УДК 004.932.2
doi:10.15217/issn1684-8853.2016.2.11

АВТОМАТИЧЕСКОЕ АННОТИРОВАНИЕ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ОДНОРОДНЫХ ТЕКСТОВО-ВИЗУАЛЬНЫХ ГРУПП

А. В. Проскурин[a], *аспирант*
М. Н. Фаворская[a], *доктор техн. наук, профессор*
[a]*Сибирский государственный аэрокосмический университет им. академика М. Ф. Решетнёва, Красноярск, РФ*

Постановка проблемы: *задача автоматического аннотирования изображений нетривиальна: часто обучающие наборы несбалансированы и содержат неполные аннотации, а между визуальными признаками и текстовым описанием изображения наблюдается семантический разрыв. Существующие методы решают эти проблемы, используя для аннотации нового изображения все обучающие изображения и ключевые слова, в том числе заведомо нерелевантные, что потенциально снижает точность и требует лишних вычислений. При этом используются визуальные признаки большой размерности, что также неэффективно в вычислительном плане. В связи с этим возникает необходимость разработки компактного визуального дескриптора и метода аннотирования тестового изображения с помощью небольшой группы наиболее информативных обучающих изображений.* ***Результаты:*** *разработана методика автоматического аннотирования изображений, основанная на поиске апостериорной вероятности ассоциации ключевого слова с визуальным дескриптором изображения. Получены шесть глобальных дескрипторов, объединенных в один дескриптор, размер которого уменьшен с помощью метода главных компонент до нескольких сотен элементов. Проведенные экспериментальные исследования показали улучшение точности аннотирования на 7 % и отклика на 1 %.* ***Практическая значимость:*** *разработанный компактный визуальный дескриптор и метод автоматического аннотирования изображений на основе формирования однородных текстово-визуальных групп может быть использован в информационно-поисковых системах в сети Интернет для повышения эффективности поиска изображений.*

Ключевые слова *— автоматическое аннотирование изображений, глобальный визуальный дескриптор, текстово-визуальные группы.*

Введение

Поиск изображений в сети Интернет является распространенной функцией, реализация которой в значительной степени полагается на наличие текстового описания. Однако из-за стремительного роста количества изображений аннотирование вручную становится невозможным, а описания, полученные из текста, окружающего изображение на веб-странице, не всегда соответствуют действительности. В связи с этим становится актуальной разработка систем автоматического аннотирования изображений (ААИ), в которых на основе большого количества заранее проаннотированных изображений (обучающих изображений) определяется семантическая модель, автоматически присваивающая новому изображению текстовое описание в виде ключевых слов.

В последние десятилетия были предложены различные подходы к формированию ААИ, наиболее эффективный из которых основан на использовании метода ближайших соседей [1–4]. С его помощью для тестового изображения определяются визуально похожие обучающие изображения (ближайшие соседи), после чего аннотация генерируется путем перемещения ключевых слов от ближайших соседей к аннотируемому изображению. Однако при этом часто наблюдается проблема семантического разрыва — отсутствие связи между визуальными признаками изображения и его интерпретацией человеком. Также на точность аннотирования сильно влияет несбалансированность обучающей выборки (огромная разница в частоте встречаемости разных ключевых слов) и наличие неполных аннотаций (изображения аннотированы не всеми релевантными ключевыми словами).

Для решения этих проблем был предложен [3] двухпроходный метод K ближайших соседей (2PKNN — 2-Pass K-Nearest Neighbor), в котором для каждого ключевого слова формировалась группа обучающих изображений, соответствующих этому ключевому слову. Из каждой группы выбиралось одинаковое количество наиболее похожих изображений, после чего они использовались для формирования аннотации. Данный метод демонстрирует существенное улучшение полноты аннотирования (изображение описывается с помощью большего количества разных релевантных ключевых слов), однако использование заведомо нерелевантных слов снижает общую точность аннотирования. Также в работе [3] предложен алгоритм вычисления весов, отражающих значимость разных визуальных дескрипторов при сравнении двух изображений. Этот алгоритм повышает эффективность работы метода 2PKNN, однако используемые глобальные признаки требуют значительных вычислительных затрат.

