# Image Classification Using Spatial Pyramid Coding and Visual Word Reweighting

Chunjie Zhang[1], Jing Liu[1], Jinqiao Wang[1], Qi Tian[2], Changsheng Xu[1], Hanqing Lu[1], and Songde Ma[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
{cjzhang,jliu,jqwang,csxu,luhq}@nlpr.ia.ac.cn, masd@most.cn
[2] University of Texas at San Antonio, One UTSA Circle San Antonio Texas, 78249-USA
qitian@cs.utsa.edu

**Abstract.** The ignorance on spatial information and semantics of visual words becomes main obstacles in the bag-of-visual-words (BoW) method for image classification. To address the obstacles, we present an improved BoW representation using spatial pyramid coding (SPC) and visual word reweighting. In SPC procedure, we adopt the sparse coding technique to encode visual features with the spatial constraint. Visual features from the same spatial sub-region of images are collected to generate the visual vocabulary. Additionally, a relaxed but simple solution for semantic embedding into visual words is proposed. We relax the semantic embedding from ideal semantic correspondence to naive semantic purity of visual words, and reweight each visual word according to its semantic purity. Higher weights are given to semantically distinctive visual words, and lower weights to semantically general ones. Experiments on a public dataset demonstrate the effectiveness of the proposed method.

**Keywords:** spatial pyramid coding, bag-of-visual-words (BoW), reweighting, image classification.

## 1 Introduction

In recent years, the bag-of-visual-words (BoW) model becomes popular in image classification. This model extracts appearance descriptors from local patches and quantizes them into discrete "visual words", and then a compact histogram representation is used to represent images. The descriptive power of the BoW model is severely limited because it discards the spatial information of local descriptors. To overcome this problem, one popular extension method, called the *spatial pyramid matching* (SPM) by Lazebnik *et al* [1], has been shown to be effective for image classification. The SPM partitions an image into several segments in different scales, then computes the BoW histogram within each segment and concatenates all the histograms to form a high dimension vector representation of the image.

R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part III, LNCS 6494, pp. 239–249, 2011.

To obtain good performances, researchers have empirically found that the SPM should be used together with SVM classifier using nonlinear Mercer kernels, e.g. *Chi-square kernel* or *intersection kernel*. However, the computational complexity is $O(n^3)$ and the memory complexity is $O(n^2)$ in the training phase, where $n$ is the size of training dataset. This constrains the scalability of the SPM-based nonlinear SVM method. To reduce the training complexity, a linear spatial pyramid matching method using sparse coding (ScSPM) is proposed by Yang *et al* [2]. This method is more robust to local spatial translations and is biological plausible [3]. Inspired by this, Wang et al [4] used locality in feature space to constrain the linear sparse coding phase (LLC) of ScSPM which further reduced the computation time. However, the performance improvement of LLC over ScSPM on real world images is not obvious. In fact, there is another constraint which was neglected in [4], i.e., the spatial locality constraint. For example, 'sky' often lies on the upper side of images, while 'beach' often lies on the lower side of images. When we try to encode an image region about the upper 'sky', it is more meaningful to use the bases which are generated by the local features on the upper side of images. Similarly, it is more meaningful to encode the lower 'beach' with the bases generated from the local features on the lower side of images.

Besides, the semantic meaning of visual word has not been considered too much in literature, which has become another obstacle to affect the performance of the BoW model. Ideally, the correspondence between visual words and semantics, namely the semantic embedding into the BoW representation, will bring the more representative and discriminative description for image classification than solely on visual features. However, the well-known semantic gap becomes a natural barrier to achieve such correspondence. Some recent work appeal to various supervised learning approaches [5, 6] to learn discriminative visual vocabulary. In fact, such supervised refinement emphasizes on the discriminative abilities of visual words rather than truly embedding semantics into image representation. We believe that the semantic embedding can further enhance the discriminative ability of visual words in image classification, but not vice versa. Consequently, it is necessary to find a suitable way to obtain such a semantic embedded BoW presentation for image classification.

