Automatic Photo Annotation Based on Visual Content Analysis

Eduard Kuric[*]
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
eduard.kuric@gmail.com

Abstract

In this paper we describe a novel approach for automatic photo annotation based on visual content analysis. We combine local and global features to ensure robustness and generalization needed by complex queries. We place great emphasis to work in real-time. To cope with the huge number of extracted features, we implemented disk-based locality-sensitive hashing to index descriptors. By searching candidates to extraction of keywords, we focus on photos analysis in terms of probability, that the retrieved photos contain the right keywords for the target photo. The result is that we are able to name key objects directly in the target photo.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

automatic image annotation, image processing, local features, global features, visual content analysis

1. Introduction

Automatic photo annotation is the process by which a computer system automatically assigns metadata (keywords) to a target photo. With increasing popularization of digital and mobile phone cameras, there occurs a need of quick and exact searching, for example by general category or focusing on a specific object (object of interest). Manual creating annotations by a user is very time-consuming and results can be often subjective. Therefore, automatic photo annotation is most challenging task.

Generally, approaches for automatic annotation are categorized into two scenarios [14]: learning-based methods primary focused on determining complex categories or group of specific objects and web-based methods use crawled web image data to obtain relevant annotations.

In learning-based methods, a statistical model is built to learn a classifier. The methods are highly dependent on the size and distribution of a corpus (training dataset).

The Web space provides unlimited vocabulary for web-based methods. It is a distinct advantage over learning-based methods. However, the main problem of web-based approaches is initial query. In other words, the problem is lack of information about the target photo. Without providing further information such as key caption, searching similar photos for the target photo on the web is like finding a needle in a haystack. No less important drawbacks of the approaches are performance (annotation in real-time) and obtained annotations are often noisy.

To determine complex category, automatic annotation is based on searching the most similar photos (photo regions) for a target photo in a corpus which contains well-annotated photos or regions. Similarity of the photos is usually evaluated by comparing global descriptors which are extracted from the photos. After a retrieval process, related keywords are assigned to the target photo. A global descriptor is a single descriptor captures entire information of a photo (e.g. color, texture and shape). Main advantages of global descriptors are low computational complexity and the ability to capture complex information.

Automatic face recognition in a target photo is good example for automatic annotation of specific objects. One of possibility is that a retrieval process uses a robust dictionary of visual terms (blobs) to identify people. Similarity can be evaluated of comparing local descriptors which are computed over local features such as edges, small patches around points of interest. An advantage of the local features is their invariance to scale changes, rotation, illumination and viewpoint. The local descriptors are much

[*]Master degree study programme in field Software Engineering. Supervisor: Professor Mária Bieliková, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

© Copyright 2011. All rights reserved. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from STU Press, Vazovova 5, 811 07 Bratislava, Slovakia.

Kuric, E. Automatic Photo Annotation Based on Visual Content Analysis. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 72-75

more precise and discriminating than global descriptors. By searching a specific object, this feature is welcome, but by searching complex categories it can be an obstacle. Another obstacle is the need to store the huge number of the extracted local descriptors, typically hundreds to thousands per a photo, depending on the complexity of the photo. It is desirable to ensure responses to queries in real-time, too. The requirements are often an obstacle for huge corpus of photos.

2. Related work

Learning-based methods attempt to learn classifiers or use machine learning techniques to learn joint probabilities between photos and annotations. Typical classifiers are Support Vector Machine (SVM) [2], Hidden Markov models [8] or Bayes point machine [1].

Co-occurance model [5] is considered the main representative on learning joint probabilities between photos and words. Learning consists of two processes. First, a grid segmentation algorithm is used to divide each image into sub-images and global features of the sub-images are calculated. Second, the probability of each word for set of segments is estimated by using a vector quantization of the features of the sub-images. Other representatives are Translation model [4], Cross-media relevance model [6] or Continuous relevance model [7].

