IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Architectural Style Classification based on Feature Extraction Module

PEIPEI ZHAO[1,2], QIGUANG MIAO[1,2], JIANFENG SONG[1,2], YUTAO QI[1,2], RUYI LIU[1,2], AND DAOHUI GE[1,2]

[1]School of Computer Science and Technology, Xidian University, Xian, Shaanxi Province, 710071
[2]Xi'an Key Laboratory of Big Data and Intelligent Vision, Xian, Shaanxi Province, 710071

Corresponding author: JianFeng Song (e-mail: jfsong@mail.xidian.edu.cn).

The work was jointly supported by the National Key Research and Development Program of China under Grant No. 2018YFC0807500, the National Natural Science Foundations of China under grant No. 61772396, 61472302, 61772392, the Fundamental Research Funds for the Central Universities under grant No. JB170306, JB170304, No.JBF180301, and Xi'an Key Laboratory of Big Data and Intelligent Vision under grant No.201805053ZD4CG37.

ABSTRACT Standard classification tasks have already achieved good results in computer vision. However, the task of Architectural style classification yet faces many challenges, since the rich inter-class relationships between different styles may disturb the classification accuracy. To better classify buildings, we propose a feature extraction module based on image preprocessed with Deformable Part-based Models (DPM). Specifically, we first use DPM to remove elements that are not related to classification, and capture representative elements of buildings, then these elements are sent to our feature extraction module. In our feature extraction module, we adopt our Improved Ensemble Projection (IEP) method to maximize the inter-class distance and minimize the intra-class distance to find the common features in the same style and differences among different styles. Finally the performances of several classifiers are tested and the best one of SVM classifier is selected to output the ultimate accuracy. Experimental results show that our approach achieves promising performance and is superior to previous methods.

INDEX TERMS Architectural style classification, Deformable Part-based Models (DPM), feature extraction module, Improved Ensemble Projection (IEP), SVM classifier

I. INTRODUCTION

Architectural style classification, of which the purpose is to classify buildings by some algorithms, is of great importance in the development of a region. The generation of architectural styles evolves as a gradual process, where characteristics of the same classes exist differences at different times. It reflects the cultural development of a region.

Although architectural style classification seems just to be a classification problems, there are many challenges still associated with accuracy of it. First of all, the generation of architectural styles evolves as a gradual process. When styles from one region to another, each region has its own unique characteristics. Meanwhile, each building is unique due the personalities of different architects. Therefore, it is a challenge to find common features within a style. Secondly, when designing a building, an architect sometimes integrates several different architectural style elements. Thence, there are similar characteristics between different styles. As shown in Fig.1, the building in the bottom left corner consists of a chimney, whereas the building in the top right corner does not. They belong to the same architectural style, i.e. American craftsman style, but they have different elements. However, the buildings in the first row belong to different architectural styles, they have the same element of a triangular roof. These complicated relationships between architectural styles lead to some difficulties in classification. So it is significant for architectural styles to find common features of the same style and differences among the 25-class architectural styles.

In this paper, we propose a feature extraction module based on image preprocessed with DPM [1]. To learn the details of building better, we first conduct a preprocessing that using DPM to extract representative elements of buildings. Subsequently, the elements of these images are sent to a feature extraction module. The module consists of depth feature extraction [2] [3] model and IEP model. We use the first model to learn high-level semantic features. To find the common features in the same style and differences among different styles, we adopt our IEP model to maximize the inter-class distance and minimize the intra-class

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2869976, IEEE Access

distance. After a comparison among various classifiers, the final result is obtained with the classifier which has the best performance. The main contributions of this paper are as follows:

- Image preprocessing with DPM Models: Not all elements in the image are favorable for classification. Sometimes, only a few representative elements work on it. Therefore, it is very important to find these representative characteristics. In this paper, DPM can find representative characteristics in an image by matching it to root filter and part filters.
- The feature extraction model: Since it is a challenging to find common features within a style and differences among different architectural styles. We adopt our feature extraction module to minimize the intra-class distance and maximize the inter-class distance. The module is based on the local-consistency and the exotic-inconsistency assumptions. Thus it can capture the common characteristics of the same style and differences among different styles. The experiments prove that our extraction model is beneficial for boosting the final performance to a large extent.
- The analysis of different classifiers: We analysis two classifiers of SVM [4] [5] and LR [6] by groups of experiments, and give the final result with the classifier having best performance.

