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

Rahat Khan, Cécile Barat, Damien Muselet, Christophe Ducottet

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.

Abstract

This paper presents a novel approach to incorporate spatial information in the bag-of-visual-words model for category level and scene classification. In the traditional bag-of-visual-words 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. Therefore, we take advantage of the orientation of the segments formed by Pairs of Identical visual Words (PIW). An evenly distributed normalized histogram of angles of PIW is computed. Histograms produced by each word type constitute a powerful description of intra type visual words relationships. Experiments on challenging datasets demonstrate that our method is competitive with the concurrent ones. We also show that, our method provides important complementary information to the spatial pyramid matching and can improve the overall performance.

1 Introduction

The bag-of-visual-words (BoVW) model has shown excellent results in recent years in category level and scene classification [5, 6]. In BoVW, local features are extracted from the whole image dataset and quantized (termed as visual words). This approach employs histogram based features for image representation. This representation is orderless and does not contain any spatial information.

A range of methods has been proposed to incorporate spatial information to improve the BoVW model [4, 5, 7, 11, 16]. Lazebnik et al. [5] proposed spatial pyramid match (SPM) kernel based on the work done by Grauman et al. [3]. It has been among the most notable works in this domain. SPM combines aggregated statistics of fixed subregions from different levels of a pyramid. Interestingly, this method is not invariant to global geometric transformations. SPM only captures the information about approximate global geometric correspon-

http://dx.doi.org/10.5244/C.26.89

dence of the visual words, however, the spatial information provided by the visual words is not only limited to this. So, for better accuracy, in [18, 19] the authors combine other types of spatial information into the SPM. Zhang et al. [19] use different heuristic approaches with success to infuse additional spatial information into the SPM and Yang et al. [18] use co-occurrence information to improve the SPM. These works show that, additional spatial information is required along with global correspondence for further improvement of the accuracy.

In the context of other types of spatial information, spatial relationships among visual words have received relatively little attention. The main reason is computational. Due to the large number of visual words in an image, modeling explicit spatial relationships among visual words is computationally expensive. So, this category of techniques reduces the size of vocabulary using different methods or employs feature selection to speed up the computation. In [11], the authors employ correlogram to model relationship among the visual words. They use the method proposed by [15] to obtain a more compact vocabulary. As the correlogram is a function of distance, its success depends on the choice of the distances. Moreover, the constraint of distance makes this representation variant to scale changes. In another work, Liu et al. [7] introduces an integrated feature selection and spatial information extraction technique to reduce the number of features and also to speed up the process. Note that, all of the previous methods under this category only deal with local and semi-local information, although global spatial methods very often outperform the local ones.

In this paper, we propose a way to model the global spatial distribution of visual words. We consider the interaction among visual words regardless of their spatial distances. 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. Note that, our method eliminates a number of drawbacks from the previous approaches 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. We only consider relationships among identical visual words. It not only reduces the complexity but also adds discriminative information to improve the classification accuracy.

The rest of the article is organized in the following way: the next section provides a review of the BoVW model and introduce our principal notations. Section 3 presents our approach to incorporate spatial information into the BoVW model. Section 4 provides 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 The BoVW model

In the conventional bag-of-visual-words model, at first each image I is represented in terms of image descriptors [8, 15]:

$$I = \{d_1, d_2, d_3, d_4, \ldots d_n\} \tag{1}$$

where d_i is the shape, texture or color description of an image patch and n is the total number of patches in the image. By this way, we get numerous descriptors from all the local patches of all the images for a given dataset. Typically, K-means unsupervised clustering is

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. PIW Angle Histogram, defined in section 3.1 can capture information about these distributions.

applied on these descriptors to find clusters $W = \{w_1, w_2, w_3, w_4 \dots w_N\}$ that constitutes the visual vocabulary, where N is the predefined number of clusters. So, each patch of the image is now mapped to the nearest visual word according to the following equation:

$$w(d_k) = \arg\min_{w \in W} \text{Dist}(w, d_k) \tag{2}$$

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. Each image is now represented as a collection of patches where each patch is mapped to one visual word. 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,

$$b_i = \text{Card}(\mathcal{D}_i) \quad \text{where} \quad \mathcal{D}_i = \{d_k, k \in \{1, \dots, n\} \mid w(d_k) = w_i\} \tag{3}$$

$\mathcal{D}_i$ is the set of all the descriptors corresponding to a particular visual word w_i in the given image. $\text{Card}(\mathcal{D}_i)$ is the cardinality of the set $\mathcal{D}_i$. This is done for every word in the vocabulary to obtain the final histogram. In this final step, the spatial information of interest points is not retained. In the next section, we define angle histogram of PIW as a tool to model this information and we infuse it to the BoVW model.

