Image and Vision Computing 31 (2013) 935–948

Contents lists available at ScienceDirect

Image and Vision Computing

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

Spatially aware feature selection and weighting for object retrieval ☆

Yanzhi Chen, Anthony Dick, Xi Li *, Anton van den Hengel

School of Computer Science, The University of Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 9 August 2012
Received in revised form 13 June 2013
Accepted 23 September 2013

Keywords:
Object retrieval
Bag-of-words
Spatial expansion
Visual word re-weighting

ABSTRACT

Many recent image retrieval methods are based on the "bag-of-words" (BoW) model with some additional spatial consistency checking. This paper proposes a more accurate similarity measurement that takes into account spatial layout of visual words in an offline manner. The similarity measurement is embedded in the standard pipeline of the BoW model, and improves two features of the model: *i*) latent visual words are added to a query based on spatial co-occurrence, to improve query recall; and *ii*) weights of reliable visual words are increased to improve the precision. The combination of these methods leads to a more accurate measurement of image similarity. This is similar in concept to the combination of query expansion and spatial verification, but does not require query time processing, which is too expensive to apply to full list of ranked results. Experimental results demonstrate the effectiveness of our proposed method on three public datasets.

1. Introduction

Recent years have witnessed increasing interest in the problem of object retrieval [1–7] and its applications [8–12]. Typically, the aim of this task is to select from a collection of images which contain the same object as a query image. Many works have addressed this problem with a standard "bag-of-words" (BoW) framework [1,13,14,2,15] borrowed from text retrieval. In this framework, local features in an image are detected and then represented by *D*-dimensional descriptors. To reduce memory usage and achieve fast indexing, these descriptors are clustered and then quantized to form a vocabulary of "visual words". An image is represented by a sparse frequency histogram over the visual vocabulary.

Although the BoW model is simple to implement and robust to some types of image variation, an object in a target image still can fail to be retrieved from the dataset, mainly due to quantization used in this standard pipeline. As noted by many works over the past few years, quantization introduces two main problems:

- Loss of recall: information about the raw features is lost, so matching features may be assigned to different visual words. This can decrease the matching score of images that contain the same object and therefore reduce recall.
- Loss of precision: features that do not correspond to the same location may be assigned to the same word. This can increase the matching score of images that do not contain the same object, and therefore reduce result precision.

To address the first problem, a number of methods focus on improving the query recall. For example, soft-assignment [4,16] method maps a raw feature to multiple visual words; in [17,18] a more accurate approximate nearest neighbor assignment is implemented by combining k-means clustering and binary vector signatures; the dimension reduction and the indexing algorithm are jointly optimized in [19,20], such that the image representation provides accurate search results with low vector dimensionality. Other methods aim to learn a better metric for feature quantization [5] or group similar features when generating the vocabulary [21,6]. Query expansion [3,22,23] addresses the loss of recall in a post-processing step. It enriches the query model by adding query relevant words collected from the initial retrieval results. To address the second problem, a re-ranking process is required [2,17,24], which filters out highly ranked false positives in the initial retrieval results. These methods mainly rely on the spatial consistency of visual words in pairs of query and dataset images. Precision is increased after images that do not contain a significant number of spatially consistent matches are re-ranked lower.

In this paper, we address loss of precision and recall by proposing a novel spatial co-occurrence measure. We use an image similarity measurement that takes into account the spatial layout of visual words, but can be embedded into the standard retrieval pipeline. We do this by creating a visual thesaurus that records spatial co-occurrence information for each pair of words in a vocabulary. Based on that, we explore two types of image-dependent contextual information to improve the BoW similarity measure: a voting-based method to include latent visual words and an information theory based measurement to re-weight the

☆ This paper has been recommended for acceptance by Y. Aloimonos.

* Corresponding author. Tel.: +61 8 8313 1855; fax: +61 8 8313 4366.
E-mail address: xi.li03@adelaide.edu.au (X. Li).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.09.005

importance of each visual word. This enables us to improve both recall and precision of the standard BoW object retrieval method.

Spatial expansion. A query can be expanded by including latent words, which are highly correlated with those already present in the query. The correlation is deduced from the degree of spatial co-occurrence of the visual words. In this way, the query recall can be improved.

Visual word re-weighting. After additional words are introduced to the query, the tf-idf scheme weights each visual word according to its frequency in the individual image as well as in the corpus. However, tf-idf weights do not distinguish between word appearances in the foreground or background of an image. Therefore it might assign high weights to words that are not informative when searching for an object, or underweight words that are informative. Thus we develop a word weight that is more directly based on how often, and in what range of conditions, a word is correctly matched when it appears as part of the foreground object. Using the automatic training data collection method of [25], we select a subset of visual words that frequently occur as inliers to object matches robustly estimated between images. Under the standard tf-idf weighting, these visual words are not necessarily weighted strongly—for example they may occur in many images in the database, and therefore have a low idf weight. Based on these inlier measurements we use entropy to measure the importance of a visual word according to its spatial co-occurrence distribution. A re-weighting scheme is proposed in this paper to encourage these informative visual words. In this way, query precision can be improved.

