Pattern Recognition 47 (2014) 705–720

Contents lists available at ScienceDirect

Pattern Recognition

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

# Visual word spatial arrangement for image retrieval and classification

Otávio A.B. Penatti [a,*], Fernanda B. Silva [a], Eduardo Valle [a,b], Valerie Gouet-Brunet [c,d], Ricardo da S. Torres [a]

[a] RECOD Lab, Institute of Computing (IC), University of Campinas (Unicamp) – Av. Albert Einstein, 1251, Campinas 13083-852, SP, Brazil
[b] Department of Computer Engineering and Industrial Automation (DCA), School of Electrical and Computer Engineering (FEEC), University of Campinas (Unicamp) – Av. Albert Einstein, 400, Campinas 13083-852, SP, Brazil
[c] Paris-Est University, IGN/SR, MATIS Lab, 73 avenue de Paris, 94160 Saint-Mandé, France
[d] CNAM, CEDRIC Lab, 292 rue Saint-Martin, 75141 Paris Cedex 03, France

ARTICLE INFO

*Article history:*
Received 25 September 2012
Received in revised form
22 June 2013
Accepted 9 August 2013
Available online 23 August 2013

*Keywords:*
Visual words
Spatial arrangement
Image retrieval
Image classification

ABSTRACT

We present word spatial arrangement (WSA), an approach to represent the spatial arrangement of visual words under the bag-of-visual-words model. It lies in a simple idea which encodes the relative position of visual words by splitting the image space into quadrants using each detected point as origin. WSA generates compact feature vectors and is flexible for being used for image retrieval and classification, for working with hard or soft assignment, requiring no pre/post processing for spatial verification. Experiments in the retrieval scenario show the superiority of WSA in relation to Spatial Pyramids. Experiments in the classification scenario show a reasonable compromise between those methods, with Spatial Pyramids generating larger feature vectors, while WSA provides adequate performance with much more compact features. As WSA encodes only the spatial information of visual words and not their frequency of occurrence, the results indicate the importance of such information for visual categorization.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

Content-based image retrieval is a key technique for improving image search engines. The most effective approaches currently used are based on a vocabulary of local patches, called visual dictionaries [1]. That model is inspired in text retrieval, where a simple but effective model takes documents simply as "bags" (multi-sets) of words. In the same spirit, visual dictionary representations take images as bags of local appearances. That model has several important advantages, such as compactness (it encodes local properties into a single feature vector) and invariance to image/scene transformations.

Creating a visual dictionary takes several steps. First and foremost, local characteristics must be obtained from a set of training images, usually by extracting local patches and describing them. The patches may be taken around Points of Interest (PoI) [2] or by dense sampling [3], and image descriptors, like the popular SIFT [4], are used to extract feature vectors for each of them. Once the learning set of feature vectors is obtained, they are used to quantize the feature space (using, for example, k-means clustering) to choose a codebook of feature vectors representative of the training set. The clusters tend to contain visually similar patches and each cluster is a visual word of the dictionary. Once the dictionary is available, images are represented by statistical information about how they activate the visual words. The final image feature vector is commonly called bag of (visual) words (BoW).

When creating an image representation, one must be aware of its target application. Applications like copy detection or partial-duplicate image search, as shown in Fig. 1(a),[1] require the creation of really discriminating representations. Very small differences between images or objects must be encoded, while still being robust to specific photometric/geometrical transformations related to the domain. Therefore, the representation must be very precise. The semantic-search application, as shown in Fig. 1(b),[2] requires precise representations but, at the same time, general enough to comprise the intra-class variations. One may be interested in finding different types of the same object, like, for example, retrieving different types of chairs, instead of finding exactly the same chair [6]. Considering the topical problem of big data, more generic representations should be more interesting. Representations that are less specific to a certain application may be more suitable for addressing the big-data problem, where the big volume of data makes it more complicated to extract several categories of features dedicated to specific scenarios (retrieval, classification, etc.), in terms of extraction time and of storage.

* Corresponding author. Tel.: +55 19 3521 5887; fax: +55 19 3521 5847.
*E-mail addresses:* penatti@ic.unicamp.br (O.A.B. Penatti), fernanda@recod.ic.unicamp.br (F.B. Silva), dovalle@dca.fee.unicamp.br (E. Valle), valerie.gouet@ign.fr (V. Gouet-Brunet), rtorres@ic.unicamp.br (R.d.S. Torres).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.08.012

[1] Creative Commons images downloaded from Flickr (as of July 9, 2012).
[2] Chairs from Caltech-101 dataset [5].

