Multi-channel and multi-scale mid-level image representation for scene classification

Jinfu Yang,[a] Fei Yang,[a,*] Guanghui Wang,[b] and Mingai Li[a]
[a]Beijing University of Technology, Faculty of Information Technology, Beijing, China
[b]University of Kansas, Department of Electrical Engineering and Computer Science, Lawrence, Kansas, United States

Abstract. Convolutional neural network (CNN)-based approaches have received state-of-the-art results in scene classification. Features from the output of fully connected (FC) layers express one-dimensional semantic information but lose the detailed information of objects and the spatial information of scene categories. On the contrary, deep convolutional features have been proved to be more suitable for describing an object itself and the spatial relations among objects in an image. In addition, the feature map from each layer is max-pooled within local neighborhoods, which weakens the invariance of global consistency and is unfavorable to scenes with highly complicated variation. To cope with the above issues, an orderless multi-channel mid-level image representation on pre-trained CNN features is proposed to improve the classification performance. The mid-level image representation of two channels from the FC layer and the deep convolutional layer are integrated at multi-scale levels. A sum pooling approach is also employed to aggregate multi-scale mid-level image representation to highlight the importance of the descriptors beneficial for scene classification. Extensive experiments on SUN397 and MIT 67 indoor datasets demonstrate that the proposed method achieves promising classification performance. *© 2017 SPIE and IS&T* [DOI: 10.1117/1.JEI.26.2.023018]

Keywords: scene classification; convolutional neural network; multi-channel; mid-level representation.

Paper 16819 received Sep. 27, 2016; accepted for publication Mar. 20, 2017; published online Apr. 11, 2017.

1 Introduction

Scene classification is a challenging task in computer vision, since most of the scenes are the collections of entities organized in a highly variable layout. Many available methods for scene classification are based on the appearance of local descriptors.[1–10] Generally speaking, image representation can be categorized into three levels: low-level, mid-level, and high-level.[11] Low-level features such as color, texture, or shape information of images describe appearance at the pixel point in an image and can be retrieved directly from images without any external knowledge. Although great progress has been achieved with the invention of low-level features,[1–3] low-level features-based methods cannot provide sufficient semantic information to scene recognition since they depend mainly on corner points. As a result, some researchers have investigated high-level features that are already impregnated with semantic information to increase the generalization ability.[4–6] For instance, Li et al.[6] proposed that the scenes can be regarded as a series of goals (e.g., objects) and some combination of them can represent a certain kind of scene. They constructed feature vectors with a series of multi-scale corresponding maps of target detection descriptors and achieved promising performance. However, extracting high-level features requires a large amount of training data and semantic entities which are ambiguous and less discriminable. The problem raised considerable interest in the subject of mid-level features,[7–10] due to its superiority of the human visual mechanism over low-level features and minor training data over high-level features. Mid-level visual elements, which are clusters of image patches rich in semantic meaning, were proposed by Singh et al.[9] They adopted discriminative clustering to train support vector machines (SVM) for discovery of mid-level image patches and utilized these patches to perform scene classification. All of the aforementioned approaches based on mid-level features use rudimentary features that represent only local information about images and throw away much of the discriminative information in the image.[12] To cope with the problem, a convolutional neural network (CNN) is used to extract global representation of images and has achieved state-of-the-art results in image classification,[13–15] object detection,[16–18] and semantic segmentation.[19–21]

Most current solutions take activations of the fully connected (FC) layer as the image representation,[13–15,18] which have a general description for images. However, the activation of the FC layer loses detailed information of the objects in comparison to the features of the convolutional layers, since objects and scenes are closely related and the objects can be helpful for recognition.[22] In addition, in the architecture of CNN, feature maps from the convolutional layer are pooled within local neighborhoods. As a result, the structure of CNN reduces the invariance to geometric transformations in all cases. As shown in Ref. 23, the reconstructed image of the output of the fifth convolutional layer is similar to the original one, which will hurt performance for scene images with high spatial variability. Compared with orderless Bag of Features (BoF),[24] CNN activations are "globally ordered" spectra for image representation, which is negative for scene classification.

[*]Address all correspondence to: Fei Yang, E-mail: yangfei199217@emails.bjut.edu.cn

1017-9909/2017/$25.00 © 2017 SPIE and IS&T

Inspired by Ref. 25, to increase the invariance of global deformations, we sample a set of multi-scale mid-level image patches[22,25–27] as orderless inputs to the CNN extractor and aggregate the descriptors of patches to represent the whole image. Different from Ref. 25, we combine the deep convolutional features and FC features to obtain a more comprehensive representation. In addition, Ref. 28 emphasized objects of interest that tend to be located close to the geometrical center of an image and proposed a simple center prior based on the sum pooling method for image retrieval. However, this method is not adaptable to scene classification, since scenes can be seen as a combination of some certain objects, instead of one primary object. As a result, we propose an importance-weighted aggregation method based on sum pooling (IWA-SP) for deep convolutional features of the combined channels. The main contributions of this paper are as follows. First, we propose a solution to generate comprehensive image representation by combining multi-channel features. Second, the proposed multi-scale mid-level image representation mechanism increases the invariance of geometric transformation and prevents the representation from lacking the ability for depicting relationships of different regions in an image. Third, a pooling method, called IWA-SP, is proposed to highlight the contribution of region descriptors to be beneficial for recognition.