В связи с этим в данной статье предлагается расширенный метод 2PKNN, основанный на однородных текстово-визуальных группах (ОТВ-группах; 2PKNN-HTVG — 2PKNN based on Homogeneous Textual-Visual Groups), в котором точность и быстродействие метода 2PKNN улучшается с помощью предварительного разделения обучающих изображений на ОТВ-группы, лишь небольшое число которых используется для аннотирования тестового изображения. Также для описания изображения предложен компактный глобальный визуальный дескриптор, описывающий характеристики как сцены, так и объектов на изображении.

Математическая модель автоматического аннотирования изображений

Любой метод ААИ предполагает наличие обучающего набора TS, состоящего из изображений и соответствующих им текстовых описаний. Пусть $J = \{I_1, ..., I_M\}$ — коллекция изображений, а $K = \{k_1, ..., k_N\}$ — словарь, состоящий из N ключевых слов, тогда обучающий набор $TS = \{(I_1, K_1), ..., (I_M, K_M)\}$, где $K_M \subseteq K$. Предположим, что обучающий набор разделен на несколько непересекающихся ОТВ-групп $H = \{H_1, ..., H_L\}$, а выбор ключевых слов в процессе аннотирования тестового изображения A зависит от ассоциации изображения с той или иной группой. Обозначим вероятность ассоциации изображения A с ОТВ-группой H_l как $P(H_l|A)$. Так же как и в работе [3], введем условную вероятность $P_l(A|k_n)$ оценки распределения визуального дескриптора изображения для ключевого слова k_n внутри группы H_l. В этом случае аннотирование изображения моделируется как проблема поиска апостериорной вероятности:

$$P(k_n|A) = \sum_{H_l \in H} \left(P(H_l|A) \frac{P_l(A|k_n) P_l(k_n)}{P(A)} \right), \quad (1)$$

где $P_l(k_n)$ — априорная вероятность ключевого слова k_n внутри ОТВ-группы H_l. Поскольку вероятность $P(A)$ является константой, то для упрощения в дальнейшем она не будет учитываться.

Наилучшее ключевое слово для тестового изображения A определяется с помощью следующей формулы:

$$k^* = \arg\max_n P(k_n|A). \quad (2)$$

В качестве аннотации используется O ключевых слов с наибольшей вероятностью $P(k_n|A)$. Значение O выбирается равным среднему значению ключевых слов в описании обучающих изображений ОТВ-группы с наибольшим значением $P(H_l|A)$. Таким образом, для аннотирования тестового изображения A необходимо оценить вероятности $P(H_l|A)$, $P_l(A|k_n)$ и $P_l(k_n)$. Для этого предлагается метод ААИ, состоящий из двух частей: формирования из обучающего набора ОТВ-групп и использования метода 2PKNN внутри каждой группы.

Формирование однородных текстово-визуальных групп

На первом этапе алгоритма обучения системы ААИ необходимо разделить обучающий набор изображений на ОТВ-группы. Идея заключается в том, что обучающие изображения одной ОТВ-группы формируют контекст для аннотируемого изображения — 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, но 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);

— 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.

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 по формуле

$$w_{i,j} = \frac{N(k_i, k_j)}{N(k_i)}, \quad (3)$$

где $N(k_i, k_j)$ — количество обучающих изображений, имеющих в описании ключевые слова k_i и k_j одновременно; $N(k_i)$ — количество обучающих

изображений, проаннотированных ключевым словом k_i.

Полученный орграф разделяется на группы с помощью быстрого алгоритма [5], показывающего хорошие результаты при небольших вычислительных затратах. 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, ключевые слова которой чаще встречаются в текстовом описании изображения. Подобное разделение обучающей выборки позволяет с минимальными затратами получить инициализацию ОТВ-групп, а также показывает более стабильный результат кластеризации.

