Image Representation Learning by Deep Appearance and Spatial Coding

Bingyuan Liu[†], Jing Liu[†], Zechao Li[‡], Hanqing Lu[†]

[†] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[‡]School of Computer Science, Nanjing University of Science and Technology
[†]{byliu, jliu, luhq}@nlpr.ia.ac.cn, [‡]zechao.li@gmail.com

Abstract. The bag of feature model is one of the most successful model to represent an image for classification task. However, the discrimination loss in the local appearance coding and the lack of spatial information hinder its performance. To address these problems, we propose a deep appearance and spatial coding model to build more optimal image representation for the classification task. The proposed model is a hierarchical architecture consisting of three operations: appearance coding, max-pooling and spatial coding. Firstly, with an image as input, we extract a set of local descriptors and adopt the appearance coding to encode them into high-dimensional robust vectors. Then max-pooling is performed within the over spatial partitioned grids to incorporate spatial information. After that, spatial coding is carried out to increasingly integrate the region vectors to a global image signature. Finally, the resulting image representation are employed to train a one-versus-others SVM classifier. In the learning of the proposed model, we layerwisely pre-train the network and then perform supervised fine-tuning with image labels. The experiments on three image benchmark datasets(*i.e.* 15-Scenes, PASCAL VOC 2007 and Caltech-256) demonstrate the effectiveness of our proposed model.

1 Introduction

The task of recognizing semantic category of an image remains one of the most important but challenging problems in computer vision and machine intelligence. The crux of the problem is how to describe an image properly for the classification task. In recent years, Bag-of-Feature (BoF)[1] remains one of the most successful method. It extracts a set of local patch descriptors(*e.g.* SIFT[2] and HOG[3]), encode them into high dimensional vectors and pool to obtain an image-level signature. The standard BoF assigns each local descriptor to the closest entry in a visual codebook which is learned offline by clustering a large sampling set of descriptors with K-means. However, two major problems hinder the performance of this model, *i.e.*, the shortcomings brought by the appearance coding scheme and the lack of spatial information.

Given the simplicity, hard-assignment appearance coding scheme in BoF comes with the problem of quantization error. There have been several extensions to reduce this information loss by adopting better coding techniques as alternative. VanGemert *et al.*[4] proposed the concept of visual ambiguity and soft assign each descriptor into multiple visual words in the codebook. Yang *et al.*[5] adopt the sparse coding algorithm, which demonstrates effective in feature representation and discriminative task. Wang *et al.*[6] proposed to relax the restrictive constraint by locally-constrained linearity regularization. However, these methods are all performed in a purely unsupervised way without any high-level guidance, leading to the absence of discriminative information. Thus the resulting representation is not optimal for the classification, and a better coding model may explore both generative and discriminative properties.

Another inherent drawback is the lack of spatial layout information as the BoF model describes an image as an orderless collection of local features. To overcome this problem, one popular extension, known as Spatial Pyramid Matching(SPM)[7], has been shown effective by partitioning each image into a fixed sequence of increasingly finer grids and concatenating the BoF features in each grids to form a global image representation. It is obvious that the simple concatenation of the region features are not optimal to handle complex spatial distribution, and the spatial coding should also take advantage of both generative and discriminative characteristics, since different image classes usually have their own particular spatial distributions of local features.

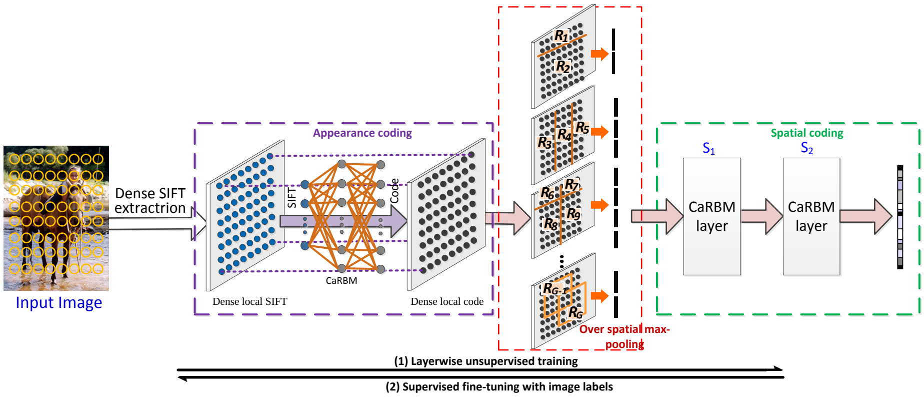

Fig. 1. The proposed deep appearance and spatial coding model. See Section 3 for details.(Better view in the color version)

