Signal Processing: Image Communication 47 (2016) 511–518

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

A multi-scene deep learning model for image aesthetic evaluation

Weining Wang, Mingquan Zhao, Li Wang, Jiexiong Huang, Chengjia Cai, Xiangmin Xu *

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

ARTICLE INFO

Article history:
Received 17 October 2015
Received in revised form
10 May 2016
Accepted 11 May 2016
Available online 13 May 2016

Keywords:
Deep learning
Image aesthetic
Multi-scene deep learning model
Pre-training

ABSTRACT

Aesthetic evaluation of images has attracted a lot of research interests recently. Previous work focused on extracting handcrafted image features or generic image descriptors to build statistical model for aesthetic evaluation. However, the effectiveness of these approaches is limited by researchers' understanding on the aesthetic rules. In this paper, we present a multi-scene deep learning model (MSDLM) to enable automatic aesthetic feature learning. This deep learning model achieves better results because it improves performance on some major problems, including limited data amount and categories, scenes dependent evaluation, unbalanced dataset, noise data etc. Major innovations are as follows. (1) We design a scene convolutional layer consist of multi-group descriptors in the network elaborately so that the model has a comprehensive learning capacity for image aesthetic. (2) We design a pre-training procedure to initialize our model. Through pre-training the multi-group descriptors discriminatively, our model can extract specific aesthetic features for various scenes, and reduce the impact of noise data when building the model. Experimental results show that our approach significantly outperforms existing methods on two benchmark datasets.

1. Introduction

Volume of images grows explosively through social network in past years. It is an arduous task for people to find and manage high quality photos in large amount of collections. Aesthetic analysis can help people to select beautiful images automatically, filter out the unappealing ones, provide aesthetic feedbacks etc. [20]. It also helps people to construct a harmony human–machine interactive system [2]. The aesthetic quality has become an important factor for image retrieval engines [1].

The aesthetic evaluation of images aims at building a computational model to simulate human's aesthetic perception. The evaluation model can give aesthetic scores to images, and classify them into groups, e.g. high quality or low quality.

In early time, researchers extracted image features, and then trained the models by using machine learning methods. In order to get good evaluation results, researchers spent huge effort on designing descriptive features based on aesthetic rules which inspired by domain knowledge from painting, photography, art, the human aesthetic feeling and visual attention mechanism [3,5,6,11–13,21] etc. Tong et al. [11] adopted many low-level features such as color histogram and image energy. Ke et al. [19] designed high-level features e.g. spatial distribution of edges and color distribution, and achieved better results with a much smaller number of features. Datta et al. [12] adopted 56-dimensional features, including both low-level features and high-level features. Tang et.al. [3] used both regional and global features to assess images into seven aesthetic categories.

These handcraft features had achieved some good results. However there are two major limitations. 1) The handcraft features are often associated with the principles of domain knowledge. But some domain knowledge is hard to be quantified mathematically. 2) The scene category is not taken into account. People have different aesthetic stimulus when seeing images with different scenes. For example, in Fig. 1 first two rows are images of high quality, and last two rows are low quality images. Each column indicates a scene category. Images in “animal” and “static” categories have clear subject regions. It means that features about clarity and contrast are more important. While in “architecture” class, features towards line and orientation are more important. Specific aesthetic features should be designed for different scenes [3]. However, it's very difficult to exhaust all kinds of aesthetic features. These two limitations are the bottleneck to further improve the performance of aesthetic evaluation.

Afterwards, researchers introduced generic image descriptors [7,13] to describe local information and particulars of images. Marchesotti et al. [13] applied generic content-based descriptors, such as Bag-of-Words, to deal with photo quality assessment. Guo et al. [7] proposed a method by fusing both the handcraft features and local features. These methods achieved positive effect in aesthetic evaluation, but have three major limitations 1) Generic image descriptors do not contain color information. In fact, colors

* Corresponding author.
E-mail address: xmxu@scut.edu.cn (X. Xu).

http://dx.doi.org/10.1016/j.image.2016.05.009
0923-5965/

Fig. 1. Images of 7 kinds of scenes in CUHKPQ dataset [3].

are very important stimulus for aesthetic perception. 2) They focus on local information, and lack the ability to describe images in global perspective. 3) They are not specially designed for aesthetic evaluation. Most of descriptors are usually generated through pooling after extracting SIFT features. The monotonicity on feature designing makes these methods hard to be improved further.

Recently, researchers began to apply deep learning in aesthetic evaluation of images. Deep learning is very successful in solving classical computer vision tasks [8–10,14,15] such as object recognition [16], image classification [17]. It makes full use of deep hidden layers to abstract image information without expert knowledge and eliminates the need for complex feature extraction. Wang et al. [10] designed a double-column CNN (convolutional neural network) for aesthetic quality categorization of images. In addition, they also employed the style and semantic attributes into network to improve classification accuracy and achieved good results. Tian et al. [8] borrowed Alex_CNN [14] to extract image features and trained a two-class SVM classifier to classify aesthetic quality. They also fused these features from CNN with some traditional handcraft features to train a SVM classifier and achieved a good accuracy [9].

