Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model

Rahat Khan, Cecile Barat, Damien Muselet, Christophe Ducottet

Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model

Rahat Khan[a], Cecile Barat[a], Damien Muselet[a], Christophe Ducottet[a]

[a]*Laboratory Hubert Curien, UMR CNRS 5516, Bâtiment F, 18 Rue du Professeur Benoît Lauras, 42000 Saint-Etienne, France*

Abstract

In the context of category level scene classification, the bag-of-visual-words model (BoVW) is widely used for image representation. This model is appearance based and does not contain any information regarding the arrangement of the visual words in the 2D image space. To overcome this problem, recent approaches try to capture information about either the absolute or the relative spatial location of visual words. In the first category, the so-called Spatial Pyramid Representation (SPR) is very popular thanks to its simplicity and good results. Alternatively, adding information about occurrences of relative spatial configurations of visual words was proven to be effective but at the cost of higher computational complexity, specifically when relative distance and angles are taken into account. In this paper, we introduce a novel way to incorporate both distance and angle information in the BoVW representation. The novelty is first to provide a computationally efficient representation adding relative spatial information between visual words and second to use a soft pairwise voting scheme based on the distance in the descriptor space. Experiments on challenging data sets MSRC-2, 15Scene, Caltech101, Caltech256 and Pascal VOC 2007 demonstrate that our method outperforms or is competitive with concurrent ones. We also show that it provides important complementary information to the spatial pyramid matching and can improve the overall performance.

Keywords: spatial information; object classification; angle and distance histograms; Bag-of-Words

1. Introduction

In category level and scene classification, the bag-of-visual-words (BoVW) method, first introduced by [1, 2], has shown excellent results in recent years [3, 4]. In this method, an image is represented as a histogram of quantized local features called visual words. However, being orderless, histogram representations do not provide any spatial information. This is considered to be one of the major drawbacks of this very successful method.

Email addresses: rahat.khan@univ-st-etienne.fr (Rahat Khan), cecile.barat@univ-st-etienne.fr (Cecile Barat), damien.muselet@univ-st-etienne.fr (Damien Muselet), ducottet@univ-st-etienne.fr (Christophe Ducottet)

Different methods have been proposed to incorporate spatial information into the BoVW representation [3, 5, 6, 7, 8]. Some of these approaches use spatial context before the vocabulary construction step to incorporate spatial information [9, 10, 11]. The spatial context is defined at the local feature level as a set of descriptors extracted form neighboring regions so as to enrich the description around a feature point. For example in [11], the authors create pairs of local neighbor SIFT descriptors, concatenate them and construct the vocabulary from these 256-D combined features instead of from the classical individual 128-D SIFT. Alternatively, most of the approaches work at the visual word level. They model the spatial arrangements of visual words on the 2D image space as an additional step [3, 5, 8, 12, 13, 14, 15, 16, 17]. These later methods are more popular as they obtain superior classification accuracies. It is due to the fact that they are able to capture both local and global relationships among the visual words.

In this context, the Spatial Pyramid Representation (SPR) [3] is probably the most notable work. Its principle relies on the division of an image into sub-windows and the computation of a local BoVW histogram in each. Two images are then compared by using an intersection kernel computed between the two corresponding sets of histograms.

Although SPR performs very well, it only captures the information about approximate absolute locations of the visual words in images. An alternative consists in extracting relative spatial interactions between the visual words [7, 17, 14, 18, 19, 20]. In these representations, the absolute location of each visual word is lost but, for example, information about visual words that are frequently co-occurring together at a certain distance is taken into account. However, the abundance of visual words in an image makes it computationally expensive to explicitly model relative spatial relationships among visual words. Thus, methods like [6, 7] employ vocabulary compression or feature selection and model only local or semi-local spatial information to speed up the computation. Nevertheless, Elfiky et al. [21] have shown that vocabulary compression before spatial information extraction results into declined classification performance. In another work, Parikh [22] examines the human vs machine performance on jumbled images and concludes that existing machine vision techniques are already effective in modeling local information from images, thus future research efforts should be focused on more advanced modeling of global information.

Based on these observations, in this work, we propose a way to model the global and local relative spatial distribution of visual words over an image. To do so, we introduce the concept of soft pairwise spatial angle-distance histograms to capture the distribution of similar descriptors. The term "soft" means that we apply a soft weighting strategy in a similar way as soft assignment techniques [23, 24].

The novelty compared with previous approaches is threefold: i) enabling infusion of pairwise relative spatial information (modeling both distances and angles between visual words while most of the approaches consider only distances), ii) adopting a simple word selection scheme that avoid combinatorial explosion, iii) combining hard assignment for visual word selection and soft assignment to weight the contribution of descriptor pairs in spatial histograms.

We experimentally evaluate our new representation on classification tasks with various challenging data sets. The aim is both to study the influence of various parameters of the method and to compare the performances over state of the art concurrent approaches.

The rest of the paper is organized in the following way: the next section describes a review of the related works. Section 3 presents our approach to incorporate spatial in-

formation into the BoVW representation. Section 4 describes the implementation details and section 5 presents the results on different benchmarks and comparisons with several other methods. Section 6 concludes the article pointing towards our future works.

2. Related works

There exist two main approaches to incorporate spatial information in the BoVW representation, the difference being the information added over the classical BoVW, i.e. either the approximate absolute positions of the visual words or the relative positions between these visual words.

The SPR [3] is the most widely used method among the ones representing absolute spatial coordinates. Its principle relies on the division of an image into a sequence of increasingly coarser grids (eg. 4 by 4, 2 by 2 and 1 by 1) and on the computation of a local BoVW in each cell. Two images are then compared using an intersection kernel computed between the two corresponding sets of histograms. Bosch et al. [25] have generalized the intersection kernel with other quasi-linear kernels like chi-square and learned weights for each pyramid level rather than using fixed weights. Another weight learning method was proposed for SPR in [26]. More recently, in [27] Cao et al. proposed to extract local BoVW from more sub-windows than in SPR with different shapes (not only rectangular) and to select the best ones with a supervised approach. Finally, in [12], Krapac et al. encode the absolute region locations of each visual word using the Fisher kernel.

Among the methods that add spatial relative interactions between the visual words, the most straightforward approaches are those based on correlograms [7, 17]. The color correlograms have been introduced by Huang et al. in [28]. The idea was to represent the distribution of color pairs as a function of the spatial distance between these colors in the image. Savarese et al. extend this representation to the distribution of pairs of visual words in the image plane. Thus, given K visual words and T distances, the dimension of one correlogram is $K \times K \times T$. Then, the authors propose to cluster the T-dimensional vectors representing the evolution of the distribution of each pair across the T distances. The cluster representatives are called correlatons and the descriptor for one image is the histogram of correlatons. Consequently, in this final representation, only the spatial interactions between the visual words are accounted, regardless of the actual identities of these words.

