Scene Recognition with Objectness

Xiaojuan Cheng, Jiwen Lu, Jianjiang Feng, Bo Yuan, Jie Zhou

PII: S0031-3203(17)30376-X
DOI: 10.1016/j.patcog.2017.09.025
Reference: PR 6292

To appear in: *Pattern Recognition*

Received date: 22 May 2017
Revised date: 6 September 2017
Accepted date: 13 September 2017

Please cite this article as: Xiaojuan Cheng, Jiwen Lu, Jianjiang Feng, Bo Yuan, Jie Zhou, Scene Recognition with Objectness, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.09.025

Highlights

- We have exploited the correlations of object configurations among different scenes to choose discriminative objects and represent image descriptors with the occurrence probabilities of discriminative objects, which eliminate the negative effects caused by common objects to enhance the inter-class discriminability.
- We have proposed a new method patch screening to prune the patches containing non-discriminative objects by the intersection of top scored objects in patches and the discriminative objects, so that we improve the generalized characteristics of the same scenes.

Scene Recognition with Objectness

Xiaojuan Cheng[a], Jiwen Lu[b,c,d,*], Jianjiang Feng[b,c,d], Bo Yuan[a], Jie Zhou[b,c,d]

[a]*Graduate School at Shenzhen, Tsinghua University, Shenzhen, 518055, China*
[b]*Department of Automation, Tsinghua University, Beijing, 100084, China*
[c]*State Key Lab of Intelligent Technologies and Systems, Beijing, 100084, China.*
[d]*Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, 100084, China.*

Abstract

In this paper, we present a feature description method called semantic descriptor with objectness (SDO) for scene recognition. Most existing scene representation methods exploit the characteristics of constituent objects in scenes with inter-class independence, which ignore the negative effects caused by the common objects among different scenes. The generic characteristics of the common objects cause some generality among different scenes, which weakens the discriminative characteristics among scenes. To address this problem, we exploit the correlations of object configurations among different scenes by the co-occurrence pattern of all objects across scenes to choose representative and discriminative objects which enhances the inter-class discriminability. Specifically, we capture the statistic information of objects appearing in each scene to compute the distribution of each object across scenes, which obtains the co-occurrence pattern of objects. Moreover, we represent the image descriptors with the occurrence probabilities of discriminative objects in image patches to eliminate the negative effects of common objects. To make image descriptors more discriminative, we discard the patches with non-discriminative objects to enhance the intra-class generalized characteristics. Experimental results on three widely used scene recognition datasets show that our method outperforms the state-of-the-art methods.

Keywords: Scene recognition, deep learning, co-occurrence pattern

*Corresponding author
Email address: lujiwen@tsinghua.edu.cn (Jiwen Lu)

1. Introduction

Scene recognition has been extensively investigated in computer vision and a variety of scene recognition methods have been proposed over the past decades [1, 2, 3, 4, 5, 6, 7, 8, 9]. Scene representation and scene classification are two important stages in a practical scene recognition system. Scene representation aims to extract discriminative features to make scene images distinguishable, while scene classification designs effective classifiers to differentiate scene categories. Compared with scene classification, scene representation significantly affects the performance of a scene recognition system, because it explores not only the generalized characteristics in the same category but also the distinctive characteristics among different categories. Specially, these characteristics are difficult to capture, due to the complexity of scene images where 1) the spatial layout of scenes often simultaneously exhibits characteristics across multiple distinct scene categories, and 2) the constituent objects vary widely in the same scene and usually occur in other scenes. Hence, how to extract more discriminative representation to enlarge the inter-class margins and reduce the intra-class variations simultaneously remains a central and challenging problem in scene recognition.

A number of scene feature representation methods have been proposed and they are mainly classified into two categories: hand-crafted features [10, 3, 11, 2, 12, 13] and learning-based features [14, 5, 15]. Typical hand-crafted features include generalized search trees (GIST) [16], oriented texture curves (OTC) [17] and census transform histogram (CENTRIST) [18], which investigate low-level visual information such as structural and textural information in scene images. However, these features are far from sufficient characteristics for complex scenes. The convolutional neural network (CNN) features are representative learning-based methods, which exploit high-level semantic information. While CNN features have achieved encouraging performance, they exploit the characteristics of the scenes with inter-class independence where the characteristics include general characteristics caused by common objects among different scenes, weakening the discriminability among scenes. Inspired by the fact that some common objects may occur simultaneously in different scenes with similar probability and some discriminative objects occur in a scene with high probabilities but

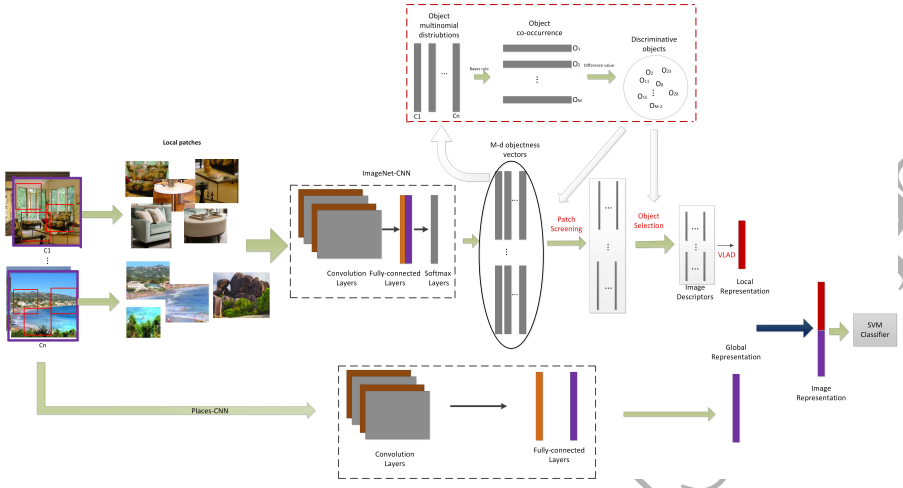


Figure 1: Pipeline of the proposed SDO approach for scene representation. We first extract object score vectors in local patches used to obtain discriminative objects. Then, we perform patch screening and object selection to make the score vectors more discriminative. Moreover, we project these vectors into low-dimensional discriminative object in the Euclidean space. We further cluster and pool these vectors to obtain local representation. Lastly, a combination of local and global representation extracted at fully-connected layer produces the final image representation.

rarely occur in others, we aim to exploit the co-occurrence pattern of objects across scenes to obtain discriminative objects for scene representation. The image local descriptors showing the occurrence probabilities of disciminative objects eliminate the negative effects of common objects.

In this paper, we propose a semantic descriptor with objectness (SDO) method for scene recognition where correlation of object configurations across scenes are exploited. We first compute object multinomial distribution based on object score vectors showing the occurrence probability of objects in local patches to obtain object configuration of each scene. Then, we derive the posterior probability of scenes given an object to exploit co-occurrence patterns of objects which is used to choose discriminative objects in scenes. Moreover, we extract image descriptors showing the appearance probabilities of discriminative objects in image local patches, which can be divided into three steps: 1) we prune the patches which do not contain discriminative objects, called

patch screening, by the intersection of top scored objects in patches and the discriminative objects; 2) we reduce the dimension of object score vectors to obtain discriminative vectors by choosing the elements representing discriminative object classes, called *object selection*; 3) we project the discriminative vectors in the non-Euclidean space into the Euclidean space, obtaining the image descriptors. We further cluster and pool image descriptors to obtain local representation. We also extract global representation on an entire image to exploit the global layout information of scenes. Lastly, we represent an image with the combination of global representation and local representation for scene recognition. Figure 1 shows the main work-flow of our proposed method. Experimental results on three widely used scene datasets, the Scene 15 [19], MIT Indoor 67 [20] and SUN 397 [21], show that our method outperform the state-of-the-art methods.

The rest of paper is organized as follows: Section 2 briefly reviews related studies. Section 3 illustrates the proposed method elaborately. Experimental results are presented and discussed in Section 4, followed by the conclusion in Section 5.

2. Related Work

In this section, we briefly review two related topics: 1) scene representation, and 2) scene classification.

2.1. Scene Representation

There have been extensive works on scene representation in the literature, and these methods can be mainly classified into two categories: hand-crafted feature representation and learning-based feature representation. A number of hand-crafted methods for scene recognition mainly extract holistic features and local features. Holistic features methods, such as GIST [16, 22], lexicographically convert an entire scene image into a high-dimensional feature vector but fail to exploit local structure information in scenes, especially the indoor scenes [18, 23] with complex spatial layouts. Unlike holistic features, local features first describe the structure pattern of each local patch and then combine the statistics of all patches into a concatenated feature vector. Typical local

features include OTC [17], CENTRIST [18] and mCENTRIST [23], which investigate structural properties and textural information in scene images. However, these hand-crafted methods only exploit the low-level visual features which are far from sufficient characteristics for the complex scenes. Moreover, they usually require strong prior knowledge to design them and some of them are computationally expensive, which may limit their practical applications.