Кластеризация обучающих изображений

Для последующей кластеризации обучающей выборки в ОТВ-группы необходимо каждое изображение I_m представить в виде текстово-визуального дескриптора $\mathbf{TV}_m = (\mathbf{T}_m, \mathbf{V}_m)$, где $\mathbf{T}_m = \{t_1, ..., t_N\}$ — текстовый дескриптор, а $\mathbf{V}_m = \{v_1, ..., v_Z\}$ — глобальный визуальный дескриптор. Длина текстового дескриптора равна размеру словаря ключевых слов, а его элементы вычисляются с помощью статистической меры TF-IDF (Term Frequency — Inverse Document Frequency):

$$t_n^m = \frac{\delta(k_n \in K_m)}{F(k_n)}, \quad (4)$$

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); $F(k_n)$ — частота встречаемости ключевого слова k_n в обучающей выборке.

Вычисление визуального дескриптора будет подробно рассмотрено ниже. При сравнении двух изображений I_i и I_j с помощью их текстово-визуальных дескрипторов сходство вычисляется по формуле

$$D(\mathbf{TV}_i, \mathbf{TV}_j) = \\ = \alpha D_T(\mathbf{T}_i, \mathbf{T}_j) + (1-\alpha)\exp\left(-D_V(\mathbf{V}_i, \mathbf{V}_j)\right); \quad (5)$$

$$D_T(\mathbf{T}_i, \mathbf{T}_j) = \frac{\sum_{n=0}^{N} \min\left(t_n^i, t_n^j\right)}{\sqrt{\sum_{n=0}^{N} t_n^i \sum_{n=0}^{N} t_n^j}}; \quad (6)$$

$$D_V(\mathbf{V}_i, \mathbf{V}_j) = \sqrt{\sum_{z=0}^{Z} \left(v_z^i - v_z^j\right)^2}, \quad (7)$$

где $D_T(\cdot)$ — косинусная метрика для сравнения текстовых дескрипторов; $D_V(\cdot)$ — евклидово расстояние между визуальными дескрипторами; α — эмпирический коэффициент, изменяющийся в пределах [0, 1].

Чем больше значение $D(\mathbf{TV}_i, \mathbf{TV}_j)$, тем более схожи изображения I_i и I_j. Полученные дескрипторы кластеризуются с использованием модификации расширенной самоорганизующейся инкрементальной нейронной сети (ESOINN — Enhanced Self-Organizing Incremental Neural Network) [6], 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 путем 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 **TV**.

3. Определяются два ближайших узла сети (победитель и второй победитель) с помощью формулы (5). Если расстояние между входным дескриптором и победителем или вторым победителем больше соответствующих порогов подобия, то входной дескриптор вставляется в сеть как первый узел нового класса, а алгоритм переходит к шагу 2 для получения нового дескриптора.

Поскольку распределение входных данных заранее неизвестно, то порог подобия s_i обновляется для каждого узла в отдельности по формуле

$$s_i = \min_{j \in R_i} D(\mathbf{W}_i, \mathbf{W}_j), \quad (8)$$

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, порог подобия вычисляется с помощью всех узлов сети:

$$s_i = \max_{j \in R \setminus \{i\}} D(\mathbf{W}_i, \mathbf{W}_j). \quad (9)$$

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_i вычисляется с помощью среднего расстояния $\overline{d_i}$ от узла до его соседей:

$$p_i = \frac{1}{\left(1+\overline{d_i}\right)^2}. \quad (10)$$

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:

$$h_i = \frac{1}{q}\sum_{Q}\sum_{\lambda} p_i, \qquad (11)$$

где q — количество периодов, в которые плотность узла r_i больше 0; Q — количество прошедших периодов обучения (можно вычислить как $Q = M/\lambda$, где M — общее количество входных дескрипторов); λ — число, обозначающее период обучения сети.