3 Encoding orientations of identical visual word pairs

Spatial relationships among visual words are typically computed for all the words in the vocabulary [7, 11]. However, we propose to characterize the relative spatial distribution of the patches associated with the same visual word. The motivation comes from the previous works [1, 12] 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. However, our work is unique and different from that of [1, 12], since our method extends the existing BoVW by encoding spatial information whereas the previous works [1, 12] extract additional self-similarity features in other domains.

Figure 1 shows an example which gives an intuition to better understand our approach. Here, images have the same visual words in similar frequencies. They are not well distinguishable based on this information but their global distribution can add additional information which is more useful to classify them to the respective categories. Moreover, considering

relationships among identical visual words can provide global shape information, specially, if the whole or parts of the object has uniform texture and/or uniform background and the image is densely sampled. Hence, for a given visual word, we propose to analyze the spatial positions of all occurrences of that visual word in the image. The PIW angle histogram presented in the following subsection is able to take into account the global distribution of visual words.

3.1 PIW angle histogram

Definition: The angle histogram we propose relies on the spatial distribution of identical visual words. 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 3), we consider all pairs of those descriptors and we build the set PIW_i constituted by the corresponding position pairs.

$$PIW_i = \{(P_k, P_l) \mid (d_k, d_l) \in \mathcal{D}_i^2, d_k \neq d_l\} \tag{4}$$

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:

$$\theta = \begin{cases} \arccos\left(\dfrac{\overrightarrow{P_kP_l}\cdot\vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{if } \overrightarrow{P_kP_l}\cdot\vec{j} > 0 \\ \pi - \arccos\left(\dfrac{\overrightarrow{P_kP_l}\cdot\vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{otherwise} \end{cases} \tag{5}$$

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 $PIWAH_i$. We propose the way to combine all $PIWAH_i$ resulting from the different words in section 3.2. In the next section, we analyze the properties of such angle histograms.

Properties: Angle histograms are interesting due to their ability to capture distinct spatial distributions of points. In our case, an angle histogram is always built on a discrete set of points on the two dimensional space. Figure 2, clearly shows that point distributions forming different geometrical shapes lead to distinct angle histograms. By definition, PIW angle histogram is invariant to translation and scaling but not directly invariant to rotation. However, our method tolerates some rotation variations controlled with the histogram bin size. Rotation invariance could be achieved using point triplets instead of point pairs, but at the price of increased complexity. Another solution would be to use proper training images. It should be noted that, in some image datasets, orientation of objects provides discriminative information.

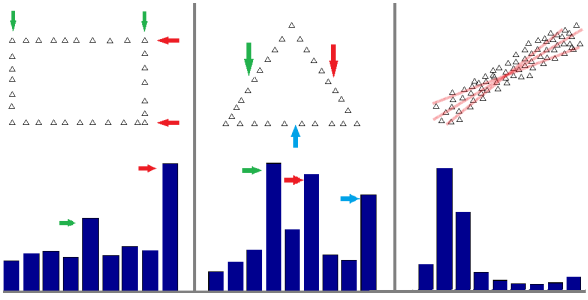

Figure 2: Some toy distributions of points and the corresponding 9-bin angle histograms. For the rectangular shape(left), the angle histogram has two main peaks, specifically, at bin 5 ($81^\circ - 100^\circ$) and bin 9 ($161^\circ - 180^\circ$), due to the vertical and horizontal sides of the rectangle. The correspondence has been shown using identical color arrows. Likewise, for the triangle(middle), the angle histogram produces 3 main peaks corresponding to the 3 sides of the triangle. For a set of points scattered around a line(right), the angle histogram shows peaks in the bins corresponding to the orientations of the fitted lines (3 examples are shown) to the distribution.

3.2 Image Representation

The $PIWAH_i$ provides the spatial distribution pattern of a visual word w_i. We assume this pattern information is specific to the visual content carried by this specific word. That is, given an object category and a visual word, the distribution of angles from the set PIW_i is stable.

To have a global representation of the image, we need a way to combine $PIWAH_i$ from all the visual words. For that, we transform the BoVW representation with a 'bin replacement' technique. Bin replacement literally means to 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 which is going to be replaced. By this way, we keep the frequency information intact and add the spatial information. Equation 6 formalizes our global representation of an image, denoted as $PIWAH$.

$$PIWAH = (\alpha_1 PIWAH_1, \alpha_2 PIWAH_2, \alpha_3 PIWAH_3, \ldots, \alpha_N PIWAH_N) \quad where \quad \alpha_i = \frac{b_i}{\|PIWAH_i\|_1} \tag{6}$$

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.