Another way to incorporate relative interactions between visual words is to use visual phrases [14, 18, 19, 20] in the final image representation. A visual phrase is a set of frequently locally co-occurring visual words. Most of the times, specific Frequent Itemset Mining (FIM) algorithms are applied in order to discover the most meaningful visual phrases since this selection is the key to these approaches. In [20], they consider only pairs of visual words located in a local neighborhood whose radius is related to the scale of the associated SIFT keypoint and use the supervised TF-IDF (Term Frequency - Inverse Document Frequency) information for the selection of the best visual phrases. In [19], they first select the 100 most frequent visual words and each visual word is paired with its 5 nearest neighbor in the image space. Unlike the previous approaches, in [14, 18, 29], the visual phrases can contain more than two visual words. Thus, specific FIM algorithms are required to select the best visual phrases among all the local visual word sets. Also, while the visual phrases from [14, 18] consists of visual words co-occurring in a local neighborhood, the ones from [29] are not constrained by any locality properties. Indeed,

in this last paper, the authors consider each image pair ($\mathbf{I}$, $\mathbf{I}'$) and for each pair of the same word in these images, they calculate their offset (Δx, Δy) which is the location of the word in the image $\mathbf{I}'$ subtracted from its location in the image $\mathbf{I}$. Then a vote is applied in the offset space in order to detect geometry-preserving visual phrases. In one cell of this offset space are falling all the words whose offset is similar between the two images, but the information about the spatial distribution in the image spaces of the visual words belonging to the same visual phrase is lost. In order to cope with this problem, the authors propose to store in memory the locations of each visual word occurrence within each image. This information is used during the image comparison step. Using a similar storage approach, Wu et al. [30] bundle visual words falling in the same local neighborhood, i.e. the same MSER regions [31]. Then, the orders of the visual words along the x and y axis within each local group is saved in memory and this information is used during the comparison step in order to check the approximate relative positions of the words within each detected MSER region. These two last approaches save more information in memory than the single histogram-like structure and this requires a particular comparison step during the test phase. Even if the approach of [32] is a bit different from the previous papers, the ideas are similar in the sense that, in this paper, the authors also exploit frequently locally co-occurring visual words to describe images. Indeed, they propose to extract several local BoVW on a grid (as done in the SPR approach) from the images of the dataset and to apply vector quantization on these local BoVW. The cluster representatives are "second order" visual words from which second order BoVW can be computed. And iteratively, the authors extend the classical "first order BoVW" to "nth order BoVW". Obviously, these BoVW contain information about the local spatial interactions between the original visual words.

All the previous papers dealing with relative visual words locations, only consider the distances between the visual words as supplementary information over the classical BoVW. In [33], we have rather proposed to incorporate angle information in the final representation. Hence, we have shown that global spatial orientations of patch pairs whose corresponding visual words are the same (called intra-type visual word pairs) are discriminative and significantly improve the classification accuracy. The approach proposed by Liu et al. in [6] has the advantage that it considers both distances and angles between the visual words and by this way represents more accurately the relative positions of the visual words in the image space. Nevertheless, this whole information requires much more memory space and computational time during both training and testing steps. Thus, the authors introduce an integrated feature selection and spatial information extraction technique to reduce the number of features and also to speed up the process. The process of feature selection and second order feature (e.g. spatial information) extraction are run alternatively at each iteration of the algorithm. At each round, feature selection selects one feature, and feature extraction pairs this feature with each of the previously selected features. The final image representation is constructed by the concatenation of the first and second order features.

Finally, these works show that both absolute and relative spatial information improve the classification accuracy over the classical BoVW. Consequently, the current trend consists in combining these two information into the final representation [16, 34, 35] in order to boost the results. For example, in [34], spatial predicates are introduced and, given a spatial predicate, a co-occurrence matrix is computed over all the visual word pairs. In their approach, an image is represented as a set of co-occurrence matrices of size

$K \times K$ (if K is the number of visual words) and thus they must limit the size of the vocabulary and the number of considered spatial configurations. Moreover, spatial distances are defined in an absolute way inside a region of a pyramid. In [35], Harada et al. incorporate local spatial information by evaluating local auto-correlations between the extracted descriptors and global information by weighting differently the contribution of each cell of a regular grid superimposed to the image (as done by SPR). These weights are learned so that they are related to the discriminative power of the considered cells, given a classification task. Thus, in most of the works combining absolute and relative spatial information, the relative spatial information is combined with the reference SPR approach. Consequently, in this paper, we propose to design a new descriptor that represents the relative spatial information and that is complementary to the SPR descriptor. Our original idea is to account both the distances and the angles between the visual words by using the notion a soft-similarity, which allows us to avoid a complex selection step as required in [6].

3. Encoding distance-orientations information of similar patches

The principle of our method is to use pairwise spatial histograms to encode spatial co-occurrence of similar word pairs. In this section we first present original pairwise spatial histograms introduced by Liu et al. [6], then we introduce pairwise spatial histograms of similar patches and finally our image representation.

3.1. Pairwise spatial histograms

In the standard BoVW model, an image is represented as a histogram of occurrences of its visual words [1]. More formally, a set of local descriptors $\{d_1, d_2, d_3, d_4, \dots d_N\}$ is first extracted from a sampled set of image patches [36]. Each descriptor d_i is a vector giving the description of the image patch i and N is the total number of patches in the image. The visual vocabulary $W = \{w_1, w_2, w_3, w_4 \dots w_K\}$ is obtained by typically applying an unsupervised K-means clustering on a large set of descriptors from training images. Each visual word w_k represents a cluster center, K being the vocabulary size, ie. the number of predefined clusters. Each patch i of the image is assigned to a visual word which corresponds to the nearest center in the descriptor space. If soft assignment or sparse coding is used [23, 24, 37], the patch is assigned to a weighted set of visual words. The final histogram is obtained by pooling the contribution of each visual word over the whole image or over spatial regions.

In Liu et al. [6], a pairwise spatial histogram is defined according to a discretization of the spatial neighborhood into several bins encoding the relative spatial position (distance and angle) of two visual words (Figure 1). Given a pair of visual words (w_k, w_l), all the pairs of patches (P_i, P_j) such that P_i is of type w_k and P_j is of type w_l are considered. Then, a pairwise spatial histogram associated to (w_k, w_l) is defined as the count of all occurrences of P_j falling into a specific spatial bin relatively to P_i, the count being averaged for all instances of P_i (Figure 1).

3.2. Motivation of considering similar cues

The number of possible pairs of visual words is potentially very large and thus, Liu et al. proposed to select only discriminative pairs with a feature selection method based

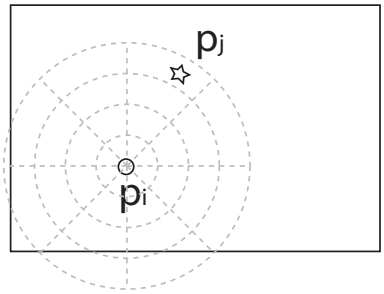

Figure 1: Spatial discretization of the image space. p_i and p_j are the 2D positions of two patches P_i and P_j respectively, P_i being of type w_k and P_j of type w_l. Concentric circles and angles are drawn centered at position p_i. The sector where pixel p_j is falling informs about the relative positions between the two patches [6].

Figure 2: Discriminative power of spatial distribution of intra type visual words pairs. Four images from Caltech101 dataset are shown. The red squares refer to identical visual words across all the images (for a dictionary of 200 words created from SIFT descriptors). For the two motorbikes in the left, the global distribution of the identical visual words is more similar than the ones in Helicopter or Bugle image.

on Discrete AdaBoost with decision stumps as in [38]. In [33], we have proposed another alternative which is to consider only pairs of identical visual words. The motivation came from the previous works [39, 40] where the authors have argued that modeling the distribution of similar cues across an image can give discriminative information about the content of that image. Figure 2 shows an example which gives an intuition to better understand the idea. In this illustration, we consider patches associated with the same visual word as similar cues.