The remaining part of this paper is organized as follows. In Sec. 2, we introduce related work on scene recognition and descriptors aggregation methods. The proposed method is elaborated in Sec. 3. Section 4 presents extensive experiments and results on the MIT 67 indoor and SUN397 datasets. Finally, this paper is concluded in Sec. 5.

2 Related Work

Learning mid-level representation using CNN has shown effective performance.[13,25,26,29,30] Oquab et al.[29] proposed learning and transferring mid-level representation by using a trained ImageNet CNN. Li et al.[13] presented a learning discriminative mid-level representation approach by extracting CNN features to recognize scenes. Generally speaking, these approaches combine multiple scales that are pooled using vector of locally aggregated descriptors (VLAD)[25] or Fisher vector (FV)[26] encoding. Gong et al.[25] proposed an orderless pooling method for the patches' output of the FC layer to increase invariance of representations for scene recognition. In general, these works use CNN to extract the FC layer's activations as generic representations of images. However, the activation of the FC layer is quite limited, which contains one-dimensional semantic information and loses detailed information of objects and destroys the property of spatial characteristics among objects. Objects and scenes are closely related, so knowledge about objects can be helpful for recognizing scenes. Features from the convolutional layers have a natural interpretation as descriptors of local image regions corresponding to receptive fields of the particular features. Therefore, in our work, we integrate two channel features from the FC layer and the convolutional layer to represent images. To aggregate patches' descriptors of convolutional layer, we propose an aggregation method based on sum pooling, which assigns weights to the local descriptors to highlight the benefits for recognition. Recently, Herranz et al.[22] proposed a model that trains a CNN architecture for each scale level of images with two types of networks (i.e., ImageNet CNN and Places CNN), to increase objects information in scene images and effectively combine two types of features. Different from the proposed method, we train one ImageNet CNN for all scales and combine the Places CNN for full images with single scale. In Sec. 4, we also adopt their idea to expand our experiments.

3 Proposed Method

3.1 *Framework*

To learn more effective image representation, in this paper, we propose a solution by constructing a CNN framework of multi-channel and multi-scale mid-level representation. Figure 1 shows the proposed framework, which consists of three stages: first, multiple scale mid-level image patches are selected as the inputs to CNN by sliding windows. Then, the deep convolutional features and the FC activations are extracted across each scale level. For deep convolutional features, the IWA-SP is utilized to form the final descriptors. For FC activations, the improved optimal VLAD is applied to obtain image representation. Finally, we concatenate the

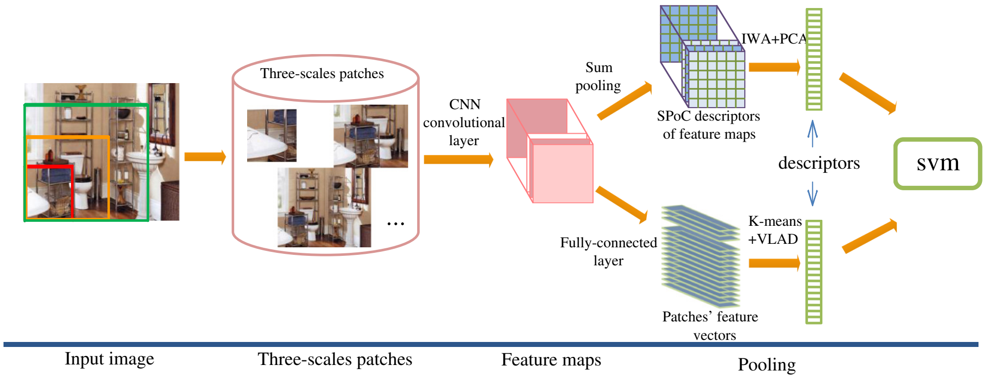

Fig. 1 The framework of our proposed method. A raw image is first divided into multi-scale levels of mid-level patches, then these patches are fed to ImageNet to obtain feature maps from the convolutional layer and one-dimensional semantic vectors of patches from the FC layer, respectively. We adopt improved optimal VLAD encoding for semantic vectors of patches from the FC layer and importance-weighted sum pooling for feature maps from the convolutional layer. Finally, two complementary representations are concatenated to train the SVM.

above two descriptors at multiple scale levels to train a linear SVM.

3.2 *Descriptors of Mid-level Patches*

As mentioned above, mid-level patches are employed as the inputs to the pre-trained CNN to extract features. The output of the convolutional layer is a set of feature maps, with each of them describing a local region (spatial unit).[31] These feature maps meet the spatial relation of objects in an image. Through the FC layer, the feature map is transformed into a feature vector, containing more semantic information in comparison to the deep convolutional features. During this process, the spatial information is lost, as each feature vector corresponding to a spatial unit is unable to be reproduced through the FC layer. Feature vectors can be regarded as semantic representation of the whole image, which is beneficial to scene classification. Therefore, to generate comprehensive image representation, we concatenate both of two kinds of features, the spatial units and the feature vectors, as the inputs to the orderless pooling.