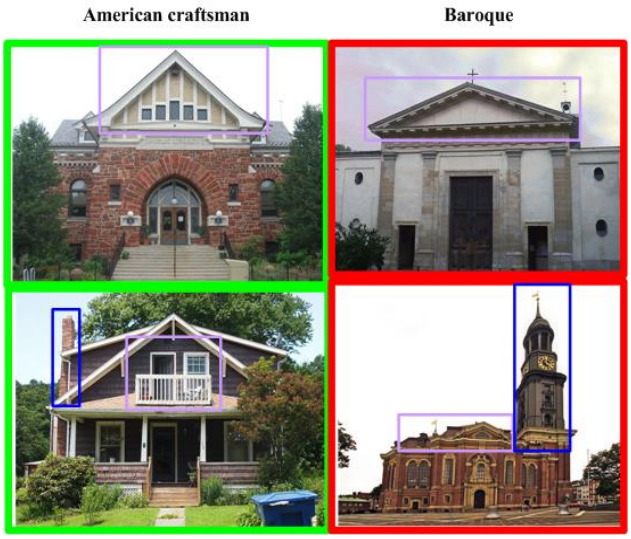


FIGURE 1. Relationships between architectural styles. There are two categories, each column as a category. The first column is American craftsman and the second column is Baroque. Purple bounding boxes in each row show the similarity between different classes. The roofs of these two classes are triangle. The blue's in each column represent the differences in the same class. For example, image in the bottom right corner has a tower. However, there isn't tower in the image at the top right.

The remainder of this paper is organized as follows. The development of Architectural style classification is briefly reviewed in Section Ⅱ. Following in section Ⅲ, image preprocessing with DPM Models and our feature extraction module with deep neural networks (DNN) and IEP method are discussed. Subsequently, extensive experiments in section Ⅳ demonstrate the effectiveness of our approach. Finally, our approach is concluded and the work for future is given in the last section.

II. RELATED WORK

Recent research in architectural style classification has obtained some success. There are various approaches to handle architectural style classification. In the early stage, providing efficient solutions to architectural style classification have a major focus on extracting elements or patterns [7] [8] [9] [10] [11] [12]. Alexander C. Berg. [7] addressed image parsing in the setting of architectural scenes by the generic recognition results bootstrap an image specific model. They approached parsing as a recognition problem both at the coarse level of street, foliage, building, sky, and at the detailed level of window, door, etc. Chu et al. [8] devised a higher-level feature representation to describe configurations of repetitive elements. The feature representation was more discriminative than visual word by modeling spatial relationships between local features, and was flexible to tackle with object scaling, rotation, and viewpoint changes. Doersch et al. [9] proposed to use a discriminative clustering approach able to take into account the weak geographic supervision. It automatically found visual elements, such as balconies, street signs and windows, that were most distinctive for a certain geo-spatial area. Meanwhile, the approach could find out which of them are both frequently occurring and geographically informative in all possible patches in all images. Philbin et al. [11] introduced a novel quantization method based on randomized trees to address a major time and performance bottleneck. One recent study [12] proposed a method that was based on Deformable Part-based Models (DPM) and Multinomial Latent Logistic Regression (MLLR). DPM could capture morphological characteristics of basic architectural components. MLLR introduced the probabilistic analysis and tackled the multi-class problem in latent variable models.

With the rapid development of deep learning and powerful hardwares like GPU, a series of successes have been achieved with the approaches based on Convolutional Neural Network (CNN) for visual task. The concept of CNN is known as LeNet5 model due to its inventor [13]. However, it has not gone further because of the limitation of hardware at that time. A few years later, Hinton et al. proposed a detailed implementation of deep learning in [14], and introduced the model of deep belief network (DBN), which made it have a number of stacked restricted Boltzmann machines (RBM). A significant application of deep CNN is the AlexNet model [15]. It achieves a great success on the ImageNet competition with 10% higher accuracy than other state-of-the-art methods in 2012. Then, a series of models have been proposed for computer vision tasks such as ZF-Net [16], Deepval-Net [17], network in network [18], and so on. On the other hand, deeper and more sophisticated CNN models have made

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2869976, IEEE Access

significant progress by increasing the number of layers [18], size of layers [19] and better activation function, e.g., Relu [20] to yield the best results on various challenges related to object classification, recognition and computer vision. In the 2014 ILSVRC [21] classification challenge, GoogLeNet [2] and VGGNet [22] yielded similarly high performance. These successes spurred a series of research that focused on finding higher performing convolutional neural networks for a task.

III. THE CLASSIFICATION OF ARCHITECTURAL STYLE WITH OUR FEATURE EXTRACTION MODULE

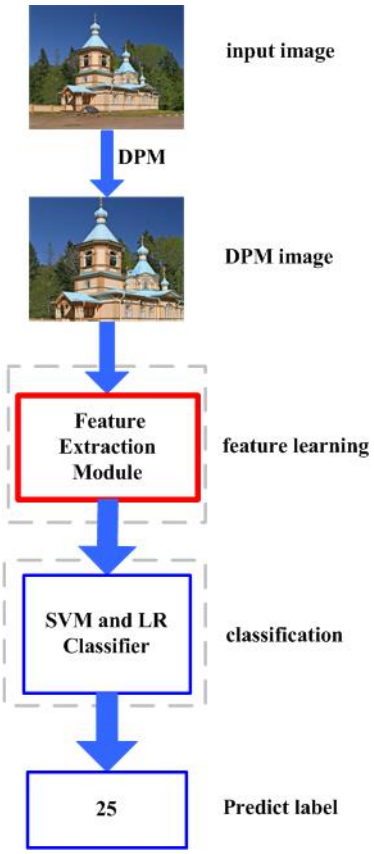


FIGURE 2. The structure of our approach.

