Pattern Recognition 47 (2014) 3466–3476

Contents lists available at ScienceDirect

Pattern Recognition

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

Visual query expansion with or without geometry: Refining local descriptors by feature aggregation

Giorgos Tolias *, Hervé Jégou

INRIA, France

A R T I C L E I N F O

Article history:
Received 3 October 2013
Received in revised form
26 March 2014
Accepted 2 April 2014
Available online 12 April 2014

Keywords:
Image retrieval
Query expansion
Hamming embedding

A B S T R A C T

This paper proposes a query expansion technique for image search that is faster and more precise than the existing ones. An enriched representation of the query is obtained by exploiting the binary representation offered by the Hamming Embedding image matching approach: the initial local descriptors are refined by aggregating those of the database, while new descriptors are produced from the images that are deemed relevant.

The technique has two computational advantages over other query expansion techniques. First, the size of the enriched representation is comparable to that of the initial query. Second, the technique is effective even without using any geometry, in which case searching a database comprising 105k images typically takes 79 ms on a desktop machine. Overall, our technique significantly outperforms the visual query expansion state of the art on popular benchmarks. It is also the first query expansion technique shown effective on the UKB benchmark, which has few relevant images per query.

1. Introduction

This paper considers the problem of image and object retrieval in image databases comprising up to millions of images. The goal is to retrieve the images describing the same visual object(s) as the query. In many applications, the query image is submitted by a user and must be processed in interactive time.

Most of the state-of-the-art approaches derive from the seminal Video-Google technique [1]. It describes an image by a bag-of-visual-words (BOVW) representation, in the spirit of the bag-of-words frequency histograms used in text information retrieval. This approach benefits from both the powerful local descriptors [2,3] such as the SIFT, and from indexing techniques inherited from text information retrieval such as inverted files [4,5]. Exploiting the sparsity of the representation, BOVW is especially effective for large visual vocabularies [6,7].

This analogy with text representation is a long-lasting source of inspiration in visual matching systems, and many image search techniques based on BOVW have their counterparts in text retrieval. For instance, some statistical phenomenons such as burstiness or co-occurrences appear both in texts [8,9] and images [10–12] and are addressed in similar ways.

One of the most successful techniques in information retrieval is the query expansion (QE) principle [13], which is a kind of automatic relevance feedback. The general idea is to exploit the reliable results returned by an initial query to produce an enriched representation, which is re-submitted in turn to the search engine. If the initial set of results is large and accurate enough, the new query retrieves some additional relevant elements that were not present in the first set of results, which dramatically increases the recall.

Query expansion has been introduced to the visual domain by Chum et al. [14], who proposed a technique implementing the QE principle and specifically adapted to visual search. Several extensions have been proposed to improve this initial QE scheme [15–17]. Although these variants have improved the accuracy, they suffer from two inherent drawbacks which severely affect the overall complexity and quality of the search:

- First, they require a costly geometrical verification step, which provides the automatic annotation of the relevant set and is typically performed on hundreds of images.
- Second, the augmented query representation contains significantly more non-zero components than the original one, which severely slows down the search. It is reported [17] that typically ten times more components are non-zeros. Since querying the inverted file has linear complexity in the number of features contained in the query vector, the second query is therefore one order of magnitude slower than the first.

Expansion methods that do not use any costly geometrical verification are typically based on an off-line stage with quadratic

* Corresponding author.

http://dx.doi.org/10.1016/j.patcog.2014.04.007
0031-3203/

complexity in the number of database images [18–20]. They are thus limited to collections of small and fixed size.

In another line of research, several techniques address the loss in quantization underpinning BOVW, such as the use of multiple assignment [21] or soft quantization [22]. In a complementary manner, the Hamming Embedding (HE) technique [23] dramatically improves the matching quality by refining the descriptors with binary signatures. HE is not compatible with existing QE techniques because these assume a vector representation of the images. A noticeable exception is the transitive QE, which does not explicitly exploit the underlying image representation. However, this variant is not satisfactory with respect to query time and performance.

This paper, for the first time, proposes a novel way to exploit the QE principle in a system that individually matches the local descriptors, namely the HE technique. The new query expansion technique is both efficient and precise, thanks to the following two contributions:

- First, we modify the selection rule for the set of relevant images so that it does not involve any spatial verification. The images deemed relevant provide additional descriptors that are employed to improve the original query representation. Unlike other QE methods, it is done *on a per-descriptor basis* and not on the global BOVW vector. Fig. 1 depicts an example of images and features that are selected by our method to refine the original query.
- The second key property of our method is that the set of local features is aggregated to produce new binary vectors defining the new query image representation. This step drastically reduces the number of individual features to be considered when submitting the enriched query.

