A Small Scale Multi-Column Network for Aesthetic Classification Based on Multiple Attributes

Chaoqun Wan and Xinmei Tian(✉)

CAS Key Laboratory of Technology in Geo-spatial Information Processing
and Application System, University of Science and Technology of China,
Anhui 230027, China
wancq14@mail.ustc.edu.cn, xinmei@ustc.edu.cn

Abstract. Image aesthetic quality assessment, which devotes to distinguishing whether an image is beautiful or not, has drawn a lot of attention in recent years. Recently deep learning has shown great power in data analysis and has been widely used in this field. However, on the one hand, deep learning is an end-to-end learning method that can be easily influenced by noisy data. On the other hand, prior information concluded from the experience of human perception of aesthetics, which widely applied in traditional aesthetic assessment methods, has not been effectively utilized in deep learning based aesthetic quality assessment methods. Therefore, in this paper we embed these prior information in deep learning as guidance for aesthetic quality assessment. Firstly, we design an extremely small network with only 38 K parameters for better training. Then we propose a multi-column network architecture to embed prior information into our deep learning model. We train our proposed network on AVA dataset, which is widely used for aesthetic assessment. The experimental results show that prior information indeed guides our network to learn better.

Keywords: Aesthetic quality assessment · Deep learning · Multi-Column · Prior information

1 Introduction

Image quality assessment from the aspect of aesthetics has been a hot topic for a long time in computer vision. It aims to search inner factors of aesthetic, which will help computer perceive beauty like what human do. Figure 1 shows a group of examples to assess whether an image is beautiful or not. Image aesthetic quality assessment has a wide range of applications. For example, it can help human to automatically analyse other kinds of mental phenomenons, guide people to take more beautiful pictures and automatically manage their albums. However, as aesthetic is a highly subjective, experiential and mentality-related perception, there are no specific rules for computer even human to make accurate decisions. Thus, it is a tough but attractive challenge for researchers.

D. Liu et al. (Eds.): ICONIP 2017, Part I, LNCS 10634, pp. 922–932, 2017.
https://doi.org/10.1007/978-3-319-70087-8_94

To address this challenging problem, researchers have done a lot of work through analysis for image aesthetics [1–3,5,7,11–13,15,16,19,20]. Image aesthetic quality assessment is usually simplified as a binary classification problem, i.e., we aim to divide images into two classes: high quality (beautiful) or low quality (unbeautiful). In traditional ways, researchers searched aesthetic-related attributes and modeled the relations between attributes and aesthetics. They made a lot of efforts to analyse through photography as well as psychology, and obtained experience from human intuition. Thus, attributes like color, layout, clarity etc. [1–3,5,7,12,15,16,19] which would influence the task in a large scale, are regarded highly related to aesthetics. We consider these attributes as significant prior information for our task. Because, compared to the abstract conception of "aesthetics", attributes are proved more intuitive and easier to represent. Researchers further designed features from varies perspectives, which will describe those attributes in mathematical ways.

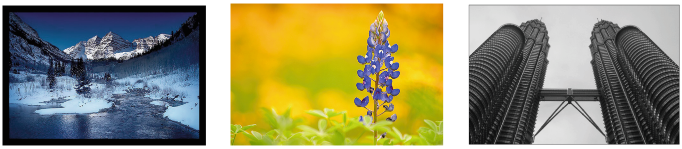

(a) (b) (c)

Fig. 1. Examples for beautiful images. Three images use different color pattern, layout structure and clarity contrast. (a) is a landscape image using cold tone with middle clarity. (b) is a single object image using warm tone with low depth of field. (c) is a symmetrical multi-object image using black-white tone with good clarity.

However, traditional methods have critical drawbacks due to the limit of human cognition. There are no specific rules to describe how these attributes influence aesthetics. Figure 1 illustrates the problem with a group of concrete examples. It's hard for researchers to design accurate features for the sake of modeling the relations between attributes and aesthetics. Recently, based on the structure of neurons in brain, Deep Convolutional Neural Network (DCNN) is designed to train data directly. Through DCNN, deep learning shows powerful ability to analyse inner relations among large scale of data. Dong et al. first introduced DCNN-AlexNet into aesthetic assessment in place of the generic features. Their results proved that DCNN outperformed traditional methods based on handcraft features [4]. Nevertheless, the characteristic of end-to-end learning in DCNN is both its advantage and disadvantage. This kind of data-driven method indeed has strong ability to model the relations between data and task, but it will also exposure DCNN under noisy data, which will lead DCNN to learn astray.

