Computer Vision and Image Understanding 116 (2012) 238–249

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Visual synonyms for landmark image retrieval

Efstratios Gavves *, Cees G.M. Snoek, Arnold W.M. Smeulders

Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 December 2010
Accepted 16 October 2011
Available online 7 November 2011

Keywords:
Image representation
Image retrieval

ABSTRACT

In this paper, we address the incoherence problem of the visual words in bag-of-words vocabularies. Different from existing work, which assigns words based on closeness in descriptor space, we focus on identifying pairs of independent, distant words – the visual synonyms – that are likely to host image patches of similar visual reality. We focus on landmark images, where the image geometry guides the detection of synonym pairs. Image geometry is used to find those image features that lie in the nearly identical physical location, yet are assigned to different words of the visual vocabulary. Defined in this way, we evaluate the validity of visual synonyms. We also examine the closeness of synonyms in the $L2$-normalized feature space. We show that visual synonyms may successfully be used for vocabulary reduction. Furthermore, we show that combining the reduced visual vocabularies with synonym augmentation, we perform on par with the state-of-the-art bag-of-words approach, while having a 98% smaller vocabulary.

1. Introduction

In recent years several local visual features have been proposed, which encode the richness of localized visual patches [1,2]. Although these features perform well in object and concept recognition as exemplified in the advances of TRECVID and PASCAL [3,4], the detection and transformation of the visual reality of an image patch into a feature vector is far from perfect [5,6]. Despite this fact and to the best of our knowledge, there has been so far limited research of the high dimensional visual feature space formed and its properties.

For their ability to capture local visual information well enough, local feature detectors and descriptors are mostly used. Feature detectors and descriptors operate directly on the raw visual data of image patches, which are affected by common image deformations. These image deformations affect either image appearance, which accounts for the way the image content is displayed, or image geometry, which accounts for the spatial distribution of the image content inside the image. Image appearance variations include the abrupt changes of illumination, shading and color constancy [7]. Image geometry variations are related to viewpoint changes, non-linear scale variations and occlusion [8–12]. Several feature descriptors that provide invariance against image appearance deformations have been proposed [7]. However, there are no specific features that deal adequately with image geometry deformations. Instead, this level of invariance is partly reached on the next level of image representation, using for example the bag-of-words model [13–16]. Despite this *a posteriori* acquired invariance under geometric deformations, feature vectors of similar visual reality are still erroneously placed in very different parts of the feature space. Thus, the image feature space spanned by local feature detectors and descriptors is fuzzily populated.

Moreover, to be sufficiently rich to capture any local concept the visual feature space has to be of high dimensionality. However, distance measures in high dimensional spaces exhibit a more sensitive nature [17]. Thus distance measures, a cornerstone of most machine learning algorithms, are less indicative of the true similarity of two vectors, which as a result disturbs the image retrieval process. Therefore, error-prone distance measures also contribute to the fuzzily populated feature space.

By treating local image descriptors as orderless words, images in the bag-of-words model may be classified in a class on the basis of word histograms. In effect, bag-of-words hopes for large number statistics to even out the consequences of the aforementioned image deformations. Words are obtained by clustering in the descriptor space [18], implicitly assuming that all patches covered by one word represent the same part of reality. And, that different clusters correspond to different parts of reality. These clusters lie inside the fuzzily populated feature space, resulting in visual words that have little coherence in the semantics of the patches they contain, see Fig. 1. For standard challenges, like PASCAL which targets at general object recognition, visual word incoherence does not affect the performance drastically and vocabularies of size up to 4 K clusters suffice. However, for more challenging datasets, like *Oxford5k* [14] or [19], image appearance and geometry deformations start to have a much greater impact. Hence techniques that make better use of the feature space are needed. For complex datasets, larger vocabularies have proven to operate more effectively [14,19].

* Corresponding author.
E-mail address: egavves@uva.nl (E. Gavves).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.10.004

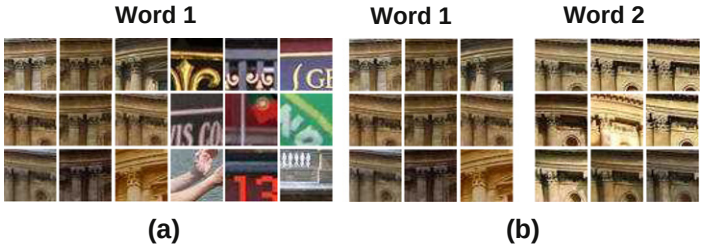

Fig. 1. (a) Image patches mapped to one visual word of the bag-of-words vocabulary. Note the visual incoherence. (b) Comparison between image patches from two different words. Note their perceptual similarity.

