Received November 13, 2018, accepted December 5, 2018, date of publication December 13, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886597

An Attribute-Based High-Level Image Representation for Scene Classification

WENHUA LIU[1], **YIDONG LI**[1], **AND QI WU**[2]
[1]Beijing Jiaotong University, Beijing 100044, China
[2]The University of Adelaide, Adelaide, SA 5005, Australia

Corresponding author: Yidong Li (ydli@bjtu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61300175, and in part by the Fundamental Research Funds for the Central Universities under Grant 2016JBM022 and Grant 2017YJS045.

ABSTRACT Scene classification is increasingly popular due to its extensive usage in many real-world applications such as object detection, image retrieval, and so on. Traditionally, the low-level hand-crafted image representations are adopted to describe the scene images. However, they usually fail to detect semantic features of visual concepts, especially in handling complex scenes. In this paper, we propose a novel high-level image representation which utilizes image attributes as features for scene classification. More specifically, the attributes of each image are firstly extracted by a deep convolution neural network (CNN), which is trained to be a multi-label classifier by minimizing an element-wise logistic loss function. The process of generating attributes can reduce the ‘‘semantic gap’’ between the low-level feature representation and the high level scene meaning. Based on the attributes, we then build a system to discover semantically meaningful descriptions of the scene classes. Extensive experiments on four large-scale scene classification datasets show that our proposed algorithm considerably outperforms other state-of-the-art methods.

INDEX TERMS Scene classification, attribute representation, convolutional neural network, high-level image representation.

I. INTRODUCTION

Scene classification is one of fundamental tasks in computer vision and has great practical significance. As an instance of image semantic classification, scene classification aims to organize images and categorize them into different scene classes such as indoor, outdoor, mountain, river. Recently, scene classification has attracted more and more attention due to its wide applications in the real-world, such as image retrieval, video retrieval, behavior detection and target recognition [1]–[3]. Although many scene classification models have been proposed [4], [5], the performance of these models is still not satisfactory for complex scene datasets. In general, two key components of scene classification are *image representation* and *robust classifier*. Image representation is the process which transforms pixel information into a vectorized representation. It is the first step for the sequential classification tasks. The performance of scene classification models largely depends on the image representation or image feature. Thus, in this work, we mainly focus on how to extract high-level image representation for scene classification.

During the past two decades, lots of algorithms have been proposed to extract image features. For example, the GIST operator [6] uses visual feature which describes the property of scene space to estimate the global space properties. Spatial pyramid (SP) representation of scale-invariant feature transform (SIFT) [7] firstly detects feature in scale space and identifies the position of the key points. Then it uses the principal direction of neighborhood gradient of key points to realize the independence of the scale and direction. These classical approaches to extracting low-level image features have gained remarkable results for scene classification. However, these approaches fail to offer sufficient discriminative power because the extracted features essentially are low-level statistical information of the image which lacks of high-level semantic information. To reduce the semantic gap between low-level representations and high-level scene semantics, Li *et al.* [8] proposed a high-level image representation, Object Bank (OB). It encodes the semantic and spatial information of objects within images. Thus, an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors. Figure 1 shows

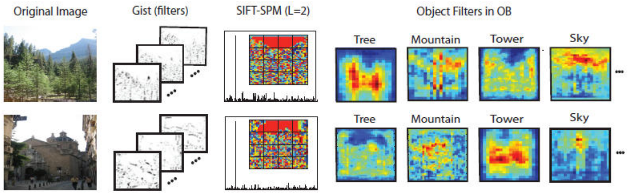


FIGURE 1. Comparison of low-level and high level image representations on two types of scene classes, mountain vs. city street. From left to right: GIST (low-level) [6], spatial pyramid (SP) representation of SIFT (low-level) [7] and OB (high-level) [8].

the selected filter responses in the GIST representation, a histogram of the SP representation of SIFT patches, and a selected number of OB responses on two typical scene images. The performance of OB is much better than most existing low-level feature-based algorithms but the OB features only have the object information and do not consider the attribute information in the image. To enrich image information, we resort to the attributes extracted from scene images to represent complex images. More specifically, a novel attribute-based high-level image representation is proposed, which reduces the burden of sophisticated models for bridging the ''semantic gap'' between high-level scene classification tasks and low-level image representation.

In summary, our main contributions are as follows.

- We construct a novel attribute vocabulary by extracting the semantic attributes from the COCO dataset [9] with the image captions. Each semantic attribute corresponds to a word mined from the image description.
- To extract semantic attributes, our training data is composed of images and attribute set, instead of object vocabulary constructed by object sensing filters. Each training image corresponds to an attribute vector which can represent objects information and descriptive information hidden in image.
- For predicting the attribute vector, we train a multi-label classification network with element-wise logistic loss function and treat the attribute set as the label set.
- Extensive experiments on four scene classification benchmarks demonstrate that the proposed approach achieves superior performance and outperforms the very recent state-of-the-art methods with a large margin.

The rest of this paper is organized as follows. In Section II, we briefly review the existing scene classification methods and visual attribute based representations. Then we introduce the proposed model for constructing attributes in Section III. We present the extensive experimental results in Section IV. Finally, we make conclusions in Section V.

II. RELATED WORK

During the past decades, lots of scene classification algorithms have been proposed [10]–[13]. However, the performance of these classification algorithms are not satisfactory in classifying complex scene images. In fact, the accuracy of scene classification mainly relies on the image descriptors or image representations [14]–[16].

A. LOW-LEVEL IMAGE REPRESENTATION

Generally, low-level image representation employs the global or local statistical information (*e.g.*, global color or texture histograms) to represent images [17], [18]. Although these algorithms are useful and have low complexity, they are restricted to the type of images and often exhibit poor performance. To resolve this issue, Lazebnik *et al.* [7] propose the SP representation of SIFT which computes histograms of local features inside each sub-region of images. Liu *et al.* [19] utilize the exactly matched visual parts and their geometric relationships to improve discriminative ability of global features. Wu and Rehg [20] propose a structural visual descriptor, which can extract the structural properties and suppresses detailed textural information.

A saliency-guided sampling strategy to extract a representative set of patches from a image was proposed by Zhang *et al.* [4]. This unsupervised method can obtain the representative information in the image contained in salient parts of the image. To relieve the scarcity of labeled data, Deng *et al.* [21] propose a multi-task feature hashing algorithm, which can not only utilize the inherent relatedness but also consider the fine-grained clustering among images. Cheriyadat [5] encoded the low-level unlabeled feature descriptors in terms of the basis functions to generate new sparse representation for the feature descriptors. Dense low-level feature descriptors were extracted to characterize the local spatial patterns. Xiao *et al.* [3] used 397 sampled categories in Scene UNderstanding (SUN) database to evaluate the state-of-the-art algorithms. This large scene dataset can solve the limited scope of currently-used databases which do not capture the full variety of scene categories. The accuracy of these methods is hard to be improved due to the lack of high-level semantic information. Recently, feature learning approaches have achieved superior performance in image classification by building advanced machine learning models which can learn high-level feature representation from raw images.