To address above issues, this paper proposes a deep appearance and spatial coding model to build representative and discriminative image representation. As shown in Fig.1, the proposed model is a hierarchical architecture, consisting of three operations: appearance coding, over spatial max-pooling and spatial coding. The base module is Cardinality Restricted Boltzmann Machine[8], which is an extension to Restricted Boltzmann Machine with the attractive properties of sparse coding by introducing competition among its hidden units. With an image as input, our model firstly extracts local patch descriptors and employ the appearance coding to encode them into high-dimensional codes. To incorporate more optimal spatial layout information, we adopt the idea of over spatial max-pooling. We create various spatial partitions covering very flexible spatial distributions and perform max-pooling within each grid. The resulting features of each partitioned region are then concatenated as input to the next spatial coding module. At last, the layers of spatial coding are explored in a hierarchical structure to increasingly integrate the region vectors to a global image signature. To learn the deep model, we layerwisely pre-train it in an unsupervised way and then fine-tune the parameters with image labels to enhance the discrimination. In this way, our model better explores the generative and discriminative properties, making the obtained feature more optimal to represent the image and adapt to classification task. The output image representations are employed to train a one-versus-others SVM classifier to perform classification. We evaluate our model on three widely used benchmarks (*i.e.* 15-Scenes, PASCAL VOC 2007 and Caltech-256), and the extensive experiments demonstrate the effectiveness of our method in comparison with baselines and related works.

2 Related Work

The Bag-of-Feature(BoF) model[9] is directly borrowed from text retrieval community. In spite of the simplicity, it has been proven very effective to represent an image for large number of vision tasks. The standard BoF extracts a set of local descriptors, and assigns each to the closest entry in a visual codebook, which is learned offline by clustering a large sampling set of descriptors with K-means. Then all these resulting local codes are pooled into an image-level histogram representation. Over the past few years, many efforts have been done to improve the performance of the BoF model.

To overcome the information loss in the codebook learning and feature coding process, some researchers attempted to learn discriminative visual codebooks for image classification[10][11]. Co-occurrence information of visual words was also considered in a generative framework [12]. In [13], the idea of visual word ambiguity is introduced to soft assign each local descriptor to multiple visual words in the learned codebook. As sparse coding is proven effective in feature representation and discriminative tasks, Yang *et al.*[5] utilized it to encode the local features into high-dimensional sparse codes. This method can automatically learn the optimal codebook and search for the optimal coding weights for each local feature. Inspired by this, Wang *et al.*[6] proposed to use locality to constrain the sparse coding process which may be computed faster and yield better performance. Jiang *et al.*[14] proposed to improve the discriminatingly of dictionary via a label consistent regularization. Some other works[15][16] also tried to jointly learn the optimal codebooks and appearance codes. However, how to better explore the generative and discriminative properties of the data is still a difficult problem. In this paper, a combination of both unsupervised feature learning[17][18] and supervised learning is adopted.

