Information Sciences 366 (2016) 99–120

Contents lists available at ScienceDirect

Information Sciences

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

Topic modeling and improvement of image representation for large-scale image retrieval

Nguyen Anh Tu, Dong-Luong Dinh, Mostofa Kamal Rasel, Young-Koo Lee*

Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Republic of Korea

ARTICLE INFO

Article history:
Received 24 September 2015
Revised 29 April 2016
Accepted 22 May 2016
Available online 26 May 2016

Keywords:
Topic modeling
Probabilistic graphical model
Image retrieval
Image representation
Image coding
Bag-of-visual words

ABSTRACT

In this paper, we present a new visual search system for finding similar images in a large database. However, there are a number of challenges regarding the robustness of the image representations and the efficiency of the retrieval framework. To tackle these challenges, we first propose an encoding technique based on soft-assignment of local features to convert an entire image into a single vector, which is a compact and discriminative representation. This encoded vector is suitable for most types of efficient indexing methods to produce an initial result. To compensate for the lack of incorporating geometric and object-related information during the encoding scheme, we then propose a probabilistic topic model to formalize the spatial structure among the local features. Moreover, the topic model allows us to effectively extract the object and background regions from the image. This is performed by a Markov Chain Monte Carlo algorithm for approximate inference. Finally, benefiting from the extracted objects in each image, we present a re-ranking scheme to automatically refine the initial search results. Our proposed retrieval framework has two major advantages: i) an aggregation strategy through soft-assignment improves the discriminative power of the representation, which has a determinative effect on the retrieval precision; and ii) the probabilistic latent topic model enables us to not only gain insight into the spatial structure of the image, but also handle a large variation in the object appearance. The experimental results from four benchmark datasets show that our approach provides competitive accuracy, and runs about ten times faster. Our studies also verify that proposed approach works effectively on large-scale databases of millions of images.

1. Introduction

In recent years, multimedia and networking technologies have significantly impacted our daily activities. In particular, the development of smart phones and other mobile devices have increased the demand for searching for information of the Internet, books, magazines, and reference materials. Many applications have been developed for automatic recognition of different objects of interest, such as product catalogs, landmarks, and art galleries. Therefore, the areas of visual recognition and image retrieval provide fascinating research opportunities. Typically, the aim of a retrieval task is to select from a collection of objects images that are similar to a query image. However, large-scale image retrieval poses a number of challenges regarding the desired quality of the image representations and the efficiency of the retrieval framework. Hence,

* Corresponding author. Tel.: +82 312013732.
E-mail addresses: tunguyen@khu.ac.kr (N.A. Tu), luongdd@ntu.edu.vn (D.-L. Dinh), rasel@khu.ac.kr (M.K. Rasel), yklee@khu.ac.kr (Y.-K. Lee).

http://dx.doi.org/10.1016/j.ins.2016.05.029

we seek to obtain representations of images and image regions that are discriminative and robust to the various types of data content. We also seek to design a retrieval framework that can efficiently and effectively handle large image databases with high search accuracy.

Currently, Bag-of-visual words (BoV) [48] is a seminal framework for image retrieval. In this framework, local features (e.g., SIFTs [31]), which typically achieves invariance of orientation and scale in modern visual recognition, are quantized to form a vocabulary of visual words. An image is encoded as a sparse frequency histogram over the visual vocabulary. Inheriting the characteristics of BoV, an advanced encoding scheme called Vector of Locally Aggregated Descriptors (VLAD) [20,21] has been proposed to produce a higher-order representation by including the statistics of local features. The VLAD model also encodes an image as a single vector like the BoV model by aggregating its local features. The aggregated vector can be compressed with the dimension reduction method to obtain a compact representation. Using this encoding scheme with an indexing technique like product quantization [19], an image can be represented by a small number of bytes to provide competitive search accuracy [21]. Although popular retrieval models (e.g., BoV and VLAD) work generally well, they suffer from three main issues: (1) The discriminative power of the image representation is decreased due to the hard-assignment of feature quantization used in these models, where a local feature might be assigned to a wrong visual word. This results in reducing the similarity between two images containing similar objects. (2) When encoding an image, the lack of geometric (e.g., location, scale, and orientation) and object-related information makes these approaches very sensitive to large variations in objects, such as occlusion, deformation, and viewpoint change. (3) For similarity measurement, these models focus only on estimating the global change between two images. They fail to exploit the human cognition related to object appearance, background, and their relationship. Hence, they cannot capture the local change of each object appearing in the image. This leads to decreasing the accuracy and efficiency of a retrieval task when coping with complex data content.