Recently, learning-based methods [6, 5] have been widely used in scene recognition, especially the CNN models which learn high-level features from scene images. For examples, Zhou *et al.* utilized the CNN fully-connected features of an entire image to explore the global layout information for scene recognition [5]. In order to investigate the local structure of a scene, Gong *et al.* proposed multi-scale orderless pooling (MOP) method to extract fully-connected features on image local patches [6]. However, fully-connected features lack more general characteristics than the convolutional features. Yang and Ramanan proposed directly acyclic graph CNN (DAG-CNN) by leveraging multi-layer convolutional features and two fully-connected features for scene recognition [24]. While CNN features at fully-connected layer and convolutional layer have achieved encouraging performance, most of them investigate the characteristics of scenes with inter-class independence which ignores the effects caused by the common objects among scenes.

2.2. Scene Classification

A number of methods have been proposed for scene classification and they are mainly classified into two categories: generative models and discriminative models. Generative models often resort to hierarchical Bayesian to describe a scene [25, 26, 27, 28], which express various relations in a complex scene. For example, the models exploit the relation between a scene and its parts [25, 28], and the correlation among concurrent objects [3, 27]. The typical generative classifiers include hidden markov model (HMM) [29], markov random fields (MRF) [30] and latent dirichlet allocation (LDA) [31]. However, these models need to build complex probabilistic graph model and are computationally expensive.

The discriminative models [32, 33, 34, 35] extract dense descriptors from an im-

age and encode these descriptors into a fixed length representation to design a reasonable classifier for recognition. The representative classifiers include logistic regression, boosting and support vector machine (SVM). Specially, the SVM classifier has been widely adopted for scene classification. For example, the Bag of Visual Words trained a linear SVM with captured abundant semantic meaning of scenes [13]. Moreover, spatial pyramid matching (SPM) [19], object bank (OB) [34] and deformable part based model (DPM) [36, 35] are the representative examples of this model to train SVM classifiers on scene representations. Unlike the generative models, discriminative models learn the parameters easily to obtain scene representation.

3. Proposed Approach

In this section, we first present the motivation and framework of our SDO method and then detail the generation of local semantic descriptors. Lastly, we introduce how to use the descriptors to produce representation for scene classification.

3.1. Motivation

While many CNN based methods have achieved encouraging results on scene recognition, most of existing methods extract high-level features of each scene with inter-class independence, which ignores the effects of the common objects among different scenes. We are inspired by the fact that the objects have different occurrence probabilities in various scenes, as shown in Figure 2. The images in Figure 2 from three different scenes (*shoestore, bookstore and jewelleryshop*) may have the similar global layout seen in the left images of each scene, but they contain different objects where the jewellery occurs in jewelleryshop frequently but rarely occurs in shoestore and bookstore. We consider the discriminative object, jewellery, has a high probability in jewelleryshop but a low probability in shoestore and bookstore. On the other side, the persons and shelves are common in these scenes that may have similar probabilities. By the observation, we exploit the co-occurrence pattern of objects across scenes to choose the discriminative objects. A scene image can be represented as a bag of image descriptors showing the occurrence probabilities of discriminative objects, which eliminate the negative effects caused by common objects.

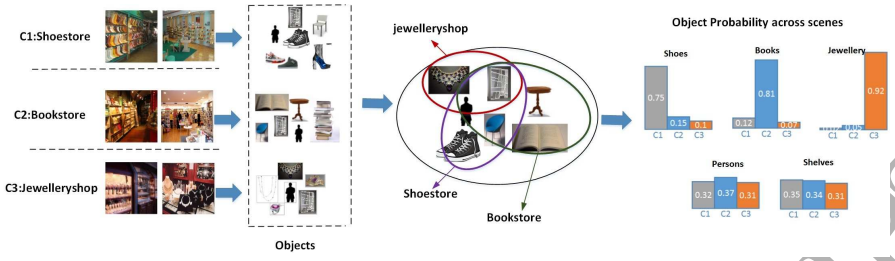


Figure 2: Illustration of different object classes with various occurrence probabilities across all the scene categories. These three scenes contain some discriminative objects (e.g., shoes in shoestore and jewellery in jewelleryshop) and common objects (e.g., persons and shelves) which have different distributions across different scenes

3.2. Framework of Our SDO

The framework of our proposed method is shown in Figure 1. Our method produces final scene representation by concatenating global representation and local representation. Both representations are built on deep CNNs, e.g., VGGNets, which are regarded as generic feature extractors for scene images. In the context of scene recognition, one CNN trained on scene-centric dataset, Places205, is used to obtain the global representation, and the other one is trained on ImageNet dataset for local representation.

To obtain the local representation, we first sample a set of patches in an image and feed them to the ImageNet-CNN, which produces a set of score vectors at softmax layer. We then compute object multinomial distributions of all scenes based on the score vectors and apply Bayes rule for posterior probabilities of scenes given objects. Finally, we choose the discriminative objects according to their discriminative power showed by posterior probabilities.

In the training phase, the discriminative object classes help us prune the patches in which the discriminative objects rarely occur. To further improve the discriminative power of the remaining patches, we perform object selection on score vectors by choosing the elements that represent discriminative object classes. Due to the non-Euclidean space nature of the score vectors which is difficult for encoding, we project the score vectors into linear Euclidean space by natural parametrization which produces the local

Algorithm 1 : SDO

Input: Training images:$\{I_i\}_{i=1}^{m}$ and test images: $\{I_i\}_{i=m+1}^{n}$. Two CNN models.

Output: Combined Representation $\{H_i\}_{i=1}^{n}$ for training and test images.

Procedure:

for $i = 1 \rightarrow n$ **do**
  Extract score vectors, each of which represents appearance probabilities of objects in a patch, on image I_i at softmax layer of the ImageNet-CNN.
  Extract features at the fully connected layer of Places-CNN on image I_i, obtain the global representation GR_i.
end for
for each scene category $c \in C$ **do**
  Process the score vectors, obtain object multinomial distribution $p(o|c)$.
  Apply Bayes rule, obtain posterior probability $p(c|o)$.
end for
Rank the object discriminative power $dis(o)$, select the top N discriminative object classes.
for $i = 1 \rightarrow n$ **do**
  Perform patch-screening and object selection, obtain image descriptors.
  Project image descriptors into Euclidean space.
  VLAD encoding, generate local representation LR_i.
  $H_i = [GR_i LR_i]$
end for

descriptors. To obtain image local representation, vector of locally aggregated descriptors (VLAD) encoding and principal component analysis (PCA) dimension reduction are applied on the local descriptors.

As the global representation, the holistic features are extracted at the second fully-connected layer of Places-CNN on entire image. The combination of the image global representation and local representation is fed into the linear SVM classifier for training. **Algorithm 1** summarizes the proposed SDO approach.

3.3. SDO Generation

Our SDO method aims to extract image descriptors that show the occurrence probabilities of discriminative objects. We compute object multinomial distributions in each

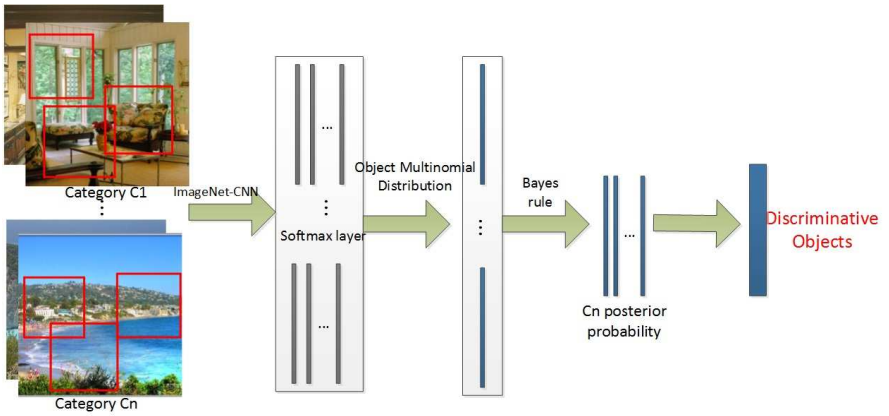


Figure 3: The procedure of discriminative objects based on the object multinomial distribution. The discriminative object classes are derived by ranking their discriminative power.

scene and derive the posterior probabilities of scenes given objects to obtain discriminative objects. In training phase, we first perform patch screening on image local patches to discard those containing non-discriminative objects. Then, object selection produces score vectors of discriminative objects in patches. Lastly, we map these vectors lied in non-Euclidean space into Euclidean space, which obtains image local descriptors.

