На правах рукописи

ПРОСКУРИН АЛЕКСАНДР ВИКТОРОВИЧ

МЕТОДЫ И АЛГОРИТМЫ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ИЗОБРАЖЕНИЙ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Красноярск – 2017

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева», г. Красноярск

Научный руководитель:	доктор технических наук, профессор Фаворская Маргарита Николаевна
Официальные оппоненты:	Сергеев Михаил Борисович, доктор технических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения», кафедра вычислительных систем и сетей, заведующий кафедрой
	Пестунов Игорь Алексеевич, кандидат физико-математических наук, доцент, Федеральное государственное бюджетное учреждение науки «Институт вычислительных технологий Сибирского отделения Российской академии наук», лаборатория обработки данных, ведущий научный сотрудник
Ведущая организация:	Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук», г. Красноярск

Защита состоится «16» июня 2017 г. в 14.00 часов на заседании диссертационного совета Д 212.099.22 на базе Сибирского федерального университета по адресу: 660074, г. Красноярск, ул. Киренского, 26, ауд. УЛК 112.

С диссертацией можно ознакомиться в библиотеке и на сайте Сибирского федерального университета по адресу http://www.sfu-kras.ru

Автореферат разослан «__» апреля 2017 г.

Ученый секретарь
диссертационного совета Покидышева Людмила Ивановна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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, S.L. Feng, M. Guillaumin, V. Lavrenko, А.С. 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, 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 и полноту аннотирования новых 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, 19-й международных конференциях и выставках «Цифровая обработка сигналов и ее применение» (Москва, 2014, 2015, 2016, 2017 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 21 печатная работа, из которых 4 изданы в журналах, рекомендованных ВАК, 2 в журналах и книгах, индексируемых в Scopus, 13 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.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулирована цель и поставлены задачи исследования, показана научная новизна и практическая

ценность выполненных исследований, представлены основные положения, выносимые на защиту.

В **первой главе** представлен обзор существующих методов ААИ, кластеризации данных и описания изображений с помощью низкоуровневых признаков. Также рассмотрен ряд программных систем, реализующих автоматическое аннотирование изображений. Приведена классификация методов ААИ, представленная в таблице 1. В настоящее время наиболее эффективными являются поисковые методы, позволяющие разрабатывать информационно-поисковые системы, способные обучаться в течение жизненного цикла и предоставлять наиболее полное текстовое описание изображений.

Таблица 1

Классификация подходов к автоматическому аннотированию изображений

Категории	Подходы
Аннотирование одним ключевым словом	– Классификационный
Аннотирование несколькими ключевыми словами	– Генеративный – Поисковый

Поисковые методы аннотирования изображений состоят из алгоритмов обучения и аннотирования. Для предварительного обучения требуется выборка изображений, уже имеющих ключевые слова. Для этих изображений дополнительно вычисляются визуальные дескрипторы, которые используются для поиска визуально похожих изображений. При аннотировании нового изображения, из него также извлекается визуальный дескриптор, с помощью которого определяется набор ближайших соседей из обучающей выборки. Используя численные значения сходства между новым изображением и изображениями полученного набора, для каждого ключевого слова определяется вероятность его принадлежности новому изображению. В качестве аннотации выбираются ключевые слова с наибольшими значениями вероятности. Таким образом, ключевыми моментами являются выбор визуального дескриптора для описания изображений, метод определения набора ближайших соседей и полнота аннотаций обучающих изображений.

Большинство программных систем, реализующих автоматическое аннотирование изображений, являются исследовательскими и предназначены для проверки работоспособности разрабатываемых методов аннотирования изображений. Для их сравнения, в основном, используют две базы изображений: IAPR TC-12 (URL: http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz) и ESP-Game (URL: http://hunch.net/~learning/ESP-ImageSet.tar.gz). Результаты работы программ для одинаковых обучающих выборок и тестовых изображений публикуются авторами исследований. В настоящее время лучшие результаты не превышают 56 % для точности и 35 % для полноты аннотирования.

Во **второй главе** представлены разработанный алгоритм быстрого извлечения набора локальных дескрипторов и алгоритм его преобразования в глобальный визуальный дескриптор. Также приведено описание разработанного метода восстановления пропущенных ключевых слов в аннотациях обу-

чающих изображений и алгоритм формирования текстового дескриптора изображения. Кроме того, описаны разработанный метод кластеризации обучающих изображений в однородные текстово-визуальные группы и алгоритм аннотирования на основе этих групп.