B. HIGH-LEVEL IMAGE REPRESENTATION

To construct high-level image representation, many researchers adopt objects as features to represent image [22], [23], [2]. For example, Li *et al.* [24] propose the object bank (OB) to characterize local image features based on the object detector [25]. The OB representation simultaneously encodes semantic and spatial information of objects, which is frequently adopted for scene classification tasks. Inspired by the same idea, Zhang *et al.* [14] propose the object-to-class (O2C) distances to build scene classification model. Because the O2C distances are based on the OB, the obtained representations can possess more semantic meanings. However, the performance of OB relies on the quality and quantity of the pre-trained object detectors. And OB has the problem of the semantic hierarchy due to the hierarchy concept of the objects in the real world. Combining the local with global features of the image to represent the image was proposed by [26].The ensemble of local visual features from earlier inception layers to the global visual features was able to describe the overall image aesthetics and it gave promising results. To overcome semantic hierarchy, attribute-based high level representations have been proposed for scene classification.

C. ATTRIBUTE-BASED IMAGE REPRESENTATION

The idea of using attributes as the basic representation of images is analogous to the approaches applying a large number of "semantic concepts" to image annotation and retrieval [27]–[29]. In facts, many researchers have treated attributes as high-level image representations to solve the complex computer vision tasks. For instance, Farhadi *et al.* [30] utilize attributes to fill the gaps in predetermined caption templates. Lampert *et al.* [31] shown that semantic attributes can be used to recognize object classes in the absence of training images. Kulkarni *et al.* [32] used the attribute detectors, which can obtain the caption of images in the complex sentence meaning. Besides, there is a significant difference in attributes obtained by different models to address computer vision tasks. The geometry information of the image was viewed as the attributes such as scale-invariant measures (e.g., homogeneity, shape descriptors, orientation, etc.), which was proposed by Cavallaro *et al.* [33]. These nonincreasing attributes have an important roles in the filtering rules and the characterization of the spatial information is performed differently due to the selected attributes and the filter rule. Wang *et al.* [34] introduced a brain-inspired deep network (BDN) which made use of style information from the AVA dataset. Fully convolutional neural networks (FCNNs) are trained for each style. Three primitive features (hue, saturation, value) of the input images are also fused with the output of the third convolutional layer of the 14 FCNNs to form an input cube for another FCNN, which predicts the overall aesthetics ratings. While their idea is inspired by neuroscience models, it is computationally heavy. It was first used to train the aesthetic classification models using only information from the image content. Kairanbay *et al.* [35] then extended [34] approach to utilize additional the style meta-information which was provided by the dataset.

Hu et.al [36] proposed that the attributes of image was transferred from low-level features by using the SVM classifiers and using the LDA topic model to extract the topic information between image samples and attributes. You et.al [37] explored a non-parametric method based on nearest neighbor image retrieval from a large collection of images with rich and unstructured textual metadata such as tags and captions. The attributes for an input image were extracted by transferring the text information from the retrieved images with similar visual appearances. The attributes of a query image was treated as the labels and can be learned as in a conventional classification problem in [38]. They used a Fully Convolutional Network (FCN) to learn attributes from local patches. For scene recognition task, some approaches obtaining attributes have been proposed, for example, [1], [27]–[29]. However, these methods were not able to localize the meaningful concepts in scene images because each semantic concept was trained with the entire images. As a result, the performance was not satisfied on cluttered scene images. To solve this problem, Vogel and Schiele [29] combined the attributes describing the image regions with local semantics. Su and Jurie [39] proposed six groups of attributes to build middle-level features for scene classification.

Different from previous works, our method views the attributes as a set of predefined categories and employs the multi-label classification network with element-wise logistic loss function to extract them.

III. SCENE CLASSIFICATION BASED ON ATTRIBUTE REPRESENTATION

In this section, we elaborate our attribute-based representation for scene classification. The overall process is illustrated in Figure 2. We formulate the attribute prediction as a multi-label classification problem.

As shown in the Figure 2, to construct the model, we first pre-train a deep CNN on the ImageNet [40] with single label images. Then the network is fine-tuned on the COCO dataset [9] with multi-label images, by minimizing an element-wise logistic loss function. Finally, we utilize max-pooling operations to extract attributes for the new image. We use the pooled attributes to classify the scene images.

A. VOCABULARY CONSTRUCTOR

We firstly introduce the semantic concept vocabulary used in the scene classification. We need to construct an attribute vocabulary before predicting the attributes. Yang *et al.* [41] employs hand-labeled training dataset which does not have semantic relations to create a vocabulary. Following [42], our semantic attributes are extracted from the image captions from the large scale COCO datatset and can be any part of the descriptions, including object names (nouns), motions (verbs) or properties (adjectives). The direct use of captions from the image guarantees that the most salient attributes for an image set are extracted. We use the c (c = 256) most common words in the COCO captions dataset to determine the attribute vocabulary $V_a = [v_1, v_2, \ldots v_c]$. The top 15 most frequent closed-class words such as "a", "on", "of" are removed since they are in nearly every caption. In contrast to [43], our vocabulary is not tense or plurality sensitive and more flexible, for instance, "ride" and "riding" are classified as the same semantic attribute, similarly "bag" and "bags". This significantly decreases the size of our attribute vocabulary. The performance of a larger vector V_a is not satisfactory because as the length of vocabulary gets larger, the accuracy of scene classification does not have obvious change.

B. ATTRIBUTES PREDICTION

The process of predicting attributes is formulated as multi-label classification. To obtain the attribute containing more discriminatory information, the shared network structure needs to be fine-tuned on COCO dataset and our image

