параметров YN_H = 200 и YN_{AN} = 5. Эти значения используются в дальнейших экспериментах.

3.3.2 Исследование параметров алгоритма восстановления ключевых слов обучающих изображений

Во втором эксперименте исследовалось влияние параметров на разработанный алгоритм восстановления ключевых слов обучающих изображений. При этом исследовано 4 значения количества изображений, выбираемых из семантических групп YN_{WL} = {1, 2, 3, 4}, а также 5 значений количества изображений, используемых для определения количества пропущенных ключевых слов YA_{WL} = {1, 2, 3, 4, 5}. Результаты вычислений для всех комбинаций параметров приведены в таблице 3.11.

Таблица 3.11

Результаты аннотирования изображений базы IAPR TC-12 при различных параметрах алгоритма восстановления ключевых слов

YN_{WL}	YA_{WL}	Точность	Полнота	$F_{0,5}$	N+
1	1	61,00	34,19	52,73	**267**
	2	**62,22**	34,05	**53,38**	**267**
	3	61,03	33,33	52,33	265
	4	61,25	33,45	52,52	264
	5	60,73	32,68	51,83	262
2	1	60,69	34,12	52,51	266
	2	61,47	34,08	52,96	266
	3	60,77	33,39	52,21	265
	4	60,91	33,62	52,40	264
	5	60,44	32,86	51,75	263
3	1	60,36	**34,27**	52,38	266
	2	61,03	34,14	52,72	264
	3	60,49	33,12	51,91	263
	4	60,99	33,37	52,32	264
	5	60,86	32,94	52,04	262
4	1	60,73	34,22	52,58	266
	2	61,44	34,23	53,01	**267**
	3	60,80	33,28	52,17	264
	4	60,57	33,27	52,03	262
	5	60,70	32,74	51,84	262

В результате исследования было выяснено, что наибольшее значение $F_{0,5}$ = 53,38 достигается при значениях параметров YN_{WL} = 1 и YA_{WL} = 2. При этом выбор значения параметра YN_{WL} отличного от 1 снижает эффективность восстановления ключевых слов, поскольку в этом случае используется

большое количество обучающих изображений, не имеющих существенного визуального сходства с исходным изображением, но разделяющих одно ключевое слово. Также выбор одного наиболее похожего обучающего изображения для определения количества отсутствующих ключевых слов является недостаточным, в то время как выбор 3 и более приводит, как правило, к добавлению излишнего количества ключевых слов.

3.3.1 Сравнение с существующими методами автоматического аннотирования изображений

Дополнительно было проведено сравнение качества работы разработанного алгоритма аннотирования с известными методами, представленными в научных статьях ученых, занимающихся проблемой автоматического аннотирования изображений. Сравнительные результаты эффективности работы разработанного алгоритма и ряда известных методов ААИ представлены в таблице 3.12. Оценки для существующих методов взяты из соответствующих статей.

Таблица 3.12

Сравнение результатов аннотирования изображений базы IAPR TC-12 с помощью разных методов ААИ

Метод	Точность, %	Полнота, %	N+
Tag Propagation with Metric-Learning (TagProp-ML) [46]	48	25	227
Tag Propagation with Metric-Learning and Logistic Discriminant Models (TagProp-σML) [46]	46	35	266
FastTag [31]	47	26	280
2-Pass K-Nearest Neighbor (2PKNN) [100]	49	32	274
Support Vector Machine and Discrete Multiple Bernoulli Relevance Model (SVM-DMBRM) [81]	56	29	283
Предлагаемая реализация	62	34	267

Предложенный метод автоматического аннотирования изображений демонстрирует повышение эффективности аннотирования в сравнении с

современным методом SVM-DMBRM по точности аннотаций на 6 % и полноте на 5 %.

3.4 Выводы по главе

В третьей главе рассматривается разработанное экспериментальное программное обеспечение для автоматического аннотирования изображений. Программное обеспечение имеет модульную организацию и состоит из семи модулей. Модули, реализующие работу алгоритмов: модуль преобразований изображения, модуль вычисления визуальных дескрипторов, модуль формирования текстового дескриптора, модуль восстановления ключевых слов, модуль формирования ОТВ-групп и часть ядра системы, отвечающая за реализацию алгоритма автоматического аннотирования изображений. Для организации взаимодействия с пользователем реализован пользовательский интерфейс. Подробно рассмотрены схемы функционирования указанных модулей.