3.3 *Pooling Methods*

3.3.1 *Improved optimal VLAD encoding*

After obtaining the above-mentioned descriptors, we combine them into a global representation. For the FC feature, we adopt the VLAD encoding (a soft assignment version[32]). Given a collection of image patches, we first learn a separate codebook $c = \{c_1, \dots, c_k\}$ through k-means. Here, we set $k = 100$ in our experiments. Assign each patch p_i to its r nearest cluster centers $r\mathrm{NN}(p_i)$ and aggregate the residuals of the patches minus the center. The VLAD descriptor is constructed using

$$X = \left[\sum_{i:c_1 \in r\mathrm{NN}(p_i)} w_{i1}(p_i - c_1), \dots, \sum_{i:c_k \in r\mathrm{NN}(p_i)} w_{ik}(p_i - c_k) \right], \tag{1}$$

where w_{ik} is the Gaussian kernel similarity between p_i and c_k as follows:

$$w_{ik} = \exp(\|p_i - c_k\|/2\sigma^2), \tag{2}$$

where σ is the width value of function. Given a set of width σ_i, $i = 1, 2, \dots, n$ randomly, train the SVM for each σ_i and compute the classification error rate, then, choose the width value corresponding to the smallest error rate. However, the values of σ_i are difficult to determine and are always determined by experience. In our experiments, we apply a scheme to determine σ_i, which is relying on the multiples of median distance between support vectors. Through training SVM, we obtain the discriminant function of an optimal separating hyperplane. The function consists of support vectors, which has no relation with other training samples. Therefore, the distance between support vectors is adopted to determine the range of width values of a Gaussian kernel. Take a binary classification as an example: first, an approximate σ_0 determined by median distance between training samples of two classes is given to train the support vector machines. Then, the median distance d_{med} between the support vectors of two classes is computed to obtain a set of width values $\sigma_i^2 \in \{1/8d_{\mathrm{med}}^2, 1/4d_{\mathrm{med}}^2, 1/2d_{\mathrm{med}}^2, d_{\mathrm{med}}^2, 2d_{\mathrm{med}}^2, 4d_{\mathrm{med}}^2, \dots\}$. Since the median distance depicts spatial characteristic of support vectors to a certain extent, in comparison to the traditional cross-validation method, the above scheme can reduce computational burden thanks to less times of validation.

The whole process for VLAD can be divided into two parts, i.e., embedding and aggregation. The first step is mapping each image patch descriptor into a higher dimensional vector. In the second step, we sum the difference between p_i and c_k. Here, we represent the descriptor by the difference of the FC feature and the clustering center in each dimension, instead of the nearest c_j representing this vector, such as BoF, which contains more detail information by considering each dimension.

Following Ref. 33, the VLAD descriptor X is subsequently L2-normalized. However, the dimension of vector obtained from this method is usually high. Given 500-dimensional patch descriptors from the FC layer after principle component analysis (PCA) and 100-dimensional k-means centers, we obtain a 50,000-dimensional vector, which is too high for large-scale patches. To increase the efficiency, we adopt PCA to reduce the pooled vectors to 4096 dimensions and obtain the image global representation. Note that applying PCA is a standard practice in previous works.[34,35]

For deep convolutional features, the dimension of which is higher than the FC features, it is not suitable to map each patch descriptor into a higher dimensional vector for its large computational cost. Babenko and Lempitsky[28] proved that a simple sum pooling aggregation applied on raw local descriptors can provide a better performance compared with high-dimensional embedding. This method simplifies the descriptors of the fifth convolutional layer, leading to more efficient and reliable compact descriptors for image retrieval.

However, their center prior method based on sum pooling is not suitable for scene classification, since a complex scene usually contains multiple objects distributed across the whole image. Therefore, we propose a weighted method for sum pooling to satisfy the importance of different descriptors to scene image.

3.3.2 *IWA-SP*

Given an image patch p with a certain number of feature maps $f(x, y)$ extracted from the last convolutional layer, where (x, y) represents the spatial position of the feature in the map stack, the sum pooling feature is represented as follows:

$$\Omega = \sum_{y=1}^{H} \sum_{x=1}^{W} f(x, y), \tag{3}$$

where H, W represent the height and the weight of feature maps, respectively.

Most existing feature encoding methods treat each local descriptor with equal importance. However, some descriptors containing background or other irrelevant objects add noise to the global image descriptor and some descriptors containing objects representative for a certain kind of scene are possibly neglected. Thus, both situations affect classification accuracy. In this section, we propose a scheme that gives a descriptor Ω an importance weight $w(\Omega)$ that would determine its contribution toward the global image descriptor. Our idea is to learn a classifier, which distinguishes discriminative features from others and derive the weight using the predicted confidence value. Random forest is used to implement the task and the number of trees is set to 500. We select one kind of scene image descriptors as positive examples and other kinds as negatives for training. Given a descriptor Ω, the posterior probability $p(\Omega)$ is predicted by the forest using the empirical distribution of the labels of the training examples assigned to the predicted leaf. Finally, a sigmoid is used to threshold this probability in a soft manner.