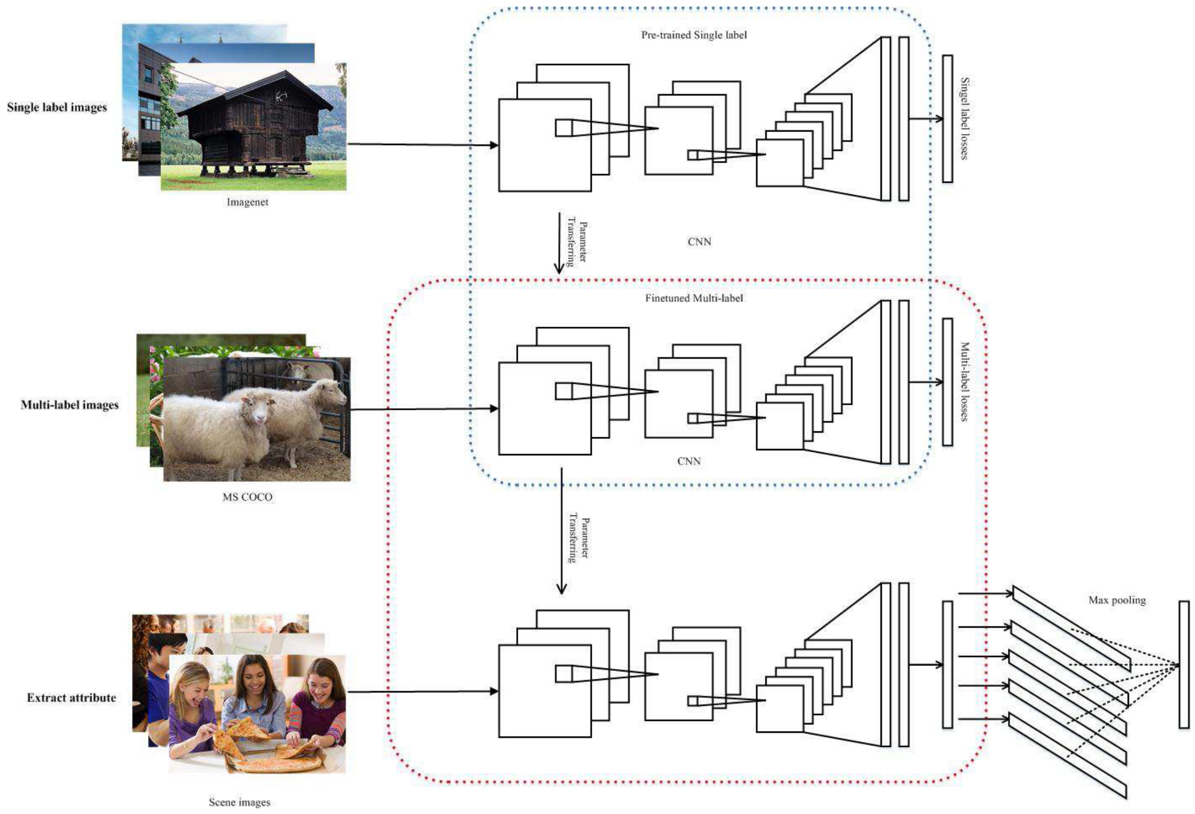


FIGURE 2. The process of predicting attributes: the model uses a VGGNet16 pre-trained on ImageNet to initialize. Secondly, it is fine-tuned on the multi-label COCO dataset by using the element-wise logistic as loss function. Given a test image is passed to this model with max pooling as loss function to produce the multi-label prediction, which gives us the high level image representation.

attribute training dataset (by fine-tuning only the fully-connected layers rather than all layers) after pre-training on ImageNet dataset. We use c-way multi-logistic as loss function during the fine-tuning on COCO dataset and c is set to 256 in this process. The predicted score presents a probability distribution over the multi-class labels. Following [44], we also take the multi-label classification framework based on regions that produce sub-region proposals. However, in our model each proposal is connected to the initialization of CNN rather than the whole image. The final prediction is obtained by combining different proposals with max pooling. For the training, assuming the training set has N samples, c multi-label attributes and the multi-label vector of the i-th image is $y_i = [y_{i1}, y_{i2}, \cdots, y_{ic}] \in \{0, 1\}$ where $y_{ij} = 1$ represents that the j-th attribute contained in the i-th image, otherwise $y_{ij} = 0$. $p_i = [p_{i1}, p_{i2}, \cdots, p_{ic}]$ is the predictive probability vector, which is corresponding to the $y_i = [y_{i1}, y_{i2}, \cdots, y_{ic}]$. The loss function of our model is

$$J = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} log(1 + \exp(-y_{ij} p_{ij})) \tag{1}$$

C. CLASSIFIER BASED ON ATTRIBUTES

For simplicity, we select two kinds of classifiers (One-vs-all and Linear binary classifier) to classify scene images based on the attribute representation. One-vs-all is a kind of SVM [45] algorithm for multi-class classification. SVM maps linearly non-separable vector in a low dimensional space to a high dimensional space and constructs a maximum separation hyperplane in this high dimensional space. Actually, the hyperplane is a real-valued function and we use 1 to represent class C_1 and -1 to represent the rest class C_{rest}. Function $f(x) = sgn()$ is used as the discriminant function, where $sgn(\cdot)$ is the symbolic function.

$$f(x) = sgn(g(x)) = \begin{cases} \text{label is } C_1 & \text{if } g(x) \geq 0 \\ \text{label is } C_2 & \text{if } g(x) < 0 \\ \text{reject} & \text{otherwise} \end{cases} \tag{2}$$

where $g(x) = wx + b$, C_1 expresses a specific class and C_2 is the rest classes. Selecting the optimal classification hyperplane from the couple of the classification hyperplane should satisfy a key condition that it needs to maximize the classification interval between two classes.

$$\max \delta = \frac{1}{||w||} |g(x)| \tag{3}$$

where $||w||$ is the normalization of w. There is an inverse relationship between δ and $||w||$. To simplify the computation, maximizing the classification interval δ is converted to min $||w||$ which is equivalent to min $\frac{1}{2}||w||^2$.

Linear binary classier [46] is used in the one-vs-many scheme for c classes. Let $X = [x_1^T; x_2^T; \ldots; x_N^T;] \in R^{N\times D}$, an $N \times D$ is a matrix, where D represents the dimensions of the attribute set and N is the number of images. And $C = (c_1, c_2, \ldots c_N) \in \{0, 1\}^N$ denotes the binary classification labels of N images. A linear classifier is a function $h_\beta : R^D \rightarrow \{0, 1\}$ defined as $h_\beta \triangleq argmax_{c\in\{0,1\}} x\beta$, where $\beta = (\beta_1, \beta_2, \ldots \beta_D) \in R^D$ is a vector of parameters and it is determined during the training. Linear binary classier is to minimize the following function.

$$min_{\beta\in R^D} \lambda R(\beta) + \frac{1}{N}\sum_{i=1}^{N} L(\beta; x_i; y_i) \tag{4}$$

where $L(\beta; x; y)$ is some non-negative and convex loss function and Log loss is commonly chosen as L; $R(\beta)$ is a regularization to avoid overfitting, and $\lambda \in R$ expresses the regularization coefficient, which can be determined by cross-validation.

$$L = \log(1/P(c_i|x_i, \beta)) \tag{5}$$

where $P(c|x, \beta) = \frac{1}{Z}\exp(\frac{1}{2}c(x \cdot \beta))$.

The detailed process of scene classification is displayed in Algorithm 1.

Algorithm 1 The Algorithm of Classifying Scene Images

```
Require: Dataset, Vocabulary;
Ensure: The classification accuracy of scene image;
    Step1: Pre-training CNN on the ImageNet with single labeled classes;
    Step2: Training pre-trained CNN on COCO dataset with multi-labels selected from vocabulary;
    Step3: Finetuning model obtained by step2 with 256-way multi-logistic loss function on target dataset;
    Step4: Predicting the attributes for test images by using the finetuned model;
    Step5: Using attributes representation of image as the input of off-the-shelf classifier to classify the scene image.
```

IV. EXPERIMENT

A. DATASET

Several datasets are commonly used to evaluate the performance of the scene classification model. We select four challenging scene datasets ranging from activity images (Sports) [47], to cluttered indoor images (Indoor) [48], outdoor images (Outdoor) [49] and natural images (15 Scene [7]) to compare our model with other models. Table 1 displays these four datasets in detail.

TABLE 1. The brief description of four scene datasets, including Sports, Indoor, Outdoor and 15 Scene.

Dataset	Description
Sports	The dataset contains 8 complex event categories and the number of images in each class ranges from 137 to 250.
Indoor	The dataset contains 67 categories and has a total of 15620 images. The number of images varies with the categories, but there are at least 100 images per category.
Outdoor	The dataset contains 8 similar scene categories and has a total of 2688 images.
15 Scene	This is a dataset of natural scene categories, which is created based on [22]. It adds two new categories of industrial and store to the thirteen category in [22].

B. EXPERIMENTS SETTINGS

1) DEEP NETWORK

Our model based on VGGNet16 has 16 layers, which includes 13 convolution layer and 3 fully connected layers. The size of all convolution filters is set to $3 * 3$ in our model because it is the minimum size for capturing the concepts on the up, down and the center. The receptive field of these convolution filters is $7*7$ which can replace a larger filter. We use multiple $3*3$ convolution layers rather than employing one larger filter in convolution layer. This is because former can enhance the capacity of decision function by presenting more nonlinear and it has fewer parameters than a larger filter. The number of units in the last layer is the length of vocabulary.