Проведено тестирование алгоритма формирования глобальных визуальных дескрипторов на основе разработанного метода быстрого вычисления набора локальных дескрипторов. Представлены результаты точности категоризации изображений с помощью предложенных дескрипторов, демонстрирующие повышение точности в среднем на 2 % при сокращении времени вычислений в 3,3 и 5,6 раз в сравнении с локальными дескрипторами SURF и G-SURF соответственно. Проанализировано влияние различных параметров алгоритма формирования глобальных визуальных дескрипторов на точность категоризации изображений. В результате были выбраны следующие значения параметров: размер словаря визуальных слов равен 128 элементам, коэффициент нормализации элементов глобального дескриптора 0,3, а длина глобального дескриптора сокращается до 384 элементов. Эти значения позволяют повысить точность категоризации изображений на 4 % при сокращении длины глобального дескриптора в 21,3

раза. Также проведены эксперименты для оценки точности определения отдельных категорий изображений с помощью цветовых локальных дескрипторов, показавшие, что наилучшие результаты в отдельных категориях достигаются с использованием трех дескрипторов, вычисленных в цветовых пространствах nRGB, HSV и LUV. Комбинация этих дескрипторов позволяет повысить среднюю точность категоризации на 3 % по сравнению с дескрипторами, вычисленными только для оттенков серого. В результате дополнительных экспериментов было выяснено, что при многопоточном вычислении локальных дескрипторов наибольший прирост производительности достигается при использовании двух потоков вычислений, в то время как дальнейшее увеличение числа потоков приводит к снижению эффективности параллельной обработки изображений.

Также проведено тестирование разработанного алгоритма автоматического аннотирования изображений на основе однородных текстово-визуальных групп. Проведен анализ влияния параметров алгоритмов формирования ОТВ-групп и ААИ без этапа восстановления ключевых слов обучающих изображений, показавший, что увеличение размеров ОТВ-групп приводит к повышению полноты формируемых аннотаций с одновременным снижением точности аннотирования. В тоже время увеличение количества изображений, выбираемых для оценки вероятностей принадлежности нового изображения семантическим и ОТВ-группам, повышает точность аннотаций и снижает полноту. Дополнительно приведены результаты оценки качества аннотирования при различных параметрах алгоритма восстановления пропущенных ключевых слов обучающих изображений, демонстрирующие повышение точности аннотирования на 2 % и полноты на 1 % при выборе по одному изображению из каждой семантической группы и использовании двух похожих изображений для определения количества пропущенных ключевых слов. Также в главе приведено сравнение разработанного алгоритма с известными методами ААИ, показавшее, что предлагаемая реализация повышает

эффективность аннотирования в сравнении с современным методом SVM-DMBRM по точности аннотаций на 6 % и полноте на 5 %.

ЗАКЛЮЧЕНИЕ

В работе представлены методы и алгоритмы автоматического аннотирования изображений в информационно-поисковых системах. Основные результаты и выводы диссертационного исследования представлены ниже:

1. Проведен анализ методов и алгоритмов автоматического аннотирования изображений. Показано, что в настоящее время наиболее эффективными являются поисковые статистические методы, в которых на основе визуально похожих обучающих изображений оценивается вероятность принадлежности ключевых слов аннотируемому изображению. При этом ключевыми моментами являются выбор визуального дескриптора для описания изображений, метод определения визуального сходства и полнота аннотаций обучающих изображений. Проведен анализ методов кластеризации данных и визуальных дескрипторов, описывающих изображения.

2. Разработан алгоритм быстрого вычисления набора локальных дескрипторов, основанный на локальном дескрипторе SURF, а также алгоритм его кодирования в глобальный дескриптор. Предложенный метод вычисления набора локальных дескрипторов позволяет повысить точность категоризации изображений по типу сцены в среднем на 2 %, затрачивая в 3,3 и 5,6 раз меньше времени, чем дескрипторы SURF и G-SURF.

3. Предложен метод расширения аннотаций обучающих изображений, позволяющий восстановить ключевые слова, пропущенные при составлении обучающих выборок. Метод автоматически определяет количество пропущенных ключевых слов и позволяет повысить точность аннотирования новых изображений на 2 % и полноту на 1 %.

4. Разработан метод кластеризации изображений с помощью самоорганизующейся нейронной сети на основе текстовых и визуальных дескрипторов. Полученные однородные текстово-визуальные группы

представляют собой контекст для аннотирования новых изображений и могут уточняться в течение жизненного цикла системы.