Discriminative Objects. Each scene category contains many objects which are also common in other scene categories (e.g., the "walls" appear frequently in *bedroom, living room* and *office*). Those common objects such as walls are less discriminative than the objects "computers", or "keyboards" for the recognition of "computer room". We consider that an object has more discriminative power if it only appears frequently in some specific classes. To choose discriminative object classes, we explore the co-occurrence pattern of all objects among scenes and achieve discriminative objects based on their discriminative ability. We propose a procedure to compute the object multinomial distributions of scenes and obtain the distribution of each object across all scene, which is illustrated in Figure 3.

We obtain the object multinomial distributions of all scenes based on object score vectors at softmax layer of a deep CNN. The object multinomial distribution for each category shows the probability statistics of all object classes in a scene category. Specif-

ically, we sample a set of patches $P = [p_1, \ldots, p_i, \ldots, p_N]$ from an image and feed them to an ImageNet-CNN, e.g., VGGNet. For one patch, we obtain a 1000-dimensional score vector at softmax layer, where each element of the score vector represents the occurrence probability of a particular object class. Each image produces a set of score vectors $S = [s_1, \ldots, s_i, \ldots, s_N]$, where s_i is the score vector of the patch p_i. To obtain object multinomial distributions, we tentatively set a confidence level θ for the score vectors S, which enables us to detect the presence of object in a patch according to

$$\delta(x|\theta) = h[s_i(x) - \theta] \tag{1}$$

where $h(x) = 1, x \geq 0$ and $h(x) = 0$ otherwise. However, this method imposes the confidence level on all the object classes, which may miss some discriminative but infrequent classes. Thus we directly use the sum of the score vectors f_o without confidence level to detect the occurrence of object o in an image x as follows:

$$f_o(x) = \sum_{p_i \in x} s_i \tag{2}$$

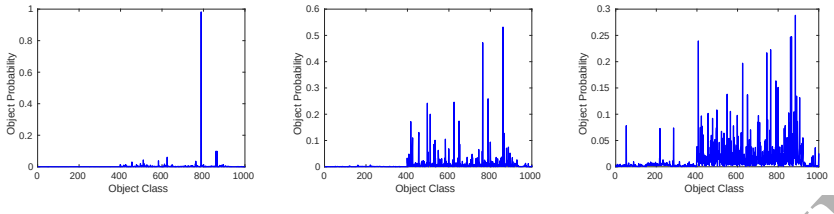
where p_i is a patch of the image x and s_i is the score vector of the patch p_i.

Generally, given a set of images I_c from a scene category c , The maximum likelihood probability of object o on class c is

$$p(o|c) = \frac{1}{N_{I_c}} \sum_{x_i \in I_c} f_o(x_i). \tag{3}$$

We refer to the probability vector $p(o|c)$ as the object multinomial distribution of c. We choose three scene categories (*shoestore, jewellery and museum*) to explore various object distributions and the results are shown in Figure 4.

Having obtained the object multinomial distributions, we obtain all normalized objects occurrence probabilities in all scene classes, which enables us to derive the pos-



(a) shoestore category (b) jewelleryshop category (c) museum category

Figure 4: Object multinomial distribution varies in different scene categories. (a) shows the object distribution in shoestore category, which is totally different from other categories, e.g., jewelleryshop in (b) and museum in (c).

terior probability of scene classes given the observation of all objects as follows:

$$P = \begin{bmatrix} p(c_1|o_1) & \dots & p(c_j|o_1) & \dots & p(c_C|o_1) \\ \vdots & & \vdots & & \vdots \\ p(c_1|o_i) & \dots & p(c_j|o_i) & \dots & p(c_C|o_i) \\ \vdots & & \vdots & & \vdots \\ p(c_1|o_N) & \dots & p(c_j|o_N) & \dots & p(c_C|o_N) \end{bmatrix} \tag{4}$$

where $p(c_j|o_i)$ is the posterior probability of a scene class c_j given the object o_i, by the application of Bayes rule

$$p(c_j|o_i) = \frac{p(o|c_j)p(c_j)}{\sum_j p(o|c_j)p(c_j)} \tag{5}$$

where $p(o|c_j)$ is the occurrence probabilities of objects given a scene class c_j shown in Equation 3 and $p(c_j)$ is a prior scene class probability $p(c = j) = 1/C$.

Given an object, the marginal probability $p(c|o_i)$ represents object co-occurrence probability which shows its discriminative power with respect to a set of scene classes C. To obtain the discriminative object classes, we try to utilize the entropy value based method. The posterior probability of non-discriminative object o tends to be uniform across classes with a low entropy value. The formula of entropy $E(o_i)$ is as follows:

$$E(o_i) = -\sum_{j=1}^{C} p(c_j|o_i) log_2(c_j|o_i) \tag{6}$$

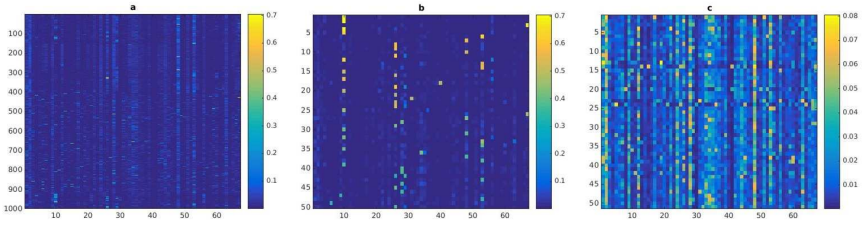


Figure 5: Object posterior probability for all scene. (a) given all objects; (b) given the least 50 discriminative objects; (c) given the most 50 discriminative objects.

where C is the number of scene classes and $p(i)$ is the posterior probability of a class. However, parts of the non-discriminative object classes may have higher entropy value than some discriminative object classes.

We choose discriminative objects based on the difference value method to find the largest difference value among objects. Having obtained the posterior probabilities based on Bayes rule, this method is performed in three steps. First, given an object o_i, the classes $c \in C$ are ranked according to the posterior probability$p(c|o_i)$. Let $r(c)$ be the ranking function, e.g., $r(1)$ for the class of largest probability and $r(C)$ for the class of lowest probability. The second step computes the discriminative power of object o by

$$dis(o_i) = \max_{r \in 1,\ldots,C-1} (r(i) - r(i+1)). \tag{7}$$

The largest probability difference of object o can be obtained by $dis(o)$ that shows the object discriminative power. The last step is to rank the $dis(o)$ and to choose top N objects, e.g., $N = 500$. Figure 5 shows the examples of posterior probability with various objects on the MIT indoor 67 dataset. The object classes in yellow illustrate higher discriminative power than those in blue. The subfigure (a) shows posterior probabilities of scenes given all objects while (b) and (c) illustrate 50 least discriminative classes and 50 most discriminative classes, respectively.

Patch Screening. We aim to discard the patches containing non-discriminative objects that ruin the generalized characteristics of a scene. For each category, there typically exists a set of representative regions that frequently appear in the images.

For example, regions with computer monitors frequently appear in the images for the "computer room" class, which we regarded as representative regions. Meanwhile, the non-representative regions, only appearing in a few images, can be viewed as outliers for a certain scene category. Conventional methods use one-class SVM to prune outliers which suppose the outliers being scattered. However, this method is computationally expensive, especially in large dataset, and it does not take some infrequent but discriminative patches into consideration. In our method, we use the intersection of top N scored object classes in a patch and the discriminative classes to help us prune outliers.

Let $S = \{s_1, s_2, \ldots, s_P\}(s_i \in R^D, D = 1000)$ be a set of the score vectors in an image. We rank the score vector s_i, choose the top N object classes with largest scores and record the set of object classes P_{io}. We define D_o as the set of discriminative object classes. We obtain the intersection I_{io} of P_{io} and D_o, as the following equation

$$I_{io} = \frac{length(P_{io} \cap D_o)}{N}. \tag{8}$$

The intersection area illustrates the discriminative power of the object classes appearing in a patch. We consider the patch as a noisy patch, when ratio I_o is quite low. In our approach, we set a threshold ϑ and select the meaningful patches by the comparison between I_o and ϑ with the following equation

$$P = \begin{cases} 1 & I_o \geq \vartheta \\ 0 & I_o < \vartheta \end{cases} \tag{9}$$

where $P = 1$ represents the patch is chosen, and otherwise is discarded .