2) THE PARAMETER SET

If the connected parameters w in the network are randomly initialized in training our model then the training process may take more long time, slow convergence speed and even lead to the poor performance on scene classification. We use values which are the same as [50] to initialize the connected parameters w in two last fully connected layers by experiments.

In the process of training our model, the learning rate is set to 0. 001 for fc6 and fc7 while the learning rate of the last fully connected layer is 0.01. To avoid the remaining local optimum, we should decrease the learning rate of all layer at the special proportion during calculating the parameter value of the connected layer w_{ij}. Like the settings of the most of the network, we respectively set the momentum and the drop rate to 0.9 and 0.5.

How many proposals are required during the process of predicting attributes? This parameter influences the accuracy and the efficiency of predicting attributes hidden in an image. In our model, we divide the group proposal bounding boxes into k clusters by using the normalized cut algorithm [44]. The top t hypotheses are selected by predictive scores for each cluster and they are input to the network. Each image has $kt + 1$ hypotheses, where 1 expresses that the whole image is treated as a hypothetical group and as an input. The parameter value of k and t is respectively set to 10 and 5.

3) EXPERIMENTAL ENVIRONMENT

Our model based on CxxNet framework runs in the server with GPU, and the configuration of this server is that the processor is Intel Xeon Haswel, the Motherboard chip set is a series of the Intel C612, the graphics card is TITANXXTREME-12GD.

C. EVALUATION

We are interested in what factors can affect the performance of our model and the comparison results between our model and other algorithms. To verify these issues, we use

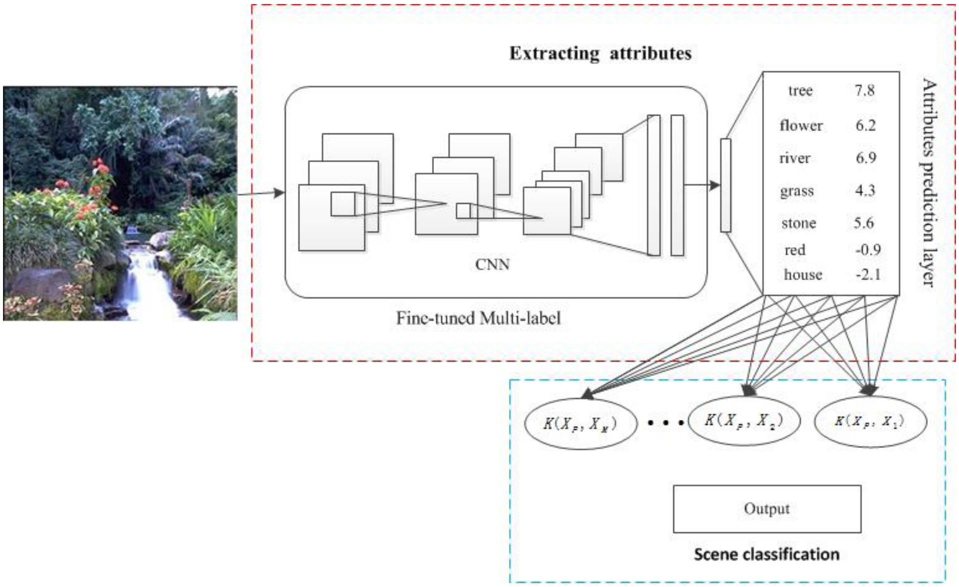


FIGURE 3. Scene classification: Firstly, our model fine-tuned with the training data is used to produce a fixed length vector representation, then, the attribute vector is fed into the classifier to obtain scene label.

average classification accuracy as evaluation indicator in experiment.

1) DIFFERENT ALGORITHMS VS OURS

In the contrast experiments, our model is defined as $\mathbf{Attributes}_{\mathbf{Finetune}}$ and we define three new models which are related to our model. The $\mathbf{Attributes}_{\mathbf{NotF}}$ is a baseline method, which is similar to our framework described in Figure 2. However, $\mathbf{Attributes}_{\mathbf{NotF}}$ does not have the process of fine-tuning on the target dataset. The $\mathbf{Attributes}_{\mathbf{NotF}} + \mathbf{Gist}$ combines the features which are respectively extracted by $\mathbf{Attributes}_{\mathbf{NotF}}$ and by the GIST [6]. The $\mathbf{Attributes}_{\mathbf{Finetune}} + \mathbf{Gist}$ also is a hybrid algorithm which is similar to $\mathbf{Attributes}_{\mathbf{NotF}} + \mathbf{Gist}$. The attributes extracted by these four models are used as the input of the Linear Binary to classify image. The results of Table 6 proves that the classification performance of Linear Binary is better than SVM in our experiment. The Figure 3 shows the process of scene classification for a new image. The attribute vector of this new image is predicted by using our model $\mathbf{Attributes}_{\mathbf{Finetune}}$ fine-tuned on target datasets and it is used as the input of the Linear Binary to obtain the scene label of this image.

We also compare our algorithm with two types of algorithms to validate the efficiency of our algorithm. The high level image representation: OB [8] employs the objects which are obtained by a large of pre-trained object detectors to represent the images; Basic O2C distances used object bank on more discriminative spaces and its variations is proposed by [14]. The low level image representation: gradient-based GIST [6], texture-based Spatial Pyramid [7] and holistic-based CENTIRST (CENsus TRansform hiSTogram) [20]. Table 2 summarizes the scene classification results of different algorithms on four scene datasets.

There are two types of image representation algorithms in Table 2. The algorithms of the high level image representation include OB, K^{NN}(searching the nearest neighbor of the *i*-th image in every class), K^{A}(employing the k-means to cluster and searching the nearest cluster anchor of the *i*-th image for every class), K^{L} (finding $k+1$ nearest neighbors, assigning different weight according to distance between *i*-th image and the class c in the first k-th nearest neighbors and treating the rest of classes equally according to $k+1$-th nearest neighbors), K^{CL} (having the varying the numbers of the nearest neighbors for different class), $\mathbf{Attributes}_{\mathbf{NotF}}$, $\mathbf{Attributes}_{\mathbf{NotF}} + \mathbf{Gist}$, $\mathbf{Attributes}_{\mathbf{Finetune}}$ and $\mathbf{Attributes}_{\mathbf{Finetune}} + \mathbf{Gist}$. $\mathbf{Attributes}_{\mathbf{NotF}} + \mathbf{Gist}$ and $\mathbf{Attributes}_{\mathbf{Finetune}} + \mathbf{Gist}$ are the hybrid method, which combines the high level and low level image representation. The algorithms of the low level image representation are Gist, Sift, CENTIRST.

Those algorithms of extracting the high level image representation have a better performance than the Gist and Sift on four datasets. This is because they can extract more semantic information while the Gist and Sift operator can only obtain the overall texture. The semantically meaningful representation can reduce the burden of sophisticated models for bridging the 'semantic gap' between high level scene classification tasks and low level image representation. This advantage of the high level image representation is more highlighted in cluttered images. Such as, the classification accuracy has the greatest improvement on Indoor dataset due to including the cluttered images.