Thus, in this paper, we consider prior information from human perception to be a kind of perfect guidance. As we mentioned above, prior information describes that color, layout, clarity etc. are highly aesthetic-related attributes.

Based on this conception, we hope to adopt DCNN to learn these attributes and further describe aesthetics of images. Classical network architectures are in a large scale. However, on the contrary, the datasets for aesthetic assessment are relative small. In order to obtain efficient learning result, we abandon these large networks but design an extremely small one especially for our task. Then, we proposed a method to "teach" our designed DCNN to learn specified attributes. Finally, we combine these networks that learn different attributes by multi-column approach. The experimental results show that introduced prior information indeed guide our network to learn in a better way. Meanwhile, without fusion, the small scale network are more excellent than those large ones.

The rest of this paper is organized as follows. In Sect. 2, we will give an overview of related works. Then, details of our proposed method will be represented in Sect. 3. Experimental details as well as results are discussed in Sect. 4. Finally, conclusions and future work will be shown in Sect. 5.

2 Related Work

In this section, we will introduce some related works. At first, some traditional methods and a few of conclusions from their results will be presented. Then, we will review some recent DCNN structures about image aesthetic quality assessment.

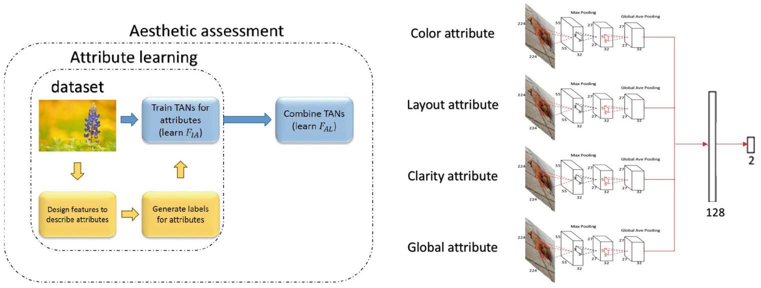
Traditional Method. In traditional ways, researchers focused on designing hand-crafted features to model attributes that are highly related to aesthetics. They adopted different methods from various perspectives to design low-level [2,19], high-level [3,5,7,12,15,16] or generic features [13]. Low-level features are a series of statistic values from original images or their transformation. Tong et al. [19] used the clarity, colorfulness, saliency map etc. to express images. Datta et al. [2] considered some classical rules in photography like "rule of thirds", "good exposure" and proposed a 56-dimensional statistical vector. As for high-level features, they are better designed based on human cognition from psychology and photography. Ke et al. [7] proposed seven kinds of well designed features to describe simplicity, contrast, brightness etc. of images. Luo et al. [12] extracted the subject region from a photo to compare with the background. Luo and Wang [16] first considered rules for aesthetics would vary based on different image content. Dong et al. proposed a 26-dimensional feature vector from five aspects [5]. Generic features could extract global information based on image content and also performed good [13].

Although traditional methods varies from researchers to researchers, there exists some generality in attributes for aesthetics. First, color is considered by most researchers. Tokumaru et al. [18] proposed eight patterns to describe color harmony. Second, there are many rules for image composition (layout), like "rule of thirds", "symmetry", "visual balance". Researchers in [1–3,5,7,16] regarded layout structure as an important aspect that influenced aesthetics. Third, clarity is an obvious indicator for image aesthetics. High resolution images are always more attractive than these low resolution ones [2,3,5,7,12,16]. In summary, color pattern, layout structure and clarity are three most significant aspects in aesthetic assessment.

Deep Learning Method. It is hard for researchers to discover all related attributes, while relations between attributes and aesthetics are intangible. Recently deep learning shows its great power in a great variety of fields. Dong et al. [4] first introduced deep learning into aesthetic assessment. They adopted AlexNet to extract generic features like what Marchesotti did [13]. After that, Lu et al. [10] attempted to train network for aesthetics. They adjusted architecture of AlexNet and achieved promising performance. Wang et al. [20] considered the distinction between different categories and proposed a multi-scene DCNN that was modified from AlexNet architecture. They replaced the fifth convolutional layer by seven sub-convolutional layers, which were pre-trained from images of predefined categories. Besides, Lu et al. [11] designed DMA-net, which was more concerned about details in images. Dong et al. [17] proposed a small network architecture, which had only two convolutional layers and three fully connected layers. The architecture in [17] gives us good reference to design better small scale network for aesthetic quality assessment.