Local Semantic Descriptors. Having obtained the representative patches, we extract the object score vectors of these patches. To make the vectors more discriminative, we perform object selection to choose N elements in score vectors which represent the discriminative objects. However, the simplex of the score vectors is non-Euclidean, which makes the image encoding difficult to learn. The problem can be addressed by natural parameterization which transforms the vectors into linear Euclidean space, yielding a natural parameter vector ν of the score vectors π. We apply different param-

eters to achieve projection and the possible projection functions are as follows:

$$\begin{cases} \nu_k^{(1)} = & \sqrt{\pi_k} \\ \nu_k^{(2)} = & \log \pi_k \\ \nu_k^{(3)} = & \log \dfrac{\pi_k}{\pi_S} \end{cases} \tag{10}$$

where π_k is the discriminative score vector in non-Euclidean space, π_S is the sum of the entries in π_k, and ν_k is the new vectors in linear Euclidean, namely image local descriptor. In our SDO, the natural parameter is obtained by a log transformation $\nu_k = log\pi_k$ that maps the high-nonlinear space into the linear Euclidean space, which makes the image embedding substantially easier and more accurate.

3.4. Scene Representation and Recognition

Our proposed SDO considers that any scene image can be represented as a bag of image local descriptors. We cluster and pool these image descriptors by VLAD to obtain the image embedding as the local representation.

Given the local descriptors, we aim to find the centers of clusters. We tentatively use three VLAD centers generation methods including object multinomial distributions $P(o|c)$, posterior probability of a scene $P(c|o)$ and K-means clustering centers C.

For the first method, we use object multinomial distributions

$$P(o|c) = \{p(o|c_1) \quad p(o|c_2) \dots p(o|c_C)\} \tag{11}$$

as VLAD centers with C centers which is the number of scene classes. For the second method, we regard the marginal probability vectors

$$P(c|o) = \{p(c|o_1) \quad p(c|o_2) \dots p(c|o_N)\} \tag{12}$$

as the VLAD centers with N centers that is the number of object classes.

We also learn a separate k-means codebook with $k = 100$ centers. Having obtained a collection of descriptors in patches and a codebook of centers $c_i, i = 1, \dots, k$, the VLAD embedding is constructed by assigning each patch ν_j to its r nearest cluster

centers $rNN(\nu_j)$ and aggregating the residuals of the patches minus the center as the following equation:

$$r_i = \sum_{j:c_i \in rNN(\nu_j)} w_{j1}(\nu_j - c_i) \tag{13}$$

where w_{jk} is the Gaussian kernel similarity between p_j and c_k. For each patch, we additionally normalize its weights to its nearest r centers. In this paper, we use $r = 5$ and kernel standard deviation of 1. The VLAD embedding is as follows:

$$x = [r_1 \quad r_2 \dots r_k] \tag{14}$$

Following [14], we normalize the pooled vectors x with $L2$ normalization. However, the resulting vectors are high dimensional: given N-dimensional patch ν_j and clustering centers C, e.g., $N = 500$ and $C = 100$ k-means centers, we end up with the embedding with $100 \times N$ dimensions which make classifier training bear computational burden. Thus we perform PCA on the pooled vectors and reduce them to 4096 dimensions.

To obtain global layout information, we extract the global feature representation of the entire image at the second fully-connected layer from Places-CNN trained on scene-centric dataset, Places205. Although the basic architecture of Places-CNN is the same with the ImageNet-CNN, the type of the learned features are very different. The convolutional units of ImageNet-CNN respond to object-like occurrences, while those in Places-CNN are selective of landscapes with more spatial features. We concatenate local representation and global representation to produce final image representation which is used for training a linear SVM classifier.

3.5. Implementation Details

In this section, we present the implementation details of our experiments. For scene recognition, we utilize two CNN models, VGG-16 nets, which are based on their caffe implementations. Specifically, one VGG-16 net is trained on ImageNet dataset that is used to obtain local representation based on score vectors extracted at softmax layer. Another VGG-16 net trained on Places205 dataset is applied for obtaining the fully-

connected feature of the entire image, regarded as global representation. In our experiments, we first resize all the images into resolution of 256×256.

For local representation, the score vectors of local $P \times P$ image patches on a uniform grid $stride$ are extracted from ImageNet-CNN. Our experiments are performed with $P = 128$, $stride = 32$ at first. To deal with the large intra-class variations, we design a multi-scale sampling strategy to extract image patches with a set of sizes $s = \{64, 96, 112, 128, 144, 160, 176, 192\}$. Given all score vectors of the local patches, we compute object multinomial distribution of each scene without confidence level and derive posterior probabilities of scenes. We choose 500 discriminative objects according to their discriminative power with difference value method. In the patch screening, We set the intersection threshold 30% of 500 discriminative objects and top 500 scored objects in test patch. Object selection produces 500-dimensional vectors by selecting 500 elements in 1000-dimensional object score vectors of representative patches which represent the discriminative object classes. We project 500-dimensional vectors into Euclidean space by log function, obtaining image local descriptors. To obtain local representation of the image, we use VLAD encoding to obtain image embedding and adopt PCA to reduce dimensionality to 4096 dimensions. Moreover, to obtain more information, we utilize the fully-connected $4096 - d$ features of the patches to concatenate the local representation.

The global features are extracted at the 2^{th} fully-connected layer of Places-CNN and L_2 normalized. The fully-connected features of the Places-CNN investigate the holistic layout information of scenes which are complementary to our local representation. The final image representations consisting of local and global representations are used to train a linear SVM classifier.

4. Experiments

We evaluate our SDO method on three widely used scene datasets including the MIT Indoor 67 [20], SUN 397 [21] and Scene 15 [19]. Specifically, the Scene 15 and SUN 397 datasets are employed to show the effectiveness of our approach for both indoor and outdoor scenes, and the MIT Indoor 67 dataset is used to show the

effectiveness of our approach for indoor scene images. The following describes the details of the experiments and results.

4.1. Datesets

MIT Indoor 67 Dataset. MIT Indoor 67 [20] is a challenging indoor scene dataset, which contains 67 scene categories and 15,620 color images. The number of images varies across categories with at least 100 images per category. Following the standard evaluation protocol of [20], we use 80 images from each category for training and another 20 images for testing. Figure 6 shows some indoor scene images from the MIT Indoor 67 dataset.

SUN 397 Dataset. SUN 397 [21] is a large-scale scene dataset, which contains 397 scene categories and 108,754 color images with at least 100 images per category. The categories include different kinds of indoor and outdoor scenes which show tremendous object and alignment variance, bringing more complexity for scene recognition. Follow the standard evaluation protocol provided by [21], we train and test our proposed approach on ten different partitions, each of which has 50 training and 50 test images. The partitions are fixed and publicly available from [21]. The average classification accuracy is selected to evaluate our approach. Figure 7 shows the images of indoor swimming pool and outdoor swimming pool from the SUN 397 dataset with different labels.

Scene 15 Dataset. Scene 15 dataset [19] contains 4485 gray images of 15 different scenes including both indoor scenes and outdoor scenes. The dataset does not provide separated training and test sets, so we use 5 random splits and compute the mean of the classification performance across splits. In each split, we use 100 training images for each category and the remaining for the test. Figure 8 shows some indoor and outdoor images from the Scene 15 dataset.

4.1.1. Comparison with the State-of-the-Arts

We compare the performance of our proposed SDO with recent approaches on MIT Indoor 67, SUN 397 and Scene 15. In our approach, the global representations are obtained from the VGG-16 net trained on Places205 database. Our local representations

Figure 6: Several indoor scene examples from the MIT Indoor 67 dataset. The images in two rows are from two different scene categories (*computerroom* and *restaurant*), respectively.

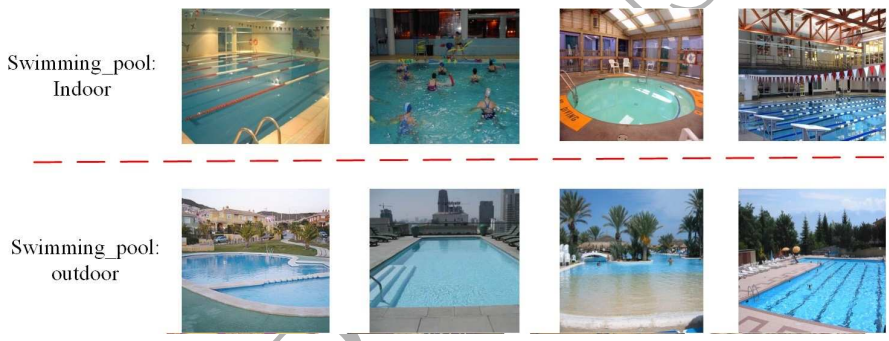


Figure 7: Several indoor and outdoor examples of swimming pool from the SUN 397 dataset.

consist of fully-connected features and probability features with multi-scale patches. The final representation of an image consists of the global representation and local representation. The results on MIT Indoor 67, SUN 397 and Scene 15 are shown in Table 1, Table 2 and Table 3, respectively.