For high level representation, the accuracy of OB algorithm is lower than those algorithms related to our model because OB just obtains the objects information hidden in

TABLE 2. Comparing the low level image representation with high level image representation on four scene datasets. Gist,Sift, CENTIRST, OB and K^{CL}, K^{NN}, K^{A} and K^{L} algorithms directly use features extracted as input to train classifier. The K* in [14] uses object bank on more discriminative spaces and different distance measurement. $\mathbf{Attributes_{Finetune}}$ needs firstly fine-tune model on the different target datasets, then it predicts the attribute set, based on words commonly found in image caption and classified by Linear Binary. Otherwise, $\mathbf{Attributes_{NotF}}$ and $\mathbf{Attributes_{NotF}}$ + Gist are direct to extract the attributes without fine-tuning on target datasets.

Dataset	Sport	Indoor	Outdoor	15 Scene
Low Level Image Representation				
Gist	82. 59%	5. 69%	81. 96%	73.28%
Sift	82. 83%	44. 21%	15. 67%	82. 36%
CENTIRST	86.22%	31.88%	89. 57%	83.88%
High Level Image Representation				
OB	77. 50%	33. 28%	88. 12%	82. 03%
K^{CL}	86.02%	32.35%	88. 83%	88.81%
K^{NN}	82.31%	35.12%	89.16%	84.85%
K^{A}	81.79%	39.90%	90. 47%	85.07%
K^{L}	82.42%	36.52%	89.92%	83.61%
Our models − CNN				
$\mathbf{Attributes_{NotF}}$	86. 26%	48. 44%	93. 05%	82. 38%
$\mathbf{Attributes_{NotF} + Gist}$	86. 54%	59. 39%	96. 83%	82. 57%
$\mathbf{Attributes_{Finetune}}$	**96. 23%**	**68. 32%**	**98. 83%**	**91. 92%**
$\mathbf{Attributes_{Finetune} + Gist}$	96. 04%	54. 39%	96. 27%	90. 54%

the image. And in [24], the issue of semantic hierarchy becomes more pronounced as the number of the objects increasing. For instance, it cannot understand when a fruit and an apple needs to be simultaneously detected. This problem is caused by assigning equal importance to each object in OB. In other words, the object vector is binary. The value is 1 when object is included in an image, otherwise is 0. However, the attributes not only contain the object information but include properties information about the image. Our attributes sought from V_a are sorted by the score and the higher the score gets, the greater importance the influence has. Therefore, unlike OB, our algorithm does not have the problem of the semantic hierarchy. It can easily detect a fruit and an apple in the same image and give two attributes (fruit and apple) different importance by the score. The algorithm K^* based on OB uses more discriminative spaces, called distance spaces, to represent the image. We can know that it can obtain more semantic meanings than OB, but its performance is obviously lower than our attributes. Three models related to our proposed model have a distinct difference in performance of extracting attributes. It is known that the process of fine-tuning on target dataset is greatly important for predicting the attributes by comparing the baseline mothed $\mathbf{Attributes_{NotF}}$ and $\mathbf{Attributes_{Finetune}}$, $\mathbf{Attributes_{NotF} + Gist}$ and $\mathbf{Attributes_{Finetune}}$. Although $\mathbf{Attributes_{NotF} + Gist}$ combines the attributes with low level features, its accuracy is slightly better than $\mathbf{Attributes_{NotF}}$ on four scene datasets. The performance of hybrid features can be worse than attributes obtained by one method when two methods have significantly difference. Apparently, table 2 compares the hybrid feaures($\mathbf{Attributes_{Finetune} + Gist}$) with the single attributes ($\mathbf{Attributes_{Finetune}}$). This experimental result proves that our model can extract more semantic information to represent the images comparing with the object representation, OB and K^*. And we know that the process of fine-tuning on target datasets greatly improves the performance of the model.

In this experiment, we compare the classification result of each class in 15 Scene dataset between our algorithm and CENTRIST [20].Confusion matrix from one run on this dataset is shown in Figure 4, where row and column names are true and predicted labels respectively. The predicted accuracy of our algorithm was mostly above 90% for each class expect the bedroom. The biggest confusion happens between category pairs such as bedroom/living room, which coincides well with the confusion distribution in CENTRIST.

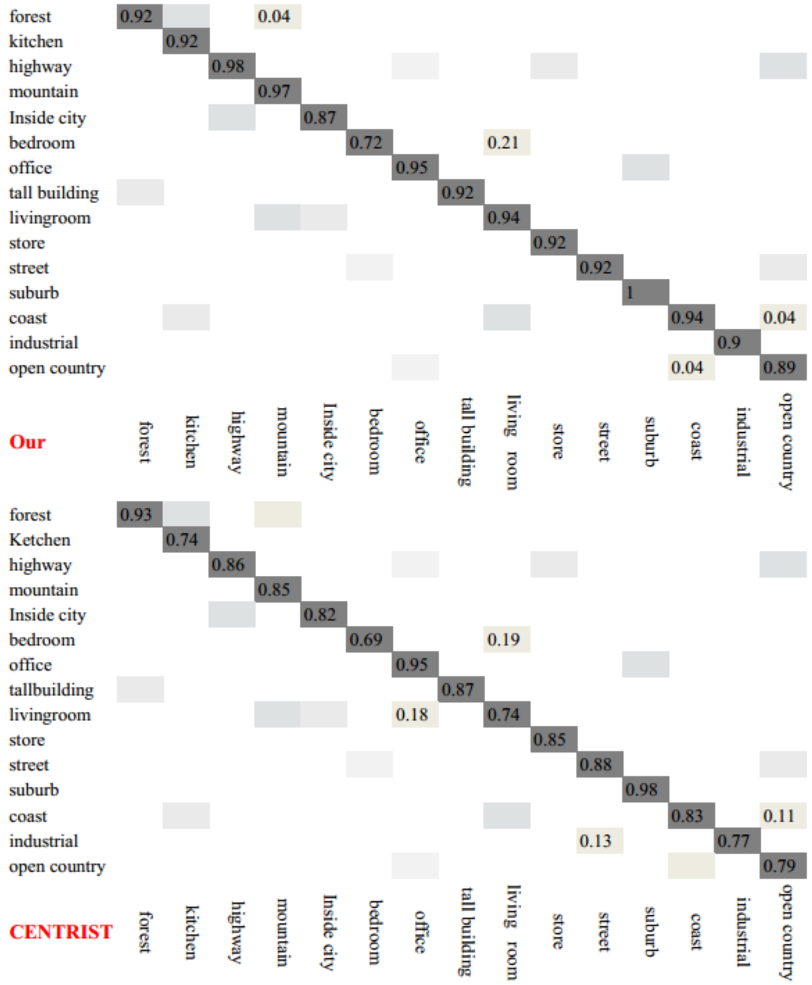
2) DIFFERENT NETWORK VS OURS

Currently, there have many different network structures which are used to extract features. We compare our network structure with some classical network on four scene datasets to demonstrate the effectiveness of our model, and these experimental results are shown in Table3.We define the **Softattribute** as the baseline method for comparing the different network structure. The **Softattribute** uses the softmax function as loss function, which is the only difference between it and our network structure. The proposed method achieves the highest classification rate on all datasets.

TABLE 3. Comparing different network structures with our model. Among, CaffeNet, AlexNet and VGGNet16 are trained on ImageNet. These models use target data to fine-tune model on four datasets.