Сложность задачи состоит в том, что в аннотациях обучающих изображений 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.

Таблица 2

Ограничения и условия, предъявляемые к обучающему набору и 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 в обучающем наборе	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:

$$P(k_n \mid A) = \frac{\sum_{H_l \in H} \left[P(H_l) P(A \mid H_l) P_l(A \mid k_n) P_l(k_n) \right]}{P(A)}, \qquad (1)$$

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. Поскольку вероятность $P(A)$ одинакова для всех ключевых слов, то будем анализировать числитель.

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.

II этап. Создание текстового дескриптора.

III этап. Формирование однородных текстово-визуальных групп.

IV этап. Автоматическое аннотирование изображений.

На первом этапе 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. Предлагаемый алгоритм быстрого вычисления набора локальных дескрипторов (Fast Dense Speeded-Up Features – FD-SUF) основан на локальном дескрипторе SURF и состоит из двух этапов: вычисления матрицы частей дескрипторов $\mathbf{M}$, в которой каждый элемент $\mathbf{M}_{rx,ry}$ является вектором из 4 чисел, и построения с ее помощью набора локальных дескрипторов. Блок-схема разработанного алгоритма представлена на изображении 1.

На первом этапе все изображение I разделяется сеткой на блоки $I_{rx,ry}$ размером $5\sigma_{sc} \times 5\sigma_{sc}$ пикселов, где σ_{sc} – масштаб. После этого в каждом блоке для 5×5 равномерно распределенных точек вычисляются свертки части изображения I с центром в точке p, имеющей координаты (x, y), с первыми производными Гауссиана $g(\sigma_{sc})$ по осям OX и OY:

$$L_x(p,\sigma_{sc}) = I(p) * \frac{\partial}{\partial x} g(\sigma_{sc}), \qquad L_y(p,\sigma_{sc}) = I(p) * \frac{\partial}{\partial y} g(\sigma_{sc}), \tag{2}$$

Используя накопленные значения первых производных, каждый элемент матрицы частей дескрипторов формируется следующим образом:

$$\mathbf{M}_{rx,ry} = \left(\sum_{p \in I_{rx,ry}} L_x(p,\sigma_{sc}), \sum_{p \in I_{rx,ry}} L_y(p,\sigma_{sc}), \sum_{p \in I_{rx,ry}} \left|L_x(p,\sigma_{sc})\right|, \sum_{p \in I_{rx,ry}} \left|L_y(p,\sigma_{sc})\right| \right). \tag{3}$$

На втором этапе алгоритма формируется набор локальных дескрипторов $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_B\}$, где $\mathbf{x}_b \in \mathbf{R}^D$. Для этого по матрице $\mathbf{M}$ перемещается скользящее окно размером 4×4 элемента. Каждый локальный дескриптор представляет собой объединение векторов $\mathbf{M}_{rx,ry}$, попавших в скользящее окно и имеет длину $D = 64$. Полученный дескриптор нормализуется с помощью l_2-нормы, после чего скользящее окно смещается. Изменяя шаг смещения скользящего окна можно регулировать количество извлекаемых локальных дескрипторов.

Базовый алгоритм FD-SUF вычисляется только на изображениях в оттенках серого, в связи с чем подвержен сильному влиянию условий освещенности и не учитывает цветовую информацию. Поэтому предложено вычислять дескрипторы в цветовых пространствах HSV и LUV, а также в цветовой модели RGB с нормализацией значений пикселов (nRGB). Для нормализации компонент RGB-пространства используется следующая формула:

$$nRGB_i = \frac{RGB_i - \mu_i}{\sigma_i}, \tag{4}$$

где RGB_i и $nRGB_i$ – исходное и нормализованное значения пиксела в i-м канале; μ_i и σ_i – среднее значение и среднеквадратичное отклонение в i-м канале, вычисляемые по выбранной области (блок или все изображение).

