J. Vis. Commun. Image R. 38 (2016) 641–657

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Leveraging similarities and structure for dense representations combination in image retrieval ☆

Tomás Mardones [a,*], Héctor Allende [a], Claudio Moraga [b]

[a] Universidad Técnica Federico Santa María, Av. España 1680, CP 110-V Valparaíso, Chile
[b] Department of Computer Science, TU Dortmund University, Germany

ARTICLE INFO

Article history:
Received 20 October 2015
Revised 15 March 2016
Accepted 17 April 2016
Available online 19 April 2016

Keywords:
Content-based image retrieval
Fisher Vector
Global dense representation
Image search
Graph based combination
Re-ranking
Product Quantization

ABSTRACT

This paper addresses the problem of content-based image retrieval in a large-scale setting. Recently several graph-based image retrieval systems have been proposed to fuse different representations, with excellent results. However, most of them use one very precise representation, which does not scale as well as global dense representations with an increasing number of images, hurting time and memory requirements as the database grows. We researched how to attain a comparable precision, while greatly reducing the memory and time requirements by avoiding the use of a main precise representation. To accomplish this objective, we proposed a novel graph-based query fusion approach—where we combined several compact representations based on aggregating local descriptors such as Fisher Vectors—using distance and neighborhood information jointly to evaluate the individual importance of each element in a query adaptive manner. The performance was analyzed in different time and memory constrained scenarios, ranging from less than a second to several seconds for the complete search process while needing only a fraction of the memory compared to other similar performing methods. Experiments were performed on 4 public datasets, namely UKBench, Holidays, Corel-5K and MIRFLICKR-1M, obtaining state-of-the-art effectiveness.

1. Introduction

Content-based image retrieval (CBIR) is an important area of research in Multimedia, linked to numerous image applications, such as web and mobile image search, duplicate detection and location finding. The advances in the area are lucidly summarized in the works of Smeulders, Datta et al. in [1,2]. Given a query image, the problem is to find, in a large database, images that represent the same object or location, normally using a similarity measure between them, and deliver these results sorted in a rank list. This is a difficult problem, since the sought objects and locations are possibly viewed from different perspectives and in the presence of clutter or occlusions. The main contribution of this work is to provide a new graph fusion method, capable of combining similarity and neighborhood structure metrics in a very efficient and robust manner.

In this paper, we addressed the problem of searching for the most similar images to a given query in a large database. The desired solution must be accurate, fast and compact. This is an important problem, since many applications have databases with over one million images, where query time is constrained. For example, if in an electronic commerce website, a user uploads a product image to access its information, having a three seconds query time could be unacceptable. The Bag of Words (BoW) image representation [3] is usually employed for moderately sized databases (up to 1 M usually), since it provides a good balance among performance, time response and memory footprint. However, it becomes impractical as databases increase their size to a hundred million (100 M), especially if a response is required in just a few seconds. For example, the basic BoW algorithm needs around 10 KB per image (considering 2500 features per image on average), needing a total of 1 TB to store this amount of images. Taking into account that everything should fit in RAM memory to avoid costly memory swapping, 1 TB is substantially more RAM memory than the usual maximum manageable by a single computer. Additionally, the time response—as it increases in a roughly linear fashion—will get over one minute [4].

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author at: Departamento de Informática, Universidad Técnica Federico Santa María, CP 110-V Valparaíso, Chile.

E-mail address: tomas.mardones@alumnos.usm.cl (T. Mardones).

http://dx.doi.org/10.1016/j.jvcir.2016.04.012
1047-3203/© 2016 Elsevier Inc. All rights reserved.

In [5,6] it was proposed to use Fisher Vectors (FV) and Vectors of Locally Aggregated Descriptors (VLAD) as an alternative to BoW. FV is a higher order representation compared to Bag of Words, being able to model non-linear relations between image features. These methods are able to use the same powerful local descriptors as BoW, but instead of storing the indices of the features using inverted files, they are aggregated into a single dense vector. These vectors can have their dimension reduced by powerful methods, such as Product Quantization (PQ) and Iterative Quantization [7,8] being able to represent an image using 32 bytes or less, while preserving a relevant part of their discriminative capacity. Therefore, by using Fisher Vectors or VLAD it is possible to store 100 M image representations in just 3.2 GB. However, in the cases where memory usage has not important constraints, the precision attained by these methods is lower than the achieved by most advanced BoW algorithms [4,6,9–12].