In this paper, we present a novel image classification method by using spatial pyramid coding (SPC) along with visual word reweighting, as shown in Figure 1. We first partition images into sub-regions on multiple scales, and adopt the sparse coding approach to encode visual features of images with the spatial constraint. Different from SPM [1], the SPC-based visual vocabulary is concatenated with each encoding results from the sub-regions which have the same spatial locality and segmentation scale. For the semantic embedding, we adopt a relaxed but simple solution to reweight the SPC-based BoW representation according to the semantic purity of each visual word, instead of the obtainment of the semantic correspondence. Specifically, we give higher weights to semantically distinctive visual words, and lower weights to semantically general visual words.

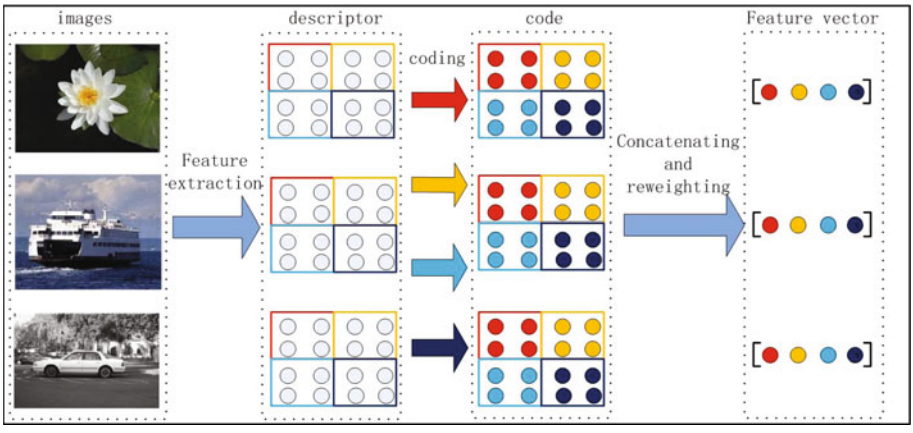


**Fig. 1.** Flowchart of the proposed spatial pyramid codebook (with two scales) and visual word reweighting methods. It is best viewed in color

Comprehensive experimental evaluations on the Scene-15 dataset demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 gives an overview of some related work. In Section 3, we present the details of the proposed spatial pyramid coding and visual word reweighting method. Experimental results and comprehensive analysis are given in Section 4. Finally, the conclusions and future research issues are discussed in Section 5.

## 2 Related Work

The bag-of-visual-words model (BoW) has been widely used due to its simplicity and good performance. Many works has been done to improve the performance of the traditional bag-of-visual-words model over the past few years. Some literatures devoted to learn discriminative visual vocabulary for object recognition [7-9]. Perronnin *et al* [7] used the Gaussian Mixture Model (GMM) to perform clustering. To alleviate the drawback of k-means clustering, Jurie and Triggs [8] tried to use a scalable acceptance-radius based clustering method instead. Moosmann *et al* [9] used random forests to construct codebook which helps to improve the classification performance. Others tried to model the co-occurrence of visual words in a generative framework [10-13]. Boiman *et al* [10] tried to classify images by nearest-neighbor classification. Bosch *et al* [11] tried to classify scene images using a hybrid generative/discriminative approach. Besides, many researchers also [1, 14-19] tried to learn more discriminative classifiers by combining the spatial and contextual information of visual words. Oliva and Torralba [15] modeled the shape of the scene by using a holistic representation. Gemert *et al* [16] proposed to learn visual word ambiguity through soft assignment. Zhang *et al* [17] utilized nearest neighbor classification for visual category recognition. Motivated by Grauman and Darrell's [19] pyramid matching in feature space,

Lazebnik *et al* [1] proposed the spatial pyramid matching (SPM) which has been proven efficient for image classification.

Although the SPM method works well for image classification, it has to be used along with nonlinear Mercer kernels for good performance. However, the computational cost is $O(n^3)$ in training phase. To improve the scalability, Yang *et al* [2] proposed a linear spatial pyramid matching method using sparse coding along with max pooling to classify images, which has been shown very effective and efficient. The approach relaxes the restrictive cardinality constraint of vector quantization in traditional BoW model and uses max spatial pooling to compute histogram which reduces the training complexity to $O(n)$. Motivated by this, many researchers [4, 20-21] proposed novel methods to further improve the performance. Wang *et al* [4] proposed to use locality constraints in feature space during the sparse coding phase of [2] and the theoretical justifications are given by Yu *et al* [20]. Boureau *et al* [21] also proposed a novel method to learn a supervised discriminative dictionary for sparse coding.

