AUTOMATIC IMAGE ANNOTATION BASED ON VISUAL CONTENT ANALYSIS

Eduard KURIC, Mária BIELIKOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{kuric,bielik}@fiit.stuba.sk

Abstract. Automatic image annotation methods require a quality training image dataset. The main problem of current methods is their low effectiveness and scalability if a large-scale training dataset is used. We proposed a method to obtain annotations for target images based on a novel combination of local and global features during search stage. We are able to ensure robustness and generalization needed by complex queries and significantly eliminate irrelevant results. We identify objects directly in target images. For each obtained annotation we estimate the probability of its relevance. We evaluated our method on the Corel5K corpus.

1. Introduction

Automatic image annotation (AIA) has been studied extensively for several years. Many of us likely has hundreds to thousands photos and apparently each of us has probably at least once thought "I would like to show her the photo, but I am unable to find it". Searching images by content only is an extremely hard and very challenging task for researchers in the content-based image retrieval (CBIR) field. With the expansion and increasing popularity of digital and mobile phone cameras, we need to search images effectively and exactly more than ever before.

Content based indexing of images is more difficult than for textual documents because they do not contain units like words. Image search is based on using annotations and semantic tags that are associated with images. However, annotations are entered by users and their manual creation for a large quantity of images is very time-consuming with often subjective results. Therefore, for more than a decade, automatic image annotation has been a most challenging task. Automatic image annotation methods are usually categorized into two categories, namely probabilistic modeling-based methods [2, 3, 5] and

classification-based methods [4]. The significant drawbacks of the presented "art" models are their performance and scalability if a large-scale image dataset is used; and/or use of only global features during search or image classification, respectively. Therefore, in our method we have focused on addressing these drawbacks.

2. Concept of Proposed Method for Automatic Image Annotation

In our approach [1], we combine global and local features to retrieve the best results. The combination is more suitable to represent complex scenes and events categories. Global and local features have limitations describing images and none of them appears to be powerful enough to represent the large amount and variety of images. Global and local features provide different kinds of information. They have their own advantages in classifying certain categories. However, they have several complementary strengths and there are many situations where the automatic image annotation should be judged based on the combination of global and local features.

We use grid segmentation for extracting the global features and efficient graph-based segmentation for extracting local features. Compared to existing methods, we are able to ensure the "robustness" and generalization needed by complex queries. In our method, in analogy with text documents, the global features represent words extracted from paragraphs of a document with the highest frequency of occurrence and the local features represent keywords extracted from the entire document.

We place great emphasis on performance and have thus tailored our method to use large-scale image training datasets. To cope with the huge number of extracted features, we have designed disk-based locality sensitive hashing for indexing and clustering descriptors. We have chosen locality sensitive hashing (LSH) for several reasons. First, it is not related to any learning corpus, it may be fast, and retrieval performance does not evolve when modifying the database while this is not true for tree-based methods. Using a K-Means approach would require updating the visual vocabulary regularly to avoid degraded performance (and to define when to do these updates). Our approach is particularly suitable solution for applications where the image corpus evolves, i.e., our solution provides a good compromise between precision and speed.

We are able to identify objects directly in target images. Our method estimates the probability that the retrieved similar images (training images) contain the right words for a given target image. We focus on the way, how people manually annotate images. We prioritize dominant objects and estimate relative importance of words in the training annotations. The estimated probability of words determines degree of accuracy with which the words describe the visual content of the target image.

Our method (see. Figure 1) consists of two main stages, namely training dataset preprocessing and processing of the target image (query). Dataset pre-processing consists of image processing (A), local and global features calculation (B) and their indexing and clustering according to similarity (C). Processing of the target image consists of image processing (1), local and global features calculation (2), querying the keypoint store and global features index (3). After queries are executed, similar images (visual terms) to the target image are retrieved as result sets (4). Subsequently, the result sets are refined (5). A final

stage of obtaining annotation is performed and relevance of assigned words is estimated (6).