5. Разработан алгоритм автоматического аннотирования изображений, основанный на разделении обучающего набора изображений на однородные текстово-визуальные группы и аннотировании нового изображения с помощью обучающих изображений визуально похожих групп. При этом возможно повышение точности аннотирования при наличии небольшого количества ключевых слов, полученных от пользователей или с помощью узкоспециализированных классификаторов.

6. Разработан экспериментальный программный комплекс, позволяющий описывать изображения с помощью глобальных визуальных признаков, восстанавливать пропущенные ключевые слова в аннотациях обучающих изображений, кластеризовать обучающие изображения на основе текстовых и визуальных дескрипторов и аннотировать новые изображения с помощью полученных групп изображений.

7. Проведены экспериментальные исследования на тестовой базе изображений IAPR TC-12, которые показали увеличение точности аннотаций на 6 % и полноты до 5 %.

Таким образом, разработанные методы и алгоритмы позволяют повысить эффективность и качество автоматического аннотирования изображений в информационно-поисковых системах.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Арнхейм Р. Искусство и визуальное восприятие / сокр. пер. с англ. В.Н. Самохина. М.: Архитектура-С, 2012. 392 с.
2. Васильева Н.С., Новиков Б.А. Построение соответствий между низкоуровневыми характеристиками и семантикой статических изображений // Труды 7-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL-2005). Ярославль: Изд-во Ярослав. гос. ун-та, 2005. С. 236–240.
3. Гонсалес Р., Вудс Р. Цифровая обработка изображений / науч. ред. П.А. Чочиа; пер. с англ. Л.И. Рубанова, П.А. Чочиа. Изд. 3-е, испр. и доп. М.: Техносфера, 2012. 1104 с.
4. Иттен И. Искусство цвета / пер. с нем. Л. Монахова. 8-е изд. М.: Издатель Д. Аронов, 2013. 96 с.
5. Проскурин А.В., Белоконь А.В. Оценка эффективности контентно-ориентированных алгоритмов поиска изображений // Материалы XVI Международной научной конференции «Решетневские чтения». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2012. Ч. 2. С. 634–635.
6. Проскурин А.В., Белоконь А.В. Оценка эффективности статистических признаков текстур первого и второго порядков при анализе ландшафтных текстур // Материалы XVI Международной научной конференции «Решетневские чтения». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2012. Ч. 2. С. 593–594.
7. Проскурин А.В., Белоконь А.В. Автоматическое аннотирование ландшафтных изображений по их содержанию // Всероссийская научно-практическая конференция студентов, аспирантов и молодых специалистов «Актуальные проблемы авиации и космонавтики». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2013. Т. 1. С. 378–379.

8. Проскурин А.В., Белоконь А.В. Оценка эффективности алгоритмов поиска похожих изображений, с индексацией на основе преобразований Хаара и значений цветовых гистограмм // Всероссийская научно-практическая конференция студентов, аспирантов и молодых специалистов «Актуальные проблемы авиации и космонавтики». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2013. Т. 1. С. 351–352.
9. Проскурин А.В. Расширенная самоорганизующаяся растущая нейронная сеть для кластеризации данных в онлайн режиме // Материалы IX Международной научно-практической конференции «Электронные средства и системы управления». Томск: В-Спектр, 2013. Ч. 2. С. 178–182.
10. Проскурин А.В. Формирование визуальных слов для категоризации изображений // Всероссийская научная конференция студентов, аспирантов и молодых ученых «Наука. Технологии. Инновации». Новосибирск : Изд-во Новосиб. гос. тех. ун-та, 2013. Ч. 2. С. 99–102.
11. Проскурин А.В. Автоматическое аннотирование ландшафтных изображений // Вестник Сибирского государственного аэрокосмического университета. 2014. Вып. 3(55). С. 120–125.
12. Проскурин А.В. Алгоритм формирования визуальных слов // Материалы I Международной научной конференции «Региональные проблемы дистанционного зондирования Земли» (РПДЗЗ-2016). Красноярск: Изд-во Сиб. федерал. ун-та, 2014. – С. 158–162.
13. Проскурин А.В. Категоризация изображений на основе самоорганизующейся нейронной сети // Материалы XVIII Международной научной конференции «Решетневские чтения». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2014. Ч. 2. С. 274–276.
14. Проскурин А.В. Формирование визуальных слов для автоматического аннотирования изображений на основе самоорганизующейся нейронной сети // Материалы 16-й международной конференции

«Цифровая обработка сигналов и ее применение» (DSPA-2014). М.: Изд-во РНТОРЭС им. А.С. Попова, 2014. Т. 2. С. 487–491.

