Spatial orientations of visual word pairs to improve Bag-of-Visual-Words model

Rahat Khan
rahat.khan@univ-st-etienne.fr
Cecile Barat
cecile.barat@univ-st-etienne.fr
Damien Muselet
damien.muselet@univ-st-etienne.fr
Christophe Ducottet
ducottet@univ-st-etienne.fr

Université de Lyon, F-42023, Saint-Etienne, France,
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France.

This paper presents a novel approach to incorporate spatial information in the bag-of-visual-words (BoVW) model [1, 3] for category level and scene classification. In the traditional BoVW model, feature vectors are histograms of visual words. This representation is appearance based and does not contain any information regarding the arrangement of the visual words in the 2D image space. In this framework, we present a simple and efficient way to infuse spatial information. Particularly, we are interested in explicit global relationships among the spatial positions of visual words. For that we first introduce the notion of Pair of Identical visual Words (PIW) defined as the set of all the pairs of visual words of the same type. Then a spatial distribution of words is represented as a histogram of orientations of the segments formed by PIW. Figure 1 shows an example which gives an intuition to better understand our approach.

Our method eliminates a number of drawbacks from the previous approaches [2, 3] by i) proposing a simpler word selection technique that supports fast exhaustive spatial information extraction, ii) enabling infusion of global spatial information, iii) being robust to geometric transformations like translation and scaling.

In the conventional BoVW model, each image is represented by a set of local descriptors $\{d_1 \dots d_n\}$ extracted from n patches around interest points or regular grids. A visual vocabulary $W = \{w_1, w_2, w_3, w_4 \dots w_N\}$ is obtained by clustering a set of descriptors from all the training images. Here, N is a predefined number and the size of the vocabulary. Each patch of the image is then mapped to the nearest visual word according to the following equation:

$$w(d_k) = \arg\min_{w \in W} \mathrm{Dist}(w, d_k) \tag{1}$$

Here, $w(d_k)$ denotes the visual word assigned to the k^{th} descriptor d_k and $Dist(w, d_k)$ is the distance between the visual word w and the descriptor d_k. In the conventional BoVW method, the final representation of the image is a histogram of visual words. The number of bins in the histogram is equal to the number of visual words in the dictionary (i.e. N). If each bin b_i represents occurrences of a visual word w_i in W, b_i is defined as:

$$b_i = \mathrm{Card}(\mathcal{D}_i) \quad \text{where} \quad \mathcal{D}_i = \{d_k, k \in \{1, \dots, n\} \mid w(d_k) = w_i\} \tag{2}$$

$\mathcal{D}_i$ is the set of all the descriptors corresponding to a particular visual word w_i in the given image. $\mathrm{Card}(\mathcal{D}_i)$ is the cardinality of the set $\mathcal{D}_i$. In this final step, the spatial information of interest points is not retained. To model this information and to infuse it to the BoVW model, we propose the angle histogram of PIW. For each visual word w_i the method is as follows: first, from the set $\mathcal{D}_i$ of descriptors assigned to w_i (Equation 2), we consider all pairs of those descriptors and we build the set PIW_i constituted by the corresponding position pairs.

$$PIW_i = \{(P_k, P_l) \mid (d_k, d_l) \in \mathcal{D}_i^2, d_k \neq d_l\} \tag{3}$$

where P_k and P_l correspond to the spatial positions in the image from which the descriptors d_k and d_l have been extracted. The spatial position of a descriptor is given by the coordinates of the top-left pixel of the corresponding patch. These coordinates vary in the range of the image spatial domain. The cardinality of the set PIW_i is $\binom{b_i}{2}$, i.e. the number of possible subsets of two distinct elements among b_i elements. Second, for each pair of points of the set PIW_i, we compute the angle θ formed with the horizontal axis using the cosine law:

Figure 1: Discriminative power of spatial distribution of intra type visual words. Four images from Caltech101 dataset are shown. The black squares refer to identical visual words across all the images. 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. Our proposal 'PIW Angle Histogram' can capture information about these distributions.

$$\theta = \begin{cases} \arccos\left(\dfrac{\overrightarrow{P_kP_l} \cdot \vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{if } \overrightarrow{P_kP_l} \cdot \vec{j} > 0 \\ \pi - \arccos\left(\dfrac{\overrightarrow{P_kP_l} \cdot \vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{otherwise} \end{cases} \tag{4}$$

where $\overrightarrow{P_kP_l}$ is the vector formed by two points P_k and P_l and i and j are orthogonal unit vectors defining the image plane. Third, the histogram of all θ angles is calculated. The bins of this histogram are equally distributed between 0° and 180°. The optimal number of bins is chosen empirically. We call this histogram the PIW angle histogram for word w_i and denote it as $PIWAH_i$.

To have a global representation, we replace each bin of the BoVW frequency histogram with the $PIWAH_i$ histogram associated to w_i. The sum of all the bins of $PIWAH_i$ is normalized to the bin-size b_i of the respective bin of the BoVW frequency histogram. By this way, we keep the frequency information intact and add the spatial information. Equation 5 formalizes our global representation of an image, denoted as $PIWAH$.

$$PIWAH = (\alpha_1 PIWAH_1, \alpha_2 PIWAH_2, \alpha_3 PIWAH_3, \dots, \alpha_N PIWAH_N) \quad \text{where} \quad \alpha_i = \frac{b_i}{\|PIWAH_i\|_1} \tag{5}$$

Here, N is the vocabulary size and α_i is the normalization term. If the number of bins in each of $PIWAH_i$ is M, the size of the $PIWAH$ representation becomes MN.

For this work, we use MSRC-v2, Caltech101, 15 Scene and Graz-01 datasets for experiments. Our method imroves classification accuracy for all the datasets. The improvement is 12% for Caltech101 and 4% for 15Scene over BoVW representation. Our method also improves accuracy for Graz-01 dataset where global information is extremly difficult to model. We show that our method is complementary to method like Spatial Pyramid [1]. We also show on the MSRC-v2 dataset that our method performs as good as the existing ones [2, 3] with the advantage of being the fastest and the simplest among all.

[1] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[2] David Liu, Gang Hua, Paul A. Viola, and Tsuhan Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[3] Silvio Savarese, John Winn, and Antonio Criminisi. Discriminative object class models of appearance and shape by correlatons. In *Computer Vision and Pattern Recognition*, pages 2033–2040, 2006.