Dependence on a training dataset is considered the main drawback of the approaches. Another problem is that the approaches do not address the issue of indexing of the huge number of features (large datasets).

On other hand, web-based methods use crawled web-image data to obtain relevant annotations. A typical web-based method is AnnoSearch [14]. With a target photo, an initial word (key caption) is provided to conduct a text-based search on a web database. Then a content-based method is used to search visually similar photos and annotations are extracted from obtained descriptions. Other related approaches [10, 13] modify and extend the proposed method.

The main problem of the existing approaches is performance (annotation in real-time) and lack of information about the target photo to conduct a web search. We propose a method which ensures robustness and generalization needed by complex queries and annotation in real-time.

3. Our approach

Our method (see Figure 1) consists of two phases, namely dataset pre-processing and processing of the target photo (query). The both phases are in principle independent.

3.1 Dataset pre-processing

Dataset pre-processing consists of

1. photo processing,
2. features extraction and
3. features indexing.

The photo processing includes photo resizing (if necessary) and dividing the photo to the fixed number of sub-images (grid segmentation). The segmentation determines the granularity with which we are able to name identified regions. The feature extraction is performed in 3 steps:

- Extraction of bounded local features: for each sub-image the fixed number of local descriptors is computed.
- Extraction of free local features: for the full photo the fixed number of local descriptors is computed and the sets of the bounded and free features are mutually disjoint.
- Extraction of global features: for each sub-image a global descriptor is computed.

To detection and extraction the local features, we use Scale Invariant Feature Transform (SIFT) [9]. The features are invariant to image scaling, translation and partially invariant to illumination changes and affine for 3D projection. They are well adapted for characterizing small details. The method consists of a detector and descriptor. The features are detected through local extreme in a Difference-of-Gaussians function and described using histograms of gradients.

The bounded features perform partially a role of the global features due to their uniform (balanced) extraction of the full photo. A drawback of the SIFT is that the features are extracted using grayscale information only. To capture complex information, we employ the global features.

To computation of the global descriptors, we use Joint Composite Descriptor (JCD) [15]. The JCD belongs to a group of Compact Composite Descriptors (CCD), which combine information about color and texture in a single histogram. They were designed with regard to dimension, but without compromising their discriminating ability.

To index the extracted descriptors, we implemented disk-based locality-sensitive hashing (LSH). It is a method for solving the near neighbor search in high dimensional spaces. The basic idea is that similar features are mapped to the same buckets with high probability (see Figure 2).

We employ the LSH scheme [3] based on p-stable distributions as follows:

$$h(v)_{(a,b)} = \left\lfloor \frac{a{\cdot}v + b}{w} \right\rfloor \qquad (1)$$

Each hash function $h(v)_{(a,b)} : R^d \to Z$ maps a d dimensional feature vector v onto the set of integers. The parameter a is a d dimensional vector with entries chosen from a p-stable distribution (Gaussian distribution), b is a real number chosen uniformly from the range $[0, w]$. The optimal value for w depends on the dataset and the feature vector. In [3] was suggested that $w = 4.0$ provides good results, therefore we chose the value. Each function g (see Figure 2) is obtained by concatenating k randomly chosen hash functions h. Furthermore, we map the set of computed integers to a single integer for bucket identification $g_i(h_1(v), ..., h_k(v)) \to N$.