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

$$\Delta \mathbf{W}_{win} = \frac{1}{U_{win}}\left(\mathbf{TV} - \mathbf{W}_{win}\right); \qquad (12)$$

$$\Delta \mathbf{W}_j = \frac{1}{100U_{win}}\left(\mathbf{TV} - \mathbf{W}_j\right), j \in R_{win}. \qquad (13)$$

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 в целях обнаружения перекрывающихся областей, после чего из нейронной сети удаляются узлы, являющиеся шумами. Такими считаются узлы r_i, имеющие двух или меньше топологических соседей и удовлетворяющие условию следующего вида:

$$h_i < b_o \sum_{j=1}^{R} \frac{h_j}{R}, \qquad (14)$$

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

10. 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 «скелетом» ОТВ-групп. Вначале для каждого изображения определяется ближайший узел сети с помощью только текстового дескриптора, 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 при помощи визуальных дескрипторов. Пример одной из ОТВ-групп, сформированной посредством базы изображений IAPR-TC12 [7], представлен на рис. 1.

Процесс оценки вероятности $P(H_l|A)$ из уравнения (1) включает следующие шаги.

1. Определяется расстояние $ds(H_l, A)$ между изображением A и ОТВ-группой H_l. Для этого с помощью уравнения (7) вычисляется расстояние между A и всеми узлами l-го кластера ESOINN и выбирается наименьшее из них.

■ ***Рис. 1.*** Пример изображений однородной текстово-визуальной группы

2. Вычисляется диаметр $da(H_{nn})$ ближайшей к изображению A ОТВ-группы H_{best} как максимальное расстояние между любыми двумя обучающими изображениями группы.

3. Оцениваются условные вероятности $P(H_1|A)$ с помощью формулы

$$P(H_l \mid A) = \begin{cases} \exp\left(-ds(H_l, A)\right), \\ \qquad \text{если } ds(H_l, A) \le da(H_{nn}). \\ 0 \text{ иначе} \end{cases} \quad (15)$$

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

Следует отметить, что эффективность предложенного метода может быть повышена, если вместе с тестовым изображением будут предоставлены некоторые ключевые слова, полученные от пользователей. В этом случае расстояние между новым изображением и ОТВ-группой вычисляется с помощью текстово-визуального дескриптора и уравнения (5).

После разделения обучающего набора на ОТВ-группы каждая из них используется в качестве исходных данных для метода 2PKNN. Пусть J — набор обучающих изображений ОТВ-группы H_l, а $J_n \subseteq J$, $n \in \{1, ..., N\}$ — набор, содержащий все изображения группы, имеющие в описании ключевое слово k_n. Поскольку изображения набора J_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. Следует отметить, что ОТВ-группы имеют ограниченный набор ключевых слов и, таким образом, некоторые семантические группы могут быть пустыми.

При аннотировании тестового изображения A из каждой семантической группы J_n с помощью уравнения (7) выбирается Y наиболее похожих изображений, формирующих набор $J_{A,n}$. Таким образом, каждый набор $J_{A,n}$ содержит изображения, наиболее информативные при оценке вероятности принадлежности ключевого слова k_n тестовому изображению A. В связи с этим в оценке вероятности $P_l(A|k_n)$ участвуют только изображения из набора $J_{A,n}$:

$$P_l(A \mid k_n) = \sum_{I_i \in J_{A,n}} \exp\left(-D_V(A, I_i)\right). \quad (16)$$

Полученные условные вероятности нормализуются для того, чтобы их сумма равнялась 1. Поскольку для оценки вероятности каждого ключевого слова используется одинаковое количество изображений, то априорная вероятность $P_l(k_n)$ в уравнении (1) одинакова для всех ключевых слов:

$$P_l(k_n) = \frac{1}{N(H_l)}, \quad (17)$$

где $N(H_l)$ — количество уникальных ключевых слов в текстовом описании изображений ОТВ-группы H_l.

Вычисление глобального визуального дескриптора

В эффективности работы предложенного метода ААИ большое значение имеет точность представления изображений в виде визуальных дескрипторов. Наиболее успешные подходы, предложенные для решения этой проблемы, включают три шага: извлечение из изображений локальных признаков (таких как SIFT [8], SURF [9] и т. д.); формирование словаря визуальных слов; кодирование локальных признаков для формирования глобального дескриптора (например, методами SC [10], LLC [11], VLAD [12]). Рассмотрим их подробнее.