As mentioned by Zhe Xu et al. [12], architectural style classification has many difficulties in feature extraction. It is a challenge to find common features within a style and highlight the specific design of an individual building. It is another challenge to find different features among different styles. In this paper, we propose a method which can extract common and different features with our feature extraction model. Firstly, to learn better features, we use DPM to capture the morphological characteristics of basic architectural components. Then the features of these images are extracted respectively by our feature extraction model. The model is based on the local-consistency and the exotic-inconsistency assumptions. It captures not only the characteristics of individual images, but also the relationships among images [23]. Finally, we test the performance of several classifiers and the extracted feature is sent to the optimal classifiers - a linear SVM classifier and LR classifier for the finally results. The flowchart of our approach is depicted in Fig.2. The details of Deformable Parts Model, our feature extraction module implementation will be introduced in following subsections.

A. IMAGE PREPROCESSING STRATEGY

Each type of building has its own style. As shown in Fig. 3, there is a line in the middle of Chicago style buildings; gothic buildings have towering spires, arches and rose windows. Therefore, we only need these representative elements to classify, instead of all elements. To this end, we employ DPM model to extract such typical elements in images.

FIGURE 3. Images in Chicago and Gothic styles

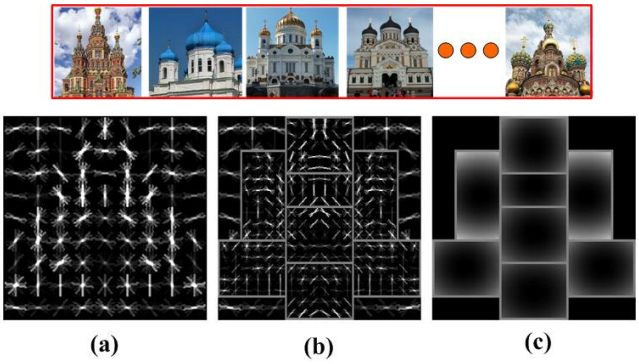


FIGURE 4. The model for Russian Style. The first line is the dataset in Russian style. The trained root filter and part filters for Russian Style are shown in (a) and (b). The root filter shows typical facade outline of Russian style buildings. The part filters captures discriminative architectural elements such as the full and round dome. (c) shows the deformation cost.

DPM describes an image by a multi-scale HOG feature pyramid [1]. The model is defined by a coarse root filter, a set of part filters and deformation costs. The root filter approximately captures the outline of the object such as the building boundary. The part filters are applied to the image with twice the resolution of the root, capturing finer resolu-

tion features of the object such as arches and rose windows. Deformation cost measures the deviation of the parts from their default locations relative to the root. We use labeled data to train models for each style. Fig.4 shows a trained model for Russian Style.

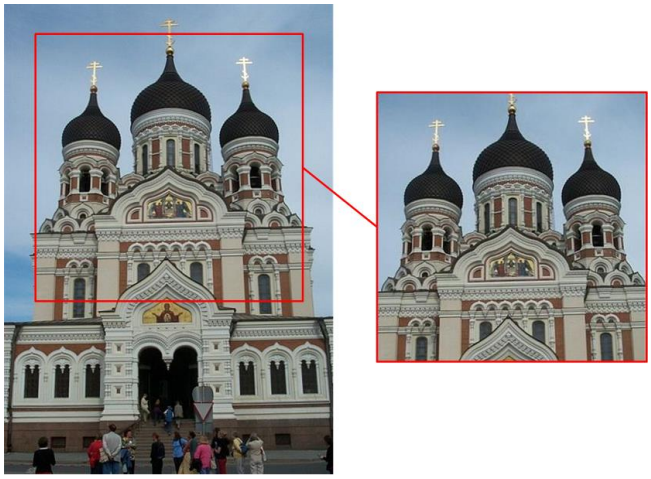


FIGURE 5. The image is decomposed into many bounding boxes of the same size. Red bounding box has the highest response. It is used to replace this image.

We use the trained models to detect representative elements in an image. A bounding box slides over the image to form a number of sub-images. In order to improve the resolution, we first conduct an up-sampling that converting sub-image into twice size with Gaussian pyramid [24]. Following, features of original sub-image and sub-image with enlarged resolution are extracted respectively by the DPM. Then the DPM features of original sub-image convolves with root filter to obtain the response diagram, which represents the match between image and root filter. Meanwhile, in order to get the response diagrams of the match between the sub-image and part filters, the DPM features of sub-image with enlarged resolution convolves with part filters. The response diagram of root filter can capture coarse resolution edges such as the structure of the building, but details like windows and doors are not visible. The response diagrams of part filters can captures the details. As these response diagrams can complement each other, these features are blended based on weighted average to obtain the final response diagram. After a comparison among the final response diagrams of all sub-images, an image is instead of the sub-image which has the highest response. As shown in Fig.5, the image is replaced by the sub-image of the highest response. The sub-image is generated from DPM model, which focuses on the representative elements, namely the full and round dome in Fig.5, and it is helpful to eliminate the extra elements, so it can boost the performance.

B. FEATURE EXTRACTION MODULE

The feature extraction module consists of DNN model and IEP model. Our first model can capture high-level features. Our IEP model can extract the common characteristics of the same style and differences among different styles.