15. Проскурин А.В., Фаворская М.Н., Зотин А.Г., Дамов М.В. Применение параллельных вычислений при расчете признаков в системах автоматического аннотирования изображений // Телекоммуникации. 2015. № 4. С. 41–47.
16. Проскурин А.В., Фаворская М.Н. Система автоматического формирования визуальных слов (ForVW). Свидетельство о государственной регистрации программы для ЭВМ №2015611845. Зарегистрировано в Реестре программ для ЭВМ г. Москва, 06.02.2015.
17. Проскурин А.В., Фаворская М.Н. Категоризация сцен на основе расширенных цветовых дескрипторов // Труды СПИИРАН. 2015. № 40. С. 203–220.
18. Проскурин А.В. Быстрый локальный дескриптор для категоризации изображений по типу сцены // Материалы XIX международной научно-практической конференции «Решетневские чтения». Красноярск: Изд-во Сиб. гос. аэрокосмич. ун-та, 2015. Ч. 2. С. 243–245.
19. Проскурин А.В. Модификация самоорганизующейся нейронной сети для автоматического аннотирования изображений // Материалы 17-й международной конференции «Цифровая обработка сигналов и ее применение» (DSPA-2015). М.: Изд-во РНТОРЭС им. А.С. Попова, 2015. Т. 2. С. 503–507.
20. Проскурин А.В., Фаворская М.Н. Автоматическое аннотирование изображений на основе однородных текстово-визуальных групп // Информационно-управляющие системы. 2016. № 2. С. 11–18.
21. Проскурин А.В., Фаворская М.Н. Система автоматического аннотирования изображений (AIA). Свидетельство о государственной регистрации программы для ЭВМ №2016611307. Зарегистрировано в Реестре программ для ЭВМ г. Москва, 29.01.2016.

22. Проскурин А.В. Формирование глобального дескриптора для классификации изображений по типу сцены и объекта // Материалы 18-й международной конференции «Цифровая обработка сигналов и ее применение» (DSPA-2016). М.: Изд-во РНТОРЭС им. А.С. Попова, 2016. Т. 2. С. 862–866.
23. Уиллиамс У.Т., Ланс Д.Н. Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. М.: Наука, 1986. С. 269–301.
24. Alcantarilla P.F., Bergasa L.M., Davison A.J. Gauge-SURF Descriptors // Image and Vision Computing. 2013. Vol. 31. N 1. pp. 103–116.
25. Arthur D. Vassilvitskii S. k-means++: the advantages of careful seeding // Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms. Philadelphia, USA. 2007. pp. 1027–1035.
26. Bay H., Ess A., Tuytelaars T., Gool L.V. Speeded-Up Robust Features (SURF) // Computer Vision and Image Understanding. 2008. Vol. 110. N 3. pp. 346–359.
27. Bell A.J., Senjnowsky T.J. The «independent components» of natural scenes are edge filters // Vision Research. 1997. Vol. 37. N 23. pp. 3327–3338.
28. Bi J., Chen Y., Wang J.Z. A sparse support vector machine approach to region-based image categorization // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2005. Vol. 1. pp. 1121–1128.
29. Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E. Fast Unfolding of Communities in Large Networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. N 10. P10008. doi:10.1088/1742-5468/2008/10/P10008
30. Chapelle O., Haffner P., Vapnik V.N. Support vector machines for histogram-based image classification // IEEE Transactions on Neural Networks. 1999. Vol. 10. pp. 1055–1064.