As shown in Table 1, Table 2 and Table 3, the performance of our SDO outperforms the previous method. We explore the complementary properties of our SDO with multi-scale patches and the fully-connected features. From the tables, we see that the multi-scale patches improve recognition accuracy by reducing the intra-class variations. Meanwhile, the fully-connected features contain more abundant information that benefit our performance.

Table 1 tabulates the performance of our method, traditional hand-crafted methods

Table 1: Comparison of our proposed approach with other methods on MIT Indoor 67 dataset.

Traditional Methods	**Accuracy**(%)
ROI[20]	26.05
DPM [36]	30.40
CENTRIST [18]	36.90
Object Bank [34]	37.60
RBOW [13]	37.93
Discriminative Patches [37]	38.10
Hybrid parts [38]	39.80
BOP [39]	46.10
Hybrid parts+GIST-color+SP [38]	47.20
ISPR [40]	50.10
Discriminative parts [41]	51.40
DSFL[42]	52.24
Discriminative Lie Group [43]	55.58
IFV [39]	60.77
IFV + BOP [39]	63.10
Mode-Seeking + IFV [44]	66.87
ISPR + IFV [40]	68.50
CNN based Methods	**Accuracy** (%)
PlaceNet [5]	68.24
MOP-CNN [6]	68.90
CNNaug-SVM [45]	69.00
HybridNet [5]	70.80
URDL + CNNaug [46]	71.90
MPP-FCR2(7 scales) [15]	75.67
DSFL + CNN[42]	76.23
MPP + DSFL[15]	80.78
CFV(VGG-19) [47]	81.00
CS(VGG-19)[48]	82.24
VSAD[49]	86.20
Our SDO $\{128\}$	83.98
Our SDO $\{64, 128, 192\}$	84.86
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$	85.12
Our SDO $\{128\}$ + fc features	85.43
Our SDO $\{64, 128, 192\}$ + fc features	86.00
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$ + fc features	**86.76**

Figure 8: Several indoor and outdoor examples from the Scene 15 dataset. The images in the first row are from kitchen scene and others are from forest scene.

and CNN based methods on MIT Indoor 67 dataset. The results show that our proposed SDO achieves the best performance. For the traditional hand-crafted methods (e.g., DPM and CENTRIST), they only exploit the low-level visual information such as the texture, color and structure information while our SDO investigates high-level semantic information built on the CNN architectures. The conventional CNN based methods mainly extract the characteristics of each scene with inter-class independence, ignoring the generality of characteristics caused by common objects among different scenes. Our method takes advantage of the co-occurrence information of constituent objects in scenes to choose discriminative objects and achieves best performance.

We then evaluate our proposed SDO on the SUN 397 dataset and compare with several traditional hand-crafted methods and CNN based methods. Table 2 records the recognition accuracy of our SDO and other methods where our SDO achieves the highest recognition rate. The traditional hand-crafted methods, such as contextBoW and OTC+HOG2$\times$2, exploit low-level visual features which are far from sufficient characteristics to represent the large scale scenes with tremendous object and alignment variance. For the methods based on CNN (e.g., MOP-CNN and CS), they mostly apply mid-level convolutional features and high-level fully-connected features. However, these features are extracted with inter-class independence without considering the correlation among different scenes. Our SDO method explores object co-occurrence patterns to eliminate the effects of common objects and achieve best performance.

Table 2: Comparison of our proposed approach with other methods on SUN397 dataset.

Traditional Methods	**Accuracy** (%)
S-manifold [50]	28.90
OTC [17]	34.56
contextBoW + semantic[51]	35.60
Xiao*et al* [21]	38.00
FV (SIFT + Local Color Statistic) [52]	47.20
OTC + HOG2×2 [17]	49.60
CNN based Methods	**Accuracy** (%)
Decaf [53]	40.94
MOP-CNN [6]	51.98
HybridNet [5]	53.86
Places-CNN [5]	54.23
Places-CNN ft [5]	56.20
CS(VGG-19)[54]	64.53
VSAD[49]	73.00
Our SDO $\{128\}$	66.98
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$	72.06
Our SDO $\{128\}$ + fc features	69.78
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$ + fc features	**73.41**

To verify the effectiveness of our method, we also compare our SDO with traditional hand-crafted methods on the Scene 15 dataset in Table 3. We see that our proposed SDO achieves highest recognition rate. The reason is that our approach is built on the top of the CNN architecture that extracts high-level semantic information, while the hand-crafted methods, such as topic models (LDA, pLSA) and contextual models (SMN, CMN), only exploit low-level visual features.

4.2. Effects of Our SDO Components

In this section, we evaluate the parameters of important components in our SDO. First, we compare the performance of different nets trained on ImageNet and Places205 datasets. Then, we evaluate the influence of patch size and the selection of discriminative objects. Having obtained the discriminative objects, we represent the image following the steps of patch-screening, natural parameterization and image encoding. Therefore, we also study the effectiveness of each step.

Table 3: Comparison of our proposed approach with other methods on Scene 15 dataset.

Methods	**Accuracy** (%)
LDA[19]	59.0
pLSA[19]	63.3
SMN [55]	71.7
BoW[19]	74.8
CMN [55]	77.2
ObjectBank [56]	80.9
Kernel descriptor [57]	82.2
SPMSM [50]	82.5
SR-LSR [56]	85.7
EMFS [58]	85.7
Object-to-Class kernels[59]	88.8
Our SDO $\{128\}$	94.37
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$	95.56
Our SDO $\{128\}$ + fc features	94.97
Our SDO $\{64, 80, 96, 112, 128, 144, 160, 176, 192\}$ + fc features	**95.88**

4.2.1. Evaluation on Different Pre-trained Nets

In our approach, we use two pre-trained VGGNets, one of which is trained on ImageNet dataset for image local representation and another trained on Places205 dataset for image global representation. To better show the advantages of the networks in our SDO, we investigate the performance of different networks including AlexNet [60], VGGNet [61] and GoogLeNet [8] trained on different datasets. We compare the performance of differents trained on ImageNet dataset and Places205 dataset for both local representation and global representation, respectively. We conduct the experiment on the Scene 15 dataset.

Table 4 lists the results that our SDO with VGGNet trained on the ImageNet dataset for image local representation and VGGNet trained on the Places205 dataset for global representation obtains the best performance. It is because the VGGNet, a 16 layers network, captures more scene patterns than AlexNet and GoogLeNet. For local representation, the network is applied on local patches that tend to be an object, so the network trained on the ImageNet dataset containing only several objects per images, extracts more accurate features of the patches. For global representation on the entire scene image, the network trained on the Places205 dataset consisting of large numbers

Table 4: Comparison of our proposed approach with different networks trained on various datasets for both local and global representation on Scene 15.

LR	GR	Accuracy
AlexNet (ImageNet)	AlexNet (ImageNet)	89.04
AlexNet (Places205)	AlexNet (Places205)	90.39
AlexNet (Places205)	AlexNet (ImageNet)	88.56
AlexNet (ImageNet)	AlexNet (Places205)	91.65
GoogLeNet (ImageNet)	GoogLeNet (Places205)	93.01
VGGNet (ImageNet)	VGGNet (Places205)	**94.37**

Table 5: Comparison of different representations on MITIndoor67, SUN397 and Scene15 datasets.

	Accuracy with GR	Accuracy with LR	Accuracy with concatenation
Scene15	92.36 %	87.34 %	94.37 %
MITIndoor67	79.76 %	68.07 %	83.98 %
SUN397	64.18 %	54.84 %	66.98 %

of scene images, improves the capacity of the scene features.

To evaluate the contribution of local and global representation, we perform experiment on Scene 15, MIT Indoor 67 and SUN 397 datasets. Table 5 records the accuracy of local representation (LR), global representation (GR) and the concatenation of them, respectively. The results show that the global representation is more accurate than the local representation. With the combination of these two types of representations, our performance become better than any one of them since these two representations are complementary.