	Sport	Indoor	Outdoor	15 Scene
SPMSM [51]	83.01%	44.15%	86.37%	82.34%
VC+VQ [52]	88.42%	52.53%	98.59%	85.38%
ISPR [53]	89.48%	50.12%	93.32%	85.06%
LPR-RBF [54]	86.17%	44.81%	92.08%	85.76%
Hybrid Parts+GIST+SP [55]	87.24%	47.21%	94.72%	86.25%
DUCA [56]	95.68%	64.82%	97.56%	90.51%
CaffeNet	94.91%	60.05%	94.54%	88.62%
AlexNet	94.06%	58.15%	93.42%	86.84%
VGGNet16	96.21%	64.13%	92.30%	91.23%
Our models − CNN				
Softattribute	93. 87%	66. 43%	98. 71%	88. 48%
$\mathbf{Attributes_{Finetune}}$	**96. 23%**	**68. 32%**	**98. 83%**	**91. 92%**

CaffeNet, AlexNet and VGGNet16 are used to compare the performance and they have different version of



Our	forest	kitchen	highway	mountain	Inside city	bedroom	office	tall building	living room	store	street	suburb	coast	industrial	open country
forest	0.92			0.04											
kitchen		0.92													
highway			0.98												
mountain				0.97											
Inside city					0.87										
bedroom						0.72			0.21						
office							0.95								
tall building								0.92							
livingroom									0.94						
store										0.92					
street											0.92				
suburb												1			
coast													0.94		0.04
industrial														0.9	
open country													0.04		0.89

CENTRIST	forest	kitchen	highway	mountain	Inside city	bedroom	office	tall building	living room	store	street	suburb	coast	industrial	open country
forest	0.93														
Ketchen		0.74													
highway			0.86												
mountain				0.85											
Inside city					0.82										
bedroom						0.69			0.19						
office							0.95								
tallbuilding								0.87							
livingroom							0.18		0.74						
store										0.85					
street											0.88				
suburb												0.98			
coast													0.83		0.11
industrial											0.13			0.77	
open country															0.79

FIGURE 4. Top: The confusion matrix of our algorithm on the 15 scene dataset. The rates higher than 0.04 are shown in the top of the figure. Bottom: The confusion matrix of CENTRIST (L = 2 spatial PACT) on the 15 scene dataset. The rates higher than 0.1 are shown in the bottom of the figure.

network structure. CaffeNet: we employ the network provided in the Zoo [57] and it uses the average pooling layer as the last layer. AlexNet: The 8 layers network structure is firstly proposed in [58] and the last layer can be viewed as the multi-label layer. VGGNet16: we use the original network structure proposed in [59] and it uses softmax function as loss function.

The similar accuracy of CaffeNet and AlexNet is significantly less than our model on four datasets in Table 3. We can know that the performance of VGGNet16 is better comparing with the former two algorithms because the VGGNet16 is a deeper network structure. However, their accuracy is lower than our model on four datasets. Obviously, we can conclude that the network structures have the different high-level image representation and the higher the abstract extent of the attributes is, the deeper the network structure has. We also know that loss function of network structure affects the performance of algorithm by comparing the **Softattribute** with **Attributes$_{\textbf{Finetune}}$** and using the element-wise logistic as loss function is better than softmax on multi-label classification. Proper loss function should be chosen for different tasks.This experiment also illustrates that attributes extracted by our network structure can carry more semantic information and treating element-wise logistic as loss function is better for multi-label classification.

3) DIFFERENT TRAINING DATASET VS OURS

Hierarchical Matching Pursuit (HMP) [60] is an recent proposed unsupervised feature learning method and trained on SUNRGBD dataset. It builds feature hierarchy layer-by-layer using matching pursuit encoder. SSCNN [61] makes use of a slightly modified AlexNet that trained with the SUNRGBD dataset. The network is divided into two branches, one for semantic segmentation, the other for image classification. Zhu *et al.* [62] proposes a novel network trained on SUNRGBD dataset, which includes RGB CNN and Depth CNN to fuse information from multimodalities. Place-CNN and Hybrid-CNN are proposed by Zhou *et al.* [63]. Place-CNN with softmax loss function is trained on the scene-centric dataset. Hybrid-CNN is trained over 700,000 iterations on Place dataset and the ImageNet dataset removed the overlapping scene categories,under the same network architecture of Place-CNN. Our model is compared with them to verify the effectiveness in table 4.

TABLE 4. Indoor scene recognition accuracy of different methods.

Model	Accuracy
HMP	25.7%
SSCNN	41.3%
[62]	41.5%
Places-CNN	39.0%
Hybrid-CNN	47.80%
$\mathbf{Attributes}_{\mathbf{Finetune}}$	**68. 32%**

From the table 4, it can be seen that the CNN trained with scene centric databases, such as SSCNN, [62], Place-CNN, Hybrid-CNN, out-perform those hand-craft GIST and HMP based on unsupervised feature learning, which proves the advantage of scene specific modeling. However, they are less effective than ours due to the difference of network structure and setting. In running the experiment, our model needs fewer iterations to convergence due to the size of the filter and the depth of network. The different interval models are saved in the process of fine-tuning. And the performance of different interval models vary with the number of fine-tune. To find the model with good performance, we compare the accuracy of the saved models in several different intervals as shown in Figure 5.

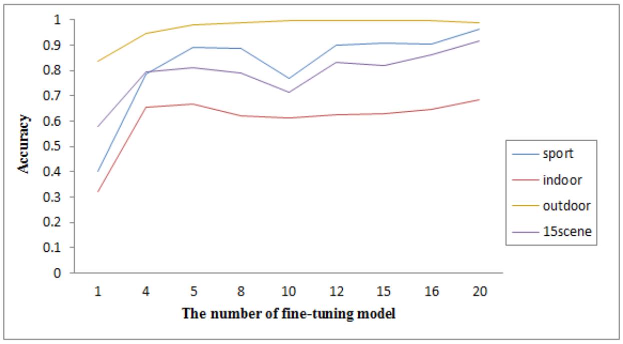


FIGURE 5. The trend of classification accuracy is obtained by our model which is fine-tuned in different interval on four scene datasets.

As seen in Figure 5, the trend of the accuracy of our model is on the rise as the increase of iteraction times from the integral view. However, The performance has the light fluctuations as the increase of interation times due to using the gradient search method to solve the parameters. From Figure 5, our model is more stable and has the best classification result on Outdoor dataset. By experiment, the performance of model saved at 20 intervals is satisfactory, which avoids overfitting and can obtain the good classification accuracy on the four datasets.

4) ROBUSTNESS WITH RESPECT TO TRAINING SAMPLE SIZE

It is an interesting question that training sample size is how to influence the performance of our model. Previously, a large number of experiments have shown that the efficiency of scene classification algorithms is closely related to training samples number [64]–[66]. In our experiment, we use the different size of the training data to fine-tune our model and employ the different size of features to train classifier for testing its robustness. Where, the features are extracted by the GIST and defined as $\mathbf{Feature}_{\mathbf{gist}}$. To guarantee the validity of the followed experiment, we use multiple sizes of training samples to fine-tune our model and to train classifier on four datasets, which ranges 50%, 70%, 100% of training data. Table 5 summaries the influence of the number of training data on fine-tuning our model and training classifier.

TABLE 5. The performance of our model fine-tuned and the classifier trained by different size of training images on four datasets. The ratio of the training data used to fine-tune model ranges 50%, 70%, 100% and the test is done on test data.