31. Chen M., Zheng A., Weinberger K.Q. Fast Image Tagging // Proceedings 30th International Conference on Machine Learning. Atlanta, USA. 2013. pp. 1274–1282.
32. Chen Y., Wang J.Z. Image categorization by learning and reasoning with regions // The Journal of Machine Learning Research. 2004. Vol. 5. pp. 913–939.
33. Deng J., Berg A., Li K., Li F.F. What does classifying more than 10,000 image categories tell us? // Proceedings of the 11th European Conference of Computer Vision (ECCV). 2010. Vol. 6315. pp 71–84.
34. Derpanis K.G., Integral image-based representations // Department of Computer Science and Engineering, York University Paper. 2007. Vol. 1. N 2, pp. 1–6.
35. Duygulu P., Barnard K., Freitas N., Forsyth D. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary // Proceedings of the 7th European Conference on Computer Vision. 2002. Vol. 2353. pp. 97–112.
36. Dubey R.S., Choubey R., Bhattacharjee J. Multi feature content based image retrieval // International Journal on Computer Science and Engineering. 2010. Vol. 2(6). pp. 2145–2149.
37. ESP-Game Image set [Электронный ресурс]. URL: http://hunch.net/~learning/ESP-ImageSet.tar.gz (дата обращения: 04.02.2017).
38. Favorskaya M.N., Proskurin A.V. Image Categorization Using Color G-SURF Invariant to Light Intensity // Procedia Computer Science. 2015. Vol. 60. pp. 681–690.
39. Favorskaya M.N., Jain L.C., Proskurin A.V. Unsupervised Clustering of Natural Images in Automatic Image Annotation Systems // New Approaches in Intelligent Image Analysis: Techniques, Methodologies and Applications / Eds. R. Kountchev, K. Nakamatsu. Switzerland: Springer International Publishing, 2016. Vol. 108. pp. 123–155.

40. Feng S.L., Manmatha R., Lavrenko V. Multiple Bernoulli relevance models for image and video annotation // In Proceedings of the International Conference on Pattern Recognition. 2004. Vol. 2. pp. 1002–1009.
41. Field D.J. Relations between the statistics of natural images and the response properties of cortical cells // Journal of the Optical Society of America. 1987. Vol. 4(12). pp. 2379–2394.
42. Flickner M., Sawhney H., Niblack W., Ashley J., Huang Q., Dom B., Gorkani M., Hafner J., Lee D., Petkovic D., Steele D., Yanker P. Query by image and video content: the QBIC system // IEEE Computer. 1995. Vol. 28(9). pp. 23–32.
43. Flickr [Электронный ресурс]. URL http://www.flickr.com/ (дата обращения: 04.02.2017).
44. Gao Y., Fan J., Xue X., Jain R. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers // Proceedings of the 14th annual ACM international conference on Multimedia. New York, USA. 2006. pp. 901–910.
45. Google Photos [Электронный ресурс]. URL: https://photos.google.com/ (дата обращения: 04.02.2017).
46. Guillaumin M., Mensink T., Verbeek J., Schmid C. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation // Proceedings of the IEEE 12th International Conference on Computer Vision. 2009. pp. 309–316.
47. Guillamet D., Vitria J. Evaluation of distance metrics for recognition based on non-negative matrix factorization // Pattern Recognition Letters. 2003. Vol. 24(9–10). pp. 1599–1605.
48. Guillamet D., Schiele B., Vitria J. Analyzing non-negative matrix factorization for image classification // Proceedings of the 16th International Conference on Pattern Recognition (ICPR). 2002. Vol. 2. pp. 116–119.

49. Haralick R.M., Shanmugam K., Dinstein I.H. Textural Features for Image Classification // IEEE Transactions on Systems, Man and Cybernetics. 1973. Vol. 3. N 6. pp. 610–621.
50. Haralick R.M. Statistical and Structural Approaches to Texture // Proceedings of the IEEE. 1979. Vol. 67(5). pp. 786–804.
51. Hong Z., Jiang Q. Hybrid content-based trademark retrieval using region and contour features // Proceedings of the 22nd International Conference on Advanced Information Networking and Applications. 2008. pp. 1163–1168.
52. Huang J., Kumar S., Mitra M., Zhu W.J., Zabih R. Image indexing using colour correlogram // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 1997. pp. 762–765.
53. IAPR TC-12 Benchmark [Электронный ресурс]. http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz (дата обращения: 04.02.2017).
54. ImageTagger [Электронный ресурс]. URL: http://attrasoft.com/products_imagetagger.asp (дата обращения: 04.02.2017).
55. Imagga Auto-Tagging API [Электронный ресурс]. URL: https://imagga.com/solutions/auto-tagging.html (дата обращения: 04.02.2017).
56. Instagram [Электронный ресурс]. https://www.instagram.com/ (дата обращения: 04.02.2017).
57. Jain A.K., Vailaya A. Image retrieval using colour and shape // Pattern Recognition. 1996. Vol. 29(8). pp. 1233–1244.
58. Jain L.C., Favorskaya M., Novikov D. Panorama Construction from Multi-view Cameras in Outdoor Scenes // Computer Vision in Control Systems-2 / Eds. M.N. Favorskaya, L.C. Jain. Switzerland: Springer International Publishing. 2015. Vol. 75. pp. 71–108.