- Depth Feature extraction

The main advantage of CNN [25] [26] [27] [28] [29] [30] [31] is the ability to learn high-level features from the low-level ones and the details which are not related with the target can be ignored. Meanwhile, CNN is robust against light, surrounding clutter and rigid transformation [32], therefore the CNN based method can achieve superior performance on architectural style classification. In this paper, we use GoogLeNet model to extract features that are beneficial to architectural style classification. Although GoogLeNet and VGGNet [22] yielded similarly high performance in architectural style classification. However, VGGNet has the compelling feature of architectural simplicity, this comes at a high cost: evaluating the network requires a lot of computation [3]. GoogLeNet was designed with computational efficiency and practicality in mind, so that the computational cost of GoogLeNet is much lower than VGGNet.

GoogleNet has a new level of organization called "Inception Module" which consists of convolutions and max-pooling operation. There are nine Inception modules in GoogLeNet architecture. Fully-connected layers are being replaced with 1×1 convolutions at the bottom of the module. The 1×1 convolutions can reduce the number of inputs and hence decreases the computation cost dramatically. It also extracts the relevant features of an input image in the same region.

However, training a deep network needs to collect a large amount of labeled data since there are millions of parameters waiting for adjusting. In architectural style classification, it is expensive to recollect the needed training data. In such cases, a pre-trained GoogLeNet model has been applied and fine-tuned using our dataset in architectural style classification. In order to train the model, we use data augmentation in this paper. The number of images can be increased by horizontal/vertical flip. Therefore, data augmentation can prevent the model from over-fitting.

- Feature extraction model with IEP

After an analysis of the wrongly classified images of the GoogLeNet model, we find one factor that hampers the accuracy is the similarities between different styles and the differences in the same styles. As shown in Fig.6, the image with red edges belongs to the Georgian style, but it is incorrectly classified into the Greek Revival style. The misclassified image has some similar features to the Greek Revival style. For example, they all have some cylinders. To distinguish the image from the Greek Revival style, we must discover differences between the image and the style. However, it is not enough just to find the differences. In order to classify the image correctly, it is also important to find out what it has in commons with the Georgian style.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2869976, IEEE Access

FIGURE 6. The similar styles. The image in the red bounding box that wrongly classified into Greek Revival style. However, it really belongs to Georgian style.

We propose an improved ensemble projection method (IEP) based on ensemble projection (EP) [34] to learn a new image representation which can solve above problems. IEP exploits the local-consistency assumption that samples with high similarity should share the same label and the exotic-inconsistency assumption that samples with low similarity are in high probability come from different classes. IEP consists of many prototypes which are inter-distinct and intra-compact, so that each one represents a different visual concept. To ensure inter-distinct and intra-compact of each prototype, we employ a two-step sampling method, called Max-Min Sampling. The Max step is designed for the inter-distinct property which depicts the differences among different styles. The Min step is designed for intra-compact property which depicts the common characteristics of the same style.

The algorithm for creating prototype sets is given in Algo.1. In particular, we first build a skeleton of the prototype set by looking for images with the large distances from each other. The distance of two images can be expressed as:

$$dis(x_i, x_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \qquad (1)$$

Where $dis(x_i, x_j)$ is the square of a distance between x_i and x_j. They are the features of two different images, which are extracted by GoogLeNet. i and j are image indexes. $x_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$, $x_j = \{x_{j1}, x_{j2}, ..., x_{jn}\}$. Therefore, the Max step guarantees that the sampled seed images are far from each other, which means it can find differences among different classes. Once the skeleton is created, we enrich the skeleton to a prototype set by looking for the closest neighbors of the skeleton images. In other words, the min step can extend each seed image to an image prototype by introducing its n closest images (including itself), which means it can extract the common characteristics of the same class. A single prototype set only defines a visual concept (image attribute). For large diversity, randomness is introduced in different trails of Max-Min Sampling to create an ensemble of diverse prototype sets, so that a rich set of image attributes are captured [23].

Algorithm 1: Max-Min Sampling in t^{th} trail

Data: Dataset D

Result: Prototype set P^t

begin

$e_1 = 0$;

While iterations $\leq$ m **do**

v = { r random image indexes};

$e = \sum_{i \in v} \sum_{j \in v} dis(X_i, X_j)$

if $e > e_1$ **then**

$e_1 = e$

$v_1 = v$

end

end

for $i \leftarrow 1$ **to** r **do**

s_i^t = indexes of the n nearest neighbors of $v(i)$ in D ;

$c_i^t = (i, i, ..., i) \in \mathbb{R}^n$

end

$s^t = (s_1^t, ..., s_r^t) \in \mathbb{R}^{rn}$

$c^t = (c_1^t, ..., c_r^t) \in \mathbb{R}^{rn}$

$P^t = \{(s_i^t, c_i^t)\}_{i=1}^{rn}$

end

After the prototype sets are established, input the image X to the prototype sets to measure similarities. The vector of all similarities is concatenated to form a new image representation which is used for the final classification.