1. На первом этапе изображения описываются с помощью набора локальных признаков $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_B\}$, где $\mathbf{x}_i \in \mathbf{R}^D$. В работе [13] для классификации изображений был предложен метод быстрого вычисления локальных признаков FDG-SUF, состоящий из двух этапов: вычисления матрицы частей локальных дескрипторов **DS** и построения с ее помощью набора локальных дескрипторов. На первом этапе все изображение разделяется сеткой на ячейки размером 5×5 пикселов. После этого в каждой из ячеек вычисляются части дескриптора G-SURF [14], которые сохраняются в матрицу **DS**. На втором этапе по матрице **DS** перемещается скользящее окно размером 4×4 ячейки. Каждый локальный дескриптор представляет собой объединение частей дескриптора, попавших в скользящее окно. Таким образом, изменяя шаг смещения скользящего окна, можно существенно увеличить количество локальных дескрипторов, извлеченных из изображения, без значительных вычислительных затрат.

2. На следующем шаге формируется словарь визуальных слов $VW = \{\mathbf{vw}_1, ..., \mathbf{vw}_S\}$, где $\mathbf{vw}_i \in \mathbf{R}^D$. Для этого с помощью алгоритма k-средних выбранные случайным образом локальные признаки кластеризуются. Количество кластеров обычно устанавливается в пределах от 16 до 256. Центр масс кластера выбирается в качестве визуального слова $\mathbf{vw}_i$.

3. По сформированному словарю локальные признаки кодируются в один глобальный вектор $\mathbf{C} \in \mathbf{R}^{S \times D}$ с помощью алгоритма VLAD (Vector of Locally Aggregated Descriptors) [12]. Суть метода заключается в том, что для каждого локального признака $\mathbf{x}_i$ находится ближайшее визуальное слово $NN(\mathbf{x}_i)$. После этого для каждого визуального слова накапливается разница $\mathbf{c}_j$ между ним и ассоциированными с ним локальными

признаками. В отличие от оригинального алгоритма, в котором для вычисления разницы используется выражение $NN(\mathbf{x}_i) - \mathbf{x}_i$, в этой работе вклад каждого локального вектора уравнивается:

$$\mathbf{c}_j = \sum_{\mathbf{x}_i : NN(\mathbf{x}_i)=\mathbf{vw}_j} \frac{\mathbf{vw}_j - \mathbf{x}_i}{\left\|\mathbf{vw}_j - \mathbf{x}_i\right\|_2}. \quad (18)$$

После вычислений всех $\mathbf{c}_j$ они нормализуются с помощью L2-нормы и объединяются, формируя глобальный дескриптор размером $S \times D$.

Полученный дескриптор показывает хорошие результаты при классификации изображений по типу сцены. Для вычисления дескриптора, описывающего объекты на изображении, используется такой же алгоритм, однако на первом этапе локальные признаки извлекаются только в особых точках, полученных с помощью матриц Гессе [14].

В работе [15] было показано, что некоторые локальные цветовые дескрипторы, имеющие высокую инвариантность к изменениям интенсивности цветов, могут повысить точность классификации изображений. В этой работе в качестве локальных дескрипторов используются G-SURF, OppG-SURF и RGBG-SURF, вычисленные на изображениях в оттенках серого, цветовых пространствах Opponent и нормализованном RGB соответственно.

Таким образом, каждое изображение описывается с помощью шести глобальных визуальных дескрипторов. Для снижения дальнейших вычислительных затрат все дескрипторы объединяются в один, после чего его размерность сокращается по методу главных компонент (PCA) [16].

Результаты экспериментальных исследований

Для экспериментов использовалась база изображений IAPR TC-12 [7], содержащая 19 627 изображений размером 480×360 пикселов, каждое из которых описано несколькими предложениями. В работе [1] для базы был предложен словарь из 291 ключевого слова, состоящий из наиболее часто встречающихся существительных. Для обучения используется 17 665 изображений, остальные 1962 изображения применяются для тестирования.