3 Multi-Column Network for Aesthetic Classification Based on Multiple Attributes

In this part, we first introduce our proposed method from an overall perspective. Then, more details about the proposed small scale network and the training procedure for different attributes will be discussed.



(a) (b)

Fig. 2. The pipeline of proposed method is illustrated in (a). First, we extract features based on traditional methods to describe attributes. Then labels for attributes can be generated by K-means. Through these labels, TANs can be trained for attributes. Finally, we combine four networks by multi-column and train with aesthetic labels. The fusion network is shown in (b).

3.1 Overview

Consider an image as an input called I, our task is to predict the label L (0 for low quality and 1 for high quality) for it. Classical DCNNs directly learn the mapping function F from I to L, which means $L = F(I)$. However, based on prior information, we would like to transfer I into combination of attributes $A = \{a_1, a_2, a_3, ...\}$ through transfer function F_{IA}, where $A = F_{IA}(I)$ and a_i is the representation of the ith attribute. And then we would learn the mapping function F_{AL} from A to L. Thus, our whole procedure can be described as $L = F_{AL}(F_{IA}(I))$, where the mapping functions F_{AL} and F_{IA} are learned through DCNNs. Meanwhile, as F is a complicated function that is hard to learn, it is obvious that splitting F apart into F_{AL} and F_{IA} is more proper for DCNN to learn. Figure 2(a) shows the whole pipeline and relations mentioned above.

3.2 Tiny Aesthetic Network

We call the proposed small scale network the Tiny Aesthetic Network (TAN). This network has only 37,760 parameters, whose architecture is schematically illustrated in Fig. 3. Two convolutional layers with two fully connected layers constitute the whole structure. There are 32 kernels in each convolutional layer. Kernels in the first convolutional layers are in size of $11 \times 11 \times 3$ with a stride of 4. For the kernels in the second convolutional layers, they are in size of $5 \times 5 \times 32$ with a stride of 1. There are one normalized layer and one pooling layer behind each convolutional layer like AlexNet [8]. However, we adopt global average pooling [9] to replace max pooling in the second pooling layer. The two fully connected layers have 16 and 2 neurons respectively.

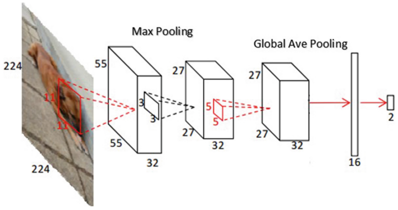


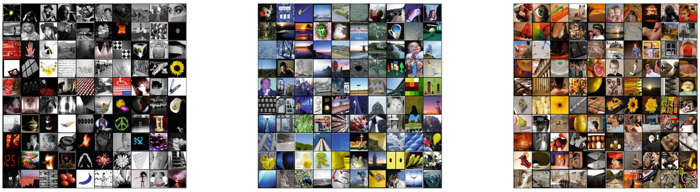
Fig. 3. Overview of the small network structure. It consists of two convolutional layers and two fully connected layers with global average pooling between the second convolutional layer and the first fully connected layer.

3.3 Networks for Different Attributes

Since prior works regard color, layout and clarity as three most important attributes that are most relevant to aesthetics, we hope that TAN can automatically learn corresponding features to represent these attributes. Here we proposed an approach to generate the labels for attributes in the basis of traditional method. Then, we can "teach" TAN to learn features related to these attributes through the attribute labels.

In the traditional method, researchers design hand-crafted features to represent these attributes. We imitate their method in [5] and obtain features for color, layout and clarity. The feature vector for color and clarity are the same, while for layout, we adopt canny detection to locate subject area [5]. And then we compute the center location of the bounding box as well as the width-ratio and length-ratio of box and image. These constitute a 4-dimensional vector to represent layout.

We consider these features are able to reflect some inner factors about corresponding attributes. So, unsupervised K-means is a proper method to reveal these relations hidden among data. For each attribute, images are clustered into K different classes through K-means and generate attribute labels. Images in the same class will be similar in the aspect of corresponding attribute. Figure 4 shows the clustering result of color attribute, from which we can observe that different class has different color pattern. In our experiment, K is set to 3. In other words, for each attribute we cluster all training images in to 3 classes and use the cluster label as attribute label to train TAN (the number of neurons in the last layer of TAN is 3 here). In this way, TAN can automatically learn features that are closely related to the corresponding attribute.