Deep learning methods are useful in image aesthetic evaluation and develop rapidly. However, it's still a new area and there is great space for further improvements. In our opinion, there are three major challenges to apply deep learning on aesthetic evaluation of images.

First, existing datasets for aesthetic evaluation have limited images amount and categories, but deep learning needs huge data for training. Previous researches have pointed that deep learning network is suitable for solving complex problems with large quantity of data [15]. For example, the ImageNet dataset [18] has 1.2 million images covering 1000 categories. After using CNN, the classification accuracy is significantly improved. Regarding the effort of aesthetic ranking by human, it is difficult to expand the dataset size and categories for aesthetic evaluation in near future. We need to consider how to take advantage of deep learning method in such circumstance.

Second, the aesthetic stimulus are varies in different scenes. With handcraft features methods, researchers can design specific features for different scenes. With deep learning method, we need to consider how to design and train a network that adapts to various scenes.

Third, there are limitations of current image datasets. The most popular datasets for image aesthetic are CUHKPQ [3] and AVA [4]. In CUHKPQ, the amounts of high quality images and low quality images are unbalanced. It leads to the problem of unbalanced classification. In AVA, there is a lot of noise in aesthetic scores. The noise data can increase the generalization ability of machine learning model, but it will also decrease the description ability in the network and affect the accuracy of evaluation. We need to find a method to reduce the influence of unbalance and noisy from dataset.

Above three challenges motivate us to design a multi-scene deep learning framework for aesthetic evaluation of images. The main contribution of this paper can be summarized as follows.

1. In order to mitigate the problems of small dataset, we introduce the network designed by Alex Krizhevsky [14] into our model. Alex_CNN is trained in advance by ImageNet dataset that contains large number of images and complex classes. The learned deep representations will be transferred to our subsequent processing.
2. To make the network a strong adaptability to different scenes, we design a scene convolutional layer which has multiple groups for different scenes.
3. To improve the model performance, we initialize the model via pre-training procedure, which decreases the influence of the noise, and balances the number of high aesthetic and low aesthetic quality images.

2. The algorithm

In this paper, we design a multi-scene deep learning model (MSDLM) for aesthetic evaluation of images. The architecture overview is shown in Fig. 2.

2.1. Main ideas for the model

2.1.1. The first 4 convolutional layers of the neural network

We set Alex_CNN [14] as the first 4 layers of our network. Alex_CNN is trained in advance by ImageNet dataset that contains large number of images and complex classes. The learned deep representations of Alex_CNN can be transferred to other task according to previous researches [8,9]. With this design, we can achieve satisfactory result with a smaller dataset.

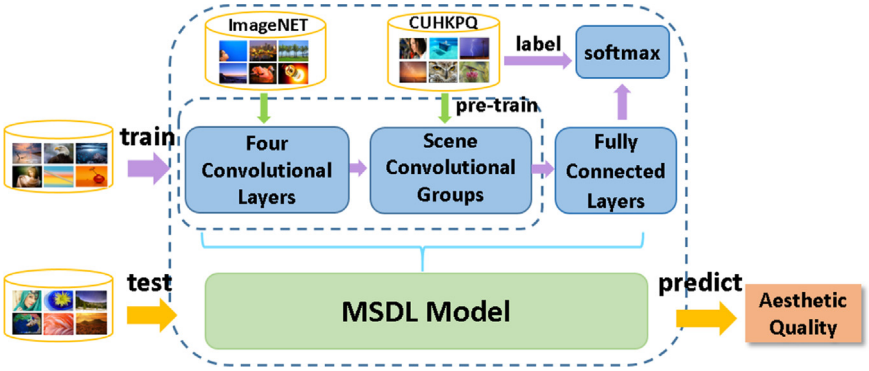

Fig. 2. The overview of proposed MSDLM.

Alex_CNN is used to solve ImageNet [18] classification and has positive results. We choose this network for following reasons. 1) ImageNet [18] is a huge dataset consists of 1.2 million images which are divided into 1000 classes. The deep network trained to classify such a large dataset can accumulate enough knowledge to understand various images deeply. 2) There is close relevance between aesthetic classification and general content classification. Alex_CNN may elicit image features useful for aesthetic evaluation. 3) The Alex_CNN is the simplest network among the ImageNet task networks. It is convenient to expand the network for image aesthetic task.

We set the first four convolution layers of Alex_CNN as the first 4 layers of our network and keep their origin weights as the initial weights of our model.

