Hierarchical Spatial Matching Kernel for Image Categorization

Tam T. Le[1], Yousun Kang[2], Akihiro Sugimoto[2],
Son T. Tran[1], and Thuc D. Nguyen[1]

[1] University of Science, VNU-HCMC, Vietnam
{lttam,ttson,ndthuc}@fit.hcmus.edu.vn
[2] National Institute of Informatics, Tokyo, Japan
{yskang,sugimoto}@nii.ac.jp

Abstract. Spatial pyramid matching (SPM) has been one of important approaches to image categorization. Despite its effectiveness and efficiency, SPM measures the similarity between sub-regions by applying the bag-of-features model, which is limited in its capacity to achieve optimal matching between sets of unordered features. To overcome this limitation, we propose a hierarchical spatial matching kernel (HSMK) that uses a coarse-to-fine model for the sub-regions to obtain better optimal matching approximations. Our proposed kernel can robustly deal with unordered feature sets as well as a variety of cardinalities. In experiments, the results of HSMK outperformed those of SPM and led to state-of-the-art performance on several well-known databases of benchmarks in image categorization, even when using only a single type of feature.

Keywords: kernel method, hierarchical spatial matching kernel, image categorization, coarse-to-fine model.

1 Introduction

Image categorization is the task of classifying a given image into a suitable semantic category. The semantic category can be defined as the depicting of a whole image such as a forest, a mountain or a beach, or of the presence of an interesting object such as an airplane, a chair or a strawberry. Among existing methods for image categorization, the bag-of-features (BoF) model is one of the most popular and efficient. It considers an image as a set of unordered features extracted from local patches. The features are quantized into discrete visual words, with sets of all visual words referred to as a dictionary. A histogram of visual words is then computed to represent an image. One of the main weaknesses in this model is that it discards the spatial information of local features in the image. To overcome it, spatial pyramid matching (SPM) [9], an extension of the BoF model, utilizes the aggregated statistics of the local features on fixed sub-regions. It uses a sequence of grids at different scales to partition the image into sub-regions, and then computes a BoF histogram for each sub-region. Thus,

M. Kamel and A. Campilho (Eds.): ICIAR 2011, Part I, LNCS 6753, pp. 141–151, 2011.

the representation of the whole image is the concatenation vector of all the histograms.

Empirically, it is realized that to obtain good performances, the BoF model and SPM have to be applied together with specific nonlinear Mercer kernels such as the intersection kernel or χ^2 kernel. This means that a kernel-based discriminative classifier is trained by calculating the similarity between each pair of sets of unordered features in the whole images or in the sub-regions. It is also well known that numerous problems exist in image categorization such as the presence of heavy clutter, occlusion, different viewpoints, and intra-class variety. In addition, the sets of features have various cardinalities and are lacking in the concept of spatial order. SPM embeds a part of the spatial information over the whole image by partitioning an image into a sequence of sub-regions, but in order to measure the optimal matching between corresponding sub-regions, it still applies the BoF model, which is known to be confined when dealing with sets of unordered features.

In this paper, we propose a new kernel function based on the coarse-to-fine approach and we call it a hierarchical spatial matching kernel (HSMK). HSMK allows not only capturing spatial order of local features, but also accurately measuring the similarity between sets of unordered local features in sub-regions. In HSMK, a coarse-to-fine model on sub-regions is realized by using multi-resolutions, and thus our feature descriptors capture not only the local details from fine resolution sub-regions, but also global information from coarse resolution ones. In addition, matching based on our coarse-to-fine model involves a hierarchical process. This indicates that a feature that does not find its correspondence in a fine resolution still has a possibility of having its correspondence in a coarse resolution. Accordingly, our proposed kernel can achieve a better optimal matching approximation between sub-regions than SPM.

2 Related Work

Many recent methods have been proposed to improve the traditional BoF model. Generative methods [1,2] model the co-occurrence of visual words while discriminative visual words learnings [13,20] or sparse coding methods [11,19] improve the dictionary in terms of discriminative ability or lower reconstruction error instead of using the quantization by K-means clustering. On the other hand, SPM captures the spatial layout of features ignored in the BoF model. Among these improvements, SPM is particularly effective as well as being easy and simple to construct. It is utilized as a major part in many state-of-the-art frameworks in image categorization [3].