One approach to further improve the precision of large-scale image retrieval systems is the fusion of multiple feature types, such as visual and textual ones. In this work, only features extracted from the image are considered, i.e. unimodal CBIR systems. By fusing several features, CBIR systems are able to improve their performance to state-of-the-art levels [4,13–18] at the expense of additional time and memory. The problem of predicting each feature effectiveness is of great relevance to every fusion algorithm, since if a good feature is ignored or a bad feature is given decision power, the precision may not reach the highest possible value. Or even, in some cases, it may get lower than the best individual feature's precision. An additional issue of most fusion algorithms is their memory and time efficiency, which are important elements in large-scale image retrieval. On one hand, feature extraction is usually one of the most time consuming tasks, and fusion methods need several features. On the other hand, commonly, several image representations must be stored for each image in the database impacting negatively the memory usage and query time. Moreover, some fusion processes can be comparatively costly time-wise and memory-wise [14,16,19].

To cope with the feature effectiveness problem, an unsupervised graph based late fusion mechanism is proposed, whose novelty resides in the manner it combines a similarity measure in tandem with neighborhood structure metrics. The proposed graph will be called Structure And Similarity Aware Graph (SASAG). Also, as part of SASAG, a new query adaptive similarity measure is described that enhances the proposed method's robustness to noisy and inadequate features. For this method to be scalable, special care is given to the feature extraction methods' configuration, to maintain the feature extraction time similar to the one used in works considering only one relatively costly feature. Moreover, the final representation memory usage and the query time (without considering feature extraction) is several times lower compared to methods using BoW.

It should be mentioned that recently, methods using Convolutional Neural Networks (CNN) have been proposed for the CBIR problem [12,13,17,18,20,21]. Their results are very promising, being CNNs very efficient. However, they still do not obtain state-of-the-art performance, although they have been used in several fusion frameworks with a great outcome [13,17,18,20]. They are perfectly compatible with this work's proposal, yet they are not used, to highlight how other individually worse performing methods can fuse successfully.

The rest of the paper is organized as follows: Section 2 reviews relevant work regarding re-ranking and fusion methods, and Section 3 details the proposal. Finally, in Section 4 the experimental results are presented and discussed and the concluding remarks are given in Section 5.

2. Related work

Fusion methods can be divided into two main categories: early and late fusion [22]. In early fusion, different descriptors are combined at feature level. This combination is used to produce a single representation. On the other hand, late fusion methods produce one representation for every feature, doing the fusion at score or decision levels. The proposed method, in essence, belongs to the later category. However, both categories will be reviewed, as they are relevant in the state-of-the-art of CBIR.

2.1. Early fusion in CBIR

For dense vector features (like FV and VLAD), one of the simplest fusion methods is vector concatenation [23–25]. It is very easy to implement and relatively efficient. Douze et al. [23] concatenate attributes scores, obtained through classifiers based on different features, with a Fisher Vector to obtain a final representation. Gordo et al. [24] proposed to use category-level labels of image classification datasets, to learn sub-spaces to reduce the dimension of two concatenated Fisher Vectors based on SIFT (Scale Invariant Feature Transform) [26] and statistical color features respectively. On the other hand, Mardones et al. [25] used three Fisher Vectors based only on SIFT descriptors, varying the sampling method used to obtain them, demonstrating that the use of different sampling methods is an important way to introduce diversity among the representations, knowledge that is harnessed here. Each one of these works achieved an important boost in performance compared to the use of the best individual feature. The main advantages of these methods are their usually low memory footprint (depending only on the individual representations) and their near-zero fusion cost, though the feature extraction process can still be costly. Their disadvantage is that they achieve lower precision than other fusion methods and they are very susceptible to false positives, since every feature is given equal weight. Similarly, the algorithm we propose, uses several dense vectors to maintain the memory usage to a minimum, but they are fused in a late fusion framework, obtaining additional neighborhood and similarity information in the process, to enhance its precision and robustness.

Other methods—using a BoW or similar frameworks—fuse different features in the indexing level. Bag of Colors [27] and coupled Multi-Index (c-MI) [15] combine local features using complementary cues to filter out false positive SIFT matches. Zhang et al., in [13], propose a co-indexing approach to jointly embed low-level image contents and semantic attributes from large-scale object recognition. Alternative related methods focus on vocabulary fusion [28,29]. Jégou and Chum merged multiple BoW vocabularies to alleviate the quantization effect [28], while Zheng et al. [29] used a probabilistic framework to take into account the different vocabularies correlation. The differences between these works and this proposal, without taking into account the methodological differences due to the fusion type, lie in the memory usage and time response linked with the use of inverted indices, that grows with the number of features. In contrast, by using solely Fisher Vectors (VLADs or CNNs), the representations' size is fixed. Additionally, the precision obtained in this work, compared to the best precision reported in [13] (using a CNN as part of the fusion), is superior in the datasets used for comparison.

2.2. Late fusion in CBIR

Late fusion methods, as previously stated, combine different representations at score or decision levels. In CBIR, most works employ—in the fusion process—one of the two following data