	Sport	Indoor	Outdoor	15 Scene
Our models – CNN				
50%	85. 32%	46. 07%	92. 61%	14. 53%
70%	89. 73%	57. 72%	93. 68%	89. 92%
100%	**96. 23%**	**68. 32%**	**98. 83%**	**91. 92%**
$\mathbf{Feature}_{\mathbf{gist}}$				
50%	59.95%	4.89%	75.68%	58.93%
70%	67.59%	5.69%	81.96%	67.68%
100%	68.86%	5.92%	83.53%	69.27%

Table 5 shows that the number of images which are used to fine-tune our model significantly influences the performance of our model. The classification accuracy of our network structure is greatly improved when the proportion of the data increases from 50% to 100% and it is the highest when all of the training data is employed to fine-tune our model. Similarly, Table 5 also illustrates that the size of the training set (attributes or features) makes the accuracy of classifier greatly increase on four datasets due to increasing the proportion of the training data from 50% to 100%. However, the efficiency of the classifier is not sensitive to the size of training data when GIST features are used to train classifier on dataset having inherited features. Such as, Indoor dataset in our experiment. It proves the limitation of the GIST operator in another perspective. The performance of attributes obtained by the fine-tuned model with 50% of training data is much better than the features obtained by Gist using 100% of training data to train classifier except 15 Scene dataset. This result demonstrates that the attribute representation carries more discriminated information which is hidden in a lower dimensional 'informative' feature space.

5) DIFFERENT SEMANTIC GAP VS OURS

The algorithms of the low level image representation just only extract single vision features such as color, texture, shape and spatial relations. These single features result in the weak connection between features and words described the image. This also is the reason why 'semantic gap' is not be well addressed by those algorithms of the low level image representation. Some algorithms employing the statistics of the local appearance of an image are often used to bridge 'semantic gap'. This partly foster the popularity of the bag-of-words (BoW) model, in which local features extracted from an image are first mapped to a set of visual words by vector quantization [20]. An image is then represented as a histogram of visual word occurrences, which naturally encodes the statistics of local features. Also, some models represent an image as a scale-invariant response map of a large number of pre-trained generic object detectors to extract the semantic contents [47]. The Figure 6 shows the effect of eliminating 'semantic gap' by comparing our model with [20] and [47].

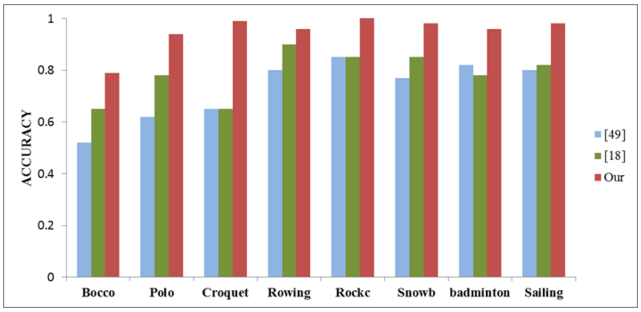


FIGURE 6. The classification accuracy of each class is compared on sport dataset. The accuracy is respectively our model, [20] and [47].

The accuracy of Sport scene classification directly validate the effect of bridging the 'semantic gap' of our model in Figure 6. The performance of our model is apparently higher than [20] and [47] on each class. It proves that constructing attribute vectors is excellent in the process of training our model on COCO dataset. Also, we can conclude that the image representation obtained [20] has more semantic information than [47].

6) CLASSIFICATION WITH STATE OF THE ART CLASSIFIER

Classification is the second step in scene classification. Different algorithms have an influence on the accuracy and efficiency of scene classification. In this paper, we treat the attributes of image as the input of simple off-the-shelf classifier to classify scene images and compare two common classifier (One-against-all algorithm and Linear binary classifier) in the experiment. Table 6 shows the comparison between One-against-all algorithm (Linear SVM) and Linear Binary with multi-logistic on four scene datasets. The results present that the classification performance of Linear SVM is less than Linear Binary on classifying the high dimensional data. That is partly because One-against-all algorithm needs to compute the distance between the point to the hyperplane in the process of solving the optimal hyperplane, but this distance may be invalid in high dimension space. In the One-against-all algorithm, the kernel function $K(\cdot, \cdot)$ is used to map the linear non-separable data to the high dimension space.

TABLE 6. Comparison of classification performance of different classifier (One-against-all vs. Linear Binary) on Sport, Indoor, Outdoor and 15Scene datasets. The input of two classifiers is attributes obtained by our model which is fine-tuned by using training data. Two classifiers have the same parameters on four datasets, respectively.

Classification method	One-against-all	Linear Binary
Sport	92.95%	96.23%
Indoor	63.38%	68.32%
Outdoor	93.38%	98.83%
15Scene	88.36%	91.92%

7) THE TIME COMPLEXITY

The time complexity is also an evaluation index. The time of the models extracting the high level features mostly includes: the time of pre-training detector/model (pre-training generic object detectors/pre-training model on ImageNet dataset), the time of fine-tuning model, the time of extracting objects/attributes and the time of classification. Among, the pre-training time takes up the majority of total time and is not in the running time due to its invariance. The time of the models extracting the low level features is just composed of the time of extracting features and the time of classification. The time complexity of each part depends on the size of the dataset expect the time of pre-training model. Table 7 compares the running time of some models.

TABLE 7. The running time of SIFT, GIST and our model fine-tuned. In this comparison, 70% data is used to fine-tune our model on each test dataset.

Runningtime	Fine-tuning	Extracting	Classifying
Sport			
GIST		316.1s	3.5s
SIFT		1023.1s	4.7s
Ours	**57.3s**	**18.2s**	**4.2s**
Outdoor			
GIST		435.4s	4.0s
SIFT		1747.2s	6.1s
Ours	**68.1s**	**20.9s**	**5.3s**
15Scene			
GIST		731.1s	7.3s
SIFT		2915.2s	10.1s
Ours	**81.9s**	**27.6s**	**6.27s**
Indoor			
GIST		2709.2s	30.5s
SIFT		10107.5s	39.4s
Ours	**148.7s**	**43.5s**	**34.8s**

From the running time, these algorithms have approximately the same classifying time. The extracting time of SIFT is highest due to obtaining the feature matrix for each image. The extracting time of GIST is more long than Ours. The running time of Ours has fine-tuning time, however, our running time is far less than GIST and SIFT shown in Table 7.

V. CONCLUSION

The attributes representation described in this paper is a novel high level image representation and has a good performance on scene classification task. It has significant advantages on classifying scene datasets containing many cluttered images. Our attributes are not only effectively narrow the 'semantic gap' between scene classification (high level visual tasks) and the low level representation, but also solve the problem of semantic hierarchy. The technical contribution of our paper is that the probability vector of attribute set is proposed to represent the images. The experimental results show that the performance of our algorithm is significantly better than the current state-of-the-art approaches on four scene datasets. The time complexity of the high level image representation is higher than the low level image representation due to the complexity. In the future, we will further test attributes representation in other visualization applications and study the more efficient network structure.

REFERENCES

[1] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 702–709.

[2] C. Wojek and B. Schiele, "A dynamic conditional random field model for joint labeling of object and scene classes," in *Proc. Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 733–747.

[3] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[4] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[5] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[6] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.