SPM is often applied with a nonlinear kernel such as the intersection kernel or χ^2 kernel. This requires high computation and large storage. Maji *et al.* [12] proposed an approximation to improve efficiency in building the histogram intersection kernel, but efficiency can be attained merely by using pre-computed auxiliary tables which are considered as a type of pre-trained nonlinear support vector machines (SVM). To give SPM the linearity needed to deal with large datasets, Yang [19] proposed a linear SPM with spare coding (ScSPM), in which a linear kernel is chosen instead of a nonlinear kernel due to the more linearly separable property of sparse features. Wang & Wang [18] proposed a multiple scale learning (MSL) framework in which multiple kernel learning (MKL) is employed to learn the optimal weights instead of using predefined weights of SPM.

Our proposed kernel concentrates on improvement of the similarity measurement between sub-regions by using a coarse-to-fine model instead of the BoF model used in SPM. We consider the sub-regions on a sequence of different resolutions as the pyramid matching kernel (PMK) [4]. Futhermore, instead of using the pre-defined weight vector for basic intersection kernels to penalize across different resolutions, we reformulate the problem into a uniform MKL to obtain it more effectively. In addition, our proposed kernel can deal with different cardinalities of sets of unordered features by applying the square root diagonal normalization [17] for each intersection kernel, which is not considered in PMK.

3 Hierarchical Spatial Matching Kernel

In this section, we first describe the original formulation of SPM and then introduce our proposed HSMK, which uses a coarse-to-fine model as a basic for improving SPM.

3.1 Spatial Pyramid Matching

Each image is represented by a set of vectors in the D-dimensional feature space. Features are quantized into discrete types called visual words by using K-means clustering or sparse coding. The matching between features turns into a comparison between discrete corresponding types. This means that they are matched if they are in the same type and unmatched otherwise.

SPM constructs a sequence of different scales with $l = 0, 1, 2, ..., L$ on an image. In each scale, it partitions the image into $2^l \times 2^l$ sub-regions and applies the BoF model to measure the similarity between sub-regions. Let X and Y be two sets of vectors in the D-dimensional feature space. The similarity between two sets at scale l is the sum of the similarity between all corresponding sub-regions:

$$\mathcal{K}_l(X, Y) = \sum_{i=1}^{2^{2l}} \mathcal{I}(X_i^l, Y_i^l), \tag{1}$$

where X_i^l is the set of feature descriptors in the i^{th} sub-region at scale l of the image vector set X. The intersection kernel $\mathcal{I}$ between X_i^l and Y_i^l is formulated as:

$$\mathcal{I}(X_i^l, Y_i^l) = \sum_{j=1}^{V} \min(\mathcal{H}_{X_i^l}(j), \mathcal{H}_{Y_i^l}(j)), \tag{2}$$

where V is the total number of visual words and $\mathcal{H}_\alpha(j)$ is the number of occurences of the j^{th} visual word which is obtained by quantizing feature descriptors in the set α. Finally, the SPM kernel (SPMK) is the sum of weighted similarity over the scale sequence:

$$\mathcal{K}(X,Y) = \frac{1}{2^L}\mathcal{K}_0(X,Y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{K}_l(X,Y). \tag{3}$$

The weight $\frac{1}{2^{L-l+1}}$ associated with scale l is inversely proportional to the sub-region width at that scale. This weight is utilized to penalize the matching since it is easier to find the matches in the larger regions. We remark that all the matches found at scale l are also included in a finer scale $l - \zeta$ with $\zeta > 0$.

3.2 The Proposed Kernel: Hierarchical Spatial Matching Kernel

To improve efficiency in achieving the similarity measurement between sub-regions, we utilize a coarse-to-fine model on sub-regions by mapping them into a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2, ..., R$ as in [4].