IV. Experiment results

In this section, we demonstrate how our strategies work by groups of experiments. Firstly, we discuss implementation details, including the architectural style dataset that our experiments process on, the parameter setting and running environment in Section A, and then the experiments that show the effect of image preprocessing with the Deformable Parts Model, the analysis of our feature extraction module and the comparison of the performance of different classifiers are given in Section B to D. For a fair comparison, the experiments on our proposed method are conducted on the same set with [12]. We run a ten-fold experiment.

A. IMPLEMENTATION DETAILS

1) DATASET

In order to study architecture styles and model their underlying relationships, we use the architectural style dataset [12] [33]. The dataset contains 25 architecture styles. The number of images in each style varies from 50 to 300, and altogether the dataset contains about 5000 images.

2) PARAMETER SETTING

As the model we use for feature extraction is GoogLeNet and IEP, these network settings are the same as [3] and [34]. However, for a better fine-tuning result and adapting to our architectural style classification, we use batches of 100, with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2869976, IEEE Access

the learning rate of 0.5 in the GoogLeNet. As to the parameters of our IEP method, we used the following for experimental sets: T=300, r=30, n=6 and m=50.

3) EXPERIMENTAL ENVIRONMENT

Our experiments are processed on a PC with Intel Core i5-4590 CPU @ 3.30GHz 3.30 GHz, 8 GB RAM and Nvidia Tesla K40c GPU. The experiments of the GoogLeNet model training and feature extracting are processed under tensorflow framework on Linux Ubuntu 14.04 LTS, others including image preprocessing with DPM, learning a new image representation with IEP and the final classification process with various classifiers are implemented by matlab R2014a on 64-bit Windows 7.

B. EFFECT OF IMAGE PREPROCESSING STRATEGY

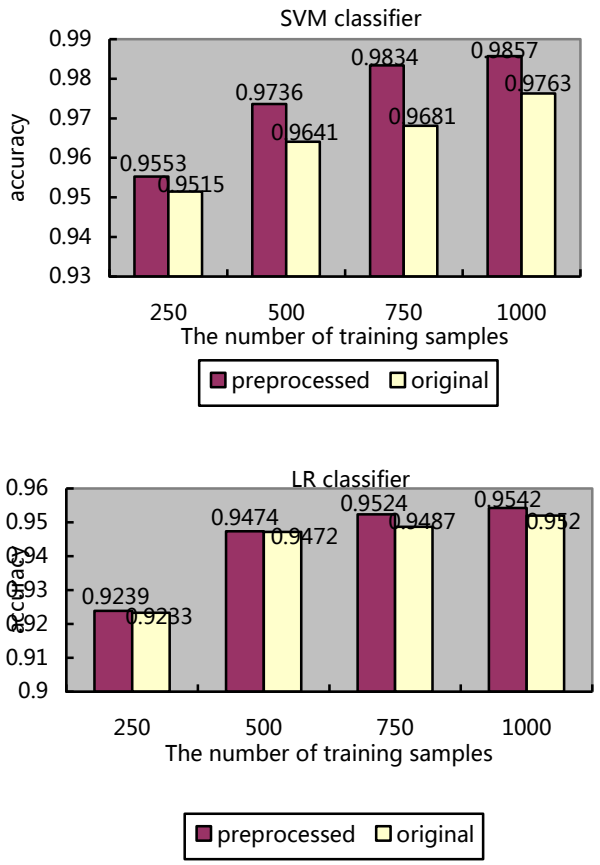


FIGURE 7. In this paper, SVM and LR are used to classify 25-class architectural style dataset. Result comparison between our feature extraction model with preprocessed and original input. The preprocessed input is superior to another on 25-class architectural style dataset.

In this subsection, we experiment on 25-class architectural style dataset and verify the effectiveness of our image preprocessing strategy. We compare the accuracy between classification results with preprocessed and original input. The results are shown in Fig.7.

As can be observed in Fig.7, our image preprocessing strategy is reasonable and achieves a great promotion. With the number of preprocessed training samples in each category increasing, the accuracy rate also increases. That is because the image preprocessing strategy can capture the representative elements which can illustrate the building better.

C. The analysis of our feature extraction module

After the correctness of our image preprocessing strategy has been proved, we verify the effect our feature extraction module. The verification of feature extraction module is in two-stage. Our feature extraction module consists of GoogLeNet and IEP. Firstly, we test the influence of high-level features with GoogLeNet, i.e., the differences between preprocessed images with/without GoogLeNet. The experiments on pre-processed images are illustrated in Fig.8.