(a) black-white tone (b) cold tone (c) warm tone

Fig. 4. Result of clustering based on color attribute. (a)–(c) are black-white tone, cold tone and warm tone respectively.

In order to assess aesthetic through these attributes, we combine TANs for color, layout and clarity through multi-column to form the fusion network named "TAN_attribute". Considering that some remaining information may exist in global conten. We also train a TAN for aesthetic to obtain remaining attribute, which we call "TAN_global". "TAN_global" is trained by using the aesthetic labels directly. Then, we further fuse "TAN_global" with "TAN_attribute" to form the final fusion network named "TAN_attribute_global". We combine networks by concatenating the outputs of global average pooling. The architecture of "TAN_attribute_global" is shown in Fig. 2(b), while architecture of "TAN_attribute" is similar but has only three columns instead (without "TAN_global").

4 Experiment

In this section, we will introduce our experiment procedure. Following the settings in previous works, we deal with aesthetic assessment as a binary classification problem. The training and testing are conducted on AVA dataset, which is widely used in this field. We first compare the proposed method with traditional ones. Then, we compare our proposed models with state-of-the-art deep learning based methods, from both performance and efficiency.

4.1 Dataset

AVA is a large-scale dataset for aesthetic assessment [14]. It consists of more than 250000 images downloaded from DPChallenge.com. Each image in AVA has 210 scores in average. A single overall score was obtained to indicate the aesthetic quality of each image by averaging all of its individual scores. Similar to what was done in [7,17], the top 10% and bottom 10% of the photos were designated as high-quality (beautiful) and low-quality (not beautiful) images, respectively, and the ambiguous images in the middle of the quality range were discarded. We randomly selected half of the images for training and the remaining images for testing.

4.2 Experimental Setting

When training for each attribute, we first randomly crop resized images to get an input of $224 * 224$ in size. Then, during training, learning rate is set to be a fixed number 0.01, while train iteration is 30,000. For other parameters, weight_decay is 0.0005, batch_size is 256 and clip_gradient is 10. Besides, we use "msra" [6] to initialize.

When training for fusion model, we initialize convolutional layers with previous trained attribute TANs, but initialize fully connected layers by "msra" [6]. The learning rate of convolutional layer is set 0.001, while learning rate of fully connected layer is 0.01. To prove that our multi-column method is indeed useful, we expand the kernel (neural) number in each layer of TAN to reach the same scale as the fusion model. We change the kernel number in TAN into 64 and 128, and neural number of fully connected layer into 64 and 128 respectively. These two TAN networks are termed "TAN_expand_64" and "TAN_expand_128". Compared with "TAN_attribute_global", "TAN_expand_64" has the approximate same amount of parameters, while "TAN_expand_128" has the same kernel (neural) number in each layer.

4.3 Experimental Results and Analysis

The experiment results are shown in the tables. We firstly compare our proposed method with traditional state-of-the-art methods based on hand-crafted features. As shown in Table 1, we can find our proposed models outperform these methods significantly. Besides, the performance of "TAN_attribute" achieves 82.12%,

Table 1. Classification accuracy (%) comparison between proposed method and traditional methods based on hand-crafted features.

Methods	Accuracy (%)
Luo [12]	61.49
Datta [2]	68.67
Ke [7]	71.06
Marchesotti [13]	68.55
Dong [5]	77.35
TAN_global	81.71
TAN_attribute	82.12
TAN_attribute_global	83.32

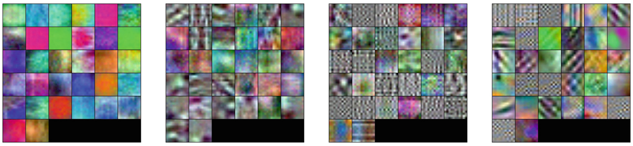
which outperforms "TAN_global". Moreover, after adding global information, the "TAN_attribute_global" achieves a better performance reaching 83.32%. This result shows that prior information can help DCNN learn better.