As illustrated above, and described in more detail in Section 4, our method captures the underlying spatial relationship among the visual words, and their importance as an indicator of a foreground object, and thereby improves both precision and recall. Our method is similar to recent methods [26,3], but inspired by those which also discover the spatial relationships among the visual words [27–30]. As we will explain in the following section in more detail, the key differences between our work and previous methods are that ours focus on improving precision and recall without the use of result re-ranking or issuing multiple queries. This allows a trade-off between effectiveness and efficiency by adjusting the query vector online to incorporate this spatial and relevance information, based on weights computed offline. In addition, our method can be applied wherever the BoW model is used.

The remainder of the paper is organized as follows: Section 2 briefly reviews related works about the retrieval methods. Section 3 outlines our method and Section 4 describes the details of methods, including the definition of the visual thesaurus and two usages of the visual thesaurus: spatial expansion (Section 4.1) and visual word re-weighting (Section 4.2). We report the experimental results in Section 5. The paper is finally concluded in Section 6.

2. Related work

The BoW model for representing images has been widely used in computer vision problems [31,13,32,1,8–10]. Here, we focus on methods that incorporate spatial context into the BoW model. These methods can be divided into two groups: *i*) spatial context derived from related images; or *ii*) spatial context derived from related visual words.

The first group of methods is based on the spatial verification of images [2]: pairwise images that pass a geometric consistency test are likely to contain the same object. Typically, a RANSAC matching method is used and those images whose inlier match count exceeds a threshold are accepted. Using spatial verification, query expansion methods refine the query model by adding words from spatially verified regions in result images [3,22] or refine the distance measure with a k-reciprocal nearest neighbor test in image space [24]. Alternatively, spatial verification can be used to detect and match visually similar features that have been assigned to different visual words [33].

The second group of methods focuses on a bag-of-phrases structure. The works proposed in [30,34,26] rely on the spatial co-occurrence of visual words in order to assemble those containing a high-order relationship into visual phrases. The visual phrases are selected from a fixed image grid in [30] or K nearest neighbors in a fixed spatial region [26]. The visual phrases are more flexibly organized in [34], which bundles local features in a randomized partition of the image. Alternatively, methods by [27,35] use spatial co-occurrence within feature space. Local features are projected to different directions and the spatial information is encoded as the ordered bag-of-features representation in [27]. In contrast, [35] use neighboring local features to softly down-weight less informative words.

Most existing methods built on the BoW model generate a large vocabulary (e.g. 1 million visual words) by clustering features from a large image dataset. Therefore, [28] argue to reduce memory requirements by selecting a small subset of features that exhibit spatial consistency. We use a similar method in this paper to select and re-weight "informative" visual words.

The idea of our method is also similar in spirit to [3,26]. Different from [3], our spatial expansion method is based on relations found in an offline step rather than mining query results online. In [26], the visual context is defined as a set of local points which are in a fixed spatial region of a visual word. These are used to expand the visual words used in the query vector. The method needs to partition the spatial region into K sectors, and for each sector a sub-histogram is constructed to record the visual context. In contrast, our method uses a spatial co-occurrence histogram (termed a visual thesaurus in [36]). As a result, the computation of our method is simpler than [26] because we only need to count votes or calculate entropy in independent visual thesaurus histograms.

3. Overview

In this paper, we propose a *visual thesaurus* data structure to record the spatial relations between visual words, and use it to refine the BoW model. The architecture of our framework is shown in Fig. 1. Our method is composed of an offline process and an online process. The offline process creates visual words and weights them in terms of their occurrence in the whole dataset. In this stage, the weights of the visual words are adjusted according to their importance in the images. The online process stage aims to retrieve the relevant images by matching the visual words vectors. In this stage, the query vector is expanded and re-weighted using the knowledge learned from the offline stage. Algorithm 1 describes the work flow of our method. For descriptive convenience, we first introduce some main notations in Table 1.

Algorithm 1 Overview of our method.

1. **Spatial expansion**
   - Offline: Build a general thesaurus. Loop over each image to find the co-occurrence of each pair of visual words (Section 4.1).
   - Online: Expand the query visual words from general visual thesaurus (Section 4.1). The expansion is based on frequent co-occurrence of visual words.
2. **Visual word re-weighting**
   - Offline: Automatic training data collection via RANSAC [5].
   - Offline: Build an object based thesaurus from the training data (Section 4.2), which only considers the points detected as *inliers* in RANSAC.
   - Offline: Calculate the entropy H from object based visual thesaurus (Section 4.2).
   - Offline: Re-weight the visual words according to their entropy (Section 4.2).

4. The visual thesaurus and applications

The visual thesaurus captures the spatial relationship among pairs of visual words. Assume that there are N visual words $W := \{w_i\}_{i=1}^{N}$ in the vocabulary. The co-occurrence of a pair of words (w_i, w_j) can be expressed as the joint probability of w_i and w_j occurring in the same spatial region, defined as $\Pr(w_i, w_j)$. For a given word w_i, we can obtain up to