J. Vis. Commun. Image R. 37 (2016) 32–39

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Local consistent hierarchical Hough Match for image re-ranking ☆

Yuanzheng Cai [a,b], Shaozi Li [a,b,*], Yun Cheng [c], Rongrong Ji [a,b]

[a] Cognitive Science Department, Xiamen University, China
[b] Brain-like Intelligent systems, Xiamen University, China
[c] Hunan University of Humanities Science and Technology, Hunan, China

ARTICLE INFO

Article history:
Received 30 December 2014
Accepted 7 June 2015
Available online 14 June 2015

Keywords:
Image retrieval
Bag-of-words
Geometry rerank
Hierarchical
Hough Space
Large scale retrieval
Local consistence
Inverted index

ABSTRACT

Geometric image re-ranking is a widely adopted phrase to refine the large-scale image retrieval systems built based upon popular paradigms such as Bag-of-Words (BoW) model. Its main idea can be treated as a sort of geometric verification targeting at reordering the initial returning list by previous similarity ranking metrics, e.g. Cosine distance over the BoW vectors between query image and reference ones. In the literature, to guarantee the re-ranking accuracy, most existing schemes requires the initial retrieval to be conducted by using a large vocabulary (codebook), corresponding to a high-dimensional BoW vector. However, in many emerging applications such as mobile visual search and massive-scale retrieval, the retrieval has to be conducted by using a compact BoW vector to accomplish the memory or time requirement. In these scenarios, the traditional re-ranking paradigms are questionable and new algorithms are urgently demanded. In this paper, we propose an accurate yet efficient image re-ranking algorithm specific for small vocabulary in aforementioned scenarios. Our idea is inspired by Hough Voting in the transformation space, where votes come from local feature matches. Most notably, this geometry re-ranking can easily been aggregated to the cutting-edge image based retrieval systems yielding superior performance with a small vocabulary and being able to store in mobile end facilitating mobile visual search systems. We further prove that its time complexity is linear in terms of the re-ranking instance, which is a significant advantage over the existing scheme. In terms of mean Average Precision, we show that its performance is comparable or in some cases better than the state-of-the-art re-ranking schemes.

1. Introduction

In this paper, we consider the problem of large scale particular object retrieval where the goal is to retrieve all images containing a specific object in a large scale image dataset, given a query image. The goal is required to be performed in near real time so that users can interactively browse the dataset or search using images from their devices. More generally, the devices will include mobile devices, such as Google Goggles and Point and Find from Nokia. Compared with traditional image search in PC end, the search in mobile device's performance will suffer by hardware performance and network limitation, hence it is quite a worthy of challenge, many works such as Refs. [1,2,28–31] have proposed the vocabulary learning methods which aim to store a small vocabulary in the mobile terminal or extract compact visual descriptors. Despite extensive works in the feature representation and ranking function design, in this paper we focus on the component of geometric image re-ranking, i.e., refine the initial top returnings by considering the geometric consistency of local feature matchings between query and reference image pairs.

General speaking, geometric image re-ranking is typically leveraged as a post processing given the initial BoW based image search results. In this stage, the top N initial images will be re-ranked to filter the wrong matches by measuring the spatial consistency between the local features extracted from the query image and reference image pairs using methods such as spatial consistency voting [3] and RANSAC [4]. For instance, the spatial consistency scheme proposed in the VideoGoogle system [5] adopts the weak geometric constrain, while the RANSAC algorithm targets at keeping the consistency matches in Hough Space. However, the geometric re-ranking retains an open problem due to the following issues:

First, to ensure the retrieval precision, the size of to Visual word Vocabulary is normally very large, e.g., one-million codewords, targeting at minimizing the quantization loss while engaging in extensive storage consumption. This is unacceptable for several emerging applications, e.g., Mobile Visual Search, in which the features are directly extracted and sent from the mobile terminal to minimize the query delivery latency (instead of sending the query

☆ This paper has been recommended for acceptance by Luming Zhang.

* Corresponding author at: Cognitive Science Department, Xiamen University, China.

E-mail address: szlig@xmu.edu.cn (S. Li).

http://dx.doi.org/10.1016/j.jvcir.2015.06.003
1047-3203/© 2015 Elsevier Inc. All rights reserved.

image). In this case, the mobile-end feature extraction system cannot afford such a large vocabulary. However, by using small vocabulary, both quantization loss [6] and burstiness [7] effect can be caused, which would seriously degenerate the search accuracy.

Second, the time consuming of re-rank phrase is linear in the number of images to match, while within the image the complexity is beyond linear to the number of local features, for instance the most popular approaches like RANSAC needs quadratic time complexity, which hesitates the re-ranking algorithms to be applied to be as many returning images as possible. Especially, the increasing time overhead is mainly caused by imposing one-to-one mapping [8] constraint to filter the substantial potential matches and this one-to-one mapping inevitably lead to the increasing of algorithm complexity.

Third, for some object retrieval, the re-rank phrase needs to explore the geometry structure between query image and dataset images, which has better estimate consist of variant about scale, orientation, etc.