[8] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[9] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[10] Y. Peng, G. Wang, G. Kou, and Y. Shi, "An empirical study of classification algorithm evaluation for financial risk prediction," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2906–2915, 2011.

[11] B. V. Ramana, M. S. P. Babu, and N. B. Venkateswarlu, "A critical study of selected classification algorithms for liver disease diagnosis," *Int. J. Database Manage. Syst.*, vol. 3, no. 2, pp. 101–114, 2011.

[12] J. R. Romero, P. F. Roncallo, P. C. Akkiraju, I. Ponzoni, V. C. Echenique, and J. A. Carballido, "Using classification algorithms for predicting durum wheat yield in the province of Buenos aires," *Comput. Electron. Agricult.*, vol. 96, no. 6, pp. 173–179, 2013.

[13] D. E. Amrine, B. J. White, and R. L. Larson, "Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease," *Comput. Electron. Agricult.*, vol. 105, pp. 9–19, Jul. 2014.

[14] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.

[15] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[17] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image classification for content-based indexing," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 117–130, Jan. 2001.

[18] N. Serrano, A. E. Savakis, and J. Luo, "Improved scene classification using efficient low-level features and semantic cues," *Pattern Recognit.*, vol. 37, no. 9, pp. 1773–1784, 2004.

[19] X. Liu, B. Lang, Y. Xu, and B. Cheng, "Feature grouping and local soft match for mobile visual search," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 239–246, 2012.

[20] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[21] C. Deng, X. Liu, Y. Mu, and J. Li, "Large-scale multi-task image labeling with adaptive relevance discovery and feature hashing," *Signal Process.*, vol. 112, pp. 137–145, Jul. 2015.

[22] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.

[23] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1365–1372.

[24] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014.

[25] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Softw. Eng.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[26] X. Jin, J. Chi, S. Peng, Y. Tian, C. Ye, and X. Li, "Deep image aesthetics classification using inception modules and fine-tuning connected layer," in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2016, pp. 1–6.

[27] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[28] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Crete, Sep. 2010, pp. 776–789.

[29] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. J. Comput. Vis.*, vol. 72, no. 2, pp. 133–157, Apr. 2007.

[30] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[31] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[32] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[33] G. Cavallaro, M. D. Mura, J. A. Benediktsson, and A. Plaza, "Remote sensing image classification using attribute filters defined over the tree of shapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3899–3911, Jul. 2016.

[34] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," *Michigan Law Rev.*, vol. 52, no. 1, pp. 123–128, 2016.

[35] M. Kairanbay, J. See, L. K. Wong, and Y.-L. Hii, "Filling the gaps: Reducing the complexity of networks for multi-attribute image aesthetic prediction," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3051–3055.

[36] C. Hu, J. Miao, Z. Su, X. Shi, Q. Chen, and X. Luo, "Precision-enhanced image attribute prediction model," in *Proc. 16th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Aug. 2017, pp. 866–872.

[37] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.

[38] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7378–7387.

[39] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 1–10, 2010.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] Y. Yang, C. L. Teo, H. Daumé, III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 444–454.

[42] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 203–212.

[43] H. Fang *et al.*, "From captions to visual concepts and back," *Comput. Sci.*, pp. 1473–1482, 2014.

[44] Y. Wei *et al.*, "CNN: Single-label to multi-label," *Comput. Sci.*, pp. 1901–1907, 2014.

[45] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[46] Y. K. Lin and K. S. Fu, "Automatic classification of cervical cells using a binary tree classifier," *Pattern Recognit.*, vol. 16, no. 1, pp. 69–80, 1983.

[47] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[48] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.

[49] C. Pavlopoulou and S. X. Yu, "Indoor-outdoor classification with human accuracies: Image or edge gist?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 41–47.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, May 2010.

[51] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 359–372.

[52] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale Internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 851–858.

[53] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3726–3733.

[54] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *Proc. Eur. Conf. Comput. Vis.*, vol. 7576, 2012, pp. 228–241.

[55] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 172–185.

[56] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel, "A discriminative representation of convolutional features for indoor scene recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3372–3383, Jul. 2016.

[57] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[58] G. Hinton, N. Srivastava, A. Krizhevsky, R. R. Salakhutdinov, and I. Sutskever, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, Jul. 2012.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, vol. 5, pp. 1–14, 2014.

[60] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2115–2123.

[61] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2016, pp. 2318–2325.

[62] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for RGBD indoor scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2969–2976.

[63] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 487–495.

[64] Y. Han and G. Liu, "Efficient learning of sample-specific discriminative features for scene classification," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 683–686, Nov. 2011.

[65] M. Lapin, B. Schiele, and M. Hein, "Scalable multitask representation learning for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1434–1441.

[66] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

WENHUA LIU was born in Yantai, Shandong, China, in 1990. She received the M.S. degrees in computer science and technology from the Shandong University of Science and Technology, China. She is currently pursuing the Ph.D. degree in computer science and technology with Beijing Jiaotong University, China.

Her research interest includes the development of computer vision using different technology, especially in scene classification and face recognition and data mining using the algorithms of analyzing big data. Her awards and honors include the National Scholarship and the title of Outstanding M.S. Graduate.

YIDONG LI was born in Shanxi, China, in 1982. He received the graduate degree from the Department of Information and Communication Engineering, Beijing Jiaotong University, in 2003, and the master's and Ph.D. degrees from the Department of Computer Science, The University of Adelaide, Australia, in 2006 and 2011, respectively. He is currently an Associate Professor and the Ph.D. Supervisor. He is also the Vice Dean of the School of Computer and Information Technology, Beijing Jiaotong University, and the Executive Director of the SAP University Competence Center, China.

His research directions are mainly in multimedia computing, privacy protection, data mining, cloud computing and high performance computing, intelligent transportation, and so on. Over 60 academic papers have been published by him in major international academic journals, including the IEEE Transactions on Information Forensics and Security, and the IEEE Transactions on Intelligent Transportation Systems and conferences.

Dr. Li was a Member of the Academic Committee of the China Computer Federation YOCSEF, the Deputy Secretary General of the Technical Committee on Control Theory, the Chinese Association of Automation (CAA TCCT), and the Committee of the Chinese Computer Society, and the High Performance Computing (CCF TCHPC) of the Computer Society of China. He served as the Chairman of the Procedure Committee of several international conferences, as the Chairman of the subcommittee, and as a Member of the Procedure Committee. He has chaired/participated in over 30 research projects, including that of the National Natural Science Foundation, the National "863", the Innovation Team of the Ministry of Education, that of the National Natural Science Foundation of Australia, the State Grid, Aviation Industry Corporation of China, and so on.

QI WU received the M.Sc. degree in global computing and media technology, the Ph.D. degree in computer science from the University of Bath, U.K., in 2011 and 2015, respectively. He was a Postdoctoral Researcher with the Australian Centre for Visual Technologies. He was an ARC Senior Research Associate with the Australian Centre for Robotic Vision, The University of Adelaide, Australia, where he is currently a Lecturer.

His research interests include cross-depictive style object modeling, object detection, and vision-to-language problems. He is especially interested in the problem of image captioning and visual question answering. His image captioning model produced the best result in the Microsoft COCO Image Captioning Challenges in the last year and his VQA model is the current state-of-the-art in the area. His work has been published in prestigious journals and conferences such as TPAMI, CVPR, ICCV, and ECCV.