2.1.2. *The scene convolutional layer*

In our research, the challenge is how to extract the scene descriptors and combine all descriptors for aesthetic classification effectively. We design a scene convolution layer in our network to let the deep network learn descriptors discriminatingly according to various scenes. This layer is consisted of multiple parallel groups. Each group is an independent convolutional network corresponding to a specific scene. We set the number of groups as 7 referencing to Tang's scene categories [3]. The descriptors separately correspond to 7 scenes which are namely “animal”, “architecture”, “human”, “landscape”, “night”, “plant” and “static”.

Subsequently, we set three fully connected layers in our multi-scene deep learning network. Too many layers will lead to overfitting while too few are not enough to get a good result. Therefore, we set the number of layer to three. The last layer corresponds to the number of classes for classification.

The weights of scene convolution layers and fully connected layers are set in the pre-training procedure.

2.1.3. *Pre-training and initialization*

In order to extract aesthetic features for different scenes and reduce the impact of data noise, we design a special pre-training process to initialize the model before formal learning starts. Pre-training procedure has two functions.

1) To reduce the impact of data noise, we chose seven kinds of scene from CHUKPQ [3].
   CUHKPQ [3] dataset and AVA [4] dataset are most popular images dataset used in aesthetic evaluation. In CUHKPQ [3], images are strictly selected and labeled by ten independent viewers after subjective evaluation. The dataset has a high coincidence and less data noise. It contains 7 kinds of scenes, which are clear and rich enough to represent the variety of image aesthetic perception. Every category has the same number of images, it is about 2500. This dataset contains 17,690 images.
   The AVA [4] dataset has more than 250,000 images and 65 kinds of categories. It is larger than CUHKPQ. However, there is much more noise in the data. For example, subjective scores of some images have big variance. In addition, the image numbers of each category are very different. For example, the “maternity” class has only 71 images, while the “nature” class has 28384 images [4].
   In order to mitigate negative influence, we choose images from CUHKPQ [3] which have lower noise to do pre-training. It can reduce the impact of data noise on the network.
2) Pre-train each group to get the aesthetic descriptor for different scenes.
   We train each group independently by using images of one scene class. During the training, the weighs of the corresponding scene's group and the fully connected layer are updated in the network. After all the groups are trained, the pre-training is finished. Then the trained weights of the groups are used as initial value of the scene convolution layer in the network.
   In general, we propose a multi-scene deep learning model, and focus on extracting various scene descriptors for image aesthetic. Our proposed framework has achieved good results in the experiments which will be introduced in the Section 3.

2.2. *The architecture of the multi-scene deep learning model*

As shown in Fig. 3, our model based on an 8-layer deep convolutional neural network, which is consist of 5 convolutional layers and 3 fully connected layers. We adopt the first 4 layers and its weights from the Alex_CNN [14] as our first 4 convolutional layers, in order to transfer the knowledge for multiple kinds of images contents. The fifth layer is considered as scene convolutional layer, which consist of 7 convolutional network groups used to learn image aesthetic descriptors for different categories. The fourth and fifth layers are linked by 7 independent branches. Each branch links one convolutional network group in the fifth layer to the fourth layer.

Different architecture and parameters settings affect the performance of deep learning network. In order to get an efficient model, we tune the number of convolutional kernel and the neurons in fully connected layer to optimize the model performance. We only focuses on the setting of parameters in the fifth convolutional layer and two fully connected layers, because the architecture of the first four convolutional layers are taken from Alex_CNN and kept unchanged. Table 1 shows the accuracy results of various network architectures based on AVA1 dataset (see Section 3.1 for details). Arch1 gets the highest accuracy. The performance of Arch5 is slightly lower than of Arch1. But the number of parameters in Arch5 is much smaller than the Arch1. Considering the size of model, we chose Arch5 as the infrastructure for