The networks trained on Places205 are applied on our target datasets to obtain global representation. We fine-tune the VGGNet on the MIT Indoor 67 dataset and extract global representation to evaluate the performance of fine-tuning. We set the learning rate $base\_lr = 0.001$ and use learning rate decay policy with *step* policy. In this policy, we return learning rate with $base\_lr * gamma^{(floor(iter/stepsize))}$ where we set $gamma = 0.1, stepsize = 2000$. We set the fixed weight decay rate with 0.0005. We extract the global representation at the second fully-connected layer. Table 6 tabulates that our SDO with the pre-trained VGGNet without fine-tuning obtains the best performance. Because the VGGNet without fine-tuning trained on a large scene-centric

Table 6: Comparison of VGGNet with fine-tuning and without fine-tuning on MIT Indoor 67 dataset.

	Accuracy with GR (%)	**Accuracy with LR + GR** (%)
VGGNet with fine-tuning	78.94	83.05
VGGNet without fine-tuning	79.76	**83.98**

Table 7: Comparison of different patch sizes with local representation and the combination of local representation (LR) and global representation (LR + GR) on MIT indoor 67 dataset.

Patch Size	**Accuracy with LR** (%)	**Accuracy with (LR + GR)**(%)
$P = 64$	64.38	81.94
$P = 96$	66.12	82.96
$\boldsymbol{P = 128}$	68.07	**83.98**
$P = 160$	66.32	83.01

dataset produces more general scene features. Moreover, our target dateset MIT Indoor 67 and the source dataset Places205 are overlapping that causes the fine-tuning network less general for scene characteristics.

4.2.2. Evaluation on Different Patch Sizes

Our SDO is implemented by first extracting image patches with 128×128 pixels. An indoor scene image from the MIT Indoor 67 dataset contains many objects with various sizes and scales, thus the object information varies with different patch sizes. We set our experiment with different patch sizes of 64×64, 96×96 ,128×128 and 160×160, and compare the performance of different patch sizes in our SDO. Table 7 shows that our SDO with the patch size of 128×128 obtains the best performance, because the 128×128 patch is the appropriate size to contain more accurate objects information. The patches with smaller size may not contain sufficient object information, resulting in the score vectors unable to reflect reasonable object classes. As for the larger patches, they contain multiple objects with different sizes and scales, which also makes the score vectors less accurate.

To evaluate the patch sampling strategy, we also consider using RPN method in [62] to extract the patches. The RPN method produces 200 region proposals at most with some very small regions and overlapping regions. We conduct our experiments with the proposals after three operations. One uses all the proposals, one uses the proposals after removing the small regions less than $1/64$ of the whole image area, and the other uses

Table 8: Comparison of different RPN regions on MIT indoor 67 dataset.

RPN	**Accuracy**(%)
all regions	82.96
without small regions	83.24
without overlapping regions	82.75
sliding windows	**83.98**

Table 9: Image numbers on MITIndoor67 training dataset containing different patch numbers with three types of regions.

Patch numbers (P)	**with all regions**	**no small regions**	**no overlapping regions**
$P \leq 50$	353	498	1553
$50 < P \leq 100$	642	2727	2167
$100 < P \leq 150$	577	1907	1461
$150 < P \leq 200$	3788	228	179
Total Images	5360	5360	5360

the proposals after removing the overlapping regions with $IoU > 0.7$, respectively.

From the results in Table 8, we see that the RPN method without small regions achieves better performance than other two operation strategies. The small regions contain some useless information which make the final representation inaccurate. After we prune the small regions, the performance becomes better than that with all regions. The overlapping regions make the region information redundant that ruin the representation. After we prune the overlapping regions, there may be less remaining regions which may discard some useful information. So this processing cause worse performance than that without small regions. Although the RPN achieves good performance, there are some reasons that RPN method is a little worse than sliding window method. Firstly, it extracts patches sparsely,leading to varying region numbers in different images as shown in Table 9. The various region numbers lead to uneven data distribution which affect the statistic information in our SDO. Secondly, the regions may not cover the entire image with small region numbers that discard some useful information while the images with large region numbers result in redundant information. Both of them affect the effective representation.

Table 10: The performance with different confidence levels on SUN 397 dataset.

confidence level θ	**Accuracy**$(\%)$
$\theta = 0.1$	59.93
$\theta = 0.2$	62.54
$\theta = 0.3$	61.79
$\theta = 0$	**66.98**

4.2.3. Effect of Discriminative Objects

Our SDO chooses 500 discriminative object classes with the difference value method described in Section 3.3. To investigate the effect of discriminative objects, we evaluate the performance of various selection methods and the numbers of discriminative object classes on SUN 397 dataset.

Computing the object multinomial distributions is the key step to choose discriminative objects. We first compare the effect of different confidence levels for obtaining object multinomial distribution. The experimental results in Table 10 show that our approach without confidence level obtains the highest recognition rate. When setting the confidence level, we impose the same confidence level on all the object classes O, which discards some discriminative but infrequent objects, leading to the negative performance.

We then compare recognition performance with different selection methods based on entropy value and difference value. Meanwhile, we conduct experiment to verify the effect of the number of discriminative object classes. Figure 9 shows the performance of different methods for obtaining discriminative objects on various classes numbers. We see that our SDO based on difference value method achieves better performance than entropy value method. This is mainly because parts of the non-discriminative object classes may have higher entropy value than some discriminative object classes. Figure 9 also demonstrates that the number of the discriminative object classes varying from 200 to 1000 also affects classification performance. The image descriptors with less object classes may discard some discriminative classes while the descriptors with more object classes may contain non-discriminative objects, which weakens the discriminability among scenes. Therefore, we choose $N = 500$ object classes to achieve the optimal result.

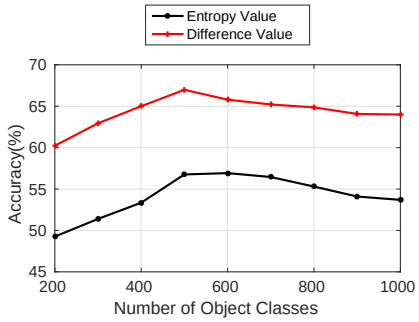


Figure 9: Recognition performance on SUN 397 with different methods at various number of discriminative object classes from 200 to 1000.

Table 11: Classification accuracy based on different discriminative object classes on SUN 397 dataset.

Methods	**Accuracy** (%)
Top	**66.98**
Random	63.54
Least	56.89
Without selection	64.73

Furthermore, we choose various strategies to pick out $N = 500$ object classes in all 1000 object classes. Three methods are conducted including top difference value based, least difference value based and random difference value based. Table 11 presents our SDO with the top difference values among object classes obtains the best performance due to its stronger discriminative ability than the others.

4.2.4. Effect of Patch Screening

In our approach, we use the intersection of top 500 scored objects in a test patch and the discriminative objects to help patch screening. We tentatively use the one-class SVM to prune the non-representative patches and set different outlier ratios $r = \{0, 10\%, 20\%, 30\%\}$. To investigate the effect of patch screening in our SDO, we also set various intersection ratios and compare their performance on MIT Indoor 67. Table 12 tabulates that the intersection ratio with 30% in our SDO achieves best performance. The one-class SVM method only prune the non-representative patches in intra-class scene and is quite time-consuming. Our method not only prune the non-

Table 12: Classification accuracy with various outlier ratios and intersection ratios on MIT indoor 67 dataset.

outlier ratio(%)	**Accuracy with LR** (%)	**Accuracy with LR + GR** (%)
0	65.53	82.27
10	66.10	82.78
20	66.72	82.89
30	65.44	82.06
Intersection Ratio	**Accuracy with GR** (%)	**Accuracy with LR + GR** (%)
$\vartheta = 0$	65.53	82.27
$\vartheta = 0.1$	65.82	82.66
$\vartheta = 0.2$	67.06	83.14
$\boldsymbol{\vartheta = 0.3}$	68.07	**83.98**
$\vartheta = 0.4$	66.32	82.91
$\vartheta = 0.5$	65.22	82.20

Table 13: Classification accuracy on the SUN 397 dataset with various projection functions.

Parameterization	**Accuracy** (%)
without parameterization	64.56
$\nu^{(1)}$ function	64.17
$\boldsymbol{\nu^{(2)}}$ function	**66.98**
$\nu^{(3)}$ function	65.46

representative patches intra-class but the non-discriminative patches inter-class. In our SDO, the recognition performance is less than 66% with low intersection ratios, because the low ratios may keep the noisy patches which do not contain the discriminative information, ruining the generalized characteristic in the same scene. Moreover, the high intersection ratios may discard some discriminative patches, which results in the recognition accuracy declining to 65% with the ratio up to 50%.

