Computers & Graphics 47 (2015) 1–15

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

ANNOR: Efficient image annotation based on combining local and global features

Eduard Kuric, Maria Bielikova *

Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Ilkovicova 2, 842 16 Bratislava 4, Slovakia

ARTICLE INFO

Article history:
Received 9 March 2014
Received in revised form
28 September 2014
Accepted 28 September 2014
Available online 12 October 2014

Keywords:
Image retrieval
Automatic annotation
Object recognition
Local features
Global features
Locality sensitive hashing

ABSTRACT

Automatic image annotation methods based on searching for correlations require a quality training image dataset. For a target image, its annotation is predicted based on a mutual similarity of the target image to the training images. One of the main problems of current methods is their low effectiveness and scalability if a relatively large-scale training dataset is used. In this paper we describe our approach "Automatic image aNNOtation Retriever" (ANNOR) for acquiring annotations for target images, which is based on a combination of local and global features. ANNOR is resistant to common transforms (cropping, scaling), which traditional approaches based on global features cannot cope with. We are able to ensure the robustness and generalization needed by complex queries and significantly eliminate irrelevant results. We identify objects directly in the target images and for each obtained annotation we estimate the probability of its relevance. We focus on the way how people manually annotate images (human aspects of image perception). We have designed ANNOR to use large-scale image training datasets. We present experimental results for three challenging (baseline) datasets. ANNOR makes an improvement as compared to the current state-of-the-art.

1. Introduction

Automatic image annotation has been studied extensively for several years. Many of us likely has hundreds to thousands photos and each of us has probably at least once thought "*I would like to show her the photo, but I am unable to find it*". With the expansion and increasing popularity of digital and mobile phone cameras, we need to search images effectively and exactly more than ever before.

Focusing on visual query forms, many content-based image retrieval methods and techniques have been proposed, but they have several limitations. On one hand, in query-by-example-based methods a query image is often absent. On the other hand, query-by-sketch approaches [1,2] are too complex for common users and a visual content interpretation of a user image concept is difficult.

A text retrieval system often helps finding rapidly related documents from a vast amount of documents containing keywords. Image search using keywords is presently the most widely used approach. Content-based indexing of images is more difficult than indexing of textual documents because they do not contain units like words. Image search is based on 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.

The goal of automatic image annotation is to assign a collection of keywords (annotation) from a given dictionary to a target (previously unseen) image. That is, the input is the target (uncaptioned) image and the output is a collection of keywords that describe the target image in a best possible way.

Why automatic image annotation is a challenge? Automatic image annotation is on the frontier of different fields such as image analysis, machine learning and information retrieval. In present, to create a general system for automatic image annotation based on object recognition is practically impossible (it is doubtful if ever at all). The Imagenet Large Scale Visual Recognition Challenge (ILSVRC)[1] is the venue for evaluating the current state-of-the-art for image classification and recognition.

To extract the semantics from data, general object recognition and scene understanding is required. This is an extremely hard task. The same object can be captured from different angles, distances or under different lightning conditions. The manual annotation is subjective and sometimes it is difficult to describe image contents by keywords. In general, an object of the real world with the same "name" may have

* Corresponding author. Tel.: +421 2 210 22 304; fax: +421 2 654 20 587.
E-mail addresses: eduard.kuric@stuba.sk (E. Kuric), maria.bielikova@stuba.sk (M. Bielikova).
URL: http://www.fiit.stuba.sk/~bielik/ (M. Bielikova).

[1] ILSVRC: http://image-net.org/

http://dx.doi.org/10.1016/j.cag.2014.09.035
0097-8493/

different visual form (e.g., shape, color). Good illustrative examples are methods for face recognition. There are several approaches for face recognition, e.g., methods based on comparing templates, which require a robust database of faces. The faces are searched based on correlating between an input (a target face) and the templates. Complex knowledge-based methods focus on analyzing morphological features such as eyes, mouth, skin and color. They are based on rules defined by the real features of human faces.

Here are some crucial questions that current automatic image annotation systems have to deal with:

- *Which image representation is appropriate to describe image*? The objects in images are often occluded and appear in poor lighting and exposure.
- *Which image features can be extracted to describe or characterize the visual content*? A feature is represented by a numerical feature vector (descriptor), by which we are able to describe a part of image content. In general, there are three essential requirements for the descriptors, their degree of robustness, discrimination ability and efficiency. The robustness represents invariance to the geometrical changes (e.g., viewpoint, zoom, object orientation) and noise-like signal distortions. The discrimination maximizes difference among non-duplicates and minimizes difference among duplicates. The feature extraction and matching requires fast computation.