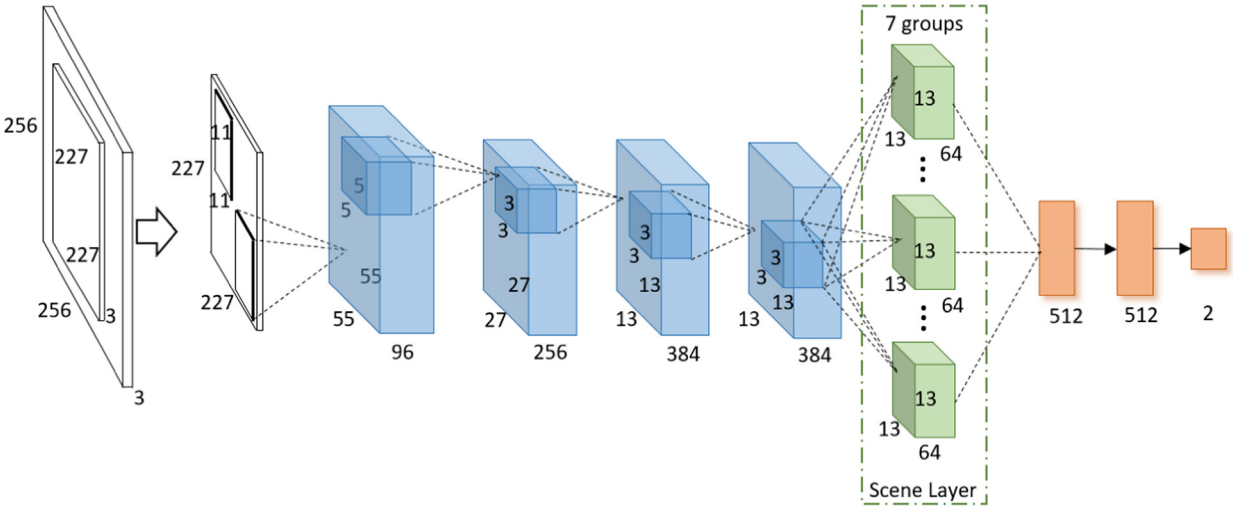

Fig. 3. The architecture of the multi-scene deep learning model (MSDLM).

Table 1
The accuracy and the number of parameters of different architectures.

Architecture		Accuracy (%)	Approximate number of parameters (M)
Arch1	**Conv5(96)_fc4096_fc1024**	**85.07**	**469**
Arch2	Conv5(96)_fc1024_fc512	84.91	117
Arch3	Conv5(96)_fc512_fc512	84.84	58
Arch4	Conv5(128)_fc512_fc512	84.93	78
Arch5	**Conv5(64)_fc512_fc512**	**84.88**	**39**

our multi-scene deep learning model (MSDLM).

Fig. 3 shows the architecture of MSDLM. At first, all images are unified to the size of $256 \times 256 \times 3$, and then cropped to $227 \times 227 \times 3$ patches as input of the network. These three dimensions represent the image width, height and number of channels. This images operation can preserve the image global information and avoid overfitting as well.

The first convolutional layer filters the $227 \times 227 \times 3$ input patch with 96 kernels of size $11 \times 11 \times 3$. The number of filters of the following four convolutional layers is set to 256, 384, 384 respectively. The fifth convolutional layer is the Scene Layer. It is consisted of 7 parallel groups which is set to 64. Max-pooling and normalization are also applied in the first two convolutional layers. The first fully connected layer has 512 neurons connected to the outputs of seven convolutional groups via mean-pooling. The second connected layer has 512 neurons and the last fully connected layer containing 2 neurons is fed to a 2-way softmax.

2.3. Pre-training for different scenes

The training network architecture for each scene is shown in Fig. 4. It is a one column deep convolutional neural network. The first 4 convolutional layers are the same as the first 4 convolutional layers in Fig. 3. The fifth layer is one group of the fifth layer in Fig. 3. The last layer of this network has 2 neurons followed by a softmax function as output, which predict the input image belongs to high quality or low quality.

Our aesthetic classification task is a typical binary classification problem with label $C=\{0,1\}$. For the i-th input image, we know the corresponding feature representation x_i extracted from fc512 layer and the corresponding label $y_i \in C$. Here, the logistic regression is very suitable for our model. We maximize the following log-likelihood function to finish model training:

$$l(\theta) = \sum_{i=1}^{m} (y_i \log h_\theta(x_i) + (1 - y_i)\log(1 - h_\theta(x_i))) \quad (1)$$

Where m is the number of images, the prediction function $h_\theta(x)$ can be expressed as:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)} \quad (2)$$

In pre-training stage, we initialize the weights of first 4 convolutional layers by the Alex's model [14]. The learning rates of Alex_CNN are set to 0.0001. Nevertheless, the learning rates of the fifth layer (feature extraction layer) and fully connected layers are set to 0.001 in our model. By setting a bigger learning rate, the weights of the layers can be mainly updated in this procedure. Moreover, the learning rate is set to be decreased 40% after every 8 epochs.

After we trained the networks and get the weights for 7 groups in the fifth layer respectively, we get 7 kinds of images scene aesthetic descriptors. Then the pre-training procedure is accomplished.

After pre-training, 7 groups are parallel linked to the previous layer together as shown in Fig. 3. The groups are independent to each other. We will show the effectiveness of using pre-training through the experiments in Section 3.

2.4. Model learning

Different from other CNN methods, the initial weights in our network are NOT set random before the formal learning. The baseline model is built after the pre-training procedure. Then, the pre-learned deep representations can be transferred across tasks. We learn the weights of each layer for the specific task (specific dataset) through supervised fine-tuning. Finally we achieve the multi-scene deep learning model for aesthetic evaluation of images.