3.3. Pairwise spatial histograms of similar patches

The notions of similar cues and similar words are not equivalent. If we consider clusters delimiting visual words in the descriptor space, two cues at the cluster borders could be very similar being in different clusters. Similar cues in this context are more related to a small inter-patch distance in the descriptor space.

This problem is similar to the visual word ambiguity problem which has been discussed in many works on the BoVW model. Indeed, the vector quantization applied at the visual word assignment step introduces ambiguity if the considered descriptor is at an intermediate position between several cluster centers [41]. To address this issue, soft assignment approaches have been introduced. Their principle is to use the distance in the descriptor space to compute a weight or a probability estimate associated to a given visual word [23, 24, 42].

Hence, we propose to analyze the spatial positions of the patches which are situated in proximity in the descriptor space. To avoid the use of a threshold to identify similar

patches (hard similarity), we consider all the pairs of patches and we weight the contribution of each pair as a decreasing function $g(x)$ of their distance x in the descriptor space (soft similarity). We propose to use a Gaussian function of standard deviation α defined as:

$$g(x) = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{x^2}{2\alpha^2}} \tag{1}$$

This parameter gives us the control to highly weight patches that are in close proximity in the descriptor space and to ignore the ones which are far. More information about the choice of this parameter can be found in section 4.2. More formally, we consider the set S_k of all the pairs of patches where at least one patch in the pair belongs to the visual word w_k. A given pair $(P_i, P_j) \in S_k$ is characterized both by a pair of descriptors (d_i, d_j) and a pair of positions in the image space denoted (p_i, p_j) (Figure 3). Note that both d_i and p_i are vectors with $d_i \in \mathbb{R}^D$ (descriptor space) and $p_i \in \mathbb{R}^2$ (image space). A pairwise spatial histogram of similar patches is then defined considering a discretization of the image space into M bins denoted $\{b_m, m = 1, \ldots, M\}$ with an angle $\theta \in [0, \pi[$ split into M_θ equal angle bins and the radius $r \in [0, R]$ split into M_r radial bins so that $M = M_\theta M_r$. For example, in the illustration at the bottom of figure 3, the total number of bins M is equal to 15 (M_θ=5 and M_r=3). The values of M_θ and M_r will be determined in the experimental section and the maximum radius R is chosen to be the diagonal of the image, in order to reduce scale sensitivity. Note that by using such maximum value for the radius, the representation may indirectly capture the absolute spatial location since patches located close from the image borders have a more limited number of possible angles and distances for pairing. In all spatial methods, such border effects exist and are difficult to quantify.

The count $H_k^{Soft-Inter}(m)$ of bin b_m of the spatial histogram of similar pairs $H_k^{Soft-Inter}$ corresponding to the visual word w_k is then given by:

$$H_k^{Soft-Inter}(m) = \sum_{(P_i,P_j)\in S_k} g\left(|d_i - d_j|_2\right) \mathbb{1}_{b_m}(p_j - p_i) \tag{2}$$

where $|d_i - d_j|_2$ is the ℓ^2 distance in the descriptor space and $\mathbb{1}_{b_m}$ is the indicator function of bin b_m such that:

$$\begin{array}{rccl} \mathbb{1}_{b_m}: & \mathbb{R}^2 & \to & \{0,1\} \\ & v & \to & \left\{ \begin{array}{ll} 1 & \text{if } v \in b_m \\ 0 & \text{otherwise} \end{array} \right. \end{array} \tag{3}$$

Note that, due to symmetric considerations, angle bins are discretized in the $[0, \pi[$ interval. Indeed, as the two members of a pair are related to the same visual word w_k (i.e. inside the corresponding cluster or not far from its border), only the absolute orientation modulo π is registered. Thus, given a pair of positions (p_i, p_j) the corresponding histogram bin is determined by taking either p_i or p_j as reference point which is equivalent to consider either $(p_j - p_i)$ or $(p_i - p_j)$ vectors (see figure 3).

To independently evaluate the benefits of using either the soft-weighting (over hard) or the inter-word pairs (over intra), we also introduce two other spatial histograms. The first one, called $H_k^{Hard-Inter}$, is using only binary weights t defined as:

$$t(x) = \left\{ \begin{array}{ll} 1 & \text{if } x < 3\alpha \\ 0 & \text{otherwise} \end{array} \right. \tag{4}$$

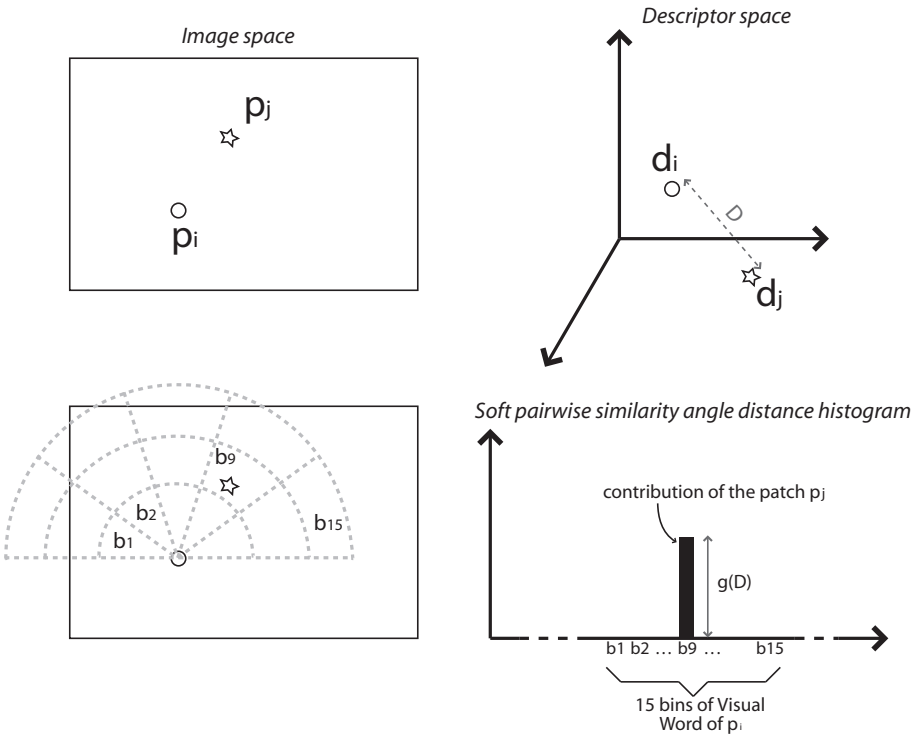

Figure 3: Pairwise spatial histogram using soft-similarity weighting. To encode spatial information, we use the distance and orientation information between pairs of patches in the image space (left) as well as their distance in the descriptor space (top-right). We consider inter and intra type word pairs based on their proximity in the descriptor space. At the left-bottom, the spatial discretization is illustrated. Translating the reference patch P_i (resp. P_j) at the center, the position of patch P_j (resp P_i) gives the bin number (b_9 here) that would be affected in the histogram (bottom-right). The contribution of the patch P_j to the histogram is related to the distance D between the two descriptors d_i and d_j by using the equation (1).

where α is still the standard deviation used in equation 1.