С помощью этих изображений проводилось сравнение предложенного метода 2PKNN-HTVG с существующими методами ААИ. Оценка эффективности заключалась в вычислении четырех параметров: средней точности (precision), среднего отклика (recall), F1-меры и количества ключевых слов с положительным откликом (N+):

$$\text{precision} = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{AA(k_n)}; \quad (19)$$

$$\text{recall} = \frac{1}{N}\sum_{n=1}^{N}\frac{CA(k_n)}{GT(k_n)}; \quad (20)$$

$$F1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (21)$$

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.

Все вычисления повторялись 5 раз, после чего выбирался лучший результат. Для работы метода 2PKNN-HTVG на разных этапах были установлены следующие параметры.

1. Для формирования каждого словаря визуальных слов из обучающего набора случайным образом выбиралось 200 000 локальных признаков, которые с помощью алгоритма k-средних кластеризовались в 128 кластеров (визуальных слов).

2. Объединенный глобальный дескриптор уменьшался с помощью метода PCA до 256 элементов.

3. При формировании ОТВ-групп текстовый дескриптор изображения имел большее значение ($\alpha = 0{,}75$), а параметры сети ESOINN устанавливались следующими: $\lambda = 50$; $age_{\max} = 25$; $b_1 = 0{,}0$; $b_2 = 0{,}1$; $b_3 = 1{,}0$.

4. Количество изображений Y, выбираемых из семантических групп в методе 2PKNN, равнялось двум.

Примеры аннотирования некоторых изображений показаны на рис. 2, *а–в*.

В таблице приведены полученные числовые оценки эффективности предложенного метода 2PKNN-HTVG. Оценки для существующих методов ААИ взяты из соответствующих статей.

■ Оценка эффективности разных методов ААИ

Метод	Точность, %	Отклик, %	F1-мера	N+
MBRM [17]	24	23	23,5	223
JEC [1]	28	29	28,5	250
GS [18]	32	29	30,4	252
TagProp(ML) [2]	48	25	32,9	227
TagProp(σML) [2]	46	35	39,8	266
FastTag [4]	47	26	33,5	**280**
2PKNN [3]	49	32	38,7	274
2PKNN-ML [3]	54	37	43,9	278
2PKNN-HTVG	**61**	**38**	**46,8**	271

а)				
б)	bush, front, rock, sign, wood	building, night, reflection, river, tree, water	boat, building, city, cloud, sky, skyline	board, car, fence, grey, people, racetrack, racing, spectator
в)	bush, sign, rock, wood, front	building, night, river, tree, reflection	boat, building, city, cloud, sky	board, car, fence, people, racetrack, racing, spectator

■ ***Рис. 2.*** Примеры изображений из базы IAPR-TC12: *а* — исходные изображения; *б* — достоверные аннотации изображений; *в* — аннотации, полученные с помощью метода 2PKNN-HTVG

Анализ результатов, приведенных в таблице, показывает, что предложенный метод эффективнее оригинального алгоритма 2PKNN по точности аннотирования на 7 %. При этом отклик увеличился на 1 % за счет более точного подбора количества ключевых слов в описании аннотируемых изображений.

Заключение

В статье представлен метод автоматического аннотирования изображений, основанный на разделении обучающего набора изображений на ОТВ-группы и аннотировании нового изображения с помощью обучающих изображений небольшого количества визуально похожих ОТВ-групп. Это позволяет при аннотировании сузить поиск наиболее информативных обучающих изображений и тем самым повысить быстродействие и точность аннотирования. Также представлен алгоритм вычисления компактного глобального визуального дескриптора, описывающего как сцену, так и объекты на изображении. Проведенные экспериментальные исследования показали, что использование предложенного метода повышает точность аннотирования на 7 %, а более точный подбор количества ключевых слов в описании увеличивает отклик на 1 %. Следует отметить, что качество аннотирования может быть повышено с помощью предоставленных пользователем неполных аннотаций.