We construct our baseline model shown in Fig. 3 using following configuration. We keep the origin weights of Alex_CNN as the initial weights in the first four convolution layers. And the weights of the fifth convolutional layer in the 7 scene group are initialized through pre-training procedure. And the initial weights of last 3 fully connected layers are set randomly.

3. Experimental results

In order to evaluate the effectiveness of our model, we carry

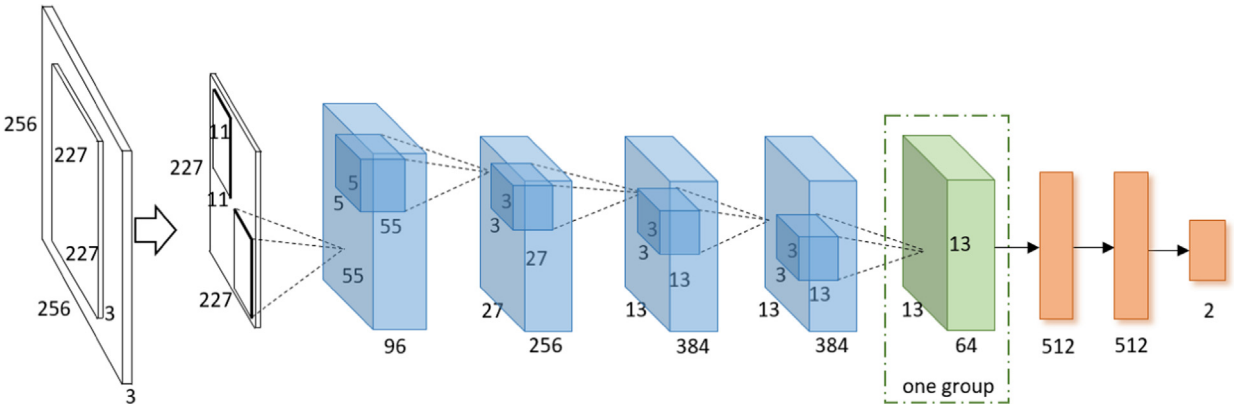

Fig. 4. The deep convolutional neural network in pre-training for one group.

out experiments on two popular datasets namely CUHKPQ [3] and AVA [4], which are frequently used in image aesthetic assessment.

3.1. Dataset

3.1.1. *CUHKPQ dataset*

CUHKPQ [3] is a dataset consists of 17,690 images, which are downloaded from professional photography website. The database is divided into 7 categories according to image scenes, such as "animal", "plant", "static", "architecture", "landscape", "human" and "night". A photo is classified as high or low quality only if eight out of ten reviewers agree on its assessment. Therefore, CUHKPQ dataset has less noise in aesthetic perception. However, this dataset is not balanced. The number of low quality images is three times more than high quality in this dataset.

We use CUHKPQ to pre-train our MSDLM model because it has less data noise. We averagely and randomly divide the dataset into 6 partitions. 4 partitions are used for training, 1 partition for validation and the remaining 1 partition for testing. Among them, training set and validation set are used for pre-training MSDLM to construct baseline model, and test set will be used for testing the accuracy of MSDLM.

In addition, we balanced the training set by adding similar images. We rotate each high quality image by 90° and 270° as new high quality images. In addition to the original images, we increased the high quality images by 3 times. Thus, the number of high quality and low quality images in training set is balanced.

Table 2 shows the number of images in each set after balancing.

3.1.2. *AVA dataset*

AVA [4] is a large scale dataset which contains more than 250,000 of images derived from www.dpchallenge.com. The scores of images are ranged from 0 to 10. The number of votes per image ranges from 78 to 549, with an average of 210. In previous work of deep learning, [9] and [10] constructed two different sub datasets of AVA for experiment. To compare with their performance, we adopted same strategy to construct two sub dataset of AVA.

(1) AVA1: We followed the experimental settings in [9], and ranked all images from high score to low score. Then we pick the images whose ranking orders are in the top 10% and last 10% of the total order. Thus, we select 51,106 images in AVA1 dataset. And all images are evenly and randomly divided into training set and test set, which contains 25,553 images respectively.

(2) AVA2: We followed the experimental settings in [10], and chose the score of 5 as the boundary to divide the dataset into high quality class and low quality class. In this way, there are 74,673 images in low quality and 180,856 images in high quality. According to the imageID list in [4], the training and test sets contain 235,599 and 19,930 images respectively.

3.2. Experiments on CUHKPQ dataset

In this section, we compared our model with state-of-the-art models, including traditional feature extraction models [3,7,19] and some other deep learning models [9,10] on CUHKPQ dataset.