Larger vocabularies fragment feature space finer yielding visual words that are more concise, albeit less populated. Despite their effectiveness, large vocabularies merely postpone rather than solve the problem of the fuzzily populated feature space. Another technique that helps to ameliorate the errors during feature acquisition is the use of soft assignment for mapping features to clusters. Unlike hard assignment that performs a crude binary assignment to a single cluster, soft assignment distributes the probability mass of the mapping to a number of adjacent clusters [20]. Unfortunately, soft assignment compensates only for the assignment errors near the cluster borders. Errors that might occur because of the misplacement of features in distant parts of the feature space remain unaffected.

In this paper we propose visual synonyms, a method for linking semantically similar words in a visual vocabulary, let them be distant in feature space or not. The bag-of-words model is used on landmark images, because their unchanged geometry allows for mapping between different images with different recording conditions, which opens the door to perspectives for linking words as synonyms. When a link to the same spot is found, it is clear the word represents nearly the identical patch in reality. However, due to the accidental recording conditions in each of the words, the features may differ significantly. Thus, this link establishes a connection between two parts of the feature space, which, despite their distance, correspond to image patches of similar visual reality. Visual synonyms comprise a vehicle for finding the parts of feature space, which are nearly identical in reality. This allows for further refinement of visual word definitions. Also, visual synonyms can be used for vocabulary reduction. By using a fraction of visual synonym words, we are able to reduce vastly the vocabulary size without a prohibitive drop in performance.

This paper extends [21] with additional experiments and a more deep analysis of the behavior of visual synonyms and visual words. The rest of the paper is organized as follows. In Section 2 we present some related work. In Section 3 we introduce the notion of visual synonyms and we propose an algorithm for their extraction. We describe our experiments in Section 4 and we present the results in Section 5. We conclude this paper with a short discussion of the acquired results.

2. Related work

The bag-of-words method is the state-of-the-art approach in landmark image retrieval [14]. The core element of the bag-of-words model is the vocabulary $W = \{w^1, \ldots, w^K\}$, which is a set of vectors that span a basis on the feature vector space. Given the vocabulary and a descriptor d, an assignment $q^r \in 1, \ldots, K$ to the closest visual word is obtained. We may construct the vocabulary W on a variety of ways, the most popular being k-means [22]. Based on the bag-of-words model, an image is represented by a histogram, with as many bins as the words in the vocabulary. The word bins are populated according to the appearance of the respective visual word in the image. Therefore, an image I is represented by $h_I = g(w_I^1), \ldots, g(w_I^K)$, where $g(\cdot)$ is a response function assigning a value usually according to the frequency of the visual word in the image. More advanced techniques have recently been proposed, better encoding the original descriptor d using the vocabulary basis W, thus yielding significant performance improvements, often at the expense of a high memory and computational cost [23] After obtaining the histogram of responses, all spatial information is lost. Following [14,24], we enrich the bag-of-words model with spatial information using homography mappings that geometrically connect pairs of images.

The most popular choice for feature extraction in the bag-of-words model is the SIFT descriptor [1]. Given a frame, usually split into a 4×4 grid, the SIFT descriptor calculates the edge gradient in eight orientations for each of the tiles in the grid. Thus resulting in a 128-D vector. Although originally proposed for matching purposes, the SIFT descriptor also dominates in image classification and retrieval. Close to SIFT follows the SURF descriptor [2]. SURF is designed to maintain the most important properties of SIFT, that is extracting edge gradients in a grid, while being significantly faster to compute due to the internal use of haar features and integral images.

An efficient and inexpensive extension to the bag-of-words model is visual augmentation [24,25]. According to visual augmentation, the retrieval of similar images is performed in three steps. In the first step the closest images are retrieved, using the bag-of-words model. In the second step, the top ranked images are verified. In the third step, the geometrically verified images lend their features to update the bag-of-words histogram of the query image and the new histogram is again used to retrieve the closest images. In the simplest case, the update in the second step averages over all verified images closest to the query [25,24]. In a more complicated scenario, the histogram update is based on a multi-resolution analysis of feature occurrences across various scenes [24]. For visual augmentation to be effective, the query's closest neighbor images have to be similar to the query image. Therefore geometric verification is applied. As expected, the top ranked images are usually very similar to the query image. However similar, these images exhibit slight differences due to their respective unique imaging conditions. These slight differences supplement the representation of the original query with the additional information that stems from the possible variations of visual reality as depicted in the image. Finally, the augmented query representation leads to a boost in performance. In this paper we draw inspiration from Chum et al. [24] and Turcot and Lowe [25], however we do not use any graph structure to connect images together.

Apart from image appearance, landmark scenes are also characterized by their unchanged geometry. Given that in a pair of images geometry changes because of translation, rotation and scale, there is a matrix that connects these two images together.