As BoF represents an image as an orderless histograms of visual words, many subsequent researches have been done to incorporate spatial information. One direction is to incorporate the local spatial layout in image, *i.e.* the relative positions or pairwise positions of local features. Savarese *et al.*[19] explored the combination of correlograms and visual words to represent spatially neighboring image regions. [20] proposed an efficient feature selection method based on boosting to mine high-order spatial features, while [21] proposed to jointly cluster feature space to build a compact local pairwise codebook capturing correlation between local descriptors and the spatial orders of local features were further considered in [22]. Since images often have spatial preferences, another direction is to incorporate global spatial layout property, *i.e.*, the absolute positions in image. Lazebnik *et al.*[7] pioneered this direction and proposed the Spatial Pyramid Matching (SPM) model. In SPM, the image is divided into uniform grids at different scales (*e.g.* $1 \times 1, 2 \times 2, 4 \times 4$), and the features are concatenated over all cells. This model is successful because it is demonstrated that the combinations of SPM with sparse coding[5], locality-constrained coding[6] and recently developed super vector[23] or fisher vector[24] models are very effective and achieved the state-of-the-art performance. However, the spatial partitions in SPM are too simple to adapt to complex nature situations and are chosen in an ad-hoc manner without any optimization[7]. To solve this problem, Harada *et al.*[25] proposed to form the image feature as a weighted sum of semi-local features over all pyramid levels and the weights are automatically selected to maximize a discriminative power. To design better spatial partition, Sharma *et al.*[26] defined a space of grids where each grid is obtained by a series of recursive axis aligned splits of cells and propose to learn the spatial partition in a maximus margin formulation. [27] formulated the problem in a multi-class fashion with structured sparse regularizer for feature selection, while [28] proposed to learn category specific spatial partition in a one-versus-others classification scheme. In this paper, we explore the idea of over spatial partition and encode it into a deep representation. The most important difference of our model with the previous works is that we take advantage of both traditional BoF models and recently developed deep feature learning framework.

The feature learning models are usually build in a hierarchical framework by stacking shallow generative models with greedy layerwise scheme. One class of feature learning algorithms is based on the encoder-decoder architecture(*e.g.* Auto-encoder)[29]. The input is fed to the encoder which produces a feature vector and the decoder module then reconstructs the input from the feature vector with the reconstruction error measured. Deep Belief Networks(DBN)[30] build multiple layers of directed sigmoid belief nets with the top layer as a Restricted Boltzmann Machines. Lee *et al.*[31] extended DBN with convolution operation for the purpose of extracting latent features from raw image pixels. Yu *et al.*[32] proposed a hierarchical sparse coding model to learn image representations from local patches. Different from these models, we apply a stacked Cardinality Restricted Boltzmann Machine[8], which is an extension to Restricted Boltzmann Machine with the attractive properties of sparse coding by introducing competition among its hidden units. . The most important difference of our model with the previous works is that we take advantage of both traditional BoF models and recently developed deep feature learning framework.

3 The Proposed Model

In this section, we describe the details of our proposed model for image representation and classification. As illustrated in Fig.1, it is a hierarchial architecture, consisting of 3 operations: appearance coding, over spatial max-pooling and spatial coding. The base module of the deep model is Cardinality Restricted Boltzmann Machine(CaRBM).

3.1 Appearance Coding

Starting with an input image I, we densely extract a set of local patch descriptors (*e.g.* SIFT and HOG) and take each descriptor as input to the appearance coding layer. The appearance coding is a deep CaRBM module, encoding the input features into high-dimensional sparse and discriminative codes.

A Restricted Boltzmann Machine(RBM)[33] is a type of bi-partite undirected graphical model that is capable of learning a dictionary of patterns from unlabeled datas. It has a two-layer structure, defining a joint probability distribution over a hidden layer $h \in \{0,1\}^{N_h}$ and a visible layer $v \in \{0,1\}^{N_v}$:

$$P(v,h) = \frac{1}{Z} exp(v^\top Wh + v^\top b_v + h^\top b_h) \tag{1}$$

where Z is the partition function, $W \in R^{N_v \times N_h}$ represents the undirected weights and $b_v \in R^{N_v}, b_h \in R^{N_h}$ are the bias terms. As RBM is a popular density model for extracting features, a desirable property, *i.e.* sparsity, is neglected when applying it to discriminative task. CaRBM is an extension to RBM with the attractive properties of sparse coding by introducing competition among its hidden units. It combines RBM with cardinality potential, which is a class of highly structured global interactions by assigning preferences to counts over subsets of binary variables. The probability of the joint configuration in CaRBM is defined as follows:

$$P(v,h) = \frac{1}{Z} exp(v^\top Wh + v^\top b_v + h^\top b_h) \cdot \psi_k(\sum_{j=1}^{N_h} h_j) \tag{2}$$

where ψ_k is a potential given by $\psi_k(c) = 1$ if $c \leq k$ and 0 otherwise. This constrains that the conditional distribution $P(h|v)$ assigns a non-zero probability mass to a vector h only if $|h| \leq k$. In other words, a data vector v can be explained by at most k hidden units.