Obviously, not all of the visual words are equally useful for image classification. [22-23] showed that the human visual system employs an effective attention mechanism and can recognize different object categories robustly by focusing on the interesting parts in an image. To choose the most discriminative visual features, Liu *et al* [24] tried to select the most discriminative visual word combination with Adaboost while Mutch and Lowe [25] used sparse, localized features for multiclass object recognition. Cai *et al* [26] also tried to learn weights for each visual word by solving a quadratic programming problem.

## 3 Spatial Pyramid Coding and Visual Word Reweighting

This section gives the details of the proposed spatial pyramid coding (SPC) and visual word reweighting method. For each image, we first densely extract local image features and then utilize the spatial pyramid principle to encode local features. Then we concatenate the BoW representation of different segments and reweight each visual word based on its semantic purity. Figure 1 shows the flowchart of the proposed spatial pyramid coding and visual word reweighting method.

### 3.1 Spatial Pyramid Coding

The idea of using spatial pyramid along with the BoW representation of images has been proven very effective for image classification by many researchers. This method partitions an image into increasingly finer spatial sub-regions and computes the histogram of local features from every sub-region [1]. Usually, $2^l \times 2^l$ subregions, with $l = 0, 1, 2$ are used. Other partition method such as $3 \times 1$ is also used to incorporate top and bottom relationships, which has been proven very useful on the PASCAL VOC Challenge. Take the $2^l \times 2^l$ for example, for $L$ levels and $M$ channels, the resulting concatenated vector for each image has a dimensionality of $M \sum_{l=0}^{L} = M\frac{1}{3}(4^{L+1} - 1)$.

To preserve the discriminative power of local image features as much as possible, researchers have tried many coding methods, among which the most popular is the k-means model. Formally, let $X$ be a set of $D$-dimensional local features. The number of local features is $N$, i.e. $X = [x_1, x_2, ..., x_N] \in R^{D \times N}$ where $x_i \in R^{D \times 1}$. Suppose we have a codebook $B$ with $M$ visual words, where $B = [b_1, b_2, ..., b_M] \in R^{D \times M}$. To convert each descriptor into a $M$-dimensional vector to represent images, $k$-means based vector quantization (VQ) method tries to solve a constrained least square fitting problem as:

$$C = argmin \sum_{i=1}^{N} = \| x_i - B \times c_i \|^2 \tag{1}$$

$$s.t. \| c_i \|_{l^0} = 1, \| c_i \|_{l^1} = 1, c_{ij} \geq 0, \forall i, j$$

where $C = [c_1, c_2, ..., c_N]$ is the codes for $X$ and $c_{ij}$ is the $j$-th element of $c_i$.

The constraints in the k-means model are very restrictive with only one element of $c_i$ is set to 1. In practice, this is often achieved by nearest neighbor search. To alleviate the discriminative power loss during vector quantization, Yang *et al* [2] proposed to use sparse coding instead. They relaxed the restrictive cardinality constraint in Eq. (1) by using a sparsity regularization term instead.$l^1$ norm of $c_i$ is used. Thus, Eq. (1) becomes a standard sparse coding problem [27] as:

$$C = argmin \sum_{i=1}^{N} = \| x_i - B \times c_i \|^2 + \lambda \| c_i \|_1 \tag{2}$$

where $\lambda$ is the regularization parameter and $\| \cdot \|_1$ is the $l^1$ norm which sums the absolute value of each element. This can be solved by optimizing over each individually.

However, as introduced in [4], locality is more essential than sparsity because locality leads to sparsity but not necessary vice versa. It allows sparse reconstruction of features in the appearance space using sparsity along with locality constraints. However, this discards the spatial information in the coding phase. This paper proposes an "orthogonal" approach: we perform pyramid coding in the two-dimensional image space and use sparse coding method [1, 27] in feature space. Specifically, we first partition the image into increasingly finer spatial subregions with $2^l \times 2^l$, $l = 0, 1, 2$. For each sub-region, the sparse coding parameters and the codebook are then jointly learned using the local image features within this sub-region. This is achieved by alternatively optimizing over the codebook $B$ and the coding parameters $C$ while keeping the other fixed. We use the alternative optimization method as did in [1, 27] to solve this problem. In our experiments, about 45,000 SIFT descriptors extracted from random patches of each segment are used to train the codebooks. Once we have learned the codebook for each sub-region, we are able to code efficiently for each local feature using Eq. (2). Max pooling [1] is then used to generate the BoW representation for each segment which has been shown very effective when combined with sparse