4.2.5. Effect of Natural Parameterization

In our approach, we project the discriminative vectors into linear Euclidean space after object selection on score vectors. We organize our experiment to verify the effect of natural parameterization with various projection functions on SUN 397. Table 13 shows that our SDO with projection function $\nu^{(2)}$ obtains better performance than other projection functions. Without natural parameterization, we achieve only 64.56% while function $\nu^{(2)}$ improve performance to 66.98%. The other projection functions perhaps miss some important information of image descriptors.

Table 14: Results of classification accuracy considering different centers on SUN 397 dataset.

Center	**Accuracy**(%)
object multinomial distribution $p(o\|c)$	61.48
posterior probability $p(c\|o)$	59.74
K-means centers $\boldsymbol{c}$	**66.98**

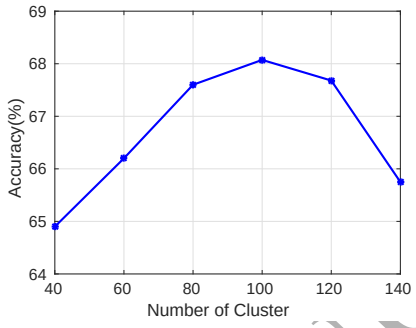


Figure 10: The line graph depicts the recognition rate on SUN 397 dataset with various number of VLAD centers.

4.2.6. Effect of Image Encoding

We apply VLAD encoding on image local descriptors to derive the embedding of an image. In our approach, we utilize K-means to obtain the clustering centers. To investigate the effect of VLAD encoding, we implement the experiment on the SUN397 dataset with different clustering centers and various center numbers. Table 14 tabulates the performance of VLAD centers generation methods including object multinomial distribution $p(o|c)$, posterior probability distribution $p(c|o)$ and K-means clustering c. K-means method produces more reasonable semantic clusters and achieves the best performance. Figure 10 shows the effect of different numbers of VLAD centers. The unreasonable cluster numbers produce poor generality of the semantic information of image local descriptors.

4.2.7. Effect of Feature Dimension

Given 500-dimensional image descriptors and 100 clusters, we end up with an image embedding with 50000 dimensions by VLAD encoding. The high dimensionality results in large computation when training a classifier, so it is necessary to adopt PCA

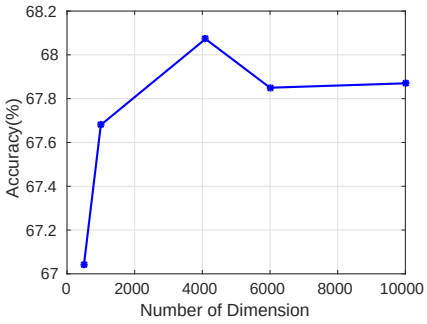


Figure 11: Different effects of dimension reduction on SUN 397 dataset

Table 15: Accuracy of representation with normalization and without normalization on MITIndoor67 dataset.

local features	**global features**	**Accuracy**(%)
without normalization	without normalization	83.04
without normalization	with normalization	83.43
with normalization	without normalization	83.72
with normalization	**with normalization**	**83.98**

to reduce the dimension. We evaluate the performance of our SDO with various feature dimensions on SUN 397. Figure 11 shows that our approach with 4096-dimensional features achieves the best performance. Because the embedding with low dimensionality possibly discards some useful information, while the high dimensionality brings the unnecessary computational burden. The best embedding with 4096-dimension not only keeps the principle information but also makes the training time acceptable.

We also evaluate the performance of normalization of local and global features. Table 15 tabulates the representation with feature normalization achieves better performance.

4.3. Evaluation on End-to-End Network

In our SDO, we obtain the representation by several steps, such as patch screening, object selection and image encoding. We tentatively design the end-to-end DNN by resorting to the idea of RPN instead of these sub-modules. The RPN method used VGG-16 network and apply a 3×3 convolutional kernel on $conv5\_3$ feature maps to

produce a low-dimensional vector with $512d$ for each receptive field. Each $512d$ vector is regarded as the feature of the corresponding region in original image. Motivated by this idea, we design our network based on VGG-16 net with the shared $conv5\_3$ feature maps which are fed into two sibling nets. One subnet is used to produce the patch labels and the other is used to produce image labels. The ground-truth patch labels are obtained by the pre-trained model on ImageNet database. To obtain the patch labels, we apply 3×3 convolutional kernel on the shared $conv5\_3$ feature maps and the fully-connected layer on each $512d$ vector to produce the patch labels matrix, which achieves patch screening and object selection procedures. We then fed the label matrix into the pooling layer and fully-connected layer to obtain the local patch representation which achieves patch encoding procedure instead of VLAD method. In another net, we utilize the pooling layer and fully-connected layer as the same with the VGG-16 network to obtain the global representation and concatenate the local representation to feed the softmax layer. In the training phase, we minimize an objective function following the multi-task loss where the loss function for an image is defined as:

$$L(p_i, I) = \sum_i L_p(p_i, p_i^*) + L_I(I, I^*) \tag{15}$$

where the p^* and I^* are the ground-truth patch labels and ground-truth image label, respectively. p_i is the patch label and I is the output of the softmax layer. Specifically, we apply 3×3 convolutional kernel with $pad = 1, stride = 1$ on the shared $conv5\_3$ feature maps to produce a convolution feature map of a size $512 \times 14 \times 14$. Each $512d$ vector on location of 14×14 map is regarded as the feature of the corresponding region in image. The location (i, j) can be seen the center (w, h) of the region in image with the correspondence $(w, h) = 16 \times (i, j)$. Having obtained the $512d$ vectors, we use fully-connected layer on each vector to produce the patch labels with the output of a size $501 \times 14 \times 14$. We design our experiment by resizing the image into 224×224 pixels and extracting 196 patches according to the patch center locations with the patch size 128×128. We feed the patches into the pre-trained ImageNet to achieve the ground-truth patch labels. Meanwhile, we compute the objects statistic information which occur in the patches to choose the 500 discriminative objects as that in our

Table 16: Accuracy with end-to-end network on MIT Indoor 67 dataset.

Method	Accuracy
end-to-end	81.02 %
our SDO	**83.98** %

SDO. The label dimension is $501d$ including 500 discriminative objects and one non-discriminative objects. If the patch label is not belong to the discriminative objects, we regard its label as non-discriminative object which means discarding this patch. Having obtained the patch labels, we utilize the pooling layer and fully-connected layer to obtain the patch representation. The network concatenates the patch representation and the global representation to predict the scene labels.

Table 16 shows the performance of end-to-end network on MIT Indoor 67 dataset. We see that our SDO is better than the end-to-end network. Because the ground-truth patch labels are regarded as the object classes with highest probability at softmax layer on the pre-trained ImageNet. The inaccurate patch labels have effects on the patch screening, which may discards some useful patches. Meanwhile, the objective function in training phase is not appropriate with the patch labels to some extent.

5. Conclusion

We have proposed in this paper a new semantic descriptor with objectness method for scene recognition to exploit the correlation of object configurations across scenes. We have chosen the discriminative objects by exploring co-occurrence pattern of objects across scenes. We have eliminated the negative effects of common objects among different scenes by representing the descriptor with the occurrence probabilities of discriminative objects. Experiments on three benchmark scene datasets are presented to demonstrate the efficiency of the proposed approach.

Acknowledgement

This work is supported by the National Key Research and Development Program of China under Grant 2016YFB1001001, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the

National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

References

[1] B. M. Luo J, Natural scene classification using overcomplete ica, PR 38 (10) (2005) 1507 – 1519.

[2] L. Cao, L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: ICCV, 2007, pp. 1–8.

[3] M. J. Choi, J. J. Lim, A. Torralba, Exploiting hierarchical context on a large database of object categories, in: CVPR, 2010, pp. 129–136.

[4] J. Yu, D. Tao, Y. Rui, Pairwise constraints based multiview features fusion for scene classification, PR 46 (2) (2013) 483 – 496.

[5] B. Zhou, A. Lapedriza, J. Xiao, Learning deep features for scene recognition using places database, in: NIPS, 2014, pp. 487–495.

[6] Y. Gong, L. Wang, R. Guo, Multi-scale orderless pooling of deep convolutional activation features, in: ECCV, 2014, pp. 392–407.

[7] R. Wu, B. Wang, W. Wang, Harvesting discriminative meta objects with deep cnn features for scene classification, in: ICCV, 2015, pp. 1287–1295.

[8] C. Szegedy, W. Liu, Y. Jia, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9.

[9] C. Y. Lee, S. Xie, P. Gallagher, Deeply-supervised nets, in: AISTATS, 2015.

[10] A. Payne, S. Singh, Indoor and outdoor scene classification in digital photographs, PR 38 (10) (2005) 1533 – 1545.