4 Experimental Protocol

Our goal is to evaluate the potential of the new $PIWAH$ representation on image classification tasks. We will describe the experiments made and the corresponding results in section 5. In this section, we present the datasets used and the implementation details.

4.1 Image Datasets

For this work, we use MSRC-v2, Caltech101, 15 Scene and Graz-01 datasets for experiments. This subsection provides short descriptions of these image datasets.

MSRC-v2: In this dataset, 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 [11, 13].

15Scene: This dataset [5, 6, 9] 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, and the average image size is 300×250 pixels.

Caltech101: The Caltech101 dataset [2] has 102 object classes. It is one of the most diverse object database available. However, this dataset 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 dataset is quite a good resource containing a lot of interclass variability.

Graz-01: The Graz-01 image dataset [10] has two object classes, bikes (373 images) and persons (460 images) and a background class (270 images). Images are presented in various scales and poses and feature high intra-class variation.

4.2 Implementation Details

For MSRC-v2 dataset we use a 15 category problem as used in [7, 11]. We use a filter-bank responses for feature extraction as in [7, 11]. The training and testing set is also chosen in accordance with those works for the sake of comparison. For the other datasets, we follow the experimental setup consistent with [5]. Thus, we use dense detector and SIFT descriptor for feature extraction. For all the datasets, 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.

Calculation of PIWAH includes a step of computing subsets of pairs from similar visual word sets. If there are n member patches in an image representing one visual word, the computational complexity of this step becomes $O(n^2)$. For a vocabulary of m visual words, the complexity of calculation of PIWAH becomes $O(mn^2)$. To speed up computation, we use a threshold and a random selection to limit the number of words of the same type used for the pairs. Taking advantage of Matlab's vectorial programming, it took only 0.5 seconds on average to compute PIWAH representation per image in a 2 GHz Pentium machine for the vocabulary size of 400. Note that, this representation does not require any quantization for 2nd order descriptor's as opposed to [11]. So, the output of our algorithm is directly fed into the classification algorithm.

We use 9-bin *PIWAH* representation for the results presented in the next section. Figure 3 shows empirical justification of our chosen number of bins on two different datasets used in our experiments. The results are obtain by a 10-fold cross validation on the training sets.

In this work, we also combine *PIWAH* representation with SPM . We denote this combination representation as *PIWAH*+. *PIWAH*+ is defined exactly as the concatenation of the finest level of the spatial pyramid with the *PIWAH* representation without any weight. Here, we only work with a 3-level spatial pyramid. That means the finest level(L=2) has 4×4 subregions. Lazebnik et al. [5] shows that 3-level pyramid performs reasonably well for all the datasets we are using for the experiments with *PIWAH*+.

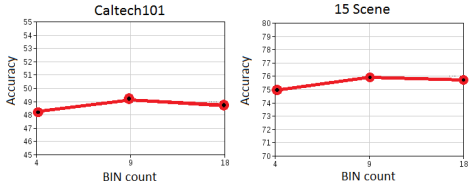

Figure 3: The influence of number of bins used in *PIWAH* representation on accuracy. 9-bin *PIWAH* gives best result for both the datasets.

Support Vector Machine (SVM) is used to perform the classification tasks. We use the intersection kernel [14] and the one-vs-all rule where multi class classification is necessary. The cost parameter C was optimized for all the experiments using a 10-fold method on the training set.

5 Results

In this section, we first study the performance of *PIWAH* and *PIWAH*+ image representation. Next, we compare *PIWAH* with other spatial methods.

5.1 Performance evaluation of PIWAH representation

Here, we first analyze the performance for *PIWAH* representation on Caltech101, 15Scene and Graz-01 datasets. We show the classification performance gain for *PIWAH* over BoVW representation and discuss the results. Next, we compare *PIWAH*, *PIWAH*+ with SPM and some other combination spatial descriptors similar to *PIWAH*+.

Table 1 shows results on Caltech101 and 15 Scene datasets for 3 different vocabulary sizes. From these results, it is clear that for each dataset the *PIWAH* representation improves the results over BoVW representation for all the vocabulary sizes. However, for smaller vocabulary sizes the improvement is more eminent than for larger vocabulary sizes. So, it seems that spatial information is more discriminative when extracted from lower vocabulary sizes. This phenomenon was also observed in some of the previous works [5, 17, 20]. A more detailed analysis of the results is shown in Figure 4. It presents the best and worst 15 classes on the Caltech101 dataset for our method.