By letting the visible layer v correspond to dimensions of the input local descriptor, the CaRBM is able to encode it into a high-dimensional codes. Different from the standard visual coding, we directly map the descriptors into a high-dimensional space by the undirected weights matrix. Thus in the appearance layer, the dimension of the visible units is denoted as $N_v = X$(128 for SIFT) and the number of hidden units is denoted as $N_h = D$ which is usually much larger than X. In the appearance coding of our model, we build a two layer model by stacking the CaRBM, where the output of the first CaRBM is regarded as the input to the next layer. These deep coding scheme is more effective according to our implementation, and we finally concatenate the resulting codes of the two layer for further improvement.

To learn the appearance coding layer, we firstly train it in an unsupervised way and then perform fine-tuning to enhance discriminative property. In the unsupervised pre-training phase, the objective is to maximize the likelihood of training data. As the calculation of conditional distribution $P(h|v)$ is tractable by the sum-product algorithm[8], we may use algorithms like Persistent Contrastive Divergence(PCD)[34]. In the fine-tuning process, we associate each input descriptor with the image label. Different from RBM, the nonlinearity of CaRBM is not clear. Therefore an approximate Jacobian multiplication method is needed to compute gradients[8]. With the learned CaRBM, each local patch is encoded into a high dimensional code during inference process.

3.2 Over Spatial Max-pooling

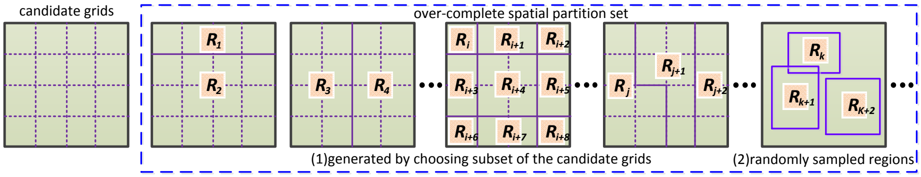

Fig. 2. The generation of over-complete spatial partition set.

In order to build a image-level signature, pooling operation is usually carried out to aggregate the local appearance codes. Traditional SPM partitions a given image into increasingly finer uniform grids (*i.e.* $1 \times 1, 2 \times 2, 4 \times 4$), and then pools within each region and concatenate all the region vectors as the final image representation. To overcome the limitation of the simple uniform partition, we propose to construct spatial partitions in a more flexible scheme incorporating as many geometric properties of the local features as possible.

As described in Fig. 2, we firstly apply uniform horizontal and vertical grids to divide the image into rectangular grids (the dotted grids). These grids are considered as the candidate grids to generate a certain kind of spatial partition. Then, a type of spatial partition is created by randomly choosing a subset of the candidate grids. By covering all the possible combinations of the grids, the spatial partition is able to present various spatial layout information. In addition, we generate some randomly sampled grids to provide more flexible spatial information. All these partitioned regions are collected as the over spatial partition set. Max-pooling is finally performed on the local appearance codes with each partitioned region and all the region vectors are concatenated as the input to the next spatial coding module. It is noted that the dimension of each region vector is D.

3.3 Spatial Coding

With the region vectors as input, spatial coding are then performed to integrate them into global image representations. If number of partitioned spatial region is k, then the input dimension of the spatial coding is kD. Since this input is highly redundant, we also turn to the sparse constrained CaRBM. As shown in Fig. 1, the spatial coding is comprised of two stacked CaRBM with the output of the first one taken as the input to the next. Instead of simply concatenating the regions vectors in SPM, our model explores the generative and discriminative properties of the spatial distribution to fuse the region vectors into a better representation. The pre-training is performed layer-by-layer by PCD algorithm to maximize the likelihood of the training dataset, while the fine-tuning is performed top down with the image label.