To our knowledge, it is the first time that a visual QE is successful without any geometrical information: the only visual QE technique [14] that we are aware of performs poorly compared with other variants such as the average query expansion (AQE). In contrast, our technique used without geometry reaches or outperforms the state of the art. Interestingly, it is effective even when a query has few corresponding images in the database, as shown by our results on the UKB image recognition benchmark [6]. Incorporating geometrical information in the pipeline further improves the accuracy. As a result, we report a large improvement compared to the state of the art. We further demonstrate the superiority of our method compared to a simple combination of HE with QE: the property of feature aggregation not only reduces the expanded query complexity, but further improves performance.

The paper is organized as follows. Section 3 introduces our core image system and Section 7 a post-processing technique for SIFT descriptors that is shown useful to improve the efficiency of the search. Section 4 introduces our Hamming Query Expansion (HQE) method and Section 5 describes our key aggregation strategy of local features. Section 6 describes how to exploit geometrical information with HQE. The experimental results presented in Section 8 demonstrate the superiority of our approach over concurrent visual QE approaches, with respect to both complexity and search quality, on the Oxford5k, Oxford105k and Paris benchmarks.

2. Related work

Chum et al. [14] were the first to translate the query expansion principle to the visual domain. Most of the variants they propose rely on a spatial verification method, which filters out the images that are not geometrically consistent with the query. The authors investigate several methods to build a new query from the images deemed relevant. The average query expansion is of particular interest and usually considered as a baseline, as it is the most efficient variant [14] and provides excellent results. It is conceptually simple: a new *term-frequency inverse document frequency* (TFIDF) vector is obtained as the average of the results assumed correct and spatially back-projected to the original image.

Following this first work, a number of QE variants and extensions have been proposed [15–17]. Using incremental spatial re-ranking, the query representation is updated by each spatially verified image and extended out of the initial query region [16]. Another extension is to learn, on-the-fly, a discriminative linear classifier [17] to define the new query instead of the average in AQE.

Other kinds of expansion have been proposed for fixed image collections [17,24]. They rely on the off-line pairwise matching of all image pairs and aim at identifying the features coming from the same object using spatial verification, which is rather costly as the complexity is quadratic in the number of images. They also assume that the image collection is fixed: the selection depends on a given set of images. These methods are also related to other methods exploiting the neighborhood of the images within a given collection [21,18], in particular by updating the comparison metric or by employing reciprocal nearest neighbors as a filtering rule. For instance, Qin et al. [18] construct a graph that links related images, and uses k-reciprocal nearest neighbors at query time to define a new similarity function that re-orders the images. Again, the cost of constructing and storing the graph in memory is impracticable for large datasets. In a similar spirit, Shen et al. [19] exploit the

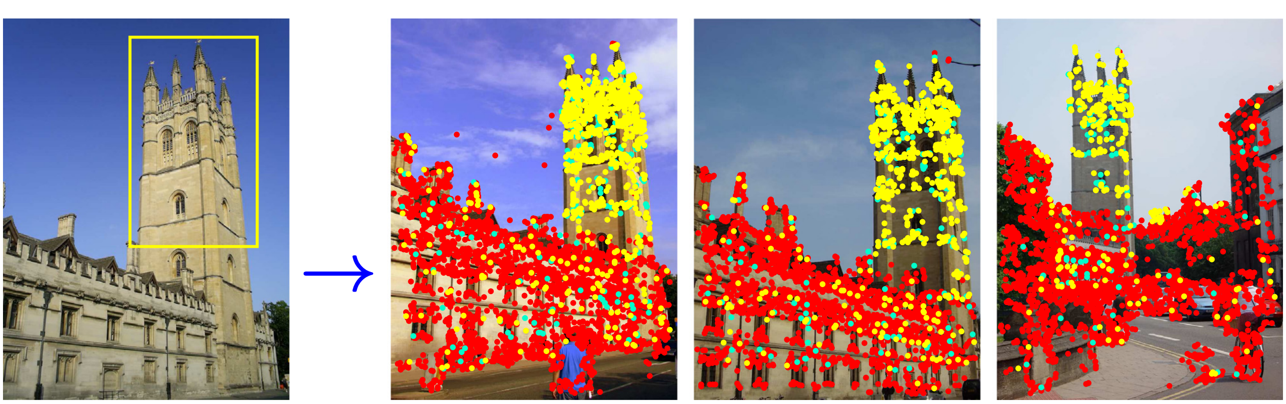

Fig. 1. Query image (left) and the features selected (yellow + cyan) from the retrieved images to refine the original query. The features in red are discarded. Cyan features correspond to visual words that appear in the query image, and yellow ones to visual words that were not present in it. The selection of the depicted images and features has not involved any geometrical information. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)