By assigning different weights to the descriptors, Eq. (3) can be rewritten as

$$\psi_1(P) = \omega(\Omega) \sum_{y=1}^{H} \sum_{x=1}^{W} f(x, y), \tag{4}$$

where we define

$$\omega(\Omega) = 1/[1 + e^{-\alpha p(\Omega)+\beta}]. \tag{5}$$

The dimension C of $\psi_1(P)$ equals the number of output maps from the fifth convolutional layer. The final representation $\psi(P)$ is subsequently L2-normalized, then PCA and whitening are employed to reduce the dimension

$$\psi_2(P) = \mathrm{diag}(s_1, s_2, \dots, s_N)^{-1} M_{\mathrm{PCA}} \psi_1(P), \tag{6}$$

where M_{PCA} is the rectangular PCA-matrix and s_i is the associated singular value. The dimension of C is much lower than the corresponding descriptors from embedding. Thus, it takes much less data to obtain the PCA matrix than the VLAD and reduces the risk of overfitting.

Finally, the whitened vector is L2-normalized

$$\psi_{\mathrm{SPOC}}(P) = \frac{\psi_2(P)}{\|\psi_2(P)\|_2}. \tag{7}$$

4 Experiments

The proposed approach is evaluated on two well-known benchmarks: SUN397 and MIT 67 indoor datasets performing on GPU980Ti. In our experiments, we further compare image representation obtained from convolutional layers. Experiments are divided into four parts: the first one is a comparison of mid-level image representation against global activations of full images; the second compares a single channel feature (FC) against two channel features (DCF +FC); the third is a comparison of two encoding methods against other competitive methods; and the last one is an experiment by introducing the Places CNN.

4.1 *Datasets*

The SUN397[36] dataset is the largest dataset to date for scene classification, which consists of 397 scene categories and a total of 108,754 images. Each category has less than 100 images. We use a subset of the dataset that contains 50 training images and 50 testing images per class as a partition, choose 10 times partitions and report the average classification accuracy. The evaluation settings are well-established over the SUN397 dataset.

The MIT dataset[4] consists of 15,620 images, with 67 scene categories, divided into subway, bathroom, closet, and so on. The standard training/test images for the dataset consists of 80 training and 20 test images per class. The evaluation settings are also well-established over the MIT dataset.

4.2 *Experimental Parameters*

For the size of mid-level patches, we choose 64, 128, and 256, respectively. FC features and deep convolutional features are extracted, respectively, relying on the ImageNet CNN. Dimensions of the FC features are first reduced to 500 dimensions. The k-means center here is 100 and the nearest cluster center is 5 in the VLAD pooling process. The final VLAD pooling dimension is set to 4096. For DCF, the number of feature maps is $C = 256$ and the spatial size of the fifth layer is $W \times H = 6 \times 6$. In the IWA-SP process, the dimension is also set to 4096. For the SUN dataset, the best result with the IWA-SP is obtained with $\alpha = 10$, $\beta = 0.3$ and for the MIT dataset, $\alpha = 20$, $\beta = 0.8$.

4.3 *Results*

4.3.1 *Mid-level image representation versus global activation of CNN*

We perform experiments using the global CNN activation from different layers for scene classification on full images. Results are listed in Table 1, where DCF represents the deep convolutional features, FC6 refers to the sixth FC ones, and FC7 denotes the seventh FC ones. As shown in Table 1, the result of FC7 features is better than that of FC6 features. Therefore, in our following experiments, we utilize FC features to represent the seventh FC layer features. It is also evident that when two channel features work together, the accuracy is better than that of the random single channel feature.

Table 2 shows the results on the mid-level image representation across the three scales. Single FC channel results are the same as Ref. 25, which is called Mop. We choose the same pooling method (VLAD) for both two channel features for the fairness of comparing the results of different channels. For the same dataset, the accuracy of the left column in Table 2 using mid-level representation of a single FC is much higher than that in Table 1 using the global activation, which proves the effectiveness of the mid-level image representation. Our method increases the invariance to geometric transformations, for example, for classes having an object in the center in Fig. 2(a), the classification results are correct using global activation of CNN.

However, for classes that have high spatial variability in Fig. 2(b), the classification results using global activation of

Table 1 The average accuracy of global activation of DCF and the FC features on SUN397 and MIT67.

Global activation	DCF (%)	FC6 (%)	FC7 (%)	DCF+FC7 (%)
SUN397	38.53	41.30	42.61	44.80
MIT67	53.80	56.51	58.40	61.52

Table 2 Results of one single FC channel and the two channel combinations (DCF+FC) on SUN397 and MIT 67 indoor datasets. For each scale level, computational time is listed in this table. VLAD encoding (-V) is employed in this experiment.