[11] X. Meng, Z. Wang, L. Wu, Building global image features for scene recognition, PR 45 (1) (2012) 373 – 380.

[12] Z. Niu, G. Hua, X. Gao, Context aware topic model for scene recognition, in: CVPR, 2012, pp. 2743–2750.

[13] S. N. Parizi, G. Oberlin, J, P. F. Felzenszwalb, Reconfigurable models for scene recognition, in: CVPR, 2012, pp. 2775–2782.

[14] H. Jegou, M. Douze, C. Schmid, Aggregating local descriptors into a compact image representation, in: CVPR, 2010, pp. 3304–3311.

[15] D. Yoo, S. Park, J.-Y. Lee, I. S. Kweon, Fisher kernel for deep neural activations, arXiv preprint arXiv:1412.1628.

[16] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, IJCV 42 (3) (2001) 145–175.

[17] R. Margolin, L. Zelnik-Manor, A. Tal, Otc: A novel local descriptor for scene classification, in: ECCV, 2014, pp. 377–391.

[18] J. Wu, J. M. Rehg, CENTRIST: A visual descriptor for scene categorization, TPAMI 33 (8) (2011) 1489–1501.

[19] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006, pp. 2169–2178.

[20] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: CVPR, 2009, pp. 413–420.

[21] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: CVPR, IEEE, 2010, pp. 3485–3492.

[22] A. Oliva, Gist of the scene, Neurobiology of attention 696 (64) (2005) 251–258.

[23] Y. Xiao, J. Wu, J. Yuan, mCENTRIST: A multi-channel feature generation mechanism for scene categorization, TIP 23 (2) (2014) 823–836.

[24] R. D. Yang S, Multi-scale recognition with dag-cnns, in: ICCV, 2015, pp. 1215–1223.

[25] L. J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in: CVPR, 2009, pp. 2036–2043.

[26] J. Qin, N. H. Yung, Scene categorization via contextual visual words, PR 43 (5) (2010) 1874 – 1888.

[27] C. Li, D. Parikh, T. Chen, Automatic discovery of groups of objects for scene understanding, in: CVPR, 2012, pp. 2735–2742.

[28] E. B. Sudderth, A. Torralba, W. T. Freeman, Learning hierarchical models of scenes, objects, and parts, in: ICCV, 2005, pp. 1331–1338.

[29] M. Stamp, A revealing introduction to hidden markov models, Department of Computer Science San Jose State University.

[30] S. Geman, C. Graffigne, Markov random field image models and their applications to computer vision, in: ICM, 1986, p. 2.

[31] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, JMLR 3 (1) (2003) 993–1022.

[32] M. E. Noha, S. K. Fahad, W. Joost, G. Jordi, Discriminative compact pyramids for object and scene recognition, PR 45 (4) (2012) 1627 – 1636.

[33] L. Zhou, Z. Zhou, D. Hu, Scene classification using a multi-resolution bag-of-features model, PR 46 (1) (2013) 424 – 433.

[34] L. J. Li, H. Su, L. Fei-Fei, Object bank: A high-level image representation for scene classification and semantic feature sparsification, in: NIPS, 2010, pp. 1378–1386.

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, Object detection with discriminatively trained part-based models, TPAMI 32 (9) (2010) 1627–1645.

[36] M. Pandey, S. Lazebnik, Scene recognition and weakly supervised object localization with deformable part-based models, in: ICCV, 2011, pp. 1307–1314.

[37] S. Singh, A. Gupta, A. A. Efros, Unsupervised discovery of mid-level discriminative patches, in: ECCV, 2012, pp. 73–86.

[38] Y. Zheng, Y. G. Jiang, X. Xue, Learning hybrid part filters for scene recognition, in: ECCV, 2012, pp. 172–185.

[39] M. Juneja, A. Vedaldi, C. V. Jawahar, Blocks that shout: Distinctive parts for scene classification, in: CVPR, 2013, pp. 923–930.

[40] D. Lin, C. Lu, R. Liao, Learning important spatial pooling regions for scene classification, in: CVPR, 2014, pp. 3726–3733.

[41] J. Sun, J. Ponce, Learning discriminative part detectors for image classification and cosegmentation, in: ICCV, 2013, pp. 3400–3407.

[42] Z. Zuo, G. Wang, B. Shuai, Learning discriminative and shareable features for scene classification, in: ECCV, 2014, pp. 552–568.

[43] C. Xu, C. Lu, J. Gao, Discriminative analysis for symmetric positive definite matrices on lie groups, CSVT 25 (10) (2015) 1576–1585.

[44] C. Doersch, A. Gupta, A. A. Efros, Mid-level visual element discovery as discriminative mode seeking, in: NIPS, 2013, pp. 494–502.

[45] R. Sharif, H. Azizpour, J. Sullivan, Cnn features off-the-shelf: an astounding baseline for recognition, in: CVPRW, 2014, pp. 806–813.

[46] B. Liu, J. Liu, J. Wang, Learning a representative and discriminative part model with deep convolutional features for scene recognition, in: ACCV, 2014, pp. 643–658.

[47] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: CVPR, 2015, pp. 3828–3836.

[48] G.-S. Xie, X.-Y. Zhang, S. Yan, C.-L. Liu, Hybrid cnn and dictionary-based models for scene recognition and domain adaptation, CSVT 27 (6) (2017) 1263–1274.

[49] Z. Wang, L. Wang, Y. Wang, B. Zhang, Y. Qiao, Weakly supervised patchnets: Describing and aggregating local patches for scene recognition, TIP 26 (4) (2017) 2028–2041.

[50] R. Kwitt, N. Vasconcelos, N. Rasiwasia, Scene recognition on the semantic manifold, in: ECCV, 2012, pp. 359–372.

[51] Y. Su, F. Jurie, Improving image classification using semantic attributes, IJCV 100 (1) (2012) 59–77.

[52] J. Sánchez, F. Perronnin, T. Mensink, Image classification with the fisher vector: Theory and practice, IJCV 105 (3) (2013) 222–245.

[53] J. Donahue, Y. Jia, O. Vinyals, Decaf: A deep convolutional activation feature for generic visual recognition, in: ICML, 2014, pp. 647–655.

[54] Y. S. Xie G S, Zhang X Y, Hybrid CNN and Dictionary-Based Models for Scene Recognition and Domain Adaptation, CSVT.

[55] N. Rasiwasia, N. Vasconcelos, Holistic context models for visual recognition, TPAMI 34 (5) (2012) 902–917.

[56] L. J. Li, H. Su, Y. Lim, Object bank: An object-level image representation for high-level visual recognition, IJCV 107 (1) (2014) 20–39.

[57] L. Bo, X. Ren, D. Fox, Kernel descriptors for visual recognition, in: NIPS, 2010, pp. 244–252.

[58] H. O. Song, R. Girshick, S. Zickler, Generalized sparselet models for real-time multiclass object recognition, TPAMI 37 (5) (2015) 1001–1012.

[59] L. Zhang, X. Zhen, L. Shao, Learning object-to-class kernels for scene classification, TIP 23 (8) (2014) 3241–3253.

[60] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (6) (2017) 84–90.

[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015.

[62] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, TPAMI 39 (6) (2017) 1137–1149.

Xiaojuan Cheng received the B.S. degree from the Department of Automation, Qingdao Technological University, Qingdao, China, in 2011, and the M.S. degree from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2014. She is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China. Her current research interests include deep learning and scene understanding.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from the Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 160 scientific papers in these areas, where more than 50 papers are published in the IEEE Transactions journals. He is an Associate Editor of Pattern Recognition, Pattern Recognition Letters, Neurocomputing, and the IEEE Access, a member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society, respectively. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He is a senior member of the IEEE.

Jianjiang Feng is an associate professor in the Department of Automation at Tsinghua University, Beijing. He received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007, respectively. From 2008 to 2009, he was a postdoctoral researcher in the PRIP lab at Michigan State University. He is an Associate Editor of Image and Vision Computing. His research interests include fingerprint recognition and computer vision.

Bo Yuan is mostly interested in Data Mining, Evolutionary Computation and Parallel Computing. He received the B.E. degree from Nanjing University of Science and Technology, China, in1998, and the M.Sc. and Ph.D. degrees from The University of Queensland, Australia, in 2002 and 2006, respectively. From 2006 to 2007, he

was a Research Officer on a project funded by the Australian Research Council at The University of Queensland. He is currently an Associate Professor in the Division of Informatics, Graduate School at Shenzhen, Tsinghua University.

Jie Zhou received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as the IEEE PAMI, TIP, and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the International Journal of Robotics and Automation and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.