Then, the count $H_k^{Hard-Inter}(m)$ of the bin b_m is given by:

$$H_k^{Hard-Inter}(m) = \sum_{(P_i,P_j)\in S_k} t\left(|d_i - d_j|_2\right) \mathbb{1}_{b_m}(p_j - p_i). \tag{5}$$

The second one, called $H_k^{Hard-Intra}$, is also using the same binary weight t but does not consider inter-words pairs (only intra). Formally, if S_k^* denotes the set of all the pairs of patches for which both patches belong to visual word w_k, the hard pairwise histogram $H_k^{Hard-Intra}$ is defined as:

$$H_k^{Hard-Intra}(m) = \sum_{(P_i,P_j)\in S_k^*} t\left(|d_i - d_j|_2\right) \mathbb{1}_{b_m}(p_j - p_i) \tag{6}$$

For a visual word w_k, equations 2, 5 and 6 can be used to compute the entire histograms $H_k^{Soft-Inter}$, $H_k^{Hard-Inter}$ and $H_k^{Hard-Intra}$ respectively. In the next section we introduce a way to combine a given set of spatial histograms H_k into one image histogram.

3.4. Image representation

As explained in the previous section, for each visual word w_k, we obtain one spatial histogram. This histogram encodes spatial information (distance and orientation) of pairwise similar patches, where at least one of the patches belongs to w_k. This modularity facilitates simple way to assemble the spatial histograms and to obtain the final representation. Starting from the proposed word specific histogram $H_k^{Soft-Inter}$, we define three different representations: the soft pairwise similarity angle-distance histogram SPS_{ad} derived from the classical BoVW histogram, $SPS_{ad}+$ its combination with the SPR and $SPS_{ad}^{1800}+$ a compact version of $SPS_{ad}+$. On the other hand, word specific histograms $H_k^{Hard-Inter}$ and $H_k^{Hard-Intra}$ are used to create hard pairwise similarity angle histograms HPS_{ad}^{Inter} and HPS_{ad}^{Intra} respectively.

3.4.1. Soft Pairwise Similarity angle distance histogram SPS_{ad} representation

To obtain the SPS_{ad} representation from the classical BoVW histogram, we use a 'bin replacement' technique. Bin replacement literally means to replace each bin of the BoVW frequency histogram with the spatial histogram $H_k^{Soft-Inter}$ associated to w_k. The sum of all the bins of the spatial histogram obtained from one visual word w_k is normalized to the number of occurrences of this word in the whole image. The final whole histogram is subsequently ℓ_1 normalized as in the BoVW method. By this way, we keep the frequency information intact and add the spatial information. The dimensionality of our representation $S = K \times M$ depends on the vocabulary size (K) and the number of angle-distance bins of the spatial histogram (M).

On the other hand, if we only use hard weighting t (eq. 4) instead of soft weighting g (eq. 1) as done in $H_k^{Hard-Inter}$ (eq. 5), we obtain a hard pairwise similarity histogram, denoted as HPS_{ad}^{Inter}. And finally, if we just consider intra-type visual words pairs and hard weighting as done in $H_k^{Hard-Intra}$ (eq. 6), the final representation is denoted HPS_{ad}^{Intra}.

3.4.2. Combination of SPS_{ad} with SPR

Since SPS_{ad} represents the relative spatial interactions between the visual words in an image, it is complementary to descriptors that represent absolute locations of the visual words such as SPR. Thus, we propose to combine SPS_{ad} with SPR. This concatenation (without any weights) is called $SPS_{ad}+$. For a vocabulary size of K the dimensionality of $SPS_{ad}+$ is $K \times (1+4+16+M)$ because the number of local histograms are respectively 1, 4 and 16 in the 0^{th}, 1^{st} and 2^{nd} levels.

3.4.3. Dimensionality reduction

One of the drawbacks of the $SPS_{ad}+$ compared to SPR is the high dimensionality of the feature vectors. Typically, for a $K = 200$ words vocabulary and $M = 45$ angle-distance bins, the dimensionality of the feature vectors is 13200. To get a more compact representation, dimension reduction techniques as feature selection or feature clustering can be applied. We propose to use a divisive information theoretic clustering (DITC) approach introduced by Dhillon et al. [43]. The DITC algorithm minimizes the within-cluster Jensen-Shannon divergence while simultaneously maximizing the between-cluster Jensen-Shannon divergence in a clustering like algorithm based on the estimation of the joint probability of a visual word in a class. This method was used by Elfiky et al. [21] to compress the SPR representation. They found that it provides the size reduction of a high dimensional pyramid representation up to an order of magnitude with little or no loss in accuracy. Alternative methods as the AdaBoost feature selection method integrated in Liu et al. approach [6] could also be used. Compared to this latter method which is based on a local optimization of the classification accuracy, DITC algorithm performs a global minimization of the loss of information due to feature clustering.

Starting from $SPS_{ad}+$ with $K = 200$ and $M = 45$ (13200 dimensions), we compress our feature vectors by DITC algorithm down to 1800 dimensions. We denote this representation as $SPS_{ad}^{1800}+$.

3.5. Comparison with related approaches

Our representation differs on many points compared to original spatial histograms introduced by Savarese et al. [7] and refined by Liu et al. [6]. We present here the differences in the methodology and in the computational point of view.

On the methodological point of view all the three versions are based on pairwise spatial histograms (called correlogram elements in [7]). Each spatial histogram is associated to a specific pair of visual words (w_k, w_l). In the original versions, all the combinations of visual words are considered leading to $K(K+1)/2$ possible histograms (K still denotes the number of visual words). In our version, only pairs of identical (or similar) visual words are considered leading to only K possible histograms. Another difference lies in the discretization of the spatial neighborhood. In [7], only distance divisions are considered whereas in [6] both distances and angle divisions are considered but in both cases, the neighborhood size is limited to a certain distance in pixels or in function of the patch size. In our case, the neighborhood size is relative to the image size and thus covers the whole image. Moreover, compared to [6], our distance divisions are linear and our angles covers only the $[0, \pi[$ due to symmetric considerations. Also note that angles are measured absolutely which does not provide invariance of the representation to image rotations but prevent the decrease of discriminative power in object categorization [44].

Concerning computation, the main point is that we avoid the combinatorial explosion by considering only similar word pairs. First, it limits the size of the final histograms (which becomes linear on K) and second it makes unnecessary to apply any additional clustering technique (as in [7]) or additional feature selection step (as in [6]). Note that we still need to consider all the patch pairs of an image (complexity $O(N^2)$), but in practice, we can fix an upper limit on the number of pairs and choose them randomly to cut down the complexity.

4. Experimental protocol

In this section, we present the data sets used and the implementation details. Then, we evaluate different aspects of the SPS_{ad} representation for image classification.

4.1. Image data sets

For this work, we use MSRC-v2, 15Scene, Caltech101, Caltech256 and Pascal VOC 2007 data sets for experiments.

This subsection provides short descriptions of these image data sets.

MSRC-v2: In this data set, there are 591 images that accommodate 23 different categories. All the categories in the images are manually segmented. Different subsets of these categories have been used by several authors to derive a classification problem in which each region or image is assigned to a class label [7, 45].

15Scene: This data set [3, 4, 46] comprises indoor (i.e. office, kitchen, bedroom, etc.) and outdoor (i.e. beach, mountain, tall building, etc.) scenes. Images were collected from different sources, predominantly from Internet, COREL collection and personal photographs. Each category has 200 to 400 images.