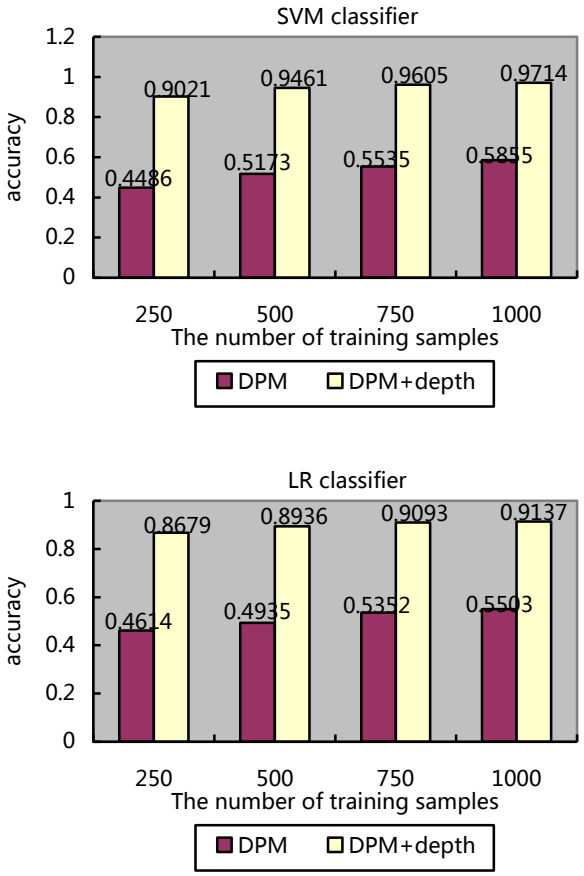


FIGURE 8. Result comparison between the accuracy of high-level feature extraction with GoogLeNet or without it. It can be find that the depth feature extraction helps to improve the classification performance.

Fig.8 reports that the depth feature extraction can really help the preprocessed image to achieve higher accuracy. The improvement on preprocessed images is about 40%. That is because the depth data helps to filter out building-irrelevant factors and it can mining high-level feature.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2869976, IEEE Access

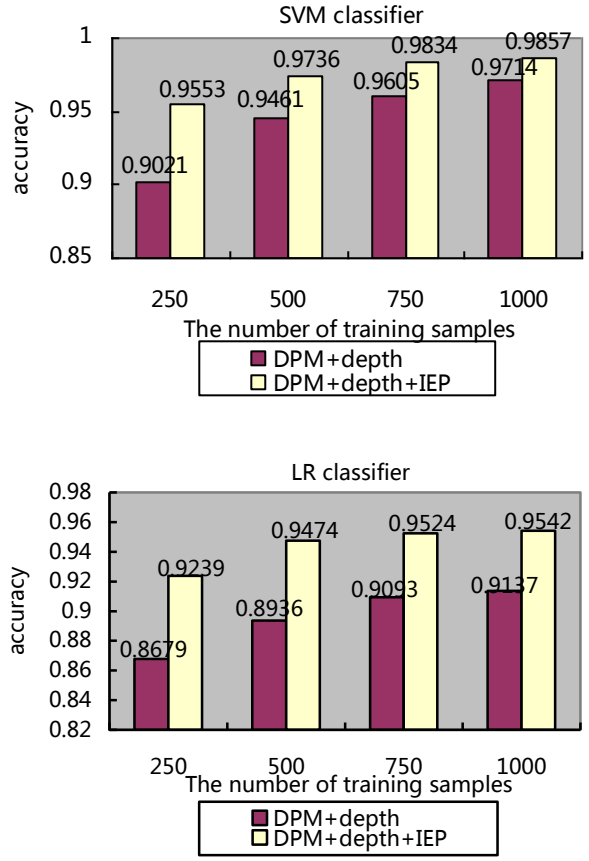
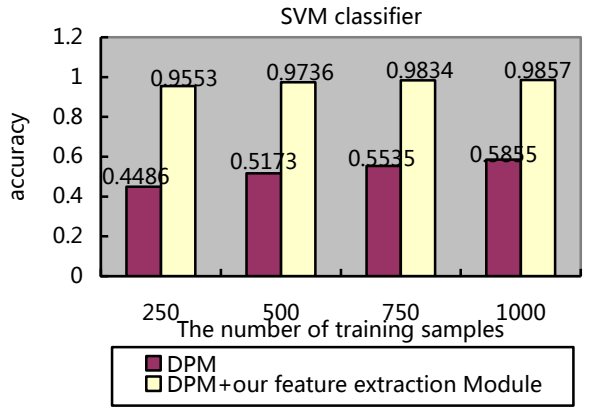


FIGURE 9. **Result comparison between the accuracy of preprocessed images with/without IEP method. Although more information is available with the depth data and the image preprocessing strategy, the addition of IEP data can still show positive influence to rise the accuracy.**

Then, in the Fig.9, we can see the IEP data is still useful for the improvement of accuracy overall. Although the participation of depth data and preprocessed images can provide more information as what the IEP data does, the addition of IEP data can still present a positive influence to improving the classification accuracy.



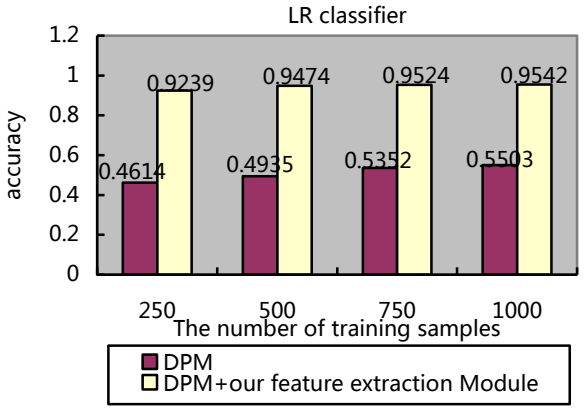


FIGURE 10. **Result comparison between the accuracy of preprocessed images with/without our feature extraction module. It can be find that the feature extraction module helps to improve the classification performance.**

As can be observed in Fig.10, our feature extraction module helps to achieve higher accuracy. That is because the depth data can build high-level features from the low-level ones and the details which are not related with the target can be ignored. Meanwhile, IEP method can capture the common characteristics of the same style and different features among different styles.