Table 3 shows the average aesthetic evaluation accuracy on total dataset and the accuracy of each kind of scene. Our model gets the highest average accuracy on total dataset, which is 92.59%. Compared with the models using handcraft features, the accuracy of our model is obviously higher than the earlier handcraft method in [19], and is 1.38% higher than Tang's method [3]. Compared with models of generic image descriptors, the accuracy of our model is 4.72% higher than Guo's model [7] which is the highest in this field. Compared with deep learning model [10], the performance of our model is still the best.

Looking over the aesthetic accuracy in seven scenes, accuracies of our model are obvious higher than other methods, except that in "human" and "landscape" scenes the accuracies of ours are about 1% less than those in Tang's method [3].

In addition, we design some experiments to demonstrate the performance enhancement by using the pre-training procedure and the paralleled groups in the scene layer. We train a model named single-column convolutional neural network (SCNN) as Fig. 4 without the paralleled groups in scene layer. We also train a

Table 2
The number of images of each scene in training, validation and test set.

Scene	Animal		Architecture		Human		Landscape		Night		Plant		Static	
Number of images	High quality	Low quality	High quality	Low quality	High quality	Low quality	High quality	Low quality	High quality	Low quality	High quality	Low quality	High quality	Low quality
Training set	1746	1539	1092	870	1398	1657	1542	1323	717	910	1200	1213	1074	1331
Validation set	180	371	112	205	99	402	142	302	55	220	95	290	84	340
Test set	191	382	119	215	113	411	164	325	59	226	99	300	89	334

Table 3
The accuracy of different models in experiments using CUHKPQ dataset.

Scene		Animal	Architecture	Human	Landscape	Night	Plant	Static	Overall
Handcrafted features	All features in [19][a]	0.7751	0.8526	0.7908	0.8170	0.7321	0.8093	0.7829	0.7944
	All features in [3]	0.8867	0.9100	**0.9662**	**0.9266**	0.8403	0.9004	0.9041	0.9121
Generic image descriptors	Semantic features in [7]	0.8623	0.8644	0.9313	0.8416	0.8742	0.8685	0.8964	0.8787
	Semantic features+handcraft features [7]	0.9033	0.8755	0.9472	0.8853	0.9052	**0.9232**	0.9094	0.9093
Deep learning methods	DCNN_Aesth_SP [9]	–	–	–	–	–	–	–	0.9193
	SCNN	0.8978	0.8988	0.9244	0.9106	0.8905	0.8919	0.8974	0.9020
	MSCNN	0.9107	0.9003	0.9271	0.8914	0.9045	0.9152	0.8893	0.9092
	MSDLM	**0.9211**	**0.9150**	0.9492	0.9177	**0.9169**	0.9192	**0.9278**	**0.9259**

[a] These results are referenced from [7].