Caltech101: The Caltech101 data set [47] has 102 object classes. It has been widely used for evaluation purpose but has some limitations. Namely, most images feature relatively little clutter and possess homogeneous backgrounds. In addition, there are very less variations among the objects of the same category. Despite the limitations, this data set is quite a good resource containing a lot of interclass variability.

Caltech256: This is a challenging set of 257 object categories containing a total of 30,607 images [48]. The minimum number of images in any category is 80. This dataset is collected the same way as Caltech101 while coping with its limitations.

Pascal VOC 2007: The challenging Pascal VOC 2007 database[1] is constituted of images downloaded from internet, containing 9,963 images split into train, val and test sets. Each of the images contains at least one occurrence of the 20 object classes. In many images there are objects of several classes present. Altogether the images contain 24,640 objects. The most common object class ("person") is present in 40% of the images, the rarest ("sheep") in 1.9%.

4.2. Implementation Details

For MSRC-v2 data set we selected the 15 classes: building, grass, tree, cow, sheep, sky, aeroplane, face, car, bike, flower, sign, bird, book, and chair as in [6, 7]. We used

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/index.html

filter-bank responses for feature extraction as in [6, 7]. The training and testing sets are also chosen in accordance with those works for the sake of comparison. For Caltech101 and 15Scene, we follow the experimental setup consistent with [3]. Thus, we use single scale dense detector (8 pixels period) and SIFT descriptor (16 pixels size) for feature extraction. To be able to compare our results with other spatial representations, we use the standard BoVW representation (hard assignment). Finally, for Caltech256 and Pascal VOC 2007, we use the experimental setup of [49] ie. a dense SIFT detector sampled every three pixels at four scales (16, 24, 32 and 40 pixels respectively). For all data sets, we apply K-means on the descriptors to construct the vocabularies. Each descriptor is then mapped to the nearest visual word based on euclidean distance. One versus all multi-class Support Vector Machine (SVM) is used to perform the classification tasks. We use the intersection kernel [50] for the first three data sets and the χ^2 kernel for Pascal VOC 2007 and Caltech256. The cost parameter C was optimized for all the experiments using a 10-fold method on the training set or using the validation set (Pascal VOC 2007). Note that, the new representation does not require any quantization for 2nd order descriptors as opposed to [7]. So, the output of our algorithm is directly fed into the classification algorithm.

4.3. Parameter tuning

In our approach, three parameters (M_θ, M_R and α) have to be set to compute classification results. We study their influence in this section. In figure 4, on the left, we plot the effect of the number of angle bins (M_θ) and distance bins (M_R) on classification accuracy on 15-scene and Caltech101 data sets for a vocabulary size of 200. The evaluation was done using 10-fold cross-validation on the training set. A 45 bins (9×5) spatial histogram appears to be a good compromise for both datasets. Considering finer quantization does not improve the accuracy significantly, but highly increases feature dimension. On the bottom of figure 4, we show the effect of the weighting parameter α on accuracy. For a very low α, not all similar patches are taken into account and for a higher α, there are patches which may not be similar and could be regarded as noise. For both data sets, the value α=0.3 gives the best results. This value seems related to the descriptor in use (SIFT in this case). We have also experimented that the previous values for the three parameters (M_θ, M_R and α) are also relevant for the filter bank descriptor used with MSRC-v2 dataset. They thus will be used in the following sections.

5. Experimental results

In this section, we present our results organized into two main parts. In the first part, we propose to analyze the performance improvements obtained with our concept of soft pairwise spatial angle-distance histograms representation, i.e. the SPS_{ad} representation. We first evaluate the performance increase brought by each of the three following contributions, namely i) infusing spatial information in the BoVW ii) considering inter type visual word pairs and iii) adding a soft weighting when inserting a new pair in the final representation. Then, we compare the SPS_{ad} representation to previous works that also deal with relative visual words positions encoding. In the second part, we focus on our key SPS_{ad}+ representation that exploits SPS_{ad} advantages and also models local and global information. We first compare it with SPR to highlight the two methods complementarity and with similar systems that also enrich the SPR representation either by

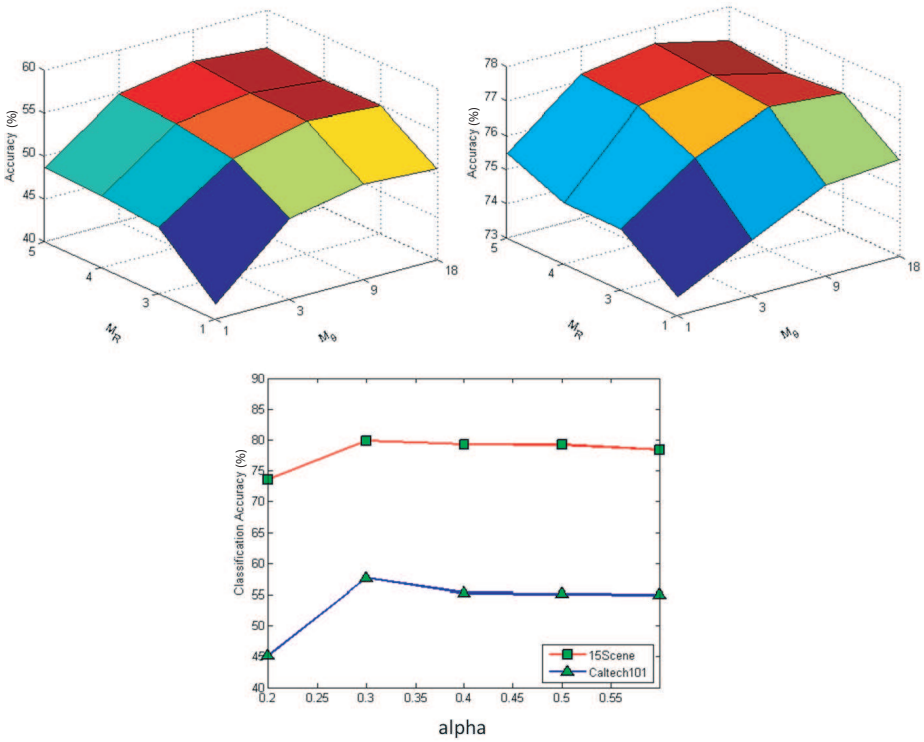

Figure 4: Parameter tuning for SPS_{ad} representation. On the top, the influence of the number of bins for Caltech101 (left) and 15Scene (right) data sets and at the bottom, the influence of α for the same data sets.

introducing relative spatial information or by using advanced coding and spatial pooling methods based on soft assignment. Second, for a complete overview of the performance of $SPS_{ad}+$, we compare it with state-of-the-art methods for image classification, that are built upon the bag-of-words pipeline or not.

5.1. The contributions of our SPS_{ad} representation

The purpose of this section is to assess the improvements provided by the proposed concept of soft pairwise spatial angle-distance histograms representation, studying first the impact of the different factors combined in this concept, and second some computational advantages over close methods.