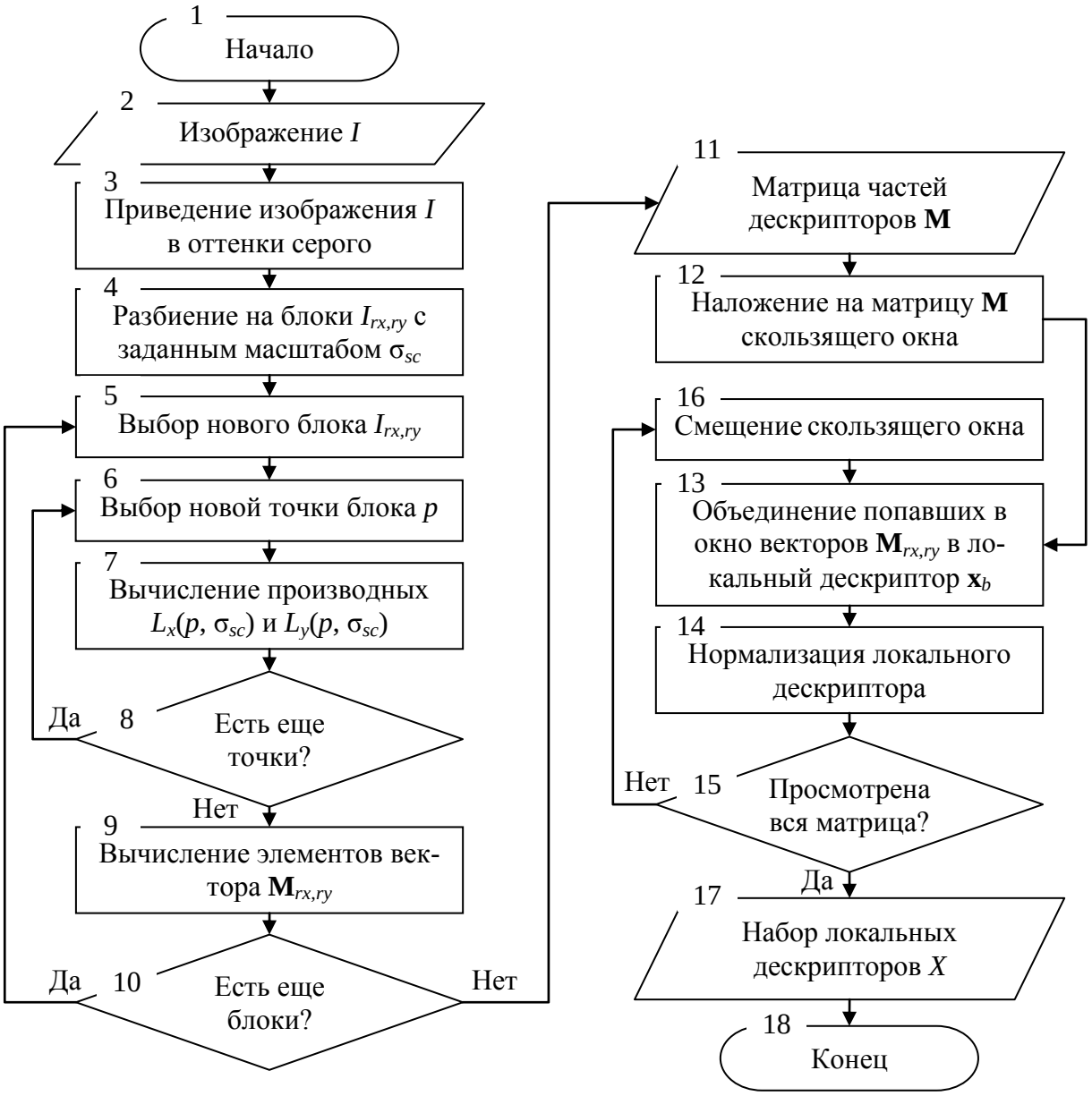

Рисунок 1. Блок-схема алгоритма вычисления набора локальных дескрипторов

При построении цветовых признаков FD-SUF дескрипторы вычисляются для каждой компоненты цветового пространства, после чего они объединяются в дескрипторы утроенной длины. Размерность полученных расширенных дескрипторов затем сокращается до 64 элементов с помощью метода главных компонент.

Извлеченный из изображения набор локальных дескрипторов кодируется с помощью словаря визуальных слов и преобразуется в глобальный дескриптор. Для создания словаря визуальных слов $VW = \{\mathbf{vw}_1, \ldots, \mathbf{vw}_S\}$, где $\mathbf{vw}_s \in \mathbf{R}^D$, из обучающих изображений случайным образом выбираются локальные признаки, которые затем кластеризуются с помощью алгоритма k-

средних. Количество кластеров обычно устанавливается в пределах от 16 до 256, а в качестве визуального слова $\mathbf{vw}_s$ выбирается центр масс s-го кластера.