Литература

1. **Makadia A., Pavlovic V., Kumar S.** A New Baseline for Image Annotation // Proc. 10th European Conf. on Computer Vision, Marseille, France, 2008. Vol. 5304. P. 316–329.
2. **Guillaumin M., Mensink T., Verbeek J., Schmid C.** TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation // Proc. IEEE 12th Intern. Conf. on Computer Vision, Kyoto, Japan, 2009. P. 309–316.
3. **Verma Y., Jawahar C. V.** Image Annotation Using Metric Learning in Semantic Neighbourhoods // Proc. 12th European Conf. on Computer Vision, Florence, Italy, 2012. Vol. 7574. P. 836–849.
4. **Chen M., Zheng A., Weinberger K. Q.** Fast Image Tagging // Proc. 30th Intern. Conf. on Machine Learning, Atlanta, USA, 2013. P. 1274–1282.
5. **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. Vol. 2008. P10008.
6. **Shen F., Ogura T., Hasegawa O.** An Enhanced Self-Organizing Incremental Neural Network for Online Unsupervised Learning // Neural Networks. 2007. Vol. 20(8). P. 893–903.
7. IAPR TC-12 Benchmark. http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz (дата обращения: 22.02.2016).
8. **Lowe D. G.** Distinctive Image Features from Scale-Invariant Keypoints // Intern. Journal of Computer Vision. 2004. Vol. 60(2). P. 91–110.
9. **Bay H., Ess A., Tuytelaars T., Gool L. V.** Speeded-Up Robust Features (SURF) // Computer Vision and Image Understanding. 2008. Vol. 110(3). P. 346–359.
10. **Yang J., Yu K., Gong Y., Huang T.** Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification // Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Miami, USA, 2009. P. 1794–1801.
11. **Wang J.**, et al. Locality-Constrained Linear Coding for Image Classification/ J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong // Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, USA, 2010. P. 3360–3367.

12. **Jegou H., Douze M., Schmid C., Perez P.** Aggregating Local Descriptors into a Compact Image Representation // Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, USA, 2010. P. 3304–3311.
13. **Проскурин А. В.** Быстрый локальный дескриптор для категоризации изображений по типу сцены // Решетневские чтения: материалы XIX Междунар. науч.-практ. конф., Красноярск, 10–14 ноября 2015 г. Красноярск, 2015. Т. 2. С. 243–245.
14. **Alcantarilla P. F., Bergasa L. M., Davison A. J.** Gauge-SURF Descriptors // Image and Vision Computing. 2013. Vol. 31(1). P. 103–116.
15. **Favorskaya M., Proskurin A.** Image Categorization Using Color G-SURF Invariant to Light Intensity // Procedia Computer Science. 2015. Vol. 60. P. 681–690.
16. **Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
17. **Feng S., Manmatha R., Lavrenko V.** Multiple Bernoulli Relevance Models for Image and Video Annotation // Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Washington, USA, 2004. Vol. 2. P. 1002–1009.
18. **Zhang S., Huang J., Li H., Metaxas D. N.** Automatic Image Annotation and Retrieval Using Group Sparsity // IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics. 2012. Vol. 42(3). P. 838–849.

UDC 004.932.2
doi:10.15217/issn1684-8853.2016.2.11

Automatic Image Annotation Based on Homogeneous Textual-Visual Groups

Proskurin A. V.[a], Post-Graduate Student, Proskurin.AV.WOF@gmail.com
Favorskaya M. N.[a], Dr. Sc., Tech., Professor, favorskaya@sibsau.ru
[a]Siberian State Aerospace University named after academician M. F. Reshetnev, 31, Krasnoyarsky Rabochy St., 660037, Krasnoyarsk, Russian Federation

Purpose: The problem of automatic image annotation is not trivial. The training images often contain unbalanced and incomplete annotations, leading to a semantic gap between the visual features and textual description of an image. The existing methods include computationally complex algorithms which optimize the visual features and annotate a new image using all the training images and keywords, potentially reducing the accuracy. A compact visual descriptor should be developed, along with a method for choosing a group of the most informative training images for each test image. **Results:** A methodology for automatic image annotation is formulated, based on searching for a posteriori probability keyword association with a visual image descriptor. Six global descriptors combined in a single descriptor were obtained. The size of this single descriptor was reduced down to several hundred elements using principal component analysis. The experimental results showed an improvement of the annotation precision by 7% and a recall by 1%. **Practical relevance:** The compact handle visual method and automatic annotation of images based on the formation of homogeneous textual-visual groups can be used in Internet retrieval systems to improve the image search quality.
Keywords — Automatic Image Annotation, Global Visual Descriptor, Textual-Visual Groups.