5.1.1. Independent contributions of each factor to the SPS_{ad} model

Three types of information are combined into the SPS_{ad} representation: spatial information, soft weighting and inter-type visual words pairs consideration. The aim, here, is to evaluate the contribution of each of these factors using the three versions introduced in section 3.4.1: pairwise spatial information with HPS_{ad}^{Intra}, inter type visual word pairs with HPS_{ad}^{Inter} and pairwise similarity weighting with SPS_{ad}. For this study, we chose to report some results obtained on two types of datasets, Caltech101 and 15

Dataset	Voc. Size	BoVW		HPS_{ad}^{Intra}		HPS_{ad}^{Inter}		SPS_{ad}	
		μ	σ	μ	σ	μ	σ	μ	σ
Caltech101	100	39.83%	1.32	53.01%	1.1	55.15%	0.77	53.91%	1.23
	200	41.12%	1.06	55.3%	0.9	55.86%	0.82	57.47%	1.00
	400	45.56 %	1.54	52.11 %	1.38	57.13%	1.32	57.62 %	1.38
	1000	48.08 %	1.42	50.82 %	0.92	58.12%	0.79	58.66 %	0.77
15 Scene Dataset	100	70.83%	0.6	76.11%	0.46	78.02%	0.43	77.96%	0.46
	200	72.2%	0.6	77.52%	0.59	79.41%	0.57	79.38%	0.67
	400	75.7 %	0.33	78.11 %	0.5	79.65%	0.56	79.58 %	0.8
	1000	76.82 %	0.61	76.52 %	0.52	80.00%	0.58	80.38 %	0.44

Table 1: Classification accuracy comparison among BoVW, HPS_{ad}^{Intra}, HPS_{ad}^{Inter} and SPS_{ad} representations. Mean (μ) and Standard Deviation (σ) over 10 individual runs are presented.

Scenes, that enable a good understanding of the analysed factors on both object and scene classification tasks.

Table 1 shows the results for four different vocabulary sizes. From these results, it is clear that for each data set the HPS_{ad}^{Intra} representation improves the results over BoVW. This means that the spatial information is very useful for image classification. For larger dictionaries, spatial information does not seem to be as effective as in the smaller ones. This was also observed in some of the previous works [3, 17, 51].

Second, by comparing HPS_{ad}^{Intra} and HPS_{ad}^{Inter}, we understand how it is important to account both intra type visual word pairs and inter type visual word pairs, since the second one outperforms the first one in most of the cases. This is explained by the new spatial information brought by pairs located near the inter visual word boundary as discussed in section 3.3. Finally, since the contribution of each pair should be different depending on the distance between the two considered descriptors, we have introduced a weighting scheme when inserting a pair in the final representation. This soft insertion is the difference between HPS_{ad}^{Inter} and our proposed representation SPS_{ad}. We notice that this final representation outperforms HPS_{ad}^{Inter} overall, specially when the number of visual words is increasing. Indeed, it is interesting to note that with the increase of vocabulary size BoVW, HPS_{ad}^{Inter} and SPS_{ad} representation improve their results whereas HPS_{ad}^{Intra} reaches an optimal and decreases with the increasing vocabulary size. The reason is that for larger dictionaries intra type words pairs become scarce (one cluster is divided into multiple clusters) and thus HPS_{ad}^{Intra} cannot provide important spatial information. On the other hand, SPS_{ad} should always be able to add spatial information into the BoVW representation regardless of the state of the vocabulary since the soft weighting contribution is not varying with the number of visual words.

5.1.2. Comparison with closely related works

Here, we compare our method with Savarese et al. [7] and Liu et al. [6]. These two works are the most notable among those which concern modeling spatial relationships among the visual words. They rely on the use of new features composed of pairs (or higher number) of words having a specific relative position in order to build spatial histograms. Note that, contrary to our method, the previous approaches do not directly incorporate the spatial information of pair of similar words. We focus on several criteria to compare our work with the mentioned ones. Table 2 shows the details of the comparisons on MSRC-v2 dataset for 400 visual words. For this dataset, SPS_{ad} representation provides the best classification results. Our method also holds different other advantages

Criteria of Comparison	SPS_{ad}	Savarese et al. [7]	Liu et al. [6]
Accuracy	83.5%	81.1%	83.1%
Feature dimensionality	18000	-	1200
Global Spatial Association	Y	N	N
2nd Order Feature Quantization	N	Y	N
Pre-processing/Feature Selection Step	N	Y	Y

Table 2: Comparison among existing methods on a 15 class problem derived from MSRC-V2 dataset. The '-' means that the dimensionality is not mentioned in the corresponding work.

over the existing methods. For example, Liu et al. [6] integrates feature selection and spatial information extraction to boost recognition rate. However, as the spatial feature extraction becomes a part of the learning step, the modification in the training set would lead to recomputation of features and thus making it difficult to generalize. Let's also note that, SPS_{ad} models only global association and unlike Savarese et al. [7], does not require a 2^{nd}-order feature quantization. On the other hand, our method has the highest feature dimensionality (the representation size of Savarese et al. method is not available in their article but it seems to be lower than ours) because, unlike the two other methods, it does not include any feature selection nor additional quantization step. Note that this dimension is still compatible with a fast SVM based classification.

5.2. SPS_{ad+} for image classification

As discussed previously, SPS_{ad+} is our novel representation, based on SPS_{ad}, that enriches the SPR model by encoding both the global and local relative distribution of visual words over an image. In this section, we first propose to study the gain of SPS_{ad+} over SPR-based methods. Then, we enlarge our comparison to a large panel of state-of-the-art methods for image classification, from BoVW-based ones to completely different ones to discuss the interest of our approach.

5.2.1. SPS_{ad+} vs $SPR-$based methods

The current trend in BoVW-based method is over the use of SPR combined with other spatial methods or advanced coding methods. The goal of this section is twofold. First, we study the complementarity of the relative spatial information introduced by SPS_{ad} over the absolute one provided by original SPR. For that purpose, we compare $SPS_{ad}+$ over SPR and over concurrent methods that also combine relative spatial information with SPR. Second, we study the performance of our combined transform over recent advanced coding methods. As previously mentioned, the word ambiguity due to hard assignment coding in the BoVW model prevents the use of large visual vocabularies and thus limits the performance of methods. Two main improvements have been proposed to overcome this problem: 1) to use a soft assignment to account for the proximity of several visual words through Kernel CodeBooks (KCB) [23] or Sparse Coding (SC) [37] methods, 2)to use a locality constraint to enable stability in the set of visual words used to represent similar descriptors through Locality-constrained Linear enCoding (LLC) [52] or Localized Soft-assignment Coding (LSC) [42]. These advanced coding methods also use several spatial pooling strategies to integrate SPR model.

Experiments reported on Table 3 were made using two different pipelines to focus on a fair comparison that enhance the benefits of our approach rather than fine-tuning and

Methods	Caltech101 30 train	15 Scenes 100 train	Pascal VOC2007	Caltech256 30 train	Dimension*
SPR (L=2)	64.6 [3]	81.1 [3]	53.42 [49]	34.1 [48]	4200
$PIWAH+$ [33]	67.1	82.5	-	-	5000
Zhang et al. [16]	65.93	81.5	-	-	13200
Yang et al. [34]	-	82.5	-	-	5565
KCB [23]	64.14	76.67	54.60 [49]	27.17	4200
SC [37]	**73.20**	80.28	-	34.02	21504
LLC [42]	**71.25**	81.53	53.79 [49]	-	21000
LSC [42]	**74.21**	82.70	-	37.2 [54]	21000
$SPS_{ad}+$	**68.4**	**83.7**	**54.97**	**39.9**	13200

Table 3: Classification accuracy (%) (or mean average precision (%) for Pascal VOC2007) provided by methods exploiting the SPR approach for four datasets. A '-' means that the result is not present in the corresponding work. * The dimension column presents the dimension of the final feature vector for the first pipeline used for Caltech101 and 15 Scenes datasets.

searching maximal performance for each dataset. The first pipeline, used for Caltech101 and 15 Scenes datasets, is consistent with [3] with a single scale SIFT sampled every 8 pixels and a small vocabulary of 200 visual words. The second one, used for Pascal VOC2007 and Caltech256, is consistent with [49] with a multiscale SIFT sampled every 3 pixels and a large vocabulary of 4000 words. Even the VLFeat open library [53] proposed by the authors is used to compute our results. The motivation is to use a relatively basic pipeline for the relatively simple datasets Caltech101 and 15 Scene and a more refined one for the more challenging Pascal VOC2007 and Caltech256. However, for advanced coding methods (rows 5-8) the results were not always available with the right parameters in related publications (particularly for Caltech101 and 15 Scenes, they are only available for 1000 visual words). In these cases, we chose publications related to the closest parameters.