Локальные признаки преобразуются в один глобальный вектор $\mathbf{V} \in \mathbf{R}^{S \times D}$ с помощью метода VLAD. Суть метода заключается в том, что для каждого локального дескриптора $\mathbf{x}_b$ находится ближайшее визуальное слово $NN(\mathbf{x}_b)$. После этого для каждого s-го визуального слова накапливается разница $\mathbf{v}_s$ между ним и ассоциированными с ним локальными признаками:

$$\mathbf{v}_s = \sum_{\mathbf{x}_b : NN(\mathbf{x}_b) = \mathbf{vw}_s} \frac{\mathbf{vw}_s - \mathbf{x}_b}{\left\| \mathbf{vw}_s - \mathbf{x}_b \right\|}. \quad (5)$$

После вычислений всех $\mathbf{v}_s$ они объединяются, формируя глобальный визуальный дескриптор $\mathbf{V}$ размером $S \times D$, который нормализуется в два этапа. Вначале модифицируется каждый элемент дескриптора: $V_z` = \text{sign}(V_z) \times |V_z|^{\gamma}$, где $\gamma \in (0, 1]$. Затем весь дескриптор нормализуются с помощью l_2-нормы. В случае использования нескольких цветовых пространств, предлагается формировать глобальные дескрипторы отдельно для каждого, после чего объединять их. Для снижения дальнейших вычислительных затрат, размерность полученного дескриптора сокращается с помощью метода главных компонент до Z элементов. При сравнении двух визуальных дескрипторов используется евклидово расстояние:

$$\mathrm{D_V}\left(\mathbf{V}_i, \mathbf{V}_j\right) = \sqrt{\sum_{z=1}^{Z} \left(V_z^i - V_z^j\right)^2}. \quad (6)$$

На втором этапе для каждого обучающего изображения также формируется текстовый дескриптор $\mathbf{T}_m = \{t_1, \ldots, t_N\}$, длина которого равна размеру словаря ключевых слов N, а элементы вычисляются с помощью статистической меры TF-IDF (Term Frequency – Inverse Document Frequency):

$$t_n^m = \frac{\delta(k_n \in K_m)}{|K_m|} \cdot \log\left(\frac{M}{\mathrm{F}(k_n)}\right), \quad (7)$$

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; $F(k_n)$ – количество обучающих изображений с ключевым словом k_n.

Сходство между двумя текстовыми дескрипторами определяется с помощью косинусной метрики:

$$\mathrm{D_T}\left(\mathbf{T}_i, \mathbf{T}_j\right) = \frac{\sum_{n=1}^{N} t_n^i \cdot t_n^j}{\sqrt{\sum_{n=1}^{N} \left(t_n^i\right)^2} \cdot \sqrt{\sum_{n=1}^{N} \left(t_n^j\right)^2}}. \quad (8)$$

Поскольку в обучающих выборках аннотации формируются вручную, то часть ключевых слов может отсутствовать. Для их восстановления предложен следующий метод. Пусть TS – набор обучающих изображений, в котором каждое изображение I_m имеет текстовый и визуальный дескрипторы, а

$TS_n \subseteq TS$, $n \in \{1, \ldots, N\}$ – набор, содержащий все изображения с ключевым словом k_n. Поскольку изображения набора 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.

При восстановлении ключевых слов изображения I_m, из каждой семантической группы TS_n с помощью уравнения (7) 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 $TS(I_m)$, выполняется оценка вероятности принадлежности каждого ключевого слова k_n изображению I_m:

$$P(k_n \mid I_m) = \sum_{(I_i, K_i) \in TS(I_m)} \mathrm{D}_T(\mathbf{T}_m, \mathbf{T}_i) \cdot \exp(-\mathrm{D}_V(\mathbf{V}_m, \mathbf{V}_i)) \cdot \delta(k_n \in K_i). \quad (9)$$

Для определения количества пропущенных ключевых слов 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 аннотация изображения I_m пополняется $Y(I_m)$ ключевыми словами с наибольшим значением вероятности $P(k_n|I_m)$. В противном случае, аннотации остаются без изменений. После восстановления пропущенных ключевых слов, текстовые дескрипторы изображений обновляются.