Then, we compared our proposed model with existing DCNN based image aesthetic quality assessment models. The results are summarized in Table 2. We can observe that "TAN_global" achieves better performance than these large scale networks. Through combination, "TAN_attribute" as well as "TAN_attribute_global" both improve the ability of original network. To avoid the influence of parameter increasement, we further compare fusion model to "TAN_expand_16" and "TAN_expand_32". Results show both expanded networks become even worse, which proves that this kind of straightforward strategy will only result in overfitting. On the contrary, based on prior information, our fusion network will perform better.

Table 2. Comparison with classical DCNN architecture from both classification accuracy and network scale.

Methods	Accuracy (%)	Number of parameters
RAPID [10]	74.54	≥47 M
DCNN [17]	75.89	124 k
DCNN_Aesth [4]	78.92	201 M
SCNN [20]	81.61	39 M
TAN_global	81.71	38 K
TAN_expand_64	80.91	130 K
TAN_expand_128	80.13	473 K
TAN_attribute	82.12	121 K
TAN_attribute_global	83.32	165 K

For the sake of better analysis, we further visualize kernels in the first convolutional layers. Figure 5 shows the results. We can observe that each kind of TAN authentically learns attribute-related kernels. For example, kernels in color TAN are mainly related to pure color. Kernels in layout are mainly related to edges with directions. Kernels in clarity are mainly related to different frequencies. Kernels in global TAN contain all types mentioned above, and some of them are more likely as a combination of different kernel types.



(a) color (b) layout (c) clarity (d) global

Fig. 5. Visualization of kernels in the first convolutional layers. (a)–(d) are kernels from TAN for color, layout, clarity and global respectively.

5 Conclusion and Future Work

In this paper, we propose a small scale multi-column network to embed prior information, for image aesthetic quality assessment. Our method has the merits of both deep learning based models and the traditional hand-crafted feature based models. By incorporating attributes into our model, the performance is successfully improved. Besides, our model is in small scale but outperforms existing large scale deep networks.

Although we propose an efficient method to introduce prior information into DCNN, not all TANs for attributes work well. For three attribute TANs, test accuracy for each attribute classification can reach 89% in color and 82% in clarity, but only 71% in layout. Besides, features from traditional method are not accurate, which will result in noisy attribute labels in K-means. Thus, our future work will focus on searching more complicated and proper network for those attributes modeling. Meanwhile, we will find more effective method to introduce prior information into deep learning.

Acknowledgments. This work is supported by the 973 project 2015CB351803, NSFC No. 61572451 and No. 61390514, Youth Innovation Promotion Association CAS CX2100060016, and Fok Ying Tung Education Foundation WF2100060004.

References

1. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 271–280. ACM (2010)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006). doi:10.1007/11744078_23
3. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1657–1664. IEEE (2011)
4. Dong, Z., Shen, X., Li, H., Tian, X.: Photo quality assessment with DCNN that understands image well. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8936, pp. 524–535. Springer, Cham (2015). doi:10.1007/978-3-319-14442-9_57
5. Dong, Z., Tian, X.: Effective and efficient photo quality assessment. In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 2859–2864. IEEE (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
7. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 419–426. IEEE (2006)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
10. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: rating pictorial aesthetics using deep learning. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 457–466. ACM (2014)
11. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 990–998 (2015)
12. Luo, Y., Tang, X.: Photo and video quality evaluation: focusing on the subject. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88690-7_29
13. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1784–1791. IEEE (2011)
14. Murray, N., Marchesotti, L., Perronnin, F.: AVA: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2408–2415. IEEE (2012)
15. Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40. IEEE (2011)
16. Tang, X., Luo, W., Wang, X.: Content-based photo quality assessment. IEEE Trans. Multimed. **15**(8), 1930–1943 (2013)

17. Tian, X., Dong, Z., Yang, K., Mei, T.: Query-dependent aesthetic model with deep learning for photo quality assessment. IEEE Trans. Multimed. **17**(11), 2035–2048 (2015)
18. Tokumaru, M., Muranaka, N., Imanishi, S.: Color design support system considering color harmony. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002), vol. 1, pp. 378–383. IEEE (2002)
19. Tong, H., Li, M., Zhang, H.-J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 198–205. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30541-5_25
20. Wang, W., Zhao, M., Wang, L., Huang, J., Cai, C., Xu, X.: A multi-scene deep learning model for image aesthetic evaluation. Signal Process. Image Commun. **47**, 511–518 (2016)