We can first see that for the four datasets, $SPS_{ad}+$ outperforms the SPR baseline (row 1). It clearly demonstrates the complementarity of the additional relative spatial information provided by our approach. Moreover, compared to other concurrent methods which combine relative and absolute spatial information (rows 2-4), the new method is also the best performing one. It is worth mentioning that the improvement from $PIWAH+$ to $SPS_{ad}+$ on 15 Scenes is representative since the standard deviations are 0.55 and 0.73 respectively.

Concerning the comparison over advanced coding methods, except for Caltech101, $SPS_{ad}+$ also outperforms other coding methods for all datasets. The worse results for Caltech101 are due to the small vocabulary used (200 words) compared to other coding methods which use a 1000 words vocabulary and thus have a higher dimensionality (21000 vs 13200). For the 15 Scenes dataset, $SPS_{ad}+$ is still the best performing even with the same low vocabulary dimensionality. The trade-off between accuracy and dimensionality is even better when using our reduced-dimension feature vector $SPS_{ad}^{1800}+$ (introduced in section 3.4.3) whose dimension is 1800 and which is providing 67.5% and 83.0% for Caltech101 and 15 Scenes respectively. These results are still very competitive with the other methods from Table 3, despite the low feature dimensionality.

On the other hand, if high accuracy is desired, the use of a refined pipeline (4000 words and denser descriptor) outperforms all improved coding methods on the two challenging

Dataset	Method	Result
Caltech101	Spatially local coding [54]	81.0
Caltech101	P-SIFT + Fisher Vectors + SPR [55]	80.13
Caltech101	Convolutional Neural Networks [56]	86.91
15 Scene	BOSSA NOVA + Fisher Vectors + SPR [57]	88.9
15 Scene	Spatial Fisher Vectors [12]	88.2
Pascal VOC2007	Fisher Vectors [58]	61.69
Pascal VOC2007	Convolutional Neural Networks [59]	77.7
Caltech256	Spatially local coding [54]	46.6
Caltech256	P-SIFT + Fisher Vectors + SPR [55]	44.86
Caltech256	Convolutional Neural Networks [60]	70.6

Table 4: Classification accuracy (%) (or mean average precision (%) for Pascal VOC2007) provided by state-of-the-art methods.

datasets Pascal VOC2007 and Caltech256. Let us also note that all results reported for Pascal VOC2007 are taken from [49] to be strictly comparable to SPS_{ad}+ computed with the same VLFeat library.

5.2.2. Comparison to state-of-the-art

In Table 4, for completeness, we give some top results for the four different datasets used so far. The aim of this table is to show where the state-of-the-art methods are today, compared with our results.

From Table 4, we can see that the most performing approaches are mainly the ones based on convolutional neural networks (CNN) and on Fisher vectors (FV). Indeed, recently, the CNN have shown to provide outstanding results on most of the datasets. However, it is worth mentioning that the CNN require huge amounts of data and time to learn the features. Nowadays, most of the CNN-based approaches are learning their features from the big ImageNet dataset because it provides sufficient data to learn and then try to adapt these pre-learned features to other smallest datasets. Consequently, the CNN provide very good results on datasets that contain classes which are not far (semantically) from the ImageNet classes (chairs, dogs, faces, ...). For specific problems with relative small datasets, such as flower or bird classification, they still not compete with handcraft features. In this case, our approach can help, since it does not require any supervised learning step.

BossaNova and FV enrich BoW representation with extra knowledge from the set of local descriptors but FV use parametric models that lead to very high dimensional image representation. For example, in Table 4, FV require more than 1,376,000-D in order to get that results while the other approaches such as BossaNova, SLC or our solution require less than 300,000-D. Furthermore, there is no guarantee that the local descriptors follow a gaussian distribution around each visual words for all the datasets. If this gaussian assumption does not hold, the representation may be unrepresentative of the local descriptor statistics.

The solutions based on FV such as [55] or [12] are adding information to the classical FV, increasing this way the huge dimension of this descriptor. And sometimes, this dimension increase does not lead to representative improvements. For example, it appears

that simple FV have a mean average precision of 59.5% for Pascal VOC20007, while the spatial-FV from [12] have 56.7%. Likewise, on 15-scenes the spatial-FV reaches 88.2% while the classical FV are at 88.1%.

BossaNova is a concatenation of a classical BoW and a histogram of distances (in the local feature space) between each cluster center (visual word) and all the descriptors. The first problem of this approach is that the range of distances (related to the number of bins in the distance histograms) has to be adapted to the dataset and the used local features. In practice, they set up differently the range bounds for each visual words. Second, the 2 histograms (BoW and distance histograms) are weighted before concatenation and the weights are learned via cross validation on a training/validation subset. Finally, the results provided in Table 4 are the ones obtained by the concatenation between BossaNova and FV. Without concatenation, BossaNova get 85.3% on the 15 Scene dataset while we get 83.7%, without supervised learning step.

Finally, the spatially local coding [54] is inserting the spatial 2D positions (in the image) of the pixels in the local feature vector before creating the dictionary. Consequently, the visual words contain information about their positions in the image space. This kind of representation is helpful only for datasets where the objects have stable positions in the images such as Caltech101 and Caltech256 ones.

6. Conclusion

In this paper, we proposed a new computationally efficient method to model global spatial distribution of visual words and improved the standard BoVW representation. This method exploits spatial orientations and distances of all pairs of similar descriptors in the image. The evaluation was made on an image classification task, using an extensive set of standard data sets.

Experiments demonstrate that: i) our approach succeeds in adding relative spatial information into the BoVW model, ii) it outperforms all other concurrent local histogram based methods, iii) it provides competitive results compared with recent systems that enrich SPR representation by using advanced coding and spatial pooling methods based on soft assignment. Moreover, it also has significant advantages over very recent state of the art approaches.

A direct extension of this work could be to try to combine advanced BoVW encoding techniques as Fisher kernel [58] to our pairwise spatial histograms. Another interesting future direction could be to include some new encoding techniques directly into our pairwise spatial histograms. For example, local soft assignment [42] could be combined with soft similarity weighting. Finally, spatial information provided by multiple cues e.g. color and shape, is also promising as a future direction.

References

[1] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, IEEE, 2003, pp. 1470–1477.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on statistical learning in computer vision, ECCV, Vol. 1, 2004, pp. 1–2.

[3] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.