D. Analysis of classifiers

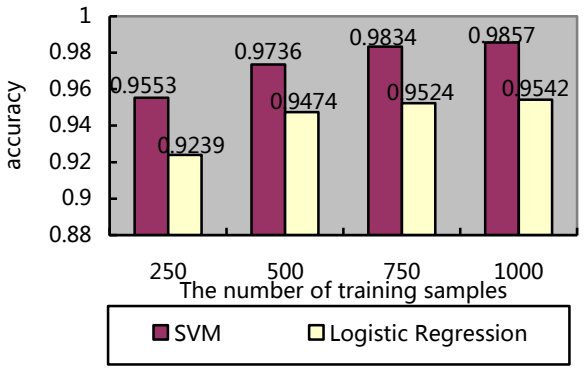


FIGURE 11. **The comparison of our proposed method with SVM classifier and Logistic Regression classifier.**

The goal of classification is using an image's characteristics to identify which class it belongs to. There are many kinds of linear classifiers like naïve Bayesian, logistic regression, support vector machine (SVM), random forest and k-Nearest Neighbors (K-NN). In this paper, we empirically test two kinds of classifiers – SVM and logistic regression in terms of the properties of the dataset we experiment on, and choose the best for obtaining the final classification result. To make a fair comparison, all the other factors, such as image preprocessing strategy and feature extraction parameters of GoogLeNet and IEP are all the same.

The result of SVM and Logistic Regression are reported with the optimal parameter as aforementioned. From Fig.11,

we can see that the classification effect of SVM is 3% higher than that of Logistic Regression. The result of SVM is the best. Thus we adopt the SVM as our final classifier.

E. The final comparison

In this subsection, we compare the proposed approach with state-of-the-art approaches. The results of the comparison are shown in table 1. It compares the classification accuracy of our method with other algorithms, including DPM-LSVM [35], DPM-MLLR [12], MLLR Spatial Pyramid (MLLR-SP) [36]. The classification of accuracy of our method has achieved the best results and our feature extraction model has made a great contribution to this.

TABLE I
RESULTS ON THE ARCHITECTURAL STYLE CLASSIFICATION.

Methods	Accuracy
DPM+LSVM	37.69%
DPM+MLLR	42.55%
MLLR+SP	46.21%
DPM+IEP+SVM	55.35%
DPM+FEATURE EXTRACTION MODEL+SVM(OUR)	98.57%

V. Conclusions

We propose a feature extraction module based on DNN and our IEP method. DNN can learn high-layer features from the architectural style images. IEP is based on the local-consistency and the exotic-inconsistency assumptions that can find common features of the same style and differences among 25-class architectural styles. The new features are used to classify the architectural style. Experimental results show that our method achieves the best performance. The method is competitive comparing with other algorithms.

REFERENCES

[1] Felzenszwalb P F, Girshick R B, McAllester D, and Ramanan D, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence. vol.32, no.9, pp. 1627-1645, Sept. 2010, DOI. 10.1109/TPAMI.2009.167.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 1–9.

[3] Szegedy C, Vanhoucke V, Ioffe S et al, "Rethinking the inception architecture for computer vision," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 2818-2826.

[4] Scholkpf B and Smola AJ, "Learning with kernels: Support vector machines, Regularization, Optimization, and Beyond," MIT Press Cambridge. MA, USA,2002,

[5] Joachims T, "Transductive support vector machines," Chapelle et al. pp. 105-118, 2006.

[6] Hosmer D W and Lemeshow S, "Introduction to the logistic regression model:testing for the significance of the coefficients," Applied Logistic Regression, Second Edition. pp. 1-30, 2000.

[7] Berg A C, Grabler F, and Malik J, "Parsing images of architectural scenes," 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1-8.

[8] Chu W T and Tsai M H, "Visual pattern discovery for architecture image classification and product image search," Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. ACM, 2012.

[9] Doersch C, Singh S, Gupta A, Sivic J, and Efros, "What makes paris look like paris?," ACM Transactions on Graphics. vol. 31, no. 4, 2012.

[10] Goel A, Juneja M, and Jawahar C V, "Are buildings only instances?: exploration in architectural style categories," Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing. ACM, 2012.

[11] Philbin J, Chum O, Isard M, Sivic J, and Zisserman A, "Object retrieval with large vocabularies and fast spatial matching," 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1-8.

[12] Xu Z, Tao D, Zhang Y, Wu J, and Tsoi A, "Architectural style classification using multinomial latent logistic regression," European Conference on Computer Vision. Springer International Publishing, 2014, pp. 600-615.

[13] Haohan Wang and Bhiksha Raj, "A survey: Time travel in deep learning space: An introduction to deep learning models and how deep learning models evolved from the initial ideas," unpublished.