Finally, the obtained representations are employed to train a one-versus-others SVM to classify an image into the category with max score. The features of the two layers in spatial coding may be combined to further improve the performance.

4 Experiments

We start our experiments with an in-depth analysis of the proposed model on 15-Scenes dataset, after which we transpose the findings to experiments on PASCAL VOC 2007 and Caltech-256 dataset. Firstly, we evaluate the effectiveness of appearance coding compared with the hard-assignment[7], sparse coding[5], LLC[6] and RBM. Then, we demonstrate the effect of our spatial coding mainly compared with the popular SPM and some other works considering spatial information.

For fair comparison, the experiments are conducted closely following the standard settings [7][5][35]. We adopt a single local descriptor, the SIFT descriptor, by densely extracting local patches of 16×16 pixels over a grid with spacing of 4 pixels. The number of hidden units in appearance coding is fixed as $D = 1024$ for fair comparison, and the dimensions of hidden units for each CaRBM in spatial coding layers are equally set and denoted as S. The number of partitioned regions in the over spatial max-pooling is 60. The target sparsity is all set to 10%. The SVM classifiers are trained with linear kernels in one-versus-others scheme and the trade-off parameters to the SVM regularization term are chosen via 5-fold cross validation on the training set.

4.1 Results on 15-Scenes

We start our experiments with a most popular scene classification benchmark, *i.e.* 15-Scenes. This dataset is complied by several researchers[36][7], including 15 scene categories(*e.g.* kitchen, coast, highway) with each class containing 200 to 400 images. Following the standard setup, 100 images per class are taken for training with the rest for testing. The performances are reported by repeating the experimental process 5 times with different randomly selected training and testing images.

Table 1. Classification rate (%) comparison of different coding methods on 15-Scenes.

Algorithms	Classification Rate
Hard+SPM[7]	81.1 ± 0.3
Soft+SPM[37]	82.7 ± 0.4
SC+SPM[5]	80.8 ± 0.9
LLC+SPM	81.8 ± 0.60
SSRBM+SPM[16]	84.1 ± 0.8
Unsupervised RBM+SPM	82.5 ± 0.5
Unsupervised CaRBM+SPM	86.9 ± 0.2
Supervised CaRBM+SPM	$\mathbf{88.3 \pm 0.3}$

We firstly evaluate the effect of our appearance coding by comparing to the baselines and related works with only difference in the feature coding phase. The final image representations are all compiled by SPM after feature coding. The detailed performance comparison is shown in Table 1, where "Unsupervised RBM+SPM" and "Unsupervised CaRBM+SPM" denote the method of adopting RBM and CaRBM module to perform feature coding respectively, and "Supervised CaRBM+SPM" denotes the results after supervised fine-tuning. It is shown that our appearance coding method outperforms traditional method, such as hard assignment, soft assignment, sparse coding and LLC. Our method also beats SSRBM[16], which is a model of applying sparse regularized RBM. Compared with SSRBM, CaRBM accomplishes sparsity in a fundamentally different way and performs better according to our experiments. It is also shown that supervised fine-tuning brings a slight improvement, demonstrating that the discrimination are enhanced by fine-tuning process.

Secondly, with the coding scheme fixed as supervised CaRBM, we evaluate the effect of our spatial coding. Fig.3 shows the performance comparison by varying the dimension of the hidden units in the spatial layers. "S1" denotes the performance of the output feature in the 1st layer of spatial coding, while "S1+2" denotes the performance of combining the features of the 2 layers. The baselines are traditional SPM and the method of directly concatenating of all the region vectors (denoted as "CAT"). It is shown that the best performance is achieved with the number of latent units set to about $9,000$, which outper-