[4] F.-F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, Vol. 2, 2005, pp. 524–531.
[5] S. Kim, X. Jin, J. Han, Disiclass: discriminative frequent pattern-based image classification, in: Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD '10, ACM, New York, NY, USA, 2010, pp. 7:1–7:10.
[6] D. Liu, G. Hua, P. A. Viola, T. Chen, Integrated feature selection and higher-order spatial feature extraction for object categorization, in: CVPR, 2008.
[7] S. Savarese, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlatons, in: Computer Vision and Pattern Recognition, 2006, pp. 2033–2040.
[8] L. Wu, M. Li, Z. Li, W. ying Ma, N. Yu, Visual language modeling for image classification, in: Multimedia Information Retrieval, 2007, pp. 115–124.
[9] J. Qin, N. H. Yung, Scene categorization via contextual visual words, Pattern Recognition 43 (2010) 1874–1888.
[10] G. Zhou, Z. Wang, J. Wang, D. Feng, Spatial context for visual vocabulary construction, in: International Conference on Image Analysis and Signal Processing, 2010, pp. 176 –181.
[11] N. Morioka, S. Satoh, Building compact local pairwise codebook with joint feature space clustering, in: Proceedings of the 11th European Conference on Computer Vision: Part I, ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 692–705.
[12] J. Krapac, J. J. Verbeek, F. Jurie, Modeling spatial layout with fisher vectors for image categorization, in: ICCV, 2011, pp. 1487–1494.
[13] P. Tirilly, V. Claveau, P. Gros, Language modeling for bag-of-visual words image categorization, in: Conference on Image and Video Retrieval, 2008, pp. 249–258.
[14] J. Yuan, Y. Wu, M. Yang, Discovery of collocation patterns: from visual words to visual phrases, in: Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[15] J. Yuan, Y. Wu, M. Yang, From frequent itemsets to semantically meaningful visual patterns, in: Knowledge Discovery and Data Mining, 2007, pp. 864–873.
[16] E. Zhang, M. Mayo, Improving bag-of-words model with spatial information, in: International Conference of Image and Vision Computing New Zealand, 2010.
[17] Y. Zheng, H. Lu, C. Jin, X. Xue, Incorporating spatial correlogram into bag-of-features model for scene categorization, in: Asian Conference on Computer Vision, 2009, pp. 333–342.
[18] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, Q. Tian, Visual synset: Towards a higher-level visual representation, in: Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[19] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, Discovering objects and their localization in images, in: International Conference on Computer Vision, IEEE, 2005, pp. 370–377.
[20] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrases for image applications, in: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, ACM, New York, NY, USA, 2009, pp. 75–84.
[21] N. M. Elfiky, F. S. Khan, J. van de Weijer, J. Gonzàlez, Discriminative compact pyramids for object and scene recognition, Pattern Recognition 45 (4) (2012) 1627–1636.
[22] D. Parikh, Recognizing jumbled images: The role of local and global information in image classification, in: ICCV, 2011, pp. 519–526.
[23] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, A. W. Smeulders, Kernel codebooks for scene categorization, in: ECCV 2008, Springer, 2008, pp. 696–709.
[24] X. Zhou, N. Cui, Z. Li, F. Liang, T. S. Huang, Hierarchical gaussianization for image classification, in: ICCV 2009, IEEE, 2009, pp. 1971–1977.
[25] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2007, pp. 401–408.
[26] T. Harada, Y. Ushiku, Y. Yamashita, Y. Kuniyoshi, Discriminative spatial pyramid, in: Computer Vision and Pattern Recognition, 2011, pp. 1617–1624.
[27] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, Spatial-bag-of-features, in: Computer Vision and Pattern Recognition, 2010, pp. 3352 – 3359.
[28] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, R. Zabih, Image indexing using color correlograms, in: Computer Vision and Pattern Recognition, CVPR '97, IEEE Computer Society, Washington, DC, USA, 1997, pp. 762–768.
[29] Y. Zhang, Z. Jia, T. Chen, Image retrieval with geometry-preserving visual phrases, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11, IEEE Computer Society, Washington, DC, USA, 2011, pp. 809–816.
[30] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR

'09, IEEE Computer Society, Washington, DC, USA, 2009, pp. 25 – 32.
[31] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from, in: In British Machine Vision Conference, 2002, pp. 384–393.
[32] A. Agarwal, B. Triggs, Hyperfeatures – multilevel local coding for visual recognition, in: Proceedings of the 9th European Conference on Computer Vision - Volume Part I, ECCV'06, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 30–43.
[33] R. Khan, C. Barat, D. Muselet, C. Ducottet, Spatial orientations of visual word pairs to improve bag-of-visual-words model, in: British Machine Vision Conference, BMVA, 2012.
[34] Y. Yang, S. Newsam, Spatial pyramid co-occurence for image classification, in: Internation Conference of Computer Vision, 2011.
[35] T. Harada, H. Nakayama, Y. Kuniyoshi, Improving local descriptors by embedding global and local spatial information, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), Proceedings of the 9th European Conference on Computer Vision, Vol. 6314 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 736–749.
[36] D. G. Lowe, Object recognition from local scale-invariant features, in: International Conference on Computer Vision, Vol. 2, 1999, pp. 1150–1157.
[37] J. Yang, K. Yu, Y. Gong, T. S. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.
[38] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2001, pp. I–511.
[39] T. Deselaers, V. Ferrari, Global and efficient self-similarity for object classification and detection, in: Computer Vision and Pattern Recognition, 2010, pp. 1633–1640.
[40] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[41] J. C. van Gemert, C. J. Veenman, A. W. Smeulders, J.-M. Geusebroek, Visual word ambiguity, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7) (2010) 1271–1283.
[42] L. Liu, L. Wang, X. Liu, In defense of soft-assignment coding, in: ICCV 2011, IEEE, 2011, pp. 2486–2493.
[43] I. Dhillon, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, JMLR 3 (2003) 1265–1287.
[44] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.
[45] Y. Su, F. Jurie, Visual word disambiguation by semantic contexts, in: International Conference of Computer Vision, 2011, pp. 311–318.
[46] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision 42 (2001) 145–175.
[47] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories., in: Workshop on Generative-Model Based Vision, 2004.
[48] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, in: Caltech Technical Report, California Institute of Technology, 2007.
[49] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: British Machine Vision Conference, 2011.
[50] M. J. Swain, D. H. Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991) 11–32.
[51] J. Yang, Y.-G. Jiang, A. G. Hauptmann, C.-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in: ACM Multimedia Information Retrieval Workshop, 2007, pp. 197–206.
[52] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 3360–3367.
[53] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, `http://www.vlfeat.org/` (2008).
[54] S. McCann, D. G. Lowe, Spatially local coding for object recognition, in: Proceedings of the 2012 Asian Conference on Computer Vision, Springer, 2013, pp. 204–217.
[55] L. Seidenari, G. Serra, A. D. Badanov, A. D. Bimbo, Local pyramidal descriptors for image recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (5) (2013) 1033–1040.
[56] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: Proceedings of the 31st International

Conference on Machine Learning, ICML 2014, 2014.
[57] S. Avila, N. Thome, M. Cord, E. Valle, A. de A. Arajo, Pooling in image representation: The visual codeword point of view, Computer Vision and Image Understanding 117 (5) (2013) 453 – 465.
[58] F. Perronnin, J. Snchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference of Computer Vision, 2010.
[59] M. Oquab, I. Laptev, L. Bottou, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014.
[60] M. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Arxiv 1311.2901, 2013.