Another question is the spatial and time complexity (computational cost). A huge number of features per image can be extracted and the dimension of the feature vector is crucial aspect, too. There is a problem how to index, store and compare the descriptors in real-time. Often in many cases, faster access to information means the need for more space allocation.

In this paper we propose a method for automatic image annotation using relatively large-scale image "training" dataset. We combine local and global features to ensure robustness and generalization needed by complex queries and therefore we focus on performance and scalability. For indexing and clustering features, we use disk-based locality sensitive hashing. To obtain annotation for a given target image, our approach is based on the way how people manually annotate images.

Compared with our previous work [3] we present completely new process of obtaining annotation called ANNOR (Automatic image aNNOtation Retriever). The evaluation part is also completely new. We have performed new experiments focused on evaluation of efficiency and quality of obtaining annotation. We have evaluated our approach on three datasets and we have compared the results of our approach with the state-of-the-art approaches.

This paper is structured as follows: Section 2 provides an overview of existing methods for automatic image annotation; Section 3 introduces our approach; Sections 4 and 5 describe in detail extracting, indexing, clustering and retrieving local features and global features, respectively. Section 6 describes in detail obtaining annotation for the target image and estimation its relevance; Section 7 presents the evaluation results of our approach; and Section 8 contains discussion and conclusion.

2. Previous work on automatic image annotation

2.1. State-of-the-art

Automatic image annotation methods are usually divided into two categories, namely probabilistic modelling-based methods and classification-based methods.

Probabilistic-based methods estimate correlations or joint probabilities between images and annotation keywords over a training image dataset (corpus).

Mori et al. [4] proposed the Co-occurrence model to capture correlations between images and keywords. The designed model is considered the main pioneer and consists of two stages. First, a grid segmentation algorithm is used to uniformly divide each image into a set of sub-images (segments) and for each segment, a global descriptor is calculated. Second, for the set of segments, the probability of each keyword is estimated by using a vector quantization of the features of the segment. The drawback of the model is a relatively low annotation performance.

Duygulu et al. [5] proposed a model of object recognition as a machine translation. A statistical translation model was used to translate keywords of an image to visual terms (blobs). A vocabulary of blobs was generated by clustering image regions segmented using the N-cut algorithm. Mapping between blobs and keywords was learned using the Expectation–Maximization algorithm. One of the key problems of the model is high computational complexity of the Expectation-Maximization algorithm and therefore it is not suitable for large-scale datasets.

Inspired by the relevance language models for text retrieval and cross-lingual retrieval, several relevance models were proposed, such as Continuous Relevance Model [6], Cross-Media Relevance Model [7], Dual Cross-media Relevance Model [8], and Multimodal Latent Binary Embedding [9]. Feng et al. proposed the Multiple Bernoulli Relevance Model [10] that takes into account image context, i.e., from training images it learns that a tiger is more often associated with *grass* and *sky* and less often with objects, such as *buildings* or *car*. In comparison with the translation model, it seems to be more effective for image annotation. However, its drawback is that only images consistent with the training images can be annotated with keywords in a limited vocabulary.

Metzler et al. [11] segment training images, connecting them and their annotations in an inference network. The inference network is based on Bayesian Network. It uses non-parametric methods to estimate probabilities within the inference network.

Yavlinsky et al. [12] proposed a framework based on non-parametric density estimation and the technique of kernel smoothing. Their results are comparable with the inference network [11] and CRM [8].

The task of classification-based methods is to construct image classifiers for annotation keywords that are trained to separate training images with the keywords from other keywords with some level of accuracy. After a classifier is trained, it is able to classify a target image into a class where the keywords in the training dataset and retrieved outputs (keywords) are used to annotate the target image. Typical representative classifiers are Support Vector Machine (SVM) [13–17], Hidden Markov Models [18], Markov Random Fields [19], Supervised Multi-class Labelling [20] or the Bayes Point Machine (BPM) [21,22].

The overall disadvantage of most classifiers is that they are designed for small-scale image datasets, i.e., classification into a small number of classes (categories). It is still an open research problem to construct large-scale learning classifiers and therefore, these methods are usually used for annotation of specific objects, such as car brands or company logos.

For all presented methods, a high quality annotated training image dataset (corpus) is crucial. There are some web-based methods, which use crawled data (images, annotations) as the training dataset such as AnnoSearch [23]. With a target photo, an initial keyword (caption) is provided to conduct a text-based search on a crawled web database. Then a content-based image retrieval method is used to search visually similar images and annotations are extracted from obtained descriptions. The notable advantage is the availability of a large-scale web image database. The main drawback is the use of only global features for the similar image search. One related approach [24] modifies the basic