To store the huge number of the extracted features, we adopted the distributed database management system Cas-

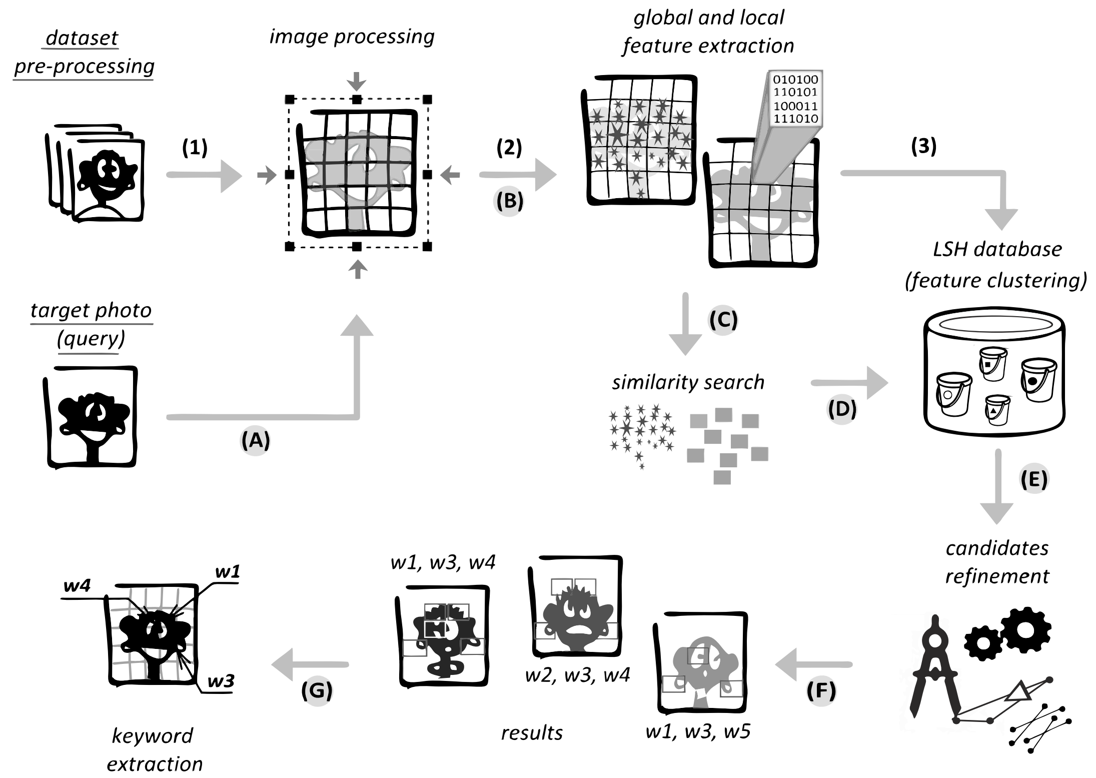

Figure 1: A scheme of our annotation method consists of two independent phases, namely dataset preprocessing (1) - (3) and processing of a target photo (A) - (G).

sandra, which is designed to handle very large amounts of data[1].

3.2 Target photo annotation

Steps of photo annotation are shown in Figure 1. After photo processing (A) and extraction of global and local features (B), similar features are retrieved from the LSH database (C, D). Each feature is associated with a photo, so they are subsequently grouped according to them.

Result of the particular steps is a list of similar photos (candidates) and their features that are grouped by similarity to sub-images of the target photo. The grouped features (into sub-image groups) are divided into a bag of bounded local features and a single global feature. Each candidate has assigned similar free local features, too. The algorithm to keyword extraction consists of the following steps (E, F, G):

1. Compute a score value SLF_c for local features as follows:
   - for the free local features compute geometric consistency using a RANSAC algorithm; inliers assign to the bag of the bounded local features according to their coordinates (into a corresponding sub-image group)
   - for each sub-image group of the candidate compute a sub-score as a proportion of the number of the similar bounded local features and the total number of the extracted local features from the corresponding sub-image of the target photo; add an extra weight for the features (objects) whose dominance or frequency is greater
   - compute the SLF_c as the sum of the computed sub-scores
2. Compute a score value SGF_c for global features as follows:
   - for each global feature in the sub-image groups compute a similarity to the corresponding sub-image of the target photo (to calculate the similarity is used Tanimoto coefficient
   - compute the SGF_c as the sum of the computed similarities
3. For each candidate compute a final score (FSc) using the equation (2), where d_l is a damping factor for the local features and d_g is a damping factor for global features. Furthermore, order the candidates.

$$FS_c = d_l * SLF_c + d_g * SGF_c \tag{2}$$

4. Merge all keywords assigned to the candidates and by frequency assign the top keywords to the target photo. From each candidate assign all its keywords to the sub-images of the target photo, provided that the corresponding sub-scores of the candidate are greater than or equal to a threshold. Words with the highest frequency are the final names for the sub-images.