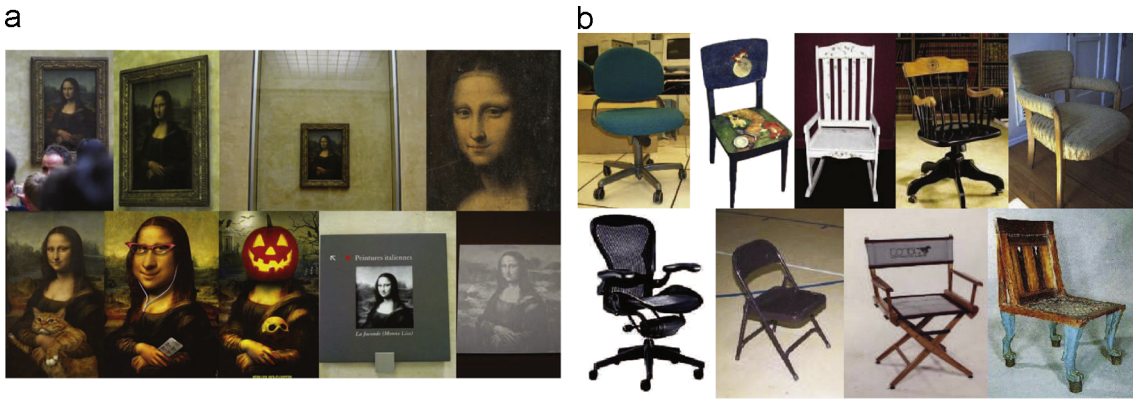


**Fig. 1.** Application examples: (a) retrieval of partial duplicates, where (parts of) the same object or scene are shared between the query and target images, possibly with transformations and noise; (b) semantic search, where query and target images share concepts (e.g., different instances coming from the same class of objects), but not necessarily objects or scenes.

The research community has been very active in the area in the latest years and many new proposals over the visual dictionary model constantly appear. Special attention has been given to the lack of geometrical information encoded by the traditional bag-of-words representation [7–13]. The spatial arrangement of visual words in images is important to understand image semantics and is often crucial to distinguish different classes of scenes or objects. In that direction, approaches are proposed for image classification [10,9] and retrieval [8,12,13,11].

In the classification scenario, usually relied on Support Vector Machines (SVMs), the high dimensionality of vectors does not degrade effectiveness, because SVMs suffer less from the curse of the dimensionality. The popular Spatial Pyramids [9] are very successful for image classification and their vectors have high dimensionality. However, for retrieval experiments, which are generally based on computing distances between vectors, with the Euclidean distance, for example, vectors should be compact, or embedded in an index structure, to avoid the curse of the dimensionality [14–16,8]. As dimensionality grows, the distribution of distances between features tends to become narrowly concentrated around an average value, reducing the contrast between similar and dissimilar features. Therefore, to create an image representation that works well in both classification and retrieval scenarios, one must be aware of the feature vector size. Many of the existing approaches for spatial pooling which are employed in the retrieval scenario leave the spatial verification as a post-processing step [12,13]. They compute a simple representation and then, after finding the matching visual words between images, they compute the spatial representation and perform a spatial consistency verification, before reranking the images. Furthermore, some of the existing approaches used in the retrieval scenario are very specific and suitable for partial-duplicate image search [12,13], thus their use for the semantic-search application is challenging.

In this paper, we present *word spatial arrangement* (WSA), a spatial pooling approach for both image retrieval and classification. Our approach adds spatial information into the feature vector having the advantages of generating more compact vectors than the popular approaches for spatial pooling. It is also more precise than the traditional bag of words but keeps the generality desired for the semantic-search application, a good property also for the big-data problem. Our approach aims at addressing both retrieval and classification scenarios. In the retrieval environment, WSA encodes the spatial information of visual words into a single feature vector prior to any filtering step with matching visual words. Most of the approaches that encode spatial information of visual words in the retrieval scenario [8,12] works solely with the assignment of a unique visual word to a point (hard assignment). WSA, however, also works with soft assignment, taking advantage of the good performance of soft assignment in classification experiments [17–19]. Soft assignment relies on considering a neighborhood around the point in analysis when assigning labels of visual words to it. In high-dimensional spaces, points tend to be in the frontier of many regions, therefore, assigning more than one visual word to them can be more robust.

WSA is simple and easy to understand. On top of that, WSA has the advantage of depending on almost no parameters, oppositely to many of the existing spatial pooling methods. The visual word spatial arrangement encoded by WSA is based on a sliding quadrant partitioning in the image space considering each point in the image as the origin of the quadrants and counting the visual word occurrences in each quadrant [7]. In this paper, we present an improvement to the original WSA method which we presented in [7]. Here, we are not concatenating the bag to the WSA information, therefore, in this paper, WSA refers to the spatial information only. Additionally, we have included several improvements over the original WSA propose. We explain how to use WSA with soft assignments introducing the threshold $t$ for very soft assignments, we evaluate the use of a weighted window, and we propose a distance function for image retrieval. Considering the experimental protocol, in this paper, we evaluate WSA for both image retrieval and classification using different datasets. We also provide an online interface based on Eva tool [20] to show the experiment results in the retrieval scenario.[3,4]

The rest of the paper is organized as follows. Section 2 presents other approaches for encoding spatial arrangement of visual words. Section 3 details the proposed method. Sections 4 and 5 show the experimental comparison in both retrieval and classification scenarios, respectively. Section 6 concludes the paper.

## 2. Related work

The related work section details the traditional representation schemes based on visual dictionaries in Section 2.1 and presents recent advances on encoding spatial information of visual words in Section 2.2.

### 2.1. Visual dictionaries

The most popular and effective approach to represent visual content nowadays is based on visual dictionaries, which generate

[3] http://www.recod.ic.unicamp.br/eva/view_images_base600.php (as of August 30, 2013).

[4] http://www.recod.ic.unicamp.br/eva/view_images_paris.php (as of August 30, 2013).

Download English Version:

https://daneshyari.com/en/article/530911

Download Persian Version:

https://daneshyari.com/article/530911

Daneshyari.com