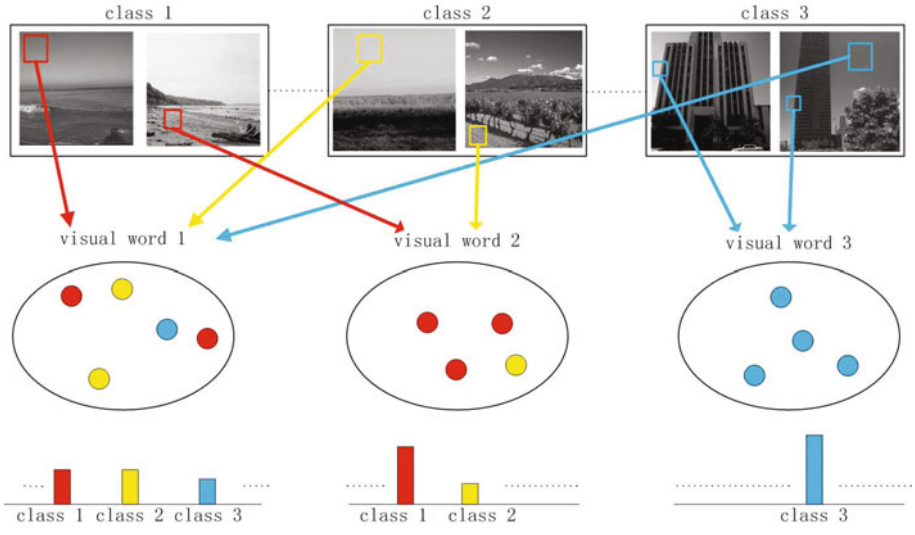


**Fig. 2.** Toy example showing the semantic meaning of visual words. Different colors represent local features extracted from different classes. Since visual word 3 is the most semantically distinctive, we believe the word is more discriminative than visual word 1 and 2 in a specific classification task. It is best viewed in color.

coding. Finally, the BoW representations of all segments are concatenated into a long vector to represent images.

### 3.2 Visual Word Reweighting

Although the bag-of-visual-words model is inspired by the bag-of-words approach to text categorization, the semantic meaning of visual word has not been considered too much in literature. We believe the semantic information of visual words can also be utilized to improve the image classification performance.

During the vector quantization of traditional BoW model or the sparse coding process, many local features are assigned to one visual word. These local features may come from different classes of images hence have different semantic meanings. Assuming each local image feature having the same semantic label as the image from which it is extracted, we can use the frequency distribution of classes of local features assigned to each visual word to represent this visual word. Formally, let $Q = [q_1, q_2, ..., q_M] \in R^{K \times M}$ is the semantic distribution of all the visual words, where $q_i \in R^{K \times 1}$ and $K$ is the number of classes. We believe that the purity of each visual word is correlated with its discriminative power. For example, sky often exists on the outdoor scene images. While classifying outdoor images of different classes, visual words representing the upper sky are often generated by local features extracted from different classes of images. These visual words are noisy for classification and should be given lower weights. On the contrary, if one visual word is generated mainly by the local features of

the same class, the discriminative power of this visual word is much stronger than visual words which are generated by local features from diverse classes of images. Figure 2 shows a toy example reflecting showing the semantic purity of visual words.

To measure the semantic purity of each visual word quantitatively, we choose to use the entropy of each visual word's semantic distribution, because it has been proven very effective and efficient to implement. The larger the entropy, the less pure the visual word and vice versa. Formally, let $e_i$ to represent the entropy of visual word $b_i$ whose semantic distribution is $q_i$. $e_i$ can then be calculated as:

$$e_i = -\sum_{k=1}^{K} q_{ik} ln(q_{ik}) \tag{3}$$

Let $w_i$ to represent the weight of visual word $i$,$i \in 1, 2, ..., M$. The weight of each visual word can then be computed as:

$$w_i = exp(-e_i/\alpha) \tag{4}$$

where $\alpha$ is the scaling parameter. In our experiments, we simply set $\alpha$ to 1. The weight of each visual word can then be computed in an efficient way as:

