ARTICLE IN PRESS
JID: NEUCOM [m5G;September 19, 2017;2:17]
Neurocomputing 000 (2017) 1–10
Contents lists available at ScienceDirect
Neurocomputing
journal homepage: www.elsevier.com/locate/neucom

Region similarity arrangement for large-scale image retrieval

Dongming Zhang[a,b,d], Jingya Tang[a,b,*], Guoqing Jin[a,b], Yongdong Zhang[a,b], Qi Tian[c]

[a] Institute of Computing Technology, Chinese Academy of Sciences, China
[b] Key Lab of Intelligent Information Processing, Chinese Academy of Sciences, China
[c] Department of Computer Science, University of Texas at San Antonio
[d] National Computer Network Emergency Response Technical Team/Coordination Center of China

ARTICLE INFO

Article history:
Received 30 May 2016
Revised 12 January 2017
Accepted 9 July 2017
Available online xxx

Communicated by Dr Xinmei Tian

Keywords:
Content based image retrieval
Geometric verification
Region property space
Spatial weighting

ABSTRACT

We propose a promising method of geometric verification to improve the precision of Bag-of-Words (BoW) model in image retrieval. Most previous methods focus on the positions of interest points or the absolute differences of regions' scales and angles. In contrast, our method, named Region Similarity Arrangement (RSA), exploits the spatial arrangement of interest regions. For each image, RSA constructs a Region Property Space, regarding each region's (scale, angle) pair as a point in a polar coordinate system, and encodes the arrangement of these points into the BoW vector. Furthermore, based on the particular distribution of points in Region Property Space, we design a Spatial Weighting to reduce the burstiness phenomenon during query. From experimental results on Holidays, Oxford5K and Paris, RSA could get comparable results with state-of-the-art methods. In addition, RSA increases no extra memory and negligible computational consumption compared with the baseline BoW approach.

1. Introduction

In image retrieval scenario, many state-of-the-art search engines rely on the Bag-of-Words (BoW) framework [1]. In this model, affine-invariant Hessian regions [2] are detected and a local descriptor such as SIFT [3] is extracted in each of these affine regions. Then local descriptors are quantized to visual words according to a pre-trained dictionary, mostly generated by K-means algorithm or approximate K-means [4] for efficiency. After that, each visual word is weighted using inverse document frequency (IDF) [1] inspired by text retrieval. An image is finally represented with a sparse BoW vector and an inverted index system is leveraged for fast accessing.

Many approaches have been proposed to improve the performance of BoW, such as query expansion [5], soft assignment [6], embedding technologies [7,8] and multi-dimensional inverted index [9,10]. What could be an amelioration of these approaches is to utilize geometric information of visual words. Therefore, geometric verification has caught a lot of attention and many works [11–17] have been investigated to improve the precision. Some methods [4,5,7,16,17] utilize geometric verification as a post-processing step, checking geometric consistency between matched local features. Other methods [11–14] encode spatial information into BoW vectors.

In this paper, we propose Region Similarity Arrangement (RSA) to improve the performance of BoW. RSA measures the similarity arrangement of different interest regions in two steps: First, a *Region Property Space* is constructed for each image. The space is based on a polar coordinate system, in which each point is determined by a (*scale, angle*) pair corresponding to a region's property. In this space, the properties of interest regions are described as a point set, which is robust to scale change due to the scale invariance property of SIFT. Second, to encode the arrangement of points in *Region Property Space*, RSA consecutively partitions the space into four quadrants considering each point as the origin and counts the number of other points appearing in each quadrant. Through these operations, each visual word has one 4-tuple RSA vector, through which the arrangement of interest regions is embedded into the BoW vector of an image.

Based on the particular distribution of points in *Region Property Space*, we propose spatial weighting to further improve the performance of RSA. The spatial weighting can reduce the influence of false matching RSA vectors in query. To resolve the memory consumption problem of RSA, each RSA vector is quantized to four unsigned chars and the experiments show that the quantization will not deteriorate its performance. The retrieval experiments are mainly performed on Holidays [7], Oxford5K [4] and Paris [6] datasets. RSA shows high performance and distinctive scalabil-

* Corresponding author at: Institute of Computing Technology, Chinese Academy of Sciences, China.
E-mail address: tangjingya@ict.ac.cn (J. Tang).

http://dx.doi.org/10.1016/j.neucom.2017.07.025
0925-2312/© 2017 Published by Elsevier B.V.

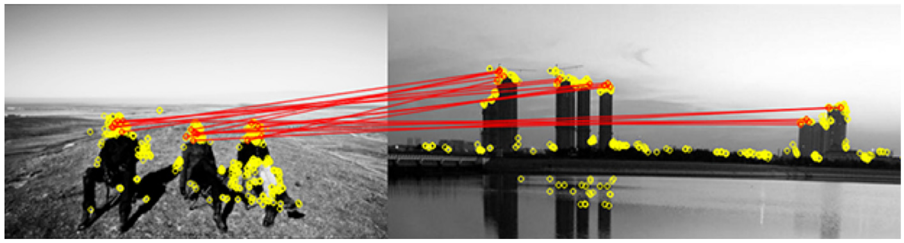

Fig. 1. False matched images returned by Word Spatial Arrangement. The yellow circles are interest points, and the red lines link the same visual words considering their positions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ity, achieving comparable performance with [18] and better results than [16] without gravity vector constrains [19]. In addition to the high performance, RSA needs no extra memory and negligible increase of matching time compared with the traditional BoW model.