После вычисления для каждого обучающего изображения I_m визуального $\mathbf{V}_m$ и текстового $\mathbf{T}_m$ дескрипторов, они объединяются в один текстово-визуальный $\mathbf{VT}_m$. С их помощью на последнем этапе обучения системы ААИ обучающий набор изображений разделяется на однородные текстово-визуальные группы. 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: все изображения одной группы включают «характерные» ключевые слова (ключевые слова, встречающиеся в описании небольшого количества изображений) и изображения одной группы имеют существенное визуальное сходство. 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 с использованием текстово-визуальных дескрипторов.

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 $w_{i,j} = Nimg(k_i, k_j) / Nimg(k_i)$ определяется как отношение количества обучающих изображений, имеющих в описании ключевые слова k_i и k_j одновременно, к общему количеству обучающих изображений проаннотированных ключевым словом k_i.

Полученный орграф разделяется на группы с помощью алгоритма Louvain, 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. После этого каждое обучающее изображение присоединяется к той группе, для которой сумма значений соответствующих элементов текстового дескриптора наибольшая. Подобное разделение обучающей выборки позволяет с минимальными затратами получить инициализацию ОТВ-групп, а также обеспечивает стабильный результат кластеризации.

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

$$\mathrm{D_{VT}}\left(\mathbf{VT}_i, \mathbf{VT}_j\right) = \alpha \cdot \mathrm{D_T}\left(\mathbf{T}_i, \mathbf{T}_j\right) + (1-\alpha) \cdot \exp\left(-\mathrm{D_V}\left(\mathbf{V}_i, \mathbf{V}_j\right)\right), \qquad (10)$$

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

Чем больше значение $\mathrm{D_{VT}}(\mathbf{VT}_i, \mathbf{VT}_j)$, тем более похожи изображения I_i и I_j. Дескрипторы кластеризуются, используя модификацию расширенной самоорганизующейся инкрементальной нейронной сети (ESOINN, Enhanced Self-Organizing Incremental Neural Network), единственный слой которой постепенно подстраивается под структуру входных данных, определяя количество кластеров и их топологию. Модифицированный алгоритм ESOINN включает следующие шаги:

1. Структура нейронной сети инициализируется с помощью первичного разделения обучающей выборки. Для этого из каждой сформированной группы выбирается по два дескриптора, с помощью которых формируются узлы сети $r_i \in R$. Узлы, принадлежащие одной группе, соединяются связями.

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

3. Определяются два ближайших узла сети (победитель и второй победитель) с помощью формулы (11). 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.

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

$$s_i = \min_{j \in R_i} \mathrm{D}_{\mathrm{VT}}(\mathbf{W}_i, \mathbf{W}_j), \tag{11}$$

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

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

$$s_i = \max_{j \in R \setminus \{i\}} \mathrm{D}_{\mathrm{VT}}(\mathbf{W}_i, \mathbf{W}_j). \tag{12}$$

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

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

$$p_{win,m} = \frac{1}{\left(1 + \overline{d_{win,m}}\right)^2}. \tag{13}$$

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, а остальным присваиваются 0. Суммарная плотность узла h_{win} определяется следующим образом:

$$h_{win} = \frac{1}{q} \cdot \sum_{dp=1}^{Q} p_{win,dp}, \tag{14}$$

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:

$$\Delta\mathbf{W}_{win} = \frac{1}{U_{win}} \cdot (\mathbf{TV} - \mathbf{W}_{win}), \qquad \Delta\mathbf{W}_j = \frac{1}{100 \cdot U_{win}} \cdot (\mathbf{TV} - \mathbf{W}_j),\ j \in R_{win}. \tag{15}$$

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

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

$$h_i < b_o \cdot \sum_{j=1}^{|R|} \frac{h_j}{|R|}, \qquad (16)$$

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 λ / 10 входных дескрипторов, ассоциированных с этим узлом.

10. Если процесс кластеризации закончен (на вход сети поданы все дескрипторы), 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. 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.