X_i^l and Y_i^l are the sets of feature descriptors in the i^{th} sub-regions at scale l of image vector sets X, Y respectively. At each resolution r, we apply the normalized intersection kernel $\mathscr{F}^r$ using the square root diagonal normalization method to measure the similarity as follows:

$$\mathscr{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathcal{I}(X_i^l(r), X_i^l(r))\mathcal{I}(Y_i^l(r), Y_i^l(r))}}, \tag{4}$$

where $X_i^l(r)$, $Y_i^l(r)$ are the sets X_i^l, Y_i^l at the resolution r respectively. Note that the histogram intersection between X and itself is equivalent with its cardinality. Thus, letting $\mathscr{N}_{X_i^l(r)}$ and $\mathscr{N}_{Y_i^l(r)}$ be the cardinality of sets $X_i^l(r)$ and $Y_i^l(r)$, the equation (4) is rewritten as:

$$\mathscr{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathscr{N}_{X_i^l(r)}\mathscr{N}_{Y_i^l(r)}}}. \tag{5}$$

The square root diagonal normalization of the intersection kernel not only satisfies Mercer's conditions [17], but also penalizes the difference in cardinality between sets as in equation (5).

To obtain the synthetic similarity measurement of the coarse-to-fine model, we define the linear combination over a sequence of local kernels, each term of which is calculated using equation (5) at each resolution. Accordingly, the kernel function $\mathscr{F}$ between two sets X_i^l and Y_i^l in the coarse-to-fine model is formulated as:

$$\begin{aligned} \mathscr{F}(X_i^l, Y_i^l) &= \sum_{r=0}^{R} \theta_r \mathscr{F}^r(X_i^l, Y_i^l) \\ \text{where} \quad \sum_{r=0}^{R} \theta_r &= 1, \theta_r \geq 0, \forall r = 0, 1, 2, ..., R. \end{aligned} \tag{6}$$

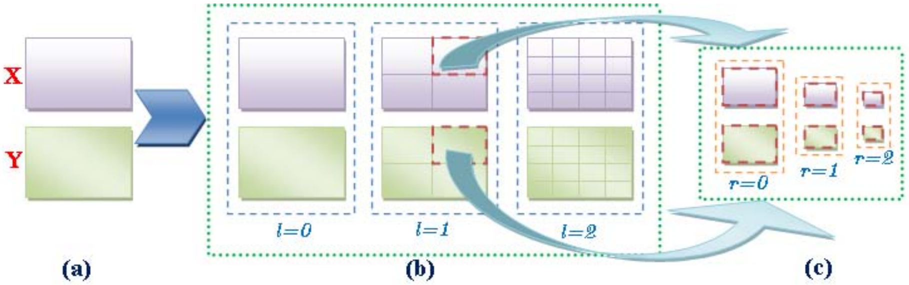

Fig. 1. An illustration for HSMK applied to images X and Y with $L = 2$ and $R = 2$ (a). HSMK first partitions the images into $2^l \times 2^l$ sub-regions with $l = 0, 1, 2$ as SPMK (b). However, HSMK applies the coarse-to-fine model for each sub-region by considering it on a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2$ (c). Equation (8) with the weight vector achieved from the uniform MKL is applied to obtain better optimal matching approximation between sub-regions instead of using the BoW model as in SPMK.

Moreover, when the linear combination of local kernels is integrated with SVM, it can be reformulated as a MKL problem where basic local kernels are defined as equation (5) across the resolutions of the sub-region as:

$$\begin{aligned} \min_{\boldsymbol{w_\alpha}, w_0, \boldsymbol{\xi}, \boldsymbol{\theta}} \quad & \frac{1}{2}\Big(\sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha \|\boldsymbol{w_\alpha}\|_2\Big)^2 + \mathcal{C}\sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i\Big(\sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha \langle \boldsymbol{w_\alpha}, \Phi_\alpha(\boldsymbol{x_i})\rangle + w_0\Big) \geq 1 - \xi_i \\ & \textstyle\sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha = 1, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\xi} \geq \mathbf{0}, \end{aligned} \tag{7}$$

where $\boldsymbol{x_i}$ is an image sample, y_i is the category label for $\boldsymbol{x_i}$, N is the number of training samples, ($\boldsymbol{w_\alpha}$, w_0, $\boldsymbol{\xi}$) are parameters of SVM, $\mathcal{C}$ is a soft margin parameter defined by users to penalize training errors in SVM, $\boldsymbol{\theta}$ is a weight vector for basic local kernels, $\mathfrak{N}$ is the number of the basic local kernels of the sub-region over the sequence of resolutions, $\boldsymbol{\theta} \geq \mathbf{0}$ means that any entry of vector $\boldsymbol{\theta}$ is nonnegative, $\Phi(\boldsymbol{x})$ is the function that maps the vector $\boldsymbol{x}$ into the reproducing Hilbert space and $< \cdot, \cdot >$ denotes the inner product. MKL solves the parameters of SVM and the weight vector for basic local kernels simultaneously.