Combined levels	SUN397 dataset (-V)				MIT67 indoor dataset (-V)			
Channel	FC (%)	Time (s)	DCF+FC (%)	Time (s)	FC (%)	Time (s)	DCF+FC (%)	Time (s)
Level 1	39.67	2.23	40.56	3.01	53.97	2.18	54.69	2.96
Level 2	45.45	2.10	48.42	2.97	65.72	2.05	67.83	2.85
Level 3	40.28	1.92	43.27	2.80	62.34	1.83	63.94	2.72
Level 1 + level 2	50.10	2.28	51.19	3.18	66.84	2.19	68.75	3.21
Level 2 + level 3	49.75	2.17	50.35	3.05	68.04	2.04	69.12	2.91
Level 1 + level 2 + level 3	52.12	2.32	52.72	3.27	68.98	2.21	70.90	3.1

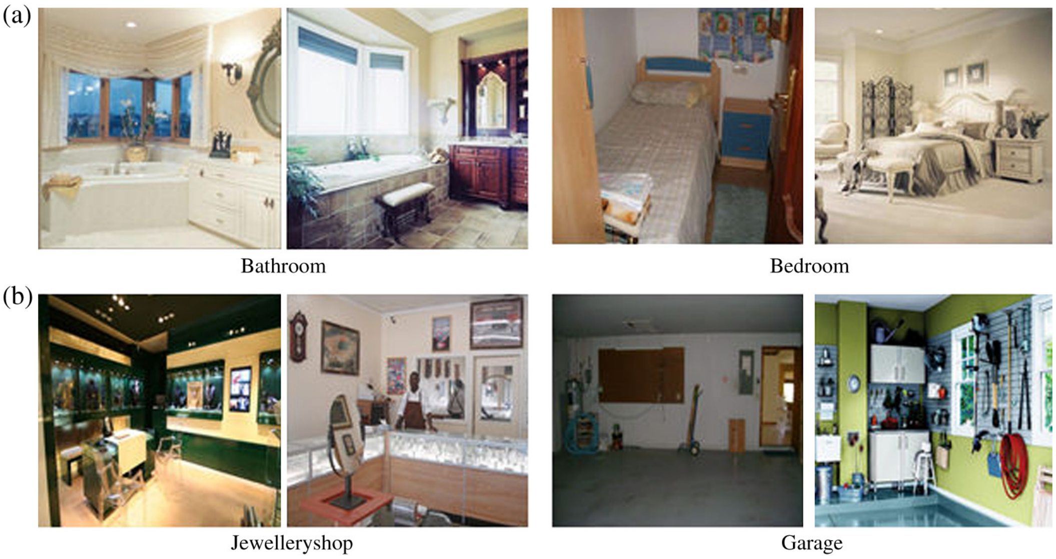

Fig. 2 MIT67 indoor classes. (a) For classes where there is a main object in the center, the classification results are correct using global activation of CNN. (b) For classes that have high spatial variability, the classification results are negative, while our method yields correct results.

CNN are incorrect in comparison to results using mid-level image representation. Notice that different scales of mid-level image representation achieve different results. The best result is achieved by combining all three scales. Our method is based on pre-trained AlexNet provided by the CAFFE package.[37]

4.3.2 *Multi-channel features versus FC feature*

We extract deep convolutional features and visualize each convolutional layer of CNN. As shown in Fig. 3, it is evident that from conv1 to conv5, the information is changed from image edge to abstract corresponding responds. The output of layer 5 corresponds to the patches of the image and it is more representative for the corresponding category. Combine some feature maps randomly sampled from the 256 feature maps in conv5 and for better visualization, we overlay them on the original images. Figure 4 shows the visualization result. It is obvious that the activated regions of the sampled feature maps (highlighted in red color) from the fifth layer also have semantically meaningful information. For example, the activated region of dinette is a part of chairs which is also a chair-part of dinette.

Compared with the FC features, the proposed method obtains better performance in terms of accuracy on the SUN397 dataset and MIT 67 indoor dataset. It is evident that a single FC feature loses sensitive information for scene classification. In order to demonstrate its performance, we list some scenes of misclassification on the MIT 67 indoor dataset utilizing single FC features, as shown in Fig. 5, where labels are listed below the images. The bottom left is the result using single FC features and the upper right is the result using two channel features if it is not the same as labels. It is obvious that for some cluster scenes, a single FC representation probably fails in recognizing the bar, focusing on the desks and chairs in general but ignoring the wines on the table and their spatial relations. The computation time of FC and DCF+FC is also listed in Table 2, from which we can

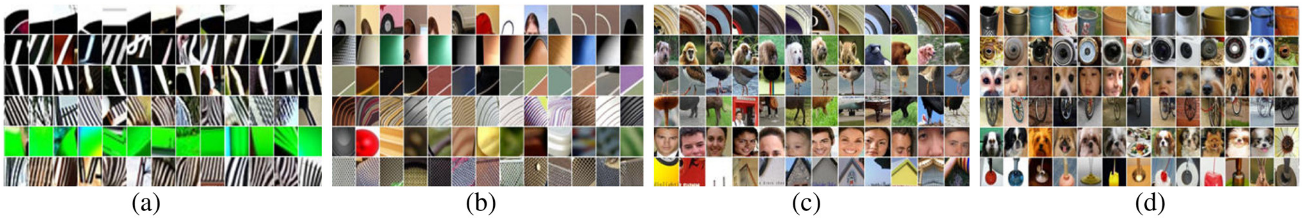

Fig. 3 Visualization of convolutional layers. From (a) conv1 to (d) conv5, information from edge, texture to more abstract and specific object. Output of (d) responds to a kind of certain image category.