На этапе аннотирования нового изображения априорная вероятность $P(H_l)$ определяется как отношение количества обучающих изображений в ОТВ-группе H_l к общему количеству обучающих изображений, а процесс оценки вероятностей $P(A|H_l)$ из уравнения (1) включает следующие шаги:

1. Для тестового изображения A с помощью уравнения (7) выбрать из всей обучающей выборки YN_H 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:

$$P(A \mid H_l) = \exp\left(-\mathrm{D_V}\left(\mathbf{V}_A, \mathbf{HV}_l\right)\right), \qquad (17)$$

где $\mathbf{V}_A$ – визуальный дескриптор тестового изображения A; $\mathbf{HV}_l$ – визуальный дескриптор наиболее похожего изображения из 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 (7) выбирается YN_{AN} изобра-

жений с минимальным расстоянием, формирующих набор $H_l(A)$. Используя его, вероятность $P_l(A|k_n)$ оценивается следующим образом:

$$P_l(A \mid k_n) = \sum_{(\mathbf{v}_i, K_i) \in H_l(A)} \exp\left(-\mathrm{D_V}(\mathbf{V}_A, \mathbf{V}_i)\right) \cdot \delta(k_n \in K_i). \quad (18)$$

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:

$$P_l(k_n) = \exp\left(-\frac{Nimg(k_n)}{M}\right). \quad (19)$$

После подстановки полученных оценок условных вероятностей в уравнение (11) и сортировки результатов формируется ранжированный список ключевых слов. В качестве аннотации нового изображения выбирается *Nkw* первых ключевых слов. Значение *Nkw* равно среднему количеству ключевых слов в описании обучающих изображений, выбранных из ОТВ-группы с наибольшей условной вероятностью $P(A|H_l)$.

Следует отметить, что в разработанном методе ОТВ-группы могут уточняться с помощью новых обучающих изображений в течение жизненного цикла системы ААИ. Также 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 сходство между новым изображением и ОТВ-группой вычисляется с помощью текстово-визуального дескриптора и уравнения (11).

Третья глава посвящена вопросам разработки приложения для автоматического аннотирования изображений на основе разработанных методов и алгоритмов. Приведены результаты экспериментальных исследований, подтверждающие эффективность предложенных методов и алгоритмов для задачи ААИ. Экспериментальные исследования разделены на две группы: вычисление визуальных дескрипторов и автоматическое аннотирование изображений на основе однородных текстово-визуальных групп.

Для проверки предложенных методов и алгоритмов вычисления визуальных дескрипторов использовался набор изображений из 8 категорий сцен OT8 (URL: http://people.csail.mit.edu/torralba/code/spatialenvelope). OT8 состоит из 2688 изображений, размер каждого изображения 256 × 256 пикселов. Тестирование заключалось в оценке точности категоризации изображений с помощью локального дескриптора FD-SUF. 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 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.

Вначале 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, коэффициент нормализации $\gamma = 1$, а длина глобального дескриптора не изменялась ($Z = 8192$). Вычисления производились с использованием одного процессорного ядра. Результаты сравнения приведены в таблице 3.

Таблица 3

Сравнение локальных дескрипторов при категоризации изображений

Тип локального дескриптора	Среднее количество вычисленных дескрипторов	Среднее время вычисления дескрипторов, мс	Средняя точность категоризации, %
SURF (Гессиан)	207	12,35	71,75
G-SURF (Гессиан)	207	15,78	68,55
SURF (Сетка)	2304	57,26	82,62
G-SURF (Сетка)	2304	96,35	82,36
FD-SUF	2304	17,18	84,65

Как видно из приведенных в таблице 3 данных, 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.

Были проведены дополнительные эксперименты для исследования влияния размера словаря визуальных слов S и величины коэффициента нормализации γ на точность категоризации изображений, а также для определения оптимальной длины глобального дескриптора Z. В результате этих исследований были выбраны следующие параметры: $S = 128$, $\gamma = 0{,}3$ и $Z = 384$. Используя эти значения, были проведены эксперименты для оценки точности определения отдельных категорий изображений с помощью цветовых локальных дескрипторов. Полученные результаты представлены в таблице 4.