These basic local kernels are analogously defined across resolutions of the sub-region. Therefore, the redundant information between them is high. The experiments in Gehler and Nowozin [3] and especially Kloft *et al.* [7] have shown that the uniform MKL, which is an approximation of MKL into traditional nonlinear kernel SVM, is the most efficient for this case in terms of both performance and complexity. Thus, formula (6) with linear combination coefficients obtained from the uniform MKL method becomes:

$$\mathscr{F}(X_i^l, Y_i^l) = \frac{1}{R+1}\sum_{r=0}^{R} \mathscr{F}^r(X_i^l, Y_i^l). \tag{8}$$

Figure 1 illustrates an application of HSMK with $L = 2$ and $R = 2$. HSMK also maps the sub-regions into a sequence of different resolutions for PMK to obtain better measurement of similarity between them. However, the weight vector is achieved from the uniform MKL. Thus, it is more efficient and theorical than predefined one in PMK. Furthermore, applying the square root diagonal normalization allows it to robustly deal with differences in cardinality that are not considered in PMK. HSMK is formulated based on SPM in the coarse-to-fine model, which is efficient with sets of unordered feature descriptors, even in the presence of differences in cardinality. Mathematically, the formulation of HSMK is as follows:

$$\begin{aligned} \mathcal{K}(X,Y) &= \frac{1}{2^L}\mathscr{F}_0(X,Y) + \sum_{l=1}^{L}\frac{1}{2^{L-l+1}}\mathscr{F}_l(X,Y) \\ \textit{with} \quad \mathscr{F}_l(X,Y) &= \sum_{i=1}^{2^{2l}}\mathscr{F}(X_i^l, Y_i^l) = \frac{1}{R+1}\sum_{i=1}^{2^{2l}}\sum_{r=0}^{R}\mathscr{F}^r(X_i^l, Y_i^l). \end{aligned} \tag{9}$$

Briefly, HSMK utilizes the kd-tree algorithm to map each feature descriptor into a discrete visual word, and then the normalized intersection kernel by the square root diagonal method is applied to the histogram of V bins to measure the similarity. We have $\mathscr{N}$ feature descriptors in the D-dimension space, and the kd-tree algorithm costs $O(\log V)$ steps to map feature descriptors. Therefore, the complexity of HSMK is $O(DM \log V)$ with $M = \max(\mathscr{N}_X, \mathscr{N}_Y)$. We note that the complexity of the optimal matching kernel [8] is $O(DM^3)$.

4 Experimental Results

Most recent approaches use local invariant features as an effective means of representating images, because they can well describe and match instances of objects or scenes under a wide variety of viewpoints, illuminations, or even background clutter. Among them, SIFT [10] is one of the most robust and efficient features. To achieve better discriminative ability, we utilize the dense SIFT by operating a SIFT descriptor of 16×16 patches computed over each pixel of an image instead of key points [10] or a grid of points [9]. In addition, to improve robustness, we convert images into gray scale ones before computing the dense SIFT. Dense features have the capability of capturing uniform regions such as sky, water or grass where key points usually do not exist. Moreover, the combination of dense features and the coarse-to-fine model allows images to be represented more exactly since feature descriptors achieves more neighbor information across many levels in resolution. We performed unsupervised K-means clustering on a random subset of SIFT descriptors to build visual words. Typically, we used two different dictionary sizes M in our experiment: $M = 400$ and $M = 800$.

We conducted experiments for two types of image categorization: object categorization and scene categorization. For object categorization, we used the Oxford Flower dataset [14]. To show the efficiency and scalability of our proposed kernel, we also used the large scale object datasets such as CALTECH-101 [2] and CALTECH-256 [5]. For scene categorization, we evaluated the proposed kernel on the MIT scene [16] and UIUC scene [9] datasets.