59. Jegou H., Douze M., Schmid C., Perez P. Aggregating local descriptors into a compact image representation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010. pp. 3304–3311.
60. Jeon J., Lavrenko V., Manmatha R. Automatic image annotation and retrieval using cross-media relevance models // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada. 2003. pp. 119–126.
61. Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley and Sons, 2005. 342 p.
62. Ke Y., Sukthankar R. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors // Proceedings of the IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR). 2004. pp. 506–513.
63. Lavrenko V., Manmatha R., Jeon J. A model for learning the semantics of pictures // Proceedings of the 17th Annual Conference on Neural Information Processing Systems. 2003. Vol. 16. pp. 553–560.
64. Lee D.D., Seung H.S. Learning the parts of objects by non-negative matrix factorization // Nature. 1999. Vol. 401(6755). pp. 788–791.
65. Leung W.H., Chen T. Trademark retrieval using contour-skeleton stroke classification // Proceedings of the IEEE International Conference on Multimedia and Expo. 2002. Vol. 2. pp. 517–520.
66. Li F.F., Perona P. A Bayesian Hierarchical Model for Learning Natural Scene Categories // Proceedings of the IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR). 2005. Vol. 2. pp. 524–531.
67. Liu W., Zheng N. Non-negative matrix factorization based methods for object recognition // Pattern Recognition Letters. 2004. Vol. 25. pp. 893–897.

68. Liu Y., Zhang J., Tjondronegoro D., Geve S. A shape ontology framework for bird classification // Proceedings of the 9th Conference on Digital Image Computing Techniques and Applications. 2007. pp. 478–484.
69. Lloyd S.P. Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28. N 2. pp. 129–136.
70. Long F., Zhang H.J., Feng D.D. Fundamentals of content-based image retrieval // Multimedia Information Retrieval and Management / Eds. D.D. Feng, W.C. Siuandg, H.J. Zhan. Springer Berlin Heidelberg. 2003. Part 1. pp. 1–26.
71. Lowe D.G. Distinctive Image Features from Scale-Invariant Keypoints // International Journal of Computer Vision. 2004. Vol. 60. N 2. pp. 91–110.
72. Makadia A., Pavlovic V., Kumar S. A New Baseline for Image Annotation // Proceedings of the 10th European Conference on Computer Vision. 2008. Vol. 5304. pp. 316–329.
73. Manjunath B.S., Salembier P., Sikora T. (Eds.) Introduction to MPEG-7: Multi-media Content Description Language. New York: John Wiley and Sons, 2002. 396 p.
74. Manning C.D., Raghavan P., Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 506 p.
75. Maree R., Geurts P., Piater J., Wehenkel L., Schmid C., Soatto S., Tomasi C. Random Subwindows for Robust Image Classification // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). 2005. Vol. 1. pp. 34–40.
76. Maron O., Lozano-Perez T. A framework for multiple-instance learning // Advances in Neural Information Processing Systems / Eds. M.I. Jordan, M.J. Kearns, S.A. Solla. The MIT Press. 1998. Vol. 10. pp. 570–576.
77. Mezaris V., Kompatsiaris I., Strintzis M.G. An ontology approach to object-based image retrieval // Proceedings of the International Conference on Image Processing. 2003. pp. 511–514.

78. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope [Электронный ресурс]. URL: http://people.csail.mit.edu/torralba/code/spatialenvelope (дата обращения: 04.02.2017).
79. Mori Y., Takahashi H., Oka R. Image-to-word transformation based on dividing and vector quantizing images with words // Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management. 1999. doi:10.1.1.31.1704
80. MUFIN Image Annotation [Электронный ресурс]. URL: http://disa.fi.muni.cz/annotation-ui (дата обращения: 04.02.2017).
81. Murthy V.N., Can E.F., Manmatha R. A Hybrid Model for Automatic Image Annotation // Proceedings of International Conference on Multimedia Retrieval. 2014. pp. 369–376.
82. Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms // Proceedings of the IEEE Conference on Information and Communication Technologies (ICT). 2013. Vol. 13. pp. 298–303.
83. Pass G., Zabith R. Histogram refinement for content-based image retrieval // Proceedings of the IEEE Workshop on Applications of Computer Vision. 1996. pp. 96–102.
84. PiXiT – an automatic image classification software in Java [Электронный ресурс]. URL: http://www.montefiore.ulg.ac.be/~maree/pixit.html (дата обращения: 04.02.2017).
85. Philbin J., Chum O., Isard M., Sivic J., Zisserman A. Object retrieval with large vocabularies and fast spatial matching // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2007. doi: 10.1109/CVPR.2007.383172
86. Profimedia [Электронный ресурс]. URL: https://www.profimedia.com/ (дата обращения: 04.02.2017).
87. Qi X., Han Y. Incorporating multiple SVMs for automatic image annotation // Pattern Recognition. 2007. Vol. 40. pp. 728–741.