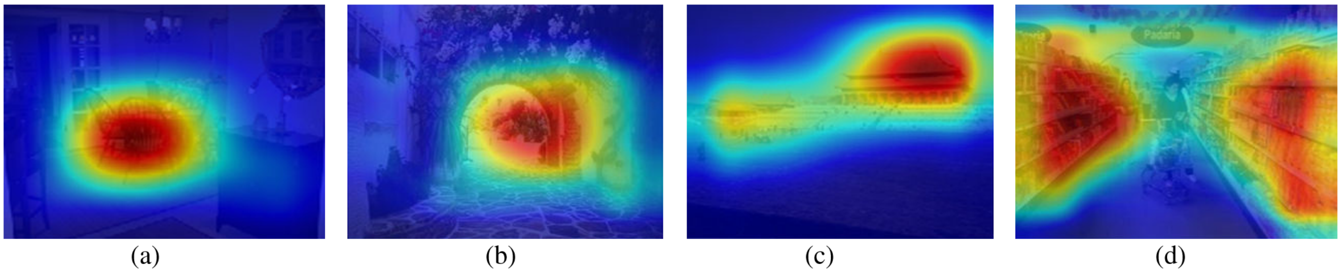

Fig. 4 Activated field of feature maps extracted from the fifth layer of a CNN on SUN397 dataset. (a) Dinette, (b) arch, (c) palace, and (d) supermarket. All of the activated regions are meaningful for scene categories.

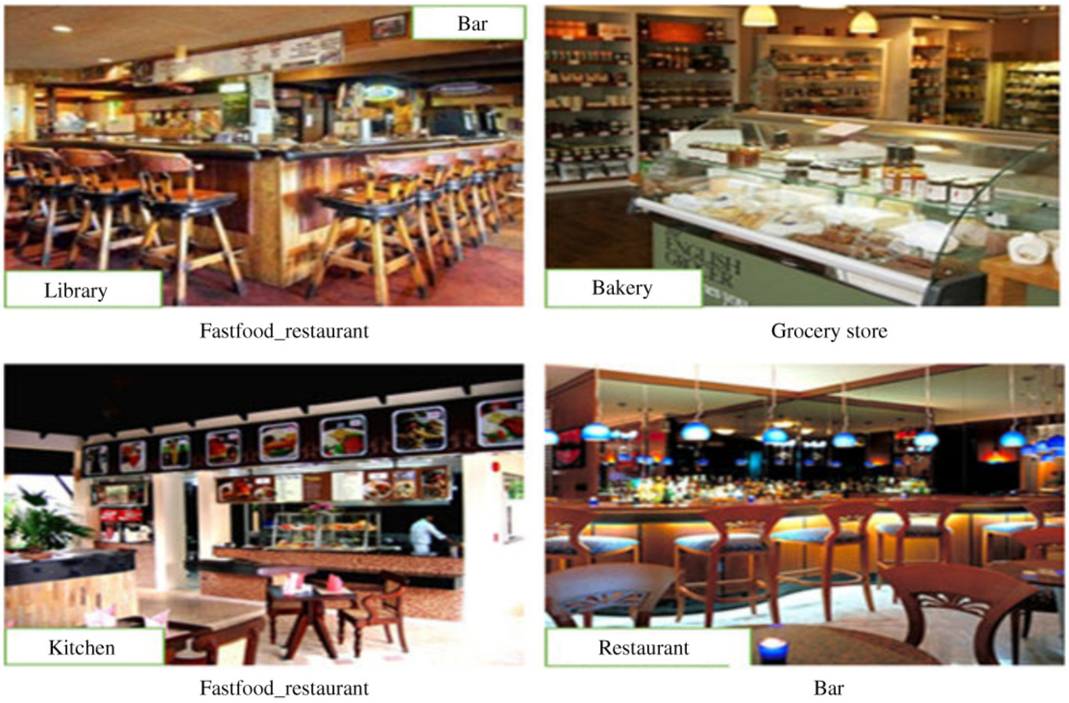

Fig. 5 Classification results using single FC features and two channel features. The ground truth labels are listed below the images; the classification results using single FC features are shown at the bottom left corner; and the results using two channel features are shown at the upper right corner if they are not the same as the ground truth.

see that DCF+FC costs more time than FC (Mop), since the dimensions of the DCF features are high.

4.3.3 *IWA-SP versus VLAD*

To verify the proposed IWA-SP method, we compare the two orderless pooling methods. Table 3 lists the results of concatenating multi-channel descriptors obtained by two pooling schemes, IWA-SP for DCF and VLAD pooling for the FC features, i.e., DCF-IWASP+FC-VLAD. From Tables 2 and 3, we can see that the performance of the proposed method (DCF-IWASP+FC-VLAD) increases 2.08% in comparison to the results on the SUN397 dataset and 1.2% on the MIT 67 indoor dataset in Table 2. The

Table 3 Results of concatenating multi-channel descriptors obtained by two pooling methods, i.e., IWA-SP for the DCF (DCF-IWASP) and VLAD pooling for the FC features (FC-VLAD).