[14] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science vol. 313, no. 5786, pp. 504–507, 2006.

[15] A. Krizhevsky and I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems," NIPS, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[16] M.D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in: European Conference on Computer Vision, Springer, Zurich, Switzerland. 2014, pp. 818–833

[17] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in: British Machine Vision Conference. Swansea, UK, 2014.

[18] M. Lin, Q. Chen, and S.C. Yan, "Network in network," International Conference on Learning Representations. Banff, Canada, 2014.

[19] Pierre Sermanet, David Eigen, Xiang Zhang, Micha ̈el Mathieu, Rob Fergus, and Yann LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," unpublished.

[20] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 2015, pp. 846-850.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al, "Imagenet large scale visual recognition challenge," vol. 115, no. 3, pp. 211-252, 2015.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," unpublished.

[23] Dai D and Van Gool L, "Unsupervised High-level Feature Learning by Ensemble Projection for Semi-supervised Image Classification and Image Clustering," unpublished.

[24] Lan Z, Lin M, Li X, et al, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp.204-212.

[25] Zhang Q, Lin G, Zhang Y, et al, "Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images," Procedia engineering. vol: 211, pp. 441-446, 2018.

[26] Ren S, He K, Girshick R, et al, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence. vol. 39, no. 6, pp. 1137-1149, 2017.

[27] Nakahara H, Yonekawa H, and Sato S, "An object detector based on multiscale sliding window search using a fully pipelined binarized CNN on an FPGA," Field Programmable Technology (ICFPT), 2017 International Conference on. IEEE. 2017, pp. 168-175.

[28] He K, Gkioxari G, Dollár P, et al, "Mask r-cnn," Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE. 2017, pp. 2980-2988.

[29] Zhang K, Zuo W, Chen Y, et al, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE Transactions on Image Processing. vol. 26, no. 7, pp. 3142-3155, 2017.

[30] Thieu N T V, Van T T, Tuan A T, et al, "An evaluation of purified Salmonella Typhi protein antigens for the serological diagnosis of acute typhoid fever," Journal of Infection. vol. 75, no. 2, pp. 104-114, 2017.

[31] Li J, Liang X, Shen S M, et al, "Scale-aware fast R-CNN for pedestrian detection," IEEE Transactions on Multimedia. vol. 20, no. 4, pp. 985-996, 2018.

[32] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2. IEEE, 2004, pp. 96–104.

[33] Miao, Qiguang, et al. "A semi-supervised image classification model based on improved ensemble projection algorithm," IEEE Access, pp. 1372-1379, 2018.

[34] Dai D and Van Gool L, "Ensemble projection for semi-supervised image classification," 2013 IEEE International Conference on Computer Vision. IEEE. 2013, pp. 2072-2079.

[35] Pandey M, and Lazebnik S, "Scene recognition and weakly supervised object localization with deformable part-based models," In: 2011 IEEE International Conference on Computer Vision (ICCV). 2011, pp. 1307–1314.

[36] Lazebnik S, Schmid C, and Ponce J, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006, vol. 2, pp. 2169–2178.

PEIPEI ZHAO received the M.E degree from the School of Computer Science and Technology, Xidian University in 2016. She is currently working toward the Ph.D. degree at Xidian University. Her research interests include pattern recognition and digital image processing.

QIGUANG MIAO is a professor and Ph.D. student supervisor of School of Computer Scienceand Technology in Xidian University. He is a member of professor committee. In 2012, he was supported by the Program for New Century Excellent Talents in University by Ministry of Education. He is a committee member of CCF, a committee member of CCF Computer Vision, the vice chairman of CCF YOCSEF.

He received his doctor degree in computer application technology from Xidian University in December 2005. His research interests include Computer Vision, machine learning and Big Data. As principal investigator, he is doing or has completed 4 projects of NSFC, 2 projects of Shaanxi provincial natural science fund; more than 10 projects of National Defence Pre-research Foundation, 863 and Weapons and Equipment fund. He has hosted 1 project supported by Fundamental Research Funds for the Central Universities by MOE. In the field of teaching, he was awarded as one of Pacemaker of Ten Excellent Teacher twice in 2008, 2011and 2014.

In recent years, He has published over 100 papers in the significant domestic and international journal or conference including IEEE Trans. On Image Processing, IEEE Trans. on Geoscience and Remote Sensing, Journal of Visual Communication and Image Representation, NeuroComputing, IET Image Processing, Knowledge Based System and so on, of which more than 30 papers are indexed by SCI and over 40 papers are included in EI. He has served as committee chairman of the first CCF Youth Elite Association, CNCC2008, CIS 2012, CCFAI 2013, CCDM2014 PC member, CIS 2013 special session chair. He is committee member of Editorial Board of the Internet of Things, assessment expert of the State Science and Technology Prizes and the National Defense Basic Scientific Research Project. He has been awarded the Prize at the ministerial and provincial level twice.

Jianfeng Song was born in 1978. He received the M.S. degree in Computer Science in 2001. His research interests include Broadband wireless network cross-layer protocol design and performance analysis, heterogeneous wireless networks, computer system security and malware analysis.