88. Singha M., Hemachandran K. Content Based Image Retrieval using Color and Texture // Signal and Image Processing: An International Journal. 2012. Vol. 3. N 1. P. 39–57.
89. Stanchev P.L., Green D. Jr., Dimitrov B. High level colour similarity retrieval // International Journal of Information Theories and Applications. 2003. Vol. 10(3). pp. 363–369.
90. Stricker M., Orengo M. Similarity of Color Images / SPIE Conference. 1995. Vol. 2420. doi:10.1117/12.205308
91. Stricker M., Dimai A. Spectral Covariance and Fuzzy Regions for Image Indexing // Machine Vision and Applications. 1997. Vol. 10(2). pp. 66–73.
92. Swain M.J., Ballard D.H. Color indexing // International Journal of Computer Vision. 1991. Vol. 7(1). pp. 11–32.
93. Shen F., Ogura T., Hasegawa O. An enhanced self-organizing incremental neural network for online unsupervised learning // Neural Networks. 2007. Vol. 20. N 8. pp. 893–903.
94. Tamura H., Mori S., Yamawaki T. Psychological and computational measurements of basic textural features and their comparison // Proceedings of the 3rd International Joint Conference of Pattern Recognition. 1976. pp. 273–277.
95. Tamura H., Mori S., Yamawaki T. Texture features corresponding to visual perception // IEEE Transactions on Systems, Man and Cybernetics. 1978. Vol. 8(6). pp. 460–473.
96. TreeTagger - a part-of-speech tagger for many languages [Электронный ресурс]. URL: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (дата обращения: 04.02.2017).
97. Tsai C.F. Bag-of-Words Representation in Image Annotation: A Review // International Scholarly Research Network ISRN Artificial Intelligence. 2012. Vol. 2012. doi:10.5402/2012/376804
98. Tsuge S., Shishibori M., Kuroiwa S., Kita K. Dimensionality reduction using non-negative matrix factorization for information retrieval // In IEEE

International Conference on Systems, Man and Cybernetics. 2001. pp. 960–965.

99. Vapnik V.N. Statistical Learning Theory. New York: John Wiley and Sons, 1998. 768 p.
100. Verma Y., Jawahar C.V. Image Annotation Using Metric Learning in Semantic Neighbourhoods // Proceedings of the 12th European Conference on Computer Vision. 2012. Vol. 7574. pp. 836–849.
101. von Ahn L., Dabbish L. Labeling images with a computer game // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2004. pp. 319–326.
102. Wang J., Yang J., Yu K., Lv F., Huang T., Gong Y. Locality-constrained linear coding for image classification // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010. pp. 3360–3367.
103. Xu R., Wunsch D. Survey of clustering algorithms // IEEE Transactions, Neural Networks. 2005. Vol. 16. N 3. pp. 645–678.
104. Xu W., Liu X., Gong Y. Document clustering based on non-negative matrix factorization // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. Toronto, Canada: ACM Press, 2003. pp. 267–273.
105. Yang C., Dong M., Fotouhi F. Region based image annotation through multiple-instance learning // Proceedings of the 13th annual ACM international conference on Multimedia. New York, USA: ACM Press, 2005. pp. 435–438.
106. Yang C., Zhang L., Lu H., Ruan X., Yang M.H., Saliency detection via graph-based manifold ranking // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013 pp. 3166–3173.
107. Yang J., Yu K., Gong Y., Huang T. Linear spatial pyramid matching using sparse coding for image classification // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009. pp. 1794–1801.

108. Zheng D., Zhao Y., Wang J. Features Extraction using Gabor Filter Family // Proceedings of the 6th IASTED International Conference Signal and Image Processing, Hawaii, USA. 2004. pp. 139–144.