Таблица 4

Результаты точности определения отдельных категорий изображений (%)

Цветовое пространство (компонент)	Coast	Forest	Highway	Inside city	Mountain	Open country	Street	Tall building	Средняя точность
Y	90,00	94,30	84,79	93,27	91,73	75,81	89,06	90,23	88,65
nRGB	**90,13**	93,86	**85,63**	**96,15**	**94,16**	**81,29**	85,42	89,84	89,56
HSV	89,23	**95,61**	80,63	90,38	85,04	74,19	88,54	87,89	86,44
LUV	86,54	94,30	85,00	88,94	88,32	76,45	**90,63**	**91,41**	87,70
nRGB+HSV+LUV	91,38	96,61	87,25	94,27	93,70	82,61	94,23	93,58	91,70

Как видно из таблицы 4, разные цветовые дескрипторы показывают хорошие результаты только для определенных категорий изображений. Таким образом, комбинируя дескрипторы, вычисленные в цветовых пространствах nRGB, HSV и LUV, можно повысить точность категоризации на 2,1 % и 3 % по сравнению с дескрипторами, вычисленными только для nRGB и оттенков серого (Y) соответственно.

Для сравнения разработанного метода автоматического аннотирования изображений с существующими методами ААИ, использовалась база изображений IAPR TC-12 (URL: http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz). Эта база содержит 19627 изображений размером 480 × 360 пикселов, каждое из которых описано несколькими ключевыми словами. Некоторые характеристики базы представлены в таблице 5.

Таблица 5

Статистические данные базы изображений IAPR TC-12

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

Оценка эффективности заключалась в вычислении трех параметров: средней точности (precision), средней полноты (recall) и количества использованных ключевых слов (N+):

$$precision = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{AA(k_n)}, \qquad recall = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{GT(k_n)}, \tag{20}$$

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.

При вычислении визуальных дескрипторов использовались значения параметров, выбранные на первом этапе экспериментов. Также для работы предложенного метода были выбраны следующие параметры:

- Для восстановления пропущенных ключевых слов из каждой семантической группы выбиралось по YN_{WL} = 1 изображению, а для определения их количества использовалось YA_{WL} = 2 изображения.
- При формировании ОТВ-групп текстовый дескриптор изображения имеет большее значение (α = 0,75), а параметры сети ESOINN устанавливаются следующими: λ = 50, age_{max} = 25, b_1 = 0,0, b_2 = 0,1, b_3 = 1,0.
- При аннотировании новых изображений для определения ОТВ-групп использовалось YN_H = 200 изображений, а из каждой группы выбиралось по YN_{AN} = 5 изображений.

В таблице 6 приведены полученные числовые оценки эффективности предложенного метода. Оценки для существующих методов ААИ взяты из соответствующих статей.

Таблица 6

Оценка эффективности разных методов ААИ

Метод	Точность, %	Полнота, %	N+
TagProp with Metric-Learning (TagProp-ML)	48	25	227
TagProp with Metric-Learning and Logistic Discriminant Models (TagProp-σML)	46	35	266
FastTag	47	26	280
2-Pass K-Nearest Neighbor (2PKNN)	49	32	274
Support Vector Machine and Discrete Multiple Bernoulli Relevance Model (SVM-DMBRM)	56	29	283
Предлагаемая реализация без восстановления ключевых слов обучающих изображений	60	33	265
Предлагаемая реализация	62	34	267

Из таблицы 6 видно, что предложенный метод эффективнее современного метода SVM-DMBRM по точности аннотаций на 6 % и полноте на 5 %.

В заключении сформулированы основные результаты и выводы, полученные в диссертационной работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

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 со-

бой 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.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях, рекомендованных ВАК:

1. **Проскурин А.В.** Автоматическое аннотирование ландшафтных изображений // Вестник Сибирского государственного аэрокосмического университета. 2014. Вып. 3(55). С. 120–125.

2. **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.

3. **Проскурин А.В.**, Фаворская М.Н. Категоризация сцен на основе расширенных цветовых дескрипторов // Труды СПИИРАН. 2015. № 40. С. 203–220.

4. **Проскурин А.В.**, Фаворская М.Н. Автоматическое аннотирование изображений на основе однородных текстово-визуальных групп // Информационно-управляющие системы. 2016. № 2. С. 11–18.

В изданиях, индексируемых Scopus:

5. 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.

6. 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.

Авторские свидетельства на программы:

7. **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.

8. **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.

Другие издания:

9. **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.

10. **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.

11. **Проскурин А.В.** Алгоритм вычисления визуальных дескрипторов для восстановления ключевых слов обучающих изображений // Материалы 19-й международной конференции «Цифровая обработка сигналов и ее применение» (DSPA-2017). М.: Изд-во РНТОРЭС им. А.С. Попова, 2017. Т. 2. С. 719–724.