4.1 Object Categorization

Oxford Flowers dataset: This dataset contains 17 classes of common flowers in the United Kingdom, collected by Nilsback *et al.* [14]. Each class has 80 images with large scale, pose and light variations. Moreover, intra-class flowers such as irises, fritillaries and pansies are also widely diverse in their colors and shapes. There are some cases of close similarity between flowers of different classes such as that between dandelion and Colts'Foot. In our experiments, we followed the set-up of Gehler and Nowozin [3], randomly choosing 40 samples from each class for training and using the rest for testing. Note that we did not use a validation set as in [14,15] for choosing the optimal parameters. Table 1 shows that our proposed kernel achieved a state-of-the-art results when using a single feature. It outperformed not only SIFT-Internal [15], the best feature for this dataset computed on a segmented image, but also the same feature on SPMK with the optimal weights by MSL [18]. In addition, Table 2 shows that the performance of HSMK also outperformed that of SPMK.

Table 1. Classification rate (%) with a single feature comparision on Oxford Flower dataset (with NN that denotes the nearest neighbour algorithm)

Method	Accuracy (%)
HSV (NN) [15]	43.0
SIFT-Internal (NN) [15]	55.1
SIFT-Boundary (NN) [15]	32.0
HOG (NN) [15]	49.6
HSV (SVM) [3]	61.3
SIFT-Internal (SVM) [3]	70.6
SIFT-Boundary (SVM) [3]	59.4
HOG (SVM) [3]	58.5
SIFT (MSL) [18]	65.3
Dense SIFT (HSMK)	**72.9**

Table 2. Classification rate (%) comparision between SPMK and HSMK on Oxford Flower dataset

Kernel	$M = 400$	$M = 800$
SPMK	68.09%	69.12%
HSMK	**71.76%**	**72.94%**

Caltech datasets: To show the efficiency and robustness of HSMK, we also evaluated its performance on large scale object datasets, i.e., the CALTECH-101 and CALTECH-256 datasets. These datasets feature high intra-class variability,

Table 3. Classification rate (%) comparision on CALTECH-101 dataset

	5 training	10 training	15 training	20 training	25 training	30 training
Grauman & Darrell [4]	34.8%	44%	50.0%	53.5%	55.5%	58.2%
Wang *et al.* [18]	-	-	61.4%	-	-	-
Lazebnik *et al.* [9]	-	-	56.4%	-	-	64.6%
Yang *et al.* [19]	-	-	67.0%	-	-	73.2%
Boimann *et al.* [1]	56.9%	-	72.8%	-	-	79.1%
Gehler & Nowozin (MKL) [3]	42.1%	55.1%	62.3%	67.1%	70.5%	73.7%
Gehler & Nowozin (LP-β) [3]	54.2%	65.0%	70.4%	73.6%	75.7%	77.8%
Gehler & Nowozin (LP-B) [3]	46.5%	59.7%	66.7%	71.1%	73.8%	77.2%
Our method (HSMK)	50.5%	62.2%	69.0%	72.3%	74.4%	77.3%

Table 4. Classification rate (%) comparision between SPMK and HSMK on CALTECH-101 dataset

	5 training	10 training	15 training	20 training	25 training	30 training
SPMK ($M = 400$)	48.18%	58.86%	65.34%	69.35%	71.95%	73.46%
HSMK(M=400)	**50.68%**	**61.97%**	**67.91%**	**71.35%**	**73.92%**	**75.59%**
SPMK ($M = 800$)	48.11%	59.70%	66.84%	69.98%	72.62%	75.13%
HSMK(M=800)	**50.48%**	**62.17%**	**68.95%**	**72.32%**	**74.36%**	**77.33%**

poses, and viewpoints. On CALTECH-101, we carried out experiments with 5, 10, 15, 20, 25, and 30 training samples for each class, including the background class, and used up to 50 samples per class for testing. Table 3 compares the classification rate results of our approach with other ones. As shown, our approach obtained the comparable result with that of state-of-the-art approaches even using only a single feature while others used many types of features and complex learning algorithms such as MKL and linear programing boosting (LP-B) [3]. Table 4 shows that the result of HSMK outperformed that of SPMK in this case as well. It should be noted that when the experiment was conducted without the background class, our approach achieved a classification rate of 78.4% for 30 training samples. This shows that our approach is efficient in spite of its simplicity.

On CALTECH-256, we performed experiments with HSMK using 15 and 30 training samples per class, including the clutter class, and 25 samples of each class for testing. We also re-implemented SPMK [5] but used our dense SIFT to enable a fair comparation of SPMK and HSMK. As shown in Table 5, the HSMK classification rate was about 3 percent higher than that of SPMK.