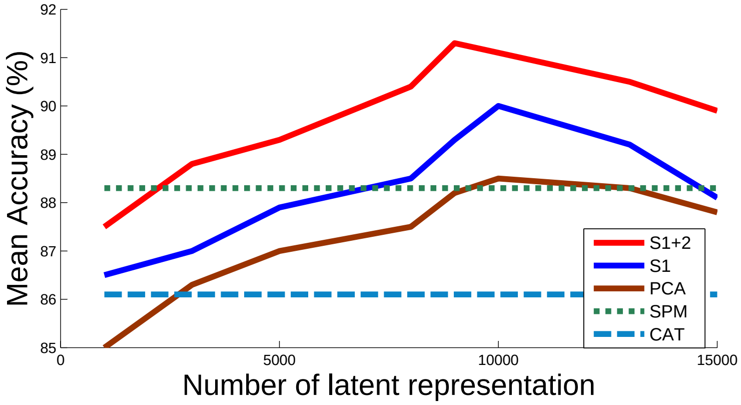

Fig. 3. Performance comparisons of the spatial layers with different dimensions of hidden units.

forms the SPM by nearly 3 percentage points. We also compare our approach to some dimension reduction technique, *e.g.* PCA. The classification accuracy is obviously improved by the CaRBM than PCA, due to the better exploration of generative and discriminative properties of the data. Table 2 shows the performance comparison between our best model with state-of-the-art results. Our best accuracy achieves 91.1%, which outperforms all the related results on this dataset. Compared with traditional BoF framework, our network needs more time to train, since both the appearance and spatial coding modules requires to be learned. However, in the test process, our model is efficient because only forward matrix multiplication operation is needed without any iteration. According to our evaluation, in “SC+SPM”[5], about 4 seconds are needed to obtain the pyramid representation for each image, while only 0.3 seconds are required in the proposed method.

Table 2. Classification rate (%) comparison on 15-Scenes.

Algorithms	Classification Rate
Lazebnik et al.[7]	81.1 ± 0.3
Boureau et al.[38]	85.6 ± 0.3
Zhou et al.[39]	85.2
Goh et al.[16]	86.0 ± 0.5
Feng et al.[40]	83.2
Our(Unsupervised)	88.1 ± 0.3
Our(Supervised)	$\mathbf{91.1 \pm 0.2}$

4.2 Results on PASCAL VOC 2007

Table 3. Recognition performance (AP in %) comparison on VOC 2007. (The table is divided into two parts due to the limitation of space.)

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	Mean
SPM[35]	68.7	57.0	39.9	64.6	22.0	58.8	73.9	53.8	52.4	38.6	49.2	36.9	75.6	61.6	81.6	20.5	40.1	50.9	73.4	49.2	53.4
LLC+SPM[35]	69.8	57.6	42.0	66.5	22.4	55.6	72.8	56.9	51.7	42.9	45.1	39.5	74.1	62.0	80.9	24.5	38.8	49.4	71.2	51.0	53.8
FK+SPM[35]	78.9	67.4	51.9	70.9	30.8	72.2	**79.9**	61.4	55.9	49.6	58.4	44.8	78.8	**70.8**	84.9	31.7	**51.0**	56.4	80.2	57.5	61.7
Object Bank[41]	68.7	53.4	34.6	61.8	19.8	49.9	75.0	42.1	48.7	28.7	50.2	31.8	71.4	53.1	79.6	15.6	29.0	44.3	67.3	49.0	48.7
Our(unsupervised)	77.3	67.3	51.2	72.1	31.8	70.5	76.0	59.3	55.8	48.4	57.9	46.1	78.5	69.0	84.1	29.8	48.3	56.9	81.0	56.9	60.9
Our(supervised)	**80.8**	**69.1**	**53.0**	**73.5**	**33.2**	**72.2**	78.3	**61.8**	**57.5**	**51.6**	**59.3**	**48.6**	**80.6**	70.3	**86.4**	**32.3**	50.5	**59.5**	**83.2**	**58.5**	**63.1**