[1] Apache Cassandra: http://cassandra.apache.org/

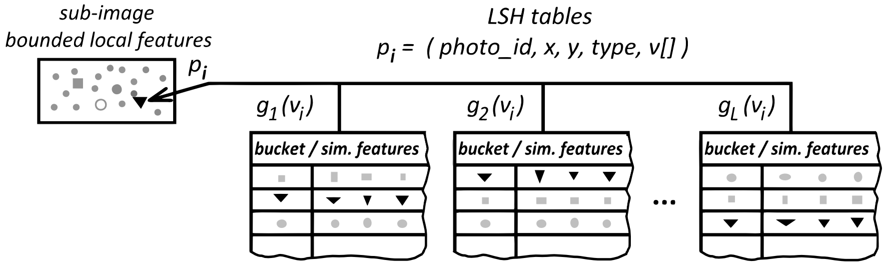

Figure 2: Locality-sensitive hashing: similar features are mapped to the same buckets with high probability.

4. Evaluation and conclusions

Our evaluation is conducted on Corel5K corpus which contains 5000 photos. This corpus is used widely in automatic photo annotation area and includes a variety of subjects, ranging from urban to nature scenes and from artificial objects to animals. It is divided into 2 sets: a training set of 4500 photos and a test set of 500 photos. Each photo is associated with 1-5 keywords.

To cope with the huge number of extracted features, we implemented disk-based locality-sensitive hashing to index descriptors. By searching similar candidates to extraction of keywords, we focus on photos analysis in terms of probability that the retrieved photos contain the right keywords for the target photo. For example, we prefer photos where extracted objects of interest from the target photo are dominant in retrieved photos or their frequency of occurrence is greater. It is more likely that such photos contain the right words. We also focus on identification of less significant objects for which we are looking for correct words. Using grid segmentation, we are able to name objects directly in the target photo.

Our approach can be effectively used to provide crucial semantic metadata required for advanced multi-paradigm exploration of image collections [12] or more general personalized exploratory search approaches [11], where traditional (text-based) metadata extraction approaches fail.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13:26–38, 2003.

[2] C. Cusano and R. Schettini. Image annotation using svm. In *Storage and Retrieval for Image and Video Databases*, volume SPIE 5304, pages 330–338, 2004.

[3] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, SCG '04, pages 253–262. ACM, 2004.

[4] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112. Springer-Verlag, 2002.

[5] Y. M. Hironobu, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. 1999.

[6] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 119–126. ACM, 2003.

[7] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *IN NIPS*. MIT Press, 2003.

[8] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:1075–1088, September 2003.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[10] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 585–594. ACM, 2007.

[11] M. Tvarožek. Exploratory search in the adaptive social semanticweb. *Inf. Sciences and Technologies Bulletin of the ACM Slovakia*, 3:42–51, 2011.

[12] M. Tvarožek and M. Bieliková. Collaborative multi-paradigm exploratory search. In *Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, pages 29–33. ACM, 2008.

[13] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Scalable search-based image annotation of personal images. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, MIR '06, pages 269–278. ACM, 2006.

[14] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1483–1490. IEEE Computer Society, 2006.

[15] K. Zagoris, S. A. Chatzichristofis, N. Papamarkos, and Y. S. Boutalis. Automatic image annotation and retrieval using the joint composite descriptor. *2010 14th Panhellenic Conference on Informatics*, pages 143–147, 2010.