2. Related work

RANSAC can achieve the state-of-the-art retrieval accuracy, but even the improved methods [4,5] based on RANSAC still occupy excessive memory and impact the efficiency of search. So typically, only top-ranked images are considered in [4,5]. Zhong et al. [16] proposed an efficient direct spatial matching (DSM) approach, which directly estimates the scale variation. It is much faster than RANSAC-based methods and achieves the state-of-the art results when gravity vector constrains [19] are satisfied.

By explicitly encoding spatial relationship among visual words, Zhou et al. [17] proposed a powerful method named Spatial Coding for partial-duplicate image search. In Spatial Coding, two binary spatial maps [17] are generated to encode the relative spatial positions between each feature pair along the horizontal and vertical directions. The consistency of spatial layout between query and gallery images will be checked according to the spatial maps of the matched features. However, the Spatial Coding is powerful for partial-duplicate detection but not general image retrieval. Liu et al. [20] proposed to embed spatial context into inverted file to speed up geometrical verification. Liu et al. [21] further encode spatial context into binary strings to boost image retrieval performance.

Li et al. achieved the new state-of-the-art on Holidays dataset by using pairwise geometric matching (PGM) [22]. In [22], both global and pairwise geometric relations are considered. In order to find correct correspondences between two matched images, Li et al. [22] proposed a heuristic algorithm to find 1 vs. 1 matching in initial correspondence set and deploys hough voting to filter out not consistently transformed pairs. However, the process of finding correct correspondences is still time consuming and PGM can only evaluate top-ranked images.

Cao et al.[11,12] split an image into specific tiles and generate one BoW vector for each. Cao et al. [11] proposed Spatial Bag-of-Features, which is a generalization of spatial pyramid matching [12], and by some operations for histogram features, Spatial Bag-of-Features could handle typical transformations of objects, e.g., translation, rotation and scaling. However both [11] and [12] confront the curse of dimensionality, and [11] needs to train particular parameters for a given dataset.

Wang et al.[14] proposed spatial contextual weighting (SCW) to enhance the discrimination of visual words by exploiting the local statistics, which include the descriptor density, the mean relative log scale and the mean orientation difference. However, Wang et al. [14] only explore simple local statistics, so SCM could get trivial improvement, about 4% on Holidays dataset.

Weak Geometrical Consistency (WGC) [7] also utilizes the properties of regions, which verifies the consistency of the properties for the matched visual words of a given image by computing the histogram of the angle and scale differences. WGC can inject a prior knowledge, mainly the orientations of images of different datasets, to improve the precision. However, as a post-processing step and depending on prior knowledge of datasets, WGC is not suitable for general image retrieval. Contrary to directly use regions' properties in WGC, our method focuses on their arrangement, and this is more robust to changes in scales and orientations of images, so it needs no prior knowledge. In addition, RSA embeds properties of interest regions into BoW vectors in a global way, by encoding the arrangement of them in the *Region Property Space* of an image, which ensures RSA can not only be applied at post-processing, but also be embedded into BoW retrieval. Penatti et al. [13] proposed Word Spatial Arrangement (WSA) to encode relative positions of visual words in a compact way. However, the discrimination of words' positions is limited. Fig. 1 shows an example of the false match based on the spatial arrangement of word. In order to encode the arrangement of different regions, RSA partitions the *Region Property Space* into four quadrants inspired by WSA [13]. Fig. 2 shows an example of how to compute WSA vectors of visual words. For each interest point p_i detected in the image, a rectangular coordinate system is built, whose origin is p_i, and the x and y axes are the horizontal and vertical directions of the image respectively; then, the amount of other detected points appearing in each quadrant are counted, and the four numbers arranged in a fixed order compose the WSA vector of point p_i. After this process is repeated for every interest point detected, the counting process finishes. RSA also builds four quadrants based on each point, but implements it in the *Region Property Space*, which encodes regions' properties of visual words. RSA has all the merits of WSA such as (1) the information is encoded into the feature vector, which needs no post processing and (2) it

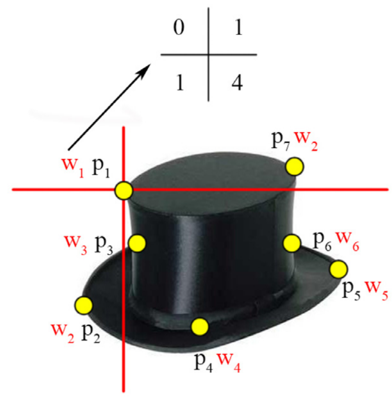

Fig. 2. Example of partitioning and counting for WSA. The yellow circles are the detected points, accompanied with their visual words. p_1 is considered as the origin in this particular example, and other points are treated in the same way. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: D. Zhang et al., Region similarity arrangement for large-scale image retrieval, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.07.025