The PSCAL Visual Object Challenge (VOC) datasets are widely used as testbeds for evaluating algorithms for image understanding tasks and provide a common evaluation platform for both object classification and detection. This dataset is considered to be an extremely challenging one because all the images are daily photos obtained from Flicker where the size, viewing angle, illumination, appearances of objects and their poses vary significantly with frequent occlusions. The PASCAL VOC 2007 dataset consists of $9,963$ images from 20 classes, which are divided into "train", "val" and "test" subsets, *i.e.* 25% for training, 25% for validation and 50% for testing. The classification performance is evaluated using the Average Precision (AP) measure, a standard metric used by PASCAL challenge. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance. We use the train and validation sets for training and report the mean average precision for the 20 classes on the test set as the performance evaluation. In the experiment setup, the size of the latent units in the appearance coding is 1024, while the size of each layer in spatial coding is set to $10,000$.

The performance comparisons of all the 20 classes are shown in Table 3. "SPM" denotes the method of applying hard assignment and SPM model, while "LLC+SPM" represents the method of using locality-constrained linear coding instead of hard assignment. "FK+SPM" denotes the method of using fisher kernel to encode the SIFT descriptors, which is the state-of-the-art feature coding method. In "Object Bank"[41], pre-trained object detectors are employed to extract image representations. As shown in the table, our method leads the performance for most categories. It is also demonstrated that the supervised fine-tuning is effective to improve the recognition performance. However, for some categories, such as car and sheep, our method decreases the accuracy. This is mainly on account of the highly diversity of the images in this challenging dataset.

4.3 Results on Caltech-256

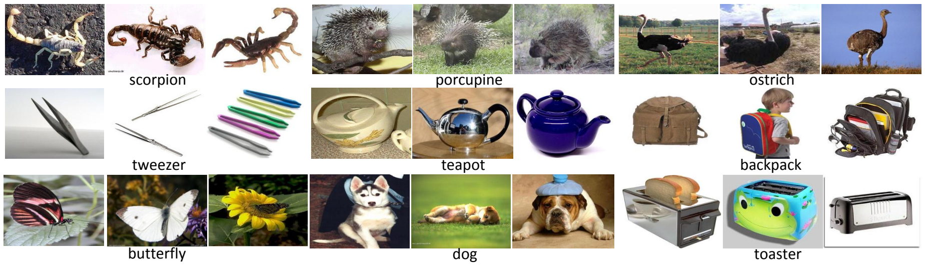

Fig. 4. Example images of Caltech-256 dataset.

The Caltech-256[42] dataset totally holds 29, 780 images in 256 object categories, where the number of images in each category varies form 31 to 800. This dataset is very much challenging as it possesses highly intra-class variability and object location variability. Some example images are shown in Fig.4. Following the standard experiment setup on this dataset, we train our model on 30 and 60 images per class and test on the rest. The other parameters setup is transposed from the former experiments on 15-Scenes.

Table 4. Classification rate (%) comparison on Caltech-256.

Algorithms	30 training	60 training
KSPM[7]	34.1	—
ScSPM[5]	34.0 ± 0.4	40.1 ± 0.9
LLCSPM[6]	41.2	47.7
GLP[40]	43.2	—
Our(Unsupervised)	46.5 ± 0.3	50.2 ± 0.4
Our(Supervised)	$\mathbf{48.7 \pm 0.2}$	$\mathbf{53.2 \pm 0.4}$

The performance comparison is shown in Table 4. In this challenging dataset, our method also consistently leads the performance on all the cases and outperforms the baseline ScSPM by more than 10%. GLP[40] is a method of using discriminatively learned pooling operation to aggregate local features and our model also behaves better than it. The reason may be that our model explores more kinds of latent spatial layout and integrate the regions beyong the simple concatenation scheme. The results on this challenging object datasets demonstrate the effectiveness of the proposed deep image representation model.