In this paper, we address these issues by proposing a novel retrieval framework that enables a robust representation and an efficient re-ranking method for measuring the similarity between the query image and candidate database image. Each image is modeled as a set of local features with a two-phase procedure: encoding and topic modeling.

Image encoding. We enhance VLAD by developing an encoding scheme called soft-VLAD. In this scheme, we use two aggregation approaches based on soft-assignment to map each local descriptor to multiple visual words and aggregate them into a single vector. The first approach is distance-based assignment, where each assignment is assigned to a weight proportional to the distance from the visual word to the local descriptor. The second approach is sparse-coding-based assignment, which uses sparse coding [38] to project local descriptors onto the learned codebook, and then computes the weight of each assignment. This approach is motivated by successful applications of sparse coding in image classification [55,57]. By using these approaches, we overcome the limitation of hard-assignment and encode each image into a highly discriminative vector for indexing and retrieving an initial list of candidates.

Topic modeling. By exploiting human knowledge about the object appearance and the background, we propose a generative latent topic model called the *spatial latent topic model with background distribution* (SLTMB) to extract the background regions and topic regions from the image. In this probabilistic model, each topic corresponds to an object or a part of an object occurring frequently in the image corpus. Specifically, an image contains the object instances with a certain spatial arrangement, while each object instance can also be represented by appearance of the relevant set of visual words. The SLTMB model is therefore intended to exploit the probability distributions over visual words for different topics. Thus, visual words co-occurring often in the same image with a particularly spatial distribution tend to belong to the same object or topic region. Naturally, the similar object instances will have similar probability distributions, and so the SLTMB enables us to also infer what unknown objects are present in those images and where they are. The spatial location of visual words integrated into our topic model is effectively used to compensate for the limitations of encoding the image, and so strengthens our image representation.

Re-ranking image. The topic regions, meanwhile, have already been extracted in the candidate images, and we can take advantage of such information to refine the search results. We observe that a database image is similar to a query image if they contain similar topic regions. Consequently, we propose a re-ranking method with a fast and efficient geometric scoring scheme for large-scale image matching. We first establish matching feature pairs between the common topics of the query and candidate images. Then, similarity scores between the common topic regions are generated based on the geometric information of the matched features. Afterward, a new score for each image pair is computed as the sum of the topic similarity scores, and re-ranking is performed using the new scores. This strategy allows us to significantly speed up re-ranking as well as perform on medium-sized datasets (e.g., a thousand images). Moreover, using extracted topic regions, our method can handle local variations of object appearance in each image.

Our main contributions are three-fold: (1) We propose an encoding scheme called soft-VLAD to produce vector representation for the whole image, which extends VLAD by using the soft-assignment approaches for aggregation. (2) Our topic model built on human knowledge about image structure formulates the appearances and locations of the different topics and background regions. This allows us to effectively extract the objects from an image. (3) We present an efficient measuring scheme in conjunction with extracted topic regions to compute the similarity between two images. Using this scheme, the proposed re-ranking method shows very promising accuracy and fast processing on publically available datasets.

The remainder of this paper is organized as follows. Section 2 provides a discussion of related works. Section 3 presents an overview of proposed framework and then describes the details of our methods, including image preprocessing (Section 3.1), an encoding scheme called soft-VLAD (Section 3.2), the SLTMB model with learning and inference procedures