References

1. Makadia A., Pavlovic V., Kumar S. A New Baseline for Image Annotation. *Proc. 10th European Conf. on Computer Vision*, Marseille, France, 2008, vol. 5304, pp. 316–329.
2. Guillaumin M., Mensink T., Verbeek J., Schmid C. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. *Proc. IEEE 12th Intern. Conf. on Computer Vision*, Kyoto, Japan, 2009, pp. 309–316.
3. Verma Y., Jawahar C. V. Image Annotation Using Metric Learning in Semantic Neighbourhoods. *Proc. 12th European Conf. on Computer Vision*, Florence, Italy, 2012, vol. 7574, pp. 836–849.
4. Chen M., Zheng A., Weinberger K. Q. Fast Image Tagging. *Proc. 30th Intern. Conf. on Machine Learning*, Atlanta, USA, 2013, pp. 1274–1282.
5. 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, vol. 2008, P10008.
6. Shen F., Ogura T., Hasegawa O. An Enhanced Self-Organizing Incremental Neural Network for Online Unsupervised Learning. *Neural Networks*, 2007, vol. 20(8), pp. 893–903.
7. IAPR TC-12 Benchmark. Available at: http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz (accessed 22 February 2016).
8. Lowe D. G. Distinctive Image Features from Scale-Invariant Keypoints. *Intern. Journal of Computer Vision*, 2004, vol. 60(2), pp. 91–110.
9. Bay H., Ess A., Tuytelaars T., Gool L. V. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 2008, vol. 110(3), pp. 346–359.
10. Yang J., Yu K., Gong Y., Huang T. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, USA, 2009, pp. 1794–1801.
11. Wang J., Yang J., Yu K., Lv F., Huang T., Gong Y. Locality-Constrained Linear Coding for Image Classification. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010, pp. 3360–3367.
12. Jegou H., Douze M., Schmid C., Perez P. Aggregating Local Descriptors into a Compact Image Representation. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010, pp. 3304–3311.
13. Proskurin A. V. Fast Local Descriptor for Scene Image Categorization. *Materialy XIX Mezhdunarodnoi nauchno-prakticheskoi konferentsii "Reshetnevskie chteniia"* [Proc. XIX Intern. Scientific Conf. "Reshetnev Readings"]. Krasnoyarsk, 2015, vol. 2, pp. 243–245 (In Russian).
14. Alcantarilla P. F., Bergasa L. M., Davison A. J. Gauge-SURF Descriptors. *Image and Vision Computing*, 2013, vol. 31(1), pp. 103–116.
15. Favorskaya M., Proskurin A. Image Categorization Using Color G-SURF Invariant to Light Intensity. *Procedia Computer Science*, 2015, vol. 60, pp. 681–690.
16. Aivazian S. A., Bukhshtaber V. M., Eniukov I. S., Meshalkin L. D. *Prikladnaia statistika. Klassifikatsiia i snizhenie razmernosti* [Applied Statistics. Classification and Reduction of Dimension]. Moscow, Finansy i statistika Publ., 1989. 607 p. (In Russian).
17. Feng S., Manmatha R., Lavrenko V. Multiple Bernoulli Relevance Models for Image and Video Annotation. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Washington, USA, 2004, vol. 2, pp. 1002–1009.
18. Zhang S., Huang J., Li H., Metaxas D. N. Automatic Image Annotation and Retrieval Using Group Sparsity. *IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics*, 2012, vol. 42(3), pp. 838–849.