5 Conclusion

In this paper, we address the issues of local feature coding and spatial information incorporation in the BoF model and propose a deep appearance and spatial coding model to build more representative and discriminative representation for image classification. We utilizes the Cardinality Restricted Boltzmann Machines, which is capable of combining generative and discriminative properties. With the CaRBM as the base module, our model includes appearance coding, over spatial max pooling and spatial coding operations, which is pre-trained in an unsupervised scheme and then fine-tuned with image labels. The extensive experiments on 15-Scenes, PASCAL VOC 2007 and Caltech-256 datasets have shown the effectiveness of our model in feature coding and spatial information integration, and the classification performances outperform the baselines and related works. Possible future work involves directly learning local patch features from the raw image pixels to make the model an end-to-end learning framework.

Acknowledgement. This work was supported by 863 Program (2014AA015104) and National Natural Science Foundation of China (61332016, 61272329, 61472422, and 61273034).

References

1. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV 2004 Workshop on statistical learning in computer vision. (2004)
2. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
4. van Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: ECCV. (2008)
5. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009)
6. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR. (2010)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
8. Swersky, K., Tarlow, D., Sutskever, I., Salakhutdinov, R., Zemel, R., Adams, R.: Cardinality restricted boltzmann machines. In: NIPS. (2012)
9. Roth, P.M., Winter, M.: Survey of Appearance-Based methods for object recognition. Technical report, Institute for Computer Graphics and Vision, Graz University of Technology (2008)

10. Perronnin, F., Dance, C., Csurka, G., Bressan, M.: Adapted vocabularies for generic visual categorization. In: ECCV. (2006)
11. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV. (2005)
12. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR. (2008)
13. van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1271–1283
14. Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent k-svd. In: CVPR. (2011)
15. Yang, J., Yu, K., Huang, T.S.: Supervised translation-invariant sparse coding. In: CVPR. (2010)
16. Goh, H., Thome, N., Cord, M., Lim, J.: Unsupervised and supervised visual codes with restricted boltzmann machines. In: ECCV. (2012)
17. Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H.: Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge and Data Engineering **26** (2014) 2138–2150
18. Li, Z., Yang, Y., Liu, J., Zhou, X., Lu, H.: Unsupervised feature selection using nonnegative spectral analysis. In: AAAI. (2012)
19. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: CVPR. (2006)
20. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher-order spatial feature extraction for object categorization. In: CVPR. (2008)
21. Morioka, N., Satoh, S.: Building compact local pairwise codebook with joint feature space clustering. In: ECCV. (2010)
22. Morioka, N., Satoh, S.: Learning directional local pairwise bases with sparse coding. In: BMVC. (2010)
23. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: ECCV. (2010)
24. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
25. Harada, T., Ushiku, Y., Yamashita, Y., Kuniyoshi, Y.: Discriminative spatial pyramid. In: CVPR. (2011)
26. Sharma, G., Jurie, F.: Learning discriminative spatial representation for image classification. In: BMVC. (2011)
27. Jia, Y., Huang, C., Darrell, T.: Beyond spatial pyramids: receptive field learning for pooled image features. In: CVPR. (2012)
28. Liu, B., Liu, J., Lu, H.: Adaptive spatial partition learning for image classification. Neurocomputing **142** (2014) 282–290
29. Huang, F.J., lan Boureau, Y., Lecun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: CVPR. (2007)
30. Hinton, G.E., Osindero, S.: A fast learning algorithm for deep belief nets. Neural Computation **18** (2006) 2006
31. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: ICML. (2009)
32. Yu, K., Lin, Y., Lafferty, J.: Learning image representations from the pixel level via hierarchical sparse coding. In: CVPR. (2011)
33. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313** (2006) 504 – 507

34. Tieleman, T.: Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. In: ICML. (2008)
35. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. (2011)
36. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR. (2005)
37. Liu, L., Wang, L., Liu, X.: In defense of soft-assignment coding. In: ICCV. (2011)
38. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. (2010)
39. Zhou, X., Cui, N., Li, Z., Liang, F., Huang, T.: Hierarchical gaussianization for image classification. In: ICCV. (2009)
40. Feng, J., Ni, B., Tian, Q., Yan, S.: Geometric lp-norm feature pooling for image classification. In: CVPR. (2011)
41. Li, L.J., Su, H., Xing, E.P., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: NIPS. (2010)
42. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)