Combined levels	SUN397 (%) (DCF-IWASP +FC-VLAD)	Time (s)	MIT67 indoor (%) (DCF-IWASP+FC-VLAD)	Time (s)
Level 1	40.90	2.53	58.47	2.23
Level 2	50.73	2.12	69.50	2.16
Level 3	44.50	2.05	65.82	1.95
Level 1 + level 2	53.70	2.37	69.87	2.30
Level 2 + level 3	52.32	2.30	70.20	2.16
Level 1 + level 2 + level 3	54.80	2.45	72.10	2.37

computational time of the concatenated method is shown in Table 3. By comparing the results in Table 2, we can find that the computational time using IWA-SP to encode the deep convolutional features is much lower than that using VLAD encoding. The computational time of DCF-IWASP +FC-VLAD is close to that of Mop using single channel features.

Tables 4 and 5 demonstrate the comparative results on the MIT 67 indoor dataset and SUN 397 dataset. On the MIT 67 indoor dataset, the method of semantic Fisher vector[25] is a little bit higher than our method, since they use the inputs to soft-max layer as image patch descriptors and then compute a semantic FV as a Gaussian mixture FV in the space of these natural parameters, increasing the semantic information of image. Dual hybrid[22] achieved the highest accuracy on both datasets, since it uses a hybrid CNN architecture for each scale level of images with two types of networks (i.e., ImageNet CNN and Places CNN); the dual hybrid is more complex than our model which utilizes only one ImageNet CNN. The Dual hybrid model can supplement more objects' information in scene images by training CNN architectures for each scale image, which is helpful for recognition. Our model is also able to capture main objects information by encoding deep convolutional features. As shown in Fig. 3, all of the activated regions are meaningful for scene categories. In the next section, similar to Ref. 22, we also present the experimental results by combining two types of networks in our experiments.

Table 4 Accuracy over MIT 67 indoor dataset. (V) represents VLAD encoding for both two channels, (V+IWA-SP) represents two encoding methods, i.e., IWA-SP for the DCF and VLAD pooling for the FC features.

Method	Accuracy (%)	Method	Accuracy (%)
Object bank[6]	37.60	Mop[25]	68.88
Patches[9]	38.10	MDPM[13]	69.69
ISPR[5]	50.10	Semantic FV[26]	72.86
SPP[14]	56.3	Dual hybrid[22]	78.28
Coarse-to-fine sparselets[15]	59.87 to 64.36	**Our method-(V)**	**70.90**
Mode seeking[10]	65.10	**Our method-(V +IWA-SP)**	**72.10**

Table 5 Classification results on SUN397.

Method	Accuracy (%)
Xiao et al.[36]	38.00
Mop[25]	51.98
Semantic FV[26]	54.40
Dual hybrid[22]	64.10
Our method-V	**52.72**
Our method-(V+IWA-SP)	**54.80**

4.3.4 *Framework with Places CNN*

Our work focuses on generating representation on objects for scene images. Zhou et al.[38] proposed a more direct CNN based on the "Places" dataset without using the ImageNet CNN, aiming at scene classification. The type of features learned from the Places CNN is different from those from the ImageNet CNN. The responds of Places CNN have more receptive fields that look like landscapes and are a holistic representation of scenes, whereas the ImageNet CNN have more receptive fields that look like object-blobs and can be seen as a specific representation of the objects. Therefore, the Places CNN is incorporated into our framework to extract the whole image features, as a complement to our method.

The accuracy can be increased by utilizing a pre-trained CNN relying on a large scale scene dataset. In Table 6, "combined I" is the results obtained by combining ImageNet CNN and Places CNN. On the SUN397 dataset, results of simple combination are almost the same as the Places CNN, whereas on the MIT 67 indoor dataset, our results are higher than those of the Places CNN with an increase of 3.86%. Inspired by Ref. 22, we also combine Places CNN and

Table 6 Comparison results of combining two types of networks.

Method	MIT67 indoor (%)	SUN397 (%)
Our method-(V+IWA-SP)	72.10	54.80
Places FC7[38]	68.24	54.32
Dual hybrid[22]	78.28	64.10
Combined I	**75.60**	**58.18**
Combined II	**79.15**	**63.90**

ImageNet CNN in a similar way to supplement our experiments, see "combined II" in Table 6. Different from combined I with single-scale for the Places CNN, we adopt a multi-scale and multi-channel mid-level representation mechanism in combined II. As listed in Table 6, combined II achieves approximately the same performance as Ref. 22 on the MIT67 indoor dataset and SUN397 dataset, which demonstrates the effectiveness of our method.

5 Conclusion

In this paper, we have proposed to use discriminative mid-level image representation to increase the performance of global activation of CNN. The FC features from the CNN FC layer contain more one-dimensional semantic information, regardless of the detail information and spatial relation among objects. Deep convolutional features have been proved to be effective. We adopt two channel features for scene classification. For the FC features, VLAD, which contains two processes of embedding and aggregation, is performed for achieving global representations of multi-scale image patches. For the DCF, embedding is not required and as a result, we adopt the proposed IWA-SP to aggregate the DCF. Finally, we combine these two kinds of descriptors and train linear SVMs for scene recognition. The proposed method achieves the comparable performance on both the MIT67 indoor dataset and SUN397 dataset, especially when the Places CNN is incorporated.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under Grant Nos. 61201362, 61573351, and 81471770, the Young Top-notch Talent Program of Beijing Educational Committee under Grant No. CIT & TCD201404039, and the Scientific Research Project of Beijing Educational Committee under Grant No. CIT & TCD201404039.