Next, on Table 2 we show the classification accuracy for bike vs no-bike and person vs no-person classifiers. In this table we only show results for vocabulary size of 200 as this vocabulary size exhibits the best results for the datasets on the Table 1. We can see that even for extremely variable object poses our method improves the classification accuracy by some margin. However, this improvement is limited by the absence of proper global cue.

Now, we compare the combination descriptor *PIWAH*+ with SPM and different other approaches [13, 19] that also propose combination descriptors based on SPM. Along with fundamental similarities with *PIWAH*+, these methods also use similar setup to us namely, dense sampling as detector, SIFT as descriptor, K-means for vocabulary construction and histogram based representation, thus facilitates fair comparison. Table 3 shows the comparison among all the mentioned methods for Caltech101 and 15 Scene datasets. We can see that the global distribution of visual words is complementary to the global correspondence and our method outperforms SPM and the other methods in most of the cases.

Dataset	Voc. Size	BoVW		PIWAH	
		μ	σ	μ	σ
Caltech101	100	39.83%	1.32	51.47%	1.49
	200	41.12%	1.06	51.86%	0.89
	400	45.56 %	1.54	49.41 %	1.08
15 Scene Dataset	100	70.83%	0.6	74.6%	0.6
	200	72.2%	0.6	76.0%	0.6
	400	75.7 %	0.33	75.9 %	0.6

Table 1: Classification accuracy comparison between BoVW representation and Angle Histogram. Mean (μ) and Standard Deviation (σ) over 10 individual runs are presented.

Class	BoVW	PIWAH
Bikes	83.9 $\pm$3.1	85.7 $\pm$2.2
People	81.3 $\pm$2.5	82.3$\pm$2.3

Table 2: Classification accuracy(%) presented for BoVW and *PIWAH* representation for Graz01 dataset. Visual vocabulary size is 200.

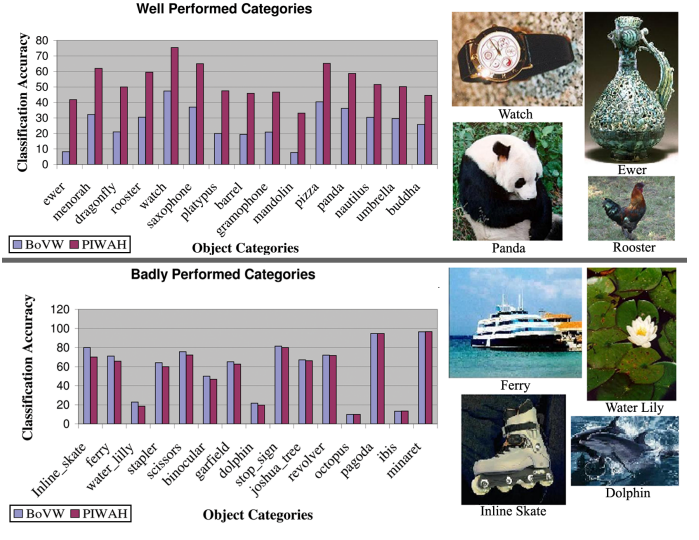

Figure 4: Object wise comparison between BoVW and *PIWAH* representation for Caltech101 dataset. On the top, the 15 categories that show the highest improvements in classification accuracy for *PIWAH* over BoVW. On the bottom, the 15 categories that show the lowest improvements. Some example images for both the cases are shown by the side.

For Graz01 dataset (Table 4), we only compare *PIWAH*+ and SPM due to the absence of these results in [18] and [19]. As discussed earlier, the classes in this dataset, present a high geometric variability thus finding useful and discriminative global cue is difficult. In this case, *PIWAH*+ improves the result although the improvement is not significant (Table 4). This is understandable because *PIWAH* and SPM both work best when the objects are in their 'canonical' pose. In case of high geometric variability, depending on the object pose and location in the image the global distribution and the global correspondence can be

Dataset	SPM Single Level (L=2)	SPM Entire Pyramid (L=2)	*PIWAH*+	Yang et al. [18]	Zhang et al. [19]
Caltech101	63.6%*	64.6%*	**67.1%**	X	65.93%
15 Scene Dataset	79.4%*	81.1%*	**82.5%**	82.5%	81.5%

Table 3: Classification accuracy(%) comparison among SPM , *PIWAH*+ and two other methods for Caltech101 and 15 scene dataset. Results with * are taken from [5]. a 'X' means that the result is not present in the corresponding work.