To overcome the above challenges, we propose a hierarchical spatial Hough Voting model which applies the analogous concept of pyramid match [9] to the transformation space. The key idea here is to use local feature shape to generate votes first, it is invariant to similarity transformations. Then we also utilize the hierarchical framework to build a non-iterative, down-top grouping process which overcome the effects of quantization loss, i.e., burstiness, we alleviate its impact by strengthening local transformation consistency of each word. We represent correspondings in the transformation space exploiting local feature shape as in [9]. In [9], local feature shape, i.e. the scale, orientation position of a key point which is detected by DoG [8], which is dominant in all local features can reflect the transformation of two images. The two main difference from [9] is that: first, inspired by pyramid match [10], we form correspondings by using visual vocabulary instead the features. The transformation is estimated by the multi-scales bins in Hough Space. Second, we do not impose a one-to-one which maps constraint such that each feature in one image is mapped to at most one feature in the other one. This is due to that the constraint mentioned above needs a very fine codebook, by using which the features are uniquely mapped to visual words. However, this condition is hard to satisfy when using a codebook of smaller size. We use an independent transformation space to approximate the one-to-one map instead.

The main contributions of this paper are summarized as follows:

1. We propose a new geometric re-ranking framework based on BoW model for image retrieval. This framework exploits corresponding matches of similarity transformation in a hierarchical top-down and local-to-global Hough Voting space. Most notably, we adopt the non-one-to-one mapping which makes it liner time complex and the performances a dramatic speedup.
2. This framework do not rely on a large or fine codebook in particular. can handle the problem such as burstiness or others effects caused by corrupt quantization loss or small size codebook. Specially, this advantage makes it being able to apply to some device limited application such as Mobile Visual Search (MVS).
3. It is a very simple algorithm that requires no learning and can easily integrated into the inverted-index files of any image retrieval process.

The rest paper is organized as follows: we first give an overview of related work of image retrieval model base on BoW in Section 2. In Section 3, we propose the hierarchical Hough Voting framework for re-rank and discuss the algorithm in Section 4. Finally, we evaluate the proposed framework by simulation experiments in Section 5, and draw conclusions in Section 6.

2. Related work

We briefly review the state-of-the-art works in both large-scale image retrieval and geometric re-ranking as below.

2.1. Large-scale image retrieval

The best existing approaches inherit from the Bag-of-Words (BoW) representation [5] [11]. It employs a visual vocabulary to quantize a set of local descriptors of each image into a single BoW vector each dimension represents the occurrence of its corresponding word. It has shown quite promising performance in various tasks such as image retrieval, object recognition, and image classification. The key consideration for using BoW based representation for these tasks lies on the design of vocabulary size and similarity function, e.g., for image classification, due the curse of dimensionality, classifiers are more suitable for relatively small visual vocabularies. But for large scale image retrieval, large vocabularies can make the search more accurate and efficient [12] with inverted file. In such a case, the sparsity of high-dimensional BoW can be exploited to assign only limited number of term frequency inverse document frequency (tf-idf), enabling query to be done in real-time even for million-scale datasets. A further improvement is proposed in [13] which utilizes machine learning to optimize the distribution of vocabulary indexing. In this scheme, the images index has been distributed to multiple servers to ensure that the online retrieval can be proceed in a parallel way.

In [5], BoW is leveraged to build a video search engine. BoW is leveraged to build a video search engine. First, trajectory descriptions and clustering are adopted to generate the visual vocabulary.

The work in [4] extends BoW model to million-scale images retrieval. Approximate-k-mean (AKM) [4] is proposed to clustering Hessian Affine interest points to an extra large vocabulary.

For BoW vector, an image can fail to be retrieved for a number of factors. One of most primary reason is due to the descriptor quantization loss, i.e., descriptors are assigned to different words, while irrelevant pairs are assigned to the same word. To overcome the quantization loss, Hamming embedding has been propose in [12], in which descriptors are first quantized into corresponding words via vocabulary. Then descriptors will be further coded into a binarization vector by a number of hash-tables. Another way to decrease the quantization loss is [14], which computes the residual of whole image's descriptors of the visual words and present image as a compact global descriptors. In particular, [2] proposes a Location Discriminative Vocabulary Coding (LDVC) method for mobile landmark search via spectral clustering and a ranking sensitive vocabulary boosting to learn LDVC. This scheme makes it possible to store vocabulary in the mobile terminal.

Ref. [15] propose a Hierarchical clustering, which build the vocabulary as tree structure. In this framework, single descriptor can be quantized to a branch of visual words. Soft-assignment proposed in [6] further assign each descriptor to multiple words with different weights which reflect the quantization confidence of each word.

Furthermore, Refs. [16,17] propose a novel product quantization which can approximately decomposes a large size vocabulary as combinations of multiple sub-vocabularies. With this approach, the overhead of vocabulary can decrease greatly, however, it mainly focus on approximate near neighbor search. In addition, Ref. [7] surveys a case called burtiness that image has large number of repeated patterns are quantized to little words in images