References

1. D. Lowe, "Distinctive image features from scale-invariant keypoints," *IEEE Int. J. Comput. Vision* **60**(2), 91–110 (2004).
2. A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," *Prog. Brain Res.: Visual Percept.* **155**(2), 23–36 (2006).
3. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–893 (2005).
4. A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 413–420 (2009).
5. D. Lin et al., "Learning important spatial pooling regions for scene classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3726–3733, (2014).
6. L. J. Li et al., "Object bank: a high-level image representation for scene classification and semantic feature sparsification," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 1378–1386 (2010).
7. C. Doersch et al., "What makes Paris look like Paris?" *ACM Trans. Graphics.* **31**(4), 13–15 (2012).
8. I. Endres et al., "Learning collections of part models for object recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 939–946 (2013).
9. S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *European Conf. on Computer Vision (ECCV)*, pp. 73–86 (2012).
10. C. Doersch, A. Gupta, and A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 494–502 (2013).
11. J. Yang et al., "Contour detection-based discovery of mid-level discriminative patches for scene classification," *Int. J. Adv. Rob. Syst.* **13**(1), 30 (2016).
12. L. Wan, D. Eigen, and R. Fergus, "End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 851–859 (2015).
13. Y. Li et al., "Mid-level deep pattern mining," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 971–980 (2015).
14. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conf. on Computer Vision (ECCV '14)*, pp. 329–344 (2014).
15. G. Cheng et al., "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1173–1181 (2015).
16. P. Sermanet et al., "Overfeat: integrated recognition, localization and detection using convolutional networks," in *Int. Conf. on Learning Representations (ICLR)* (2014).
17. T. Xiang et al., "Discriminative boosted forest with convolutional neural network-based patch descriptor for object detection," *J. Electron. Imaging* **25**(1), 013002 (2016).
18. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587 (2014).
19. J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3992–4000 (2015).
20. D. Ciresan et al., "Deep neural networks segment neuronal membranes in electron microscopy images," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 2852–2860 (2012).
21. C. Farabet et al., "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1915–1929 (2013).
22. L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: objects, scales and dataset bias," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 571–579 (2016).
23. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 1106–1114 (2012).
24. H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IEEE Int. J. Comput. Vision* **87**(3), 316–336 (2010).
25. Y. Gong et al., "Multi-scale orderless pooling of deep convolutional activation features," in *European Conf. on Computer Vision (ECCV)*, pp. 392–407 (2014).
26. M. Dixit and S. Chen, "Scene classification with semantic Fisher vectors," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2974–2983 (2015).
27. C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 2553–2561 (2013).
28. A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1401–1408 (2015).
29. M. Oquab et al., "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724 (2014).
30. D. Yoo et al., "Multi-scale pyramid pooling for deep convolutional representation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 71–80 (2015).
31. L. Liu et al., "The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 971–980 (2015).
32. A. Bergamo, S. N. Sinha, and L. Torresani, "Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 763–770 (2013).
33. H. Jegou et al., "Aggregating local descriptors into a compact image representation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311 (2010).
34. F. Perronnin et al., "Large-scale image retrieval with compressed Fisher vectors," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3384–3391 (2010).
35. F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *European Conf. on Computer Vision (ECCV)*, pp. 119–133 (2010).
36. J. Xiao et al., "SUN database: large scale scene recognition from abbey to zoo," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492 (2010).
37. Y. Jia, "Caffe: an open source convolutional architecture for fast feature embedding," Eprint Arxiv: 675-678 (2014).
38. B. Zhou et al., "Learning deep features for scene recognition using places database," in *Conf. on Neural Information Processing Systems (NIPS)*, pp. 487–495 (2014).

Jinfu Yang received his PhD in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences in 2006. He is currently an associate professor with the College of Electronic Information and Control Engineering, Beijing University of Technology. His current research interests include pattern recognition, computer vision, and robot navigation.

Fei Yang received her BSc degree in mechatronic engineering from Beijing Information Science and Technology University, Beijing, China, in 2010. She is currently a master's degree candidate with the Department of Control Science and Engineering, Beijing University of Technology, Beijing, China. Her current research interests include deep learning and computer vision.

Guanghui Wang is currently an assistant professor at the University of Kansas, USA. He is also with the Institute of Automation, Chinese Academy of Sciences, as an adjunct professor. His research interests include computer vision, image processing, and robotics.

Mingai Li received her PhD from Beijing University of Technology, Beijing, China, in 2006. She is currently a professor with the School of Electronic Information and Control Engineering, Beijing University of Technology. Her current research interests mainly include brain computer interface, intelligent control, pattern recognition, and implementation of autonomous learning control technology for flexible two-wheeled upstanding robots.