$$w_i = \prod_{k=1}^{K} q_{ik}^{q_{ik}} \tag{5}$$

## 4 Experiments

We evaluate the proposed spatial pyramid coding and visual word reweighting method on the fifteen natural scene dataset by provided Lazebnik *et al* [1]. The fifteen scene dataset composes 4,485 images, which vary from natural scenes like forests and mountains to man-made environments like offices and kitchens. Thirteen were provided by Fei-Fei and Perona [12] (eight of these were originally provided by Oliva and Torralba [15]) and two were collected by Lazebnik *et al* [1]. We perform all processing in grayscale of images even when sometimes the color images are provided. As to the feature extraction, we follow Lazebnik *et al* [1] and densely compute SIFT descriptors on overlapping 16×16 pixels with an overlap of 8 pixels. The codebook size is set to 1,024, as Yang *et al* [2] did. Multi-class classification is done via the one-versus-all rule: a SVM classifier is learned to separate each class from the rest and a test image is assigned the label of the classifier with the highest response. The average of per-class classification rates is used to quantitatively measure the performance.

We show some example images of the Scene-15 dataset in Figure 3. The major picture sources in this dataset include the COREL collection, personal photographs and Google image search. Each category has 200 to 400 images, and the average image size is 300×250 pixels. We follow the same experiment procedure of Lazebnik *et al* [2] and randomly choose 100 images per category as


**Fig. 3.** Example images of the Scene-15 dataset

**Table 1.** Classification rate comparison on the Scene-15 dataset. Numerical values in the table stand for mean and standard derivation.

| Algorithms | Classification Rate |
|---|---|
| KSPM[2] | $76.73 \pm 0.65$ |
| KC[16] | $76.67 \pm 0.39$ |
| ScSPM[2] | $80.28 \pm 0.93$ |
| ScSPM | $78.77 \pm 0.50$ |
| SPC | $81.14 \pm 0.46$ |
| SPC+Reweighting | $\mathbf{82.98 \pm 0.23}$ |

the training set and use the remaining images as the test set. This process is repeated for five times.

Table 1 gives the detailed comparison results. We compare the proposed methods with the kernel codebook proposed by Gemert *et al* [16], the ScSPM and the reimplementation of nonlinear kernel SPM by Yang *et al* [2]. Our implementation of ScSPM is not able to reproduce the results reported by Yang *et al* [2] probably due to the feature extraction process and normalization process. We can see from the results that the proposed SPC outperforms ScSPM, which shows the effectiveness of combining spatial information in the coding phase. Besides, the classification rate can be further improved by reweighting each visual word based on its semantic purity. This demonstrates the effectiveness of the proposed method.

**Table 2.** Classification rate per concept for the ScSPM, SPC and SPC+Reweighting

| Class | ScSPM | SPC | SPC+Reweighting |
|---|---|---|---|
| Bedroom | 67.24± 5.57 | 83.62± 1.16 | **84.48± 1.28** |
| CALsuburb | 99.29± 1.42 | 99.29± 0.95 | 99.29± 1.00 |
| Industrial | 56.40± 2.00 | 57.35± 2.67 | **57.82± 3.22** |
| Kitchen | 66.36± 3.44 | 65.45± 2.54 | **69.09± 4.96** |
| Livingroom | 62.43± 2.92 | 64.02± 2.55 | **65.61± 3.42** |
| MITcoast | 97.69± 1.51 | 96.15± 0.61 | **98.08± 1.87** |
| MITforest | 97.81± 0.91 | **99.12± 1.30** | 97.37± 1.00 |
| MIThighway | 86.25± 2.67 | 88.12± 4.34 | **88.12± 3.71** |
| MITinsidecity | 88.94± 1.16 | 88.94± 1.43 | **89.90± 1.50** |
| MITmountain | 84.67± 2.70 | **86.50± 2.96** | 85.77± 2.83 |
| MITopencountry | 74.19± 3.33 | 79.03± 4.55 | **100± 0.00** |
| MITstreet | 91.15± 2.29 | **94.79± 3.31** | 92.71± 3.01 |
| MITtallbuilding | 97.27± 0.35 | 98.05± 0.33 | **99.22± 0.28** |
| PARoffice | 86.96± 2.25 | **87.83± 2.84** | 83.48± 0.78 |
| Store | 69.77± 2.70 | 73.03± 3.50 | **73.95± 3.59** |