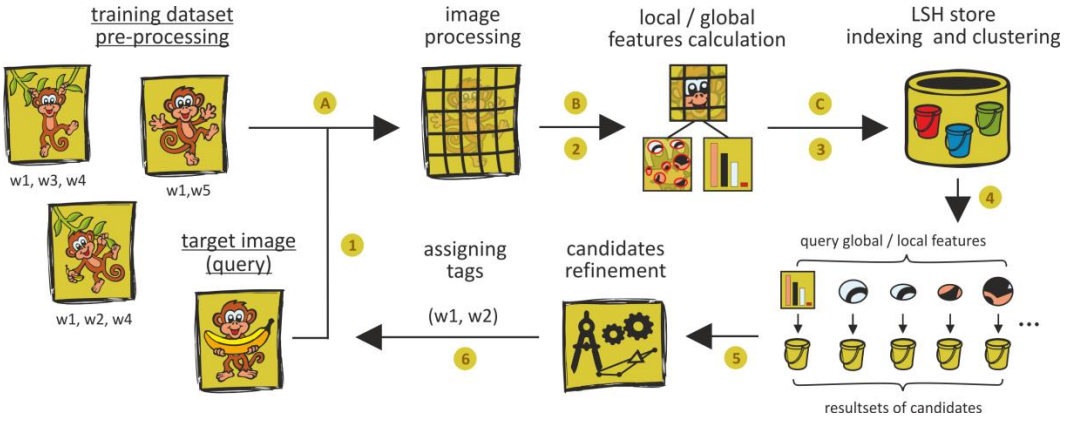

Figure 1. Scheme of our method for automatic image annotation.

3. Evaluation and Conclusions

Our evaluation was conducted over the Corel5K corpus. It consists of 5,000 images from 50 Corel Stock Photo CDs and each CD includes 100 images with the same theme. The corpus is used widely in the automatic image 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 4,500 photos and a test set of 500 photos. Each photo is associated with 1-5 keywords. We evaluate our method using 500 images from the Corel5K test set. The annotation problem can be understood as the problem of retrieving an image from the test set using words from the test vocabulary. To evaluate the annotation performance, we use the precision and recall metrics. Let A be the number of images automatically annotated with a given word, B the number of images correctly annotated with that word. C is the number of images having that word in the manual annotation.

We calculate the mean precisions and recalls for a given query set. There are overall 260 words in the test set but not all of them can be predicted by the annotation system. A word is predictable if its average recall is greater than 0. Our method is able to predict 92 words with 0.26 mean per-word precision and 0.35 mean per-word recall. We compare our method with the Translation model [3] and the Co-occurrence model [5]. The number of queries, which retrieve at least one relevant image, is dependent on the annotation models – the Translation model has 49 with 0.20 mean precision and 0.34 mean recall and the Co-occurrence model has 19 with 0.21 mean precision and 0.39 mean recall.

The local features are very successful for problems involving retrieval of target objects (objects of interest). They exhibit very good robustness to moderate scaling, brightness changes and “in-plane” rotation. The global features capture the entire information of an image (e.g. texture, color). Both have advantages and drawbacks, the local features are much more precise than global features and their discrimination ability is relatively high. When looking for a target object, this ability is welcome, however, when looking for a general category (e.g. find all yellow Ferrari), it may cause restrictions. A method for automat-

ic image annotation should be able to ensure both the requirements, namely, robustness and generalization and it was one of our goals. Via the combination of global and local features, we achieved the required robustness for effective automatic annotation.

Because we use a combination of local and global features, our approach is relatively resistant to common transforms (cropping, scaling). Traditional approaches based on global features cannot cope with. A potential drawback of using local features is that we need to store and index a huge number of extracted features and there is a need to query hundreds to thousands of features which could be slow. This "side-effect" often causes performance issues (e.g. approaches based on k Nearest Neighbors search) and it limits using large-scale image (training) datasets. For this problem, we employed efficient solution through locality sensitive hashing which is based on the idea that similar objects are stored to the same bucket. Our solution provides a good compromise between precision and speed; it allows random access to stored data (in sub-linear time); and index is generated dynamically. We have adopted the distributed database management system Cassandra, which was specially designed for storing the huge number of data. For efficient access to extracted data, we have designed data layouts for using with LSH.

Acknowledgement: This work was partially supported by the grants VG1/0675/11/2011-2014, KEGA 028-025STU-4/2010, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

to other papers publishing the results that are summarized here

[1] Bieliková, M., Kuric, E.: Automatic Image Annotation Using Global and Local Features. In: *Proc. of the 6th Int. Workshop on Semantic Media Adaptation and Personalization*. IEEE Computer Society, Washington, USA, (2011), pp. 33-38.

Other references

[2] Chang, E., et al.: CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines. *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 1, (2003), pp. 26–38.

[3] Duygulu, P., et al.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Proc of the 7th European Conf. on Computer Vision-Part IV, Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen (Eds.). Springer-Verlag, London, UK, (2002), pp. 97-112.

[4] Feng, S. L., Manmatha, R., and Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: *Proc. of the Int. Conf. on Computer vision and pattern recognition*. IEEE Computer Society, Washington, USA, (2004), pp. 1002-1009.

[5] Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: *Proc. of the Int. Workshop on Multimedia Intelligent Storage and Retrieval Management*. 1999.