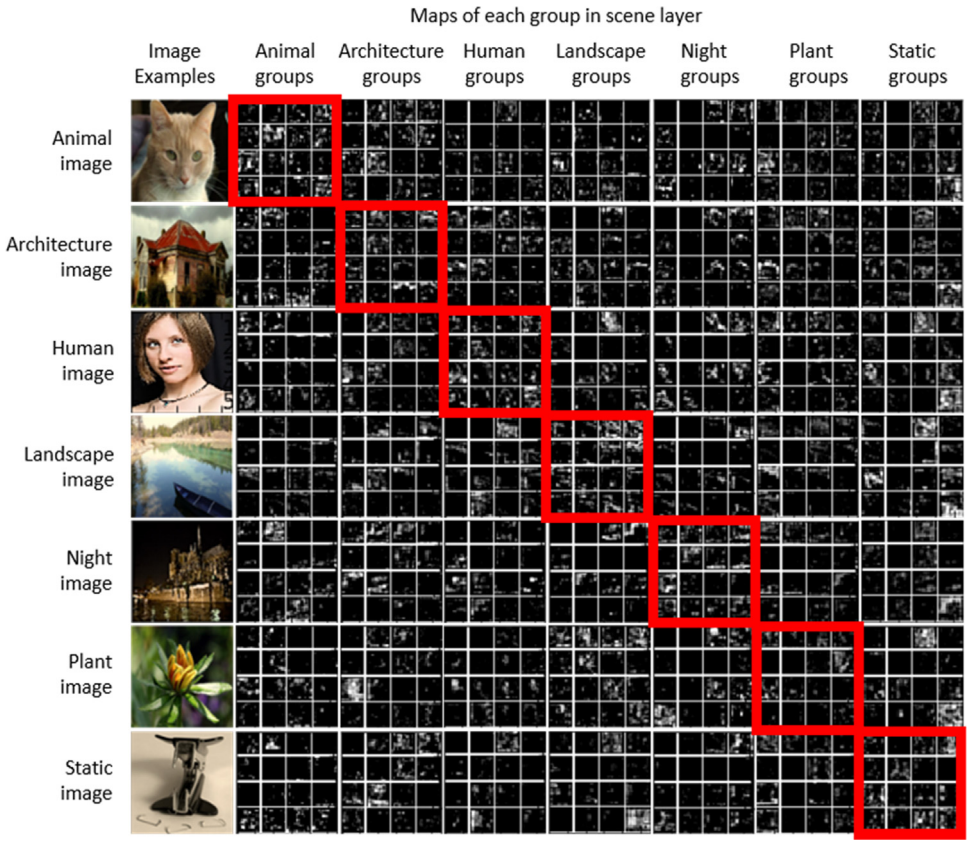

Fig. 5. The visualization of maps in the scene layers.

Table 4
The accuracy of different methods on AVA1 dataset.

	Methods in AVA1	Accuracy
Handcrafted features	Datta [12][a]	68.67%
	Ke [19][a]	71.06%
Deep learning methods	DCNN_Aesth_SP [9]	83.52%
	SCNN	81.61%
	MSCNN	83.57%
	MSDLM	**84.88%**

[a] These results are referenced from [9].

Table 5
The accuracy of different methods on AVA2 dataset.

Methods in AVA2	Accuracy
RDCNN_semantic [10]	75.42%
SCNN	72.85%
MSCNN	74.06%
MSDLM	**76.94%**

model named multi-scene convolutional neural network (MSCNN) as Fig. 3 without pre-training procedure. SCNN and MSCNN are parts of MSDLM. As shown in Table 3, the accuracy of MSCNN is higher than SCNN, which verified the effectiveness of the paralleled layer design. And the MSDLM is higher than MSCNN which verified the effectiveness of the pre-training strategy.

In order to show the effectiveness of the scene layer in our network, we visualize the scene layers of several images from each scene in Fig. 5. Each group of scene layer contains 64 maps with the size of 13*13. The maps according to a certain input image represent the descriptors introduced by the MSDLM for the image. Limited by the space of the paper, we visualize the former 16 maps of each group for an input image. The row of Fig. 5 corresponds to the responding maps of 7 scene groups in the scene layers for one image. Red boxes are the cells whose image categories and scene groups are matched. It is apparent that the maps in the read boxes contain more efficient response for the input images, comparing with other maps in the same row. This verifies our initial

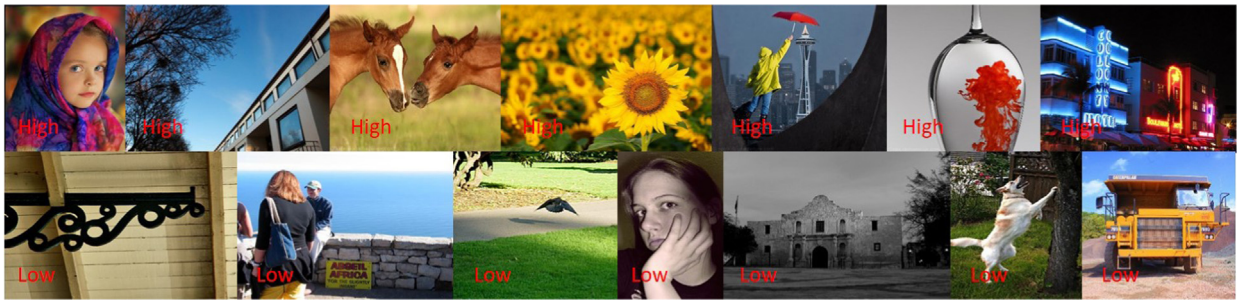

Fig. 6. Examples correctly classified by MSDLM and misclassified by MSCNN.

hypothesis that each scene category has its corresponding unique neurons. We can further confirm that the use of pre-training methods help to improve classification accuracy.

3.3. Experiments on AVA dataset

We also train the model and test its performance on AVA dataset. The results of experiments are shown in Tables 4 and 5.

The results on AVA1 are shown in Table 4. We can see that, the performance of MSDLM is the best with an accuracy of 84.88%. In general, the performance of deep learning model is better than the traditional model using handcraft features. The accuracy of MSCNN is even higher than that of DCNN_Aesth_SP [9], which is 83.57%. It shows that the network architecture is effective even without pre-training.

The results on AVA2 are shown in Table 5. We can see that, the performance of our MSDLM model is still the best with an accuracy of 76.94%. The accuracy of RDCNN_sementic [10] is 75.42%, which is 1.52% less than our model.

We also analyzed the advantage of the MSDLM using the paralleled scene layer and pre-training procedure. As shown in Tables 4 and 5, the accuracy of MSDLM is higher than SCNN and MSCNN. We present the examples in Fig. 6, where we show typical test images that have been correctly classified by MSDLM but misclassified by MSCNN. Basically we can observe that those images contain obvious objects or foregrounds are more easily classified correctly.