To analyze the detailed classification performance, we give the classification rate per concept in table 2. Generally, four conclusions can be made. First, we can have similar observation as [1] did that the indoor classes (e.g. kitchen, livingroom) are more difficult to classify than the outdoor classes (e.g. MITopencountry, MITtallbuilding). Second, the advantages of SPC over ScSPM mainly focus on indoor classes, e.g. bedroom, livingroom and store. This is because the SPC method is able to combine the spatial information into the coding process; hence helps make correct categorization of images. Third, the improvement of SPC+Reweighting over SPC mainly lies on outdoor classes, this is because images of the outdoor classes (e.g. "MITopencountry") are relative simple and with less objects compared with images of indoor classes. We believe this is the reason why the reweighting works. Finally, the proposed SPC and SPC+Reweighting methods outperform ScSPM for all the fifteen classes.

## 5 Conclusion

This paper proposes a novel method for image classification using spatial pyramid coding (SPC) and visual word reweighting. SPC is easy to compute and can incorporate spatial information in the coding phase which is lost in the sparse coding spatial pyramid matching (ScSPM). SPC applies spatial constraint in the coding phase for each sub-region of images; hence is more discriminative than ScSPM. Besides, we relax the semantic embedding from ideal semantic correspondence to semantic purity of visual words and reweight each visual word according to its semantic purity, giving higher weights to semantically distinctive visual words, and lower weights to semantically general ones. The experimental evaluations on the Scene-15 dataset demonstrate the effectiveness of the proposed spatial pyramid coding and visual word reweighting for image classification.

Our future work includes the following possible directions. First, More efficient coding methods, such as semi-supervised methods will be studied. Second, how to further reduce the computation cost will also be investigated. Third, how to integrate the spatial information of local features more efficiently will also be studied.

**Acknowledgement.** This work is supported by Major State Basic Research Development Program (2010CB327905) and the Natural Science Foundation of China (Grant No. 60835002, 60723005, 60723005).

## References

1. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
2. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. CVPR (2009)
3. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: Proc. CVPR (2005)
4. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. CVPR (2010)
5. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised Dictionary Learning. In: Proc. ECCV (2008)
6. Lazebnik, S., Raginsky, M.: Supervised learning of quantizer codebooks by information loss minimization. PAMI (2009)
7. Perronnin, F., Dance, C., Csurka, G., Bressan, M.: Adapted vocabularies for generic visual categorization. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 464–475. Springer, Heidelberg (2006)
8. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Proc. ICCV, pp. 17–21 (2005)
9. Moosmann, F., Nowak, E., Jurie, F.: Randomized clustering forests for image classification. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(9), 1632–1646 (2008)
10. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc. CVPR (2008)
11. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Trans. on Pattern Analysis and Machine Intelligence (2008)
12. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
13. 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: WGMBV (2004)
14. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical report, CalTech (2007)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42(3) (2001)
16. Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.: Visual word ambiguity. IEEE Transactions and Pattern Analysis and Machine Intelligence

17. Zhang, H., Berg, A., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: Proc. CVPR (2006)
18. Sivic, J.S., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. ICCV, vol. 2, pp. 1470–1477 (2003)
19. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: Proc. ICCV, pp.1458–1465 (2005)
20. Yu, K., Zhang, T., Gong, Y.: Nonlinear learning using local coordinate coding. In: Proc. NIPS (2009)
21. Boureau, Y.-L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: Proc. CVPR (2010)
22. Tsotsos, J.: Analyzing vision at the complexity level. Behav. Brain Sci. 13, 423–469 (1990)
23. Chen, X., Zelinsky, G.J.: Real-world visual search is dominated by top-down guidance. Vision Research 46, 4118–4133 (2006)
24. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher-order spatial feature extraction for object categorization. In: Proc. CVPR (2008)
25. Mutch, J., Lowe, D.G.: Multiclass object recognition with sparse, localized features. In: Proc. CVPR (2006)
26. Cai, H., Yan, F., Mikolajczyk, K.: Learning weights for codebook in image classification and retrieval. In: Proc. CVPR (2010)
27. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808. MIT Press, Cambridge (2007)
28. Zhang, C., Liu, J., Ouyang, Y., Tian, Q., Lu, H., Ma, S.: Category sensitive codebook construction for object category recognition. In: ICIP (2009)