Dataset	SPM Single Level (L=2, 4×4)	SPM Entire Pyramid (L=2)	PIWAH+
Bikes	85.5 $\pm$2.7	86.6 $\pm$2.5	87.35 $\pm$2.6
People	83.4 $\pm$2.1	84.5$\pm$2.4	84.6$\pm$2.4

Table 4: Classification accuracy(%) comparison between spatial pyramid and *PIWAH*+ for Graz01 dataset. Visual vocabulary size is 200.

contradictory. This problem can be dealt by learning appropriate weights for the features.

5.2 Comparison between PIWAH and other spatial methods

Here, we compare our method with Savarese et al. [11] and Liu et al. [7]. 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 pair (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 identical words. We focus on several criteria to compare our work with the mentioned ones. The table 5 shows the details of the comparisons. The classification is performed on MSRC-v2 dataset. For this dataset, *PIWAH* representation provides best result when the vocabulary size is 400. Our method gives comparable results to the existing methods being the fastest among all. Although, our method outperforms Savarese et al. [11], it provides worse results than [7]. In their work Liu et al. [7] integrate feature selection and spatial information extraction to boost recognition rate. This integrated approach can also be easily adapted to our method to further boost the performance. 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, *PIWAH* models global association only and unlike Savarese et al. [11], does not require a 2^{nd}-order feature quantization. As the previous approaches fail to incorporate the spatial information of similar pairs properly, our approach is complementary to these approaches as well.

6 Conclusion

In this work, we propose a new method to model global spatial distribution of visual words and improve the standard BoVW model. This method exploits orientation of all pairs of identical visual words in the image. The evaluation was made on an image classification task, using an extensive set of standard datasets. Experiments confirm that our model outperforms

Criteria of Comparison	*PIWAH*	Savarese et al. [11]	Liu et al. [7]
Accuracy	82.0%	81.1%	83.1%
Speed	+++	+	++
Global Spatial Association	Y	N	N
2nd Order Feature Quantization	N	Y	N
Pre-processing/Feature Selection Step	N	Y	Y

Table 5: Comparison among existing methods on a 15 class problem derived from MSRC-V2 dataset.

the standard BoVW approach and most of the existing spatial methods on all the datasets. Although, the method proposed in [7] provides better accuracy than our method on MSRC-v2 dataset, it requires a dataset specific feature selection step which can be difficult to apply in practical situations. Compared to the global correspondence methods as SPM, our model brings complementary information. In this case , we outperform most of the methods that do the same. Furthermore, the possibility to integrate local spatial information in this framework is still on the table. Spatial information provided by other cues e.g. color is also promising as a future direction.

References

[1] Thomas Deselaers and Vittorio Ferrari. Global and efficient self-similarity for object classification and detection. In *Computer Vision and Pattern Recognition*, pages 1633–1640, 2010.

[2] L. Fei-Fei, R. Fergus, and 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.

[3] Kristen Grauman and Trevor Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *International Conference of Computer Vision*, pages 1458–1465, 2005.

[4] Sangkyum Kim, Xin Jin, and Jiawei Han. Disiclass: discriminative frequent pattern-based image classification. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 7:1–7:10, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0220-3.

[5] 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.

[6] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

[7] 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.

[8] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[9] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, May 2001. ISSN 0920-5691.

[10] Andreas Opelt, Michael Fussenegger, Axel Pinz, and Peter Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV (2)*, pages 71–84, 2004.

[11] 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.

[12] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[13] Yu Su and Frédéric Jurie. Visual word disambiguation by semantic contexts. In *International Conference of Computer Vision*, pages 311–318, 2011.

[14] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[15] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *International Conference on Computer Vision*, pages 1800–1807. IEEE Computer Society, 2005.

[16] Lei Wu, Mingjing Li, Zhiwei Li, Wei ying Ma, and Nenghai Yu. Visual language modeling for image classification. In *Multimedia Information Retrieval*, pages 115–124, 2007.

[17] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval Workshop*, pages 197–206, 2007.

[18] Yi Yang and Shawn Newsam. Spatial pyramid co-occurence for image classification. In *Internation Conference of Computer Vision*, 2011.

[19] Edmond Zhang and Michael Mayo. Improving bag-of-words model with spatial information. In *International Conference of Image and Vision Computing New Zealand*, 2010.

[20] Yingbin Zheng, Hong Lu, Cheng Jin, and Xiangyang Xue. Incorporating spatial correlogram into bag-of-features model for scene categorization. In *Asian Conference on Computer Vision*, pages 333–342, 2009.