4. Conclusions

In this paper, we propose a multi-scene deep learning model (MSDLM) for image aesthetic evaluation, which adopt a novel deep neural network and pre-training strategy. In the design of network construction, we reference the first 4 convolutional layers from Alex_CNN [14] so that our network can gain strongly descriptive ability on image content. In addition, we design a scene convolutional layer with parallel arranging groups to learn related aesthetic descriptors for different scenes images. Through the carefully designed pre-training procedure, we get an initialized model with prior knowledge. Then the model could be trained for a specific task on its own dataset.

The model we proposed solves three problems when applying deep learning method on aesthetic evaluation of images. (1) Deep learning network cannot fully perform its advantage when the number and classes of images in the dataset are insufficient for training the model. (2) The common networks lack the comprehensive ability in extracting aesthetic features for different image scenes. (3) There're a lot of data noises in datasets and the datasets are usually unbalanced.

The results of experiments on two large scale benchmark datasets CUHKPQ and AVA have shown the powerful descriptive ability of our model. The results also show that the accuracy of image aesthetic evaluation can be increased after adopting pre-training procedure.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province #2015A030313212, Natural Science Foundation of China (NSFC) #61401161, and the Science and Technology Planning project of Guangdong Province of China #2014B010111003, #2014B010111006, and the State Scholarship Found of China #201506155081, and the National Engineering Technology Research Center of Mobile Ultrasonic Detection #2013FU125X02.

References

[1] W. Yin, T. Mei, C. W. Chen, Assessing photo quality with geo-context and crowd sourced photos, in: Proceedings of the IEEE Visual Communications and Image Processing (VCIP), 2012, pp. 1–6.
[2] W. Wang, W. Zhao, C. Cai, J. Huang, X. Xu, L. L, An efficient image aesthetic analysis system using Hadoop, Signal Process.: Image Commun. 39 (2015) 499–508.
[3] X. Tang, W. Luo, X. Wang, Content-based photo quality assessment, IEEE Trans. Multimed. 15 (2013) 1930–1943.
[4] N. Murray, L. Marchesotti, F. Perronnin, AVA: a large-scale dataset for aesthetic visual analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2408–2415.
[5] W. Wang, J. Yi, X. Xu, L. Wang, Computational aesthetics of image classification and evaluation, J. Comput.-Aided Des. Comput. Graph. 26 (2014) 1075–1083.
[6] W. Wang, D. Cai, L. Wang, Q. Huang, X. Xu, X. Li, Synthesized computational aesthetic evaluation of photos, Neurocomputing 172 (2016) 244–252.
[7] L. Guo, Y. Xiong, Q. Huang, X. Li, Image esthetic assessment using both hand-crafting and semantic features, Neurocomputing 143 (2014) 14-–126.
[8] Z. Dong, X. Tian, Multi-level photo quality assessment with multi-view features, Neurocomputing 168 (2015) 308-–3319.
[9] Z. Dong, X. Shen, H. Li, X. Tian, Photo quality assessment with DCNN that understands image well, in: Proceedings of the International Conference on MultiMedia Modeling (MMM), 2015, pp. 524–535.
[10] X. Lu, Z. Lin, H. Jin, J. Yang, J. Wang, Rating pictorial aesthetics using deep learning, in: Proceedings of the ACM Conference on Multimedia, 2014, pp. 457–466.
[11] H. Tong, M. Li, H.J. Zhang, J. He, C. Zhang, Classification of digital photos taken by photographers or home users, in: Proceeding of Pacific-Rim Conference on Multimedia, 2004, pp. 198–205.
[12] R. Datta, D. Joshi, J. Li, J. Wang, Studying aesthetics in photographic images using a computational approach, in: Proceeding of the European Conference on Computer Vision (ECCV), 2006, pp. 288–301.
[13] L. Marchesotti, F. Perronnin, D. Larlus, G. Csurka, Assessing the aesthetic quality of photographs using generic image descriptors, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1784–11791.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolution neural networks, in: Proceedings of the Annual Conference on Neural Information Processing System (NIPS), 2012, pp. 1097–1105.
[15] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1891–1898.
[16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2009, pp. 2146–22153.

[17] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 609–616.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2009, pp. 248–255.
[19] Y. Ke, X. Tang, F. Jing, The design of high-level features for photo quality assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 419–426.
[20] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, J. Luo, Aesthetics and emotions in images, IEEE Signal Process. Mag. 28 (2011) 94–115.
[21] W. Wang, D. Cai, X. Xu, A. Liew, Visual saliency detection based on region descriptors and prior knowledge, Signal Process.: Image Commun. 29 (2014) 424–433.