4.2 Scene Categorization

We also performed experiments using HSMK on the MIT Scene (8 classes) and UIUC Scene (15 classes) dataset. In these datasets, we set $M = 400$ as the

Table 5. Classification rate (%) comparision on CALTECH-256 dataset

Kernel	15 training	30 training
Griffin *et al.* (SPMK) [5]	28.4%	34.2%
Yang *et al.* (ScSPM) [19]	27.7%	34.0%
Gehler & Nowozin (MKL) [3]	30.6%	35.6%
SPMK	25.3%	31.3%
Our method (HSMK)	27.2%	34.1%

dictionary size. On the MIT Scene dataset, we randomly chose 100 samples per class for training and 100 other samples per class for testing. As shown in Table 6, the classification rate for HSMK was 2.5 percent higher than that of SPMK. Our approach also outperformed other local feature approaches [6] as well as local feature combinations [6] by more than 10 percent, and was better than the global feature GIST [16], an efficient feature in scene categorization.

Table 6. Classification rate (%) comparision on MIT Scene (8 classes) dataset

Method	Accuracy (%)
GIST [16]	83.7
Local features [6]	77.2
Dense SIFT (SPMK)	85.8
Dense SIFT (HSMK)	**88.3**

On the UIUC Scene dataset, we followed the experiment setup described in [9]. We randomly chose 100 training samples per class and the rest were used for testing. As shown in Table 7, the result of our proposed kernel also outperformed that of SPMK [9] as well as SPM based on sparse coding [19] for this dataset.

Table 7. Classification rate (%) comparision on UIUC Scene (15 classes) dataset

Method	Accuracy (%)
Lazebnik *et al.* (SPMK) [9]	81.4
Yang *et al.* (ScSPM) [19]	80.3
SPMK	79.9
Our method (HSMK)	**82.2**

5 Conclusion

In this paper, we proposed an efficient and robust kernel that we call the hierarchical spatial matching kernel (HSMK). It uses a coarse-to-fine model for sub-regions to improve spatial pyramid matching kernel (SPMK) and thus obtains more neighbor information through a sequence of different resolutions. In addition, the kernel efficiently and robustly handles sets of unordered features as SPMK and pyramid matching kernel as well as sets having different cardinalities.

Combining the proposed kernel with a dense feature approach was found to be sufficiently effective and efficient. It enabled us to obtain at least comparable results with those by existing methods for many kinds of datasets. Moreover, our approach is simple since it is based on only a single feature with nonlinear support vector machines, in constrast to other more complicated recent approaches based on multiple kernel learning or feature combinations.

In most well-known datasets of object and scene categorization, the proposed kernel was also found to outperform SPMK which is an important component such as a basic kernel in multiple kernel learning. This means that we can replace SPMK with HSMK to improve the performance of frameworks based on basic kernels.

Acknowledgements

This work was performed under the National Institute of Informatics international internship program and also the framework of Memorandum of Understanding between the Vietnam National University of Ho Chi Minh City and the National Institute of Informatics, Japan. This work was in part supported by JST, CREST.

References

1. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR (2008)
2. Fei-Fei, L., Fergus, R., Perona, P.: 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. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV, pp. 221–228 (2009)
4. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: ICCV, vol. 2, pp. 1458–1465 (2005)
5. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
6. Johnson, M.: Semantic Segmentation and Image Search. Ph.D. thesis, University of Cambridge (2008)
7. Kloft, M., Brefeld, U., Laskov, P., Sonnenburg, S.: Non-sparse multiple kernel learning. In: NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008)
8. Kondor, R.I., Jebara, T.: A kernel between sets of vectors. In: ICML, pp. 361–368 (2003)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
11. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML, pp. 689–696 (2009)

12. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR, pp. 1–8 (2008)
13. Moosmann, F., Triggs, B., Jurie, F.: Randomized clustering forests for building fast and discriminative visual vocabularies. In: NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008)
14. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR, vol. 2, pp. 1447–1454 (2006)
15. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42, 145–175 (2001)
17. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
18. Wang, S.C., Wang, Y.C.F.: A multi-scale learning framework for visual categorization. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 310–322. Springer, Heidelberg (2011)
19. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801 (2009)
20. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying discriminative visual codebook generation with classifier training for object category recognition. In: CVPR, Los Alamitos, CA, USA, pp. 1–8 (2008)