ПРИЛОЖЕНИЕ 1. СВИДЕТЕЛЬСТВА О РЕГИСТРАЦИИ ПРОГРАММЫ «СИСТЕМА АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ ВИЗУАЛЬНЫХ СЛОВ (FORVW)»

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2015611845

Система автоматического формирования визуальных слов (ForVW)

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева» (СибГАУ) (RU)*

Авторы: *Проскурин Александр Викторович (RU), Фаворская Маргарита Николаевна (RU)*

Заявка № 2014662846
Дата поступления 11 декабря 2014 г.
Дата государственной регистрации
в Реестре программ для ЭВМ *06 февраля 2015 г.*

Врио руководителя Федеральной службы по интеллектуальной собственности

Л.Л. Кирий

ПРИЛОЖЕНИЕ 2. СВИДЕТЕЛЬСТВА О РЕГИСТРАЦИИ ПРОГРАММЫ «СИСТЕМА АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ИЗОБРАЖЕНИЙ (AIA)»

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2016611307

Система автоматического аннотирования изображений (AIA). Версия 1.0

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева» (СибГАУ) (RU)***

Авторы: ***Проскурин Александр Викторович (RU), Фаворская Маргарита Николаевна (RU)***

Заявка № **2015661962**

Дата поступления **08 декабря 2015 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***29 января 2016 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

ПРИЛОЖЕНИЕ 3. АКТ ОБ ИСПОЛЬЗОВАНИИ РЕЗУЛЬТАТОВ В ООО «НПП «БЕВАРД»

BEWARD
IP-ВИДЕОНАБЛЮДЕНИЕ

117198, г. Москва, ул. Миклухо-Маклая, д.8 стр. 3.
660118, г. Красноярск, ул. Молокова, д.16 оф 355
Тел. (495) 502-27-29, (391) 278-92-00 www.beward.ru

ООО «НПП «Бевард»
ИНН 2465266818
КПП 246501001
ОГРН № 1122468006385
ОКПО 38587901
ОКОНХ 51.43.2
Р/счет 40702810500210010523
в Новосибирском филиале ОАО «МТС-Банк»
БИК 045003847
к/счет № 30101810200000000847

Утверждаю

Директор ООО «НПП «Бевард»
Седин Д.В. _____
«12» августа 2015г.

АКТ

об использовании (применении) результатов диссертационной работы аспиранта - **Проскурина Александра Викторовича** на тему : « Методы и алгоритмы автоматического аннотирования изображений в информационно-поисковых системах», представленной на соискание ученой степени кандидата технических наук.

Отделом программирования приняты для дальнейшего использования в работе материалы, содержащие блок-схемы алгоритмов задачи автоматического аннотирования изображений, а также разработанное программное обеспечение с соответствующей программной документацией.

Руководитель отдела программирования _____ С. В. Шевчук

М.П.

ПРИЛОЖЕНИЕ 4. АКТ ОБ ИСПОЛЬЗОВАНИИ МАТЕРИАЛОВ В СИБИРСКОМ ГОСУДАРСТВЕННОМ АЭРОКОСМИЧЕСКОМ УНИВЕРСИТЕТЕ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева» (СибГАУ)

просп. им. газеты «Красноярский рабочий», 31
г. Красноярск, 660037
тел.: +7 (391) 264-00-14 факс: +7 (391) 264-47-09
http://www.sibsau.ru e-mail: info@sibsau.ru
ОКПО 02069734, ОГРН 1022402056038
ИНН/КПП 2462003320/246201001

15 ФЕВ 2017 № 17/623

На № _____ от _____

УТВЕРЖДАЮ

Проректор по ОД

_____ Ю.В. Ерыгин

«___» _____ 2017 г.

АКТ

об использовании материалов диссертационной работы Проскурина А.В. «Методы и алгоритмы автоматического аннотирования изображений в информационно-поисковых сетях»

Мы, нижеподписавшиеся, директор Института информатики и телекоммуникаций Попов А.М., доцент кафедры информатики и вычислительной техники Зотин А.Г. составили настоящий акт о том, что материалы диссертационного исследования, выполненного Проскуриным А.В. используются в учебном процессе при проведении лекционных занятий и лабораторных работ по дисциплинам «Теоретические основы цифровой обработки изображений», «Алгоритмы обработки изображений и видеопоследовательностей» для магистрантов, обучающихся по направлению подготовки 09.04.01 «Информатика и вычислительная техника» в Сибирском государственном аэрокосмическом университете имени академика М.Ф. Решетнева (СибГАУ).

Директор института информатики
и телекоммуникаций, профессор, д.ф.-м.н. А.М. Попов

Доцент кафедры информатики
и вычислительной техники , к.т.н. А.Г. Зотин

Исполнитель:
Львова Анна Викторовна
8 (391) 213-96-22