IMAGE AESTHETIC ASSESSMENT VIA DEEP SEMANTIC AGGREGATION

Kung-Hung Lu, Kuang-Yu Chang, and Chu-Song Chen

Institute of Information Science, Academia Sinica, Taipei, Taiwan.

ABSTRACT

Aesthetic quality estimation of an image is a challenging task. In this paper, we introduce a deep CNN approach to tackle this problem. We adopt the sate-of-the-art object-recognition CNN as our baseline model, and adapt it for handling several high-level attributes. The networks capable of dealing with these high-level concepts are then fused by a learned logical connector for predicting the aesthetic rating. Results on the standard benchmark shows the effectiveness of our approach.

***Index Terms*—** Aesthetic quality assessment, deep CNN, semantic aggregation, OWA operator

1. INTRODUCTION

Over the years, there has been a growing interest in assessing the aesthetic quality (AQ) of an image. Automatic AQ assessment is useful in many applications such as picture editing softwares and image retrieval systems. In the past, AQ's evaluation was often handled by using hand-crafted feature extraction methods such as the rule of thirds, golden ratio and color distribution [1, 2, 3]. Thanks to the progress of deep convolutional neural network (CNN) that achieves a considerable improvement on image recognition, recent studies toward this direction have employed deep CNN for AQ assessment to overcome the feature representation problem and reached better performance [4, 5].

However, image AQ's assessment could be associated with many high-level concepts reflected by an image, such as the general semantics [6], the object classes, the scene categories [7, 8], and even the emotions conveyed by an image [9]. A deep CNN that is expert at recognizing objects is not necessarily the most suitable for AQ's evaluation. In this paper, we employ the current state-of-the-art deep residual network (ResNet) [10] for the AQ assessment. In particular, we adapt the ResNet to the new models capable of identifying the high-level attributes (such as semantics, scene and emotion). Then, the features extracted from different network models (a.k.a. the expert networks) are aggregated to make the final AQ decision. Our approach is illustrated in Figure 1.

Suppose that n high-level attributes are used, and each expert network generates an m-dimensional feature vector. How to fuse the expert opinions becomes an issue. We seek to pool these n feature vectors to a single m-dimensional feature vector that serves for the AQ assessment. Typical pooling methods include max, min, and averaged pooling. In this paper, instead of choosing the pooling operator in advance, we propose to learn it automatically. We employ the order-weighted-average (OWA) [11] operator that can represent a general class of logical connectors parametrically. From OWA, we then introduce a new pooling layer that finds an appropriator logical connector to aggregate the feature vectors generated from the high-level attributes extracted by the deep networks. Main characteristics of this paper include:

• We adopt the state-of-the-art deep CNN for object recognition and adapt it to the new models for handling high-level concepts including semantics, scene, and emotion. We then exploit them for the AQ assessment. The experimental results show that these deeply learned high-level attributes perform well for AQ's evaluation.

• We introduce an OWA pooling layer to aggregate the multiple features deeply extracted. The OWA layer is flexible and can automatically learn the aggregation rule, which further improves the performance.

The rest of the paper is organized as follows. We review the related work in Section 2, and introduce the high-level features extracted in Section 3. The aggregation layer is described in Section 4. Experimental results and conclusions are given in Sections 5 and 6, respectively.

2. RELATED WORK

AQ Assessment. Photo AQ assessment has been investigated for many years [2, 3, 7, 12, 13, 14]. Most of the researches focus on designing aesthetic features, e.g., the rule of thirds, sky illumination, simplicity, etc. The designed image descriptors can outperform traditional low level features, such as color histograms, hue, saturation, value, and so on. In addition to the well-design features, aesthetic image datasets are also important. Though various aesthetic images datasets have been proposed, most of them are not of large scale or contain less annotations. Murray et al. [6] introduced a large-scale aesthetic dataset, Aesthetic Visual Analysis (AVA) that contains rich annotations. We use the rich annotations for training the expert network and evaluate our approach on this dataset.

Deep CNNs. Recently, deep CNN has advanced in object classification and detection. Via modern training skills such as ReLu activation (that speeds up the error descent) and

978-1-5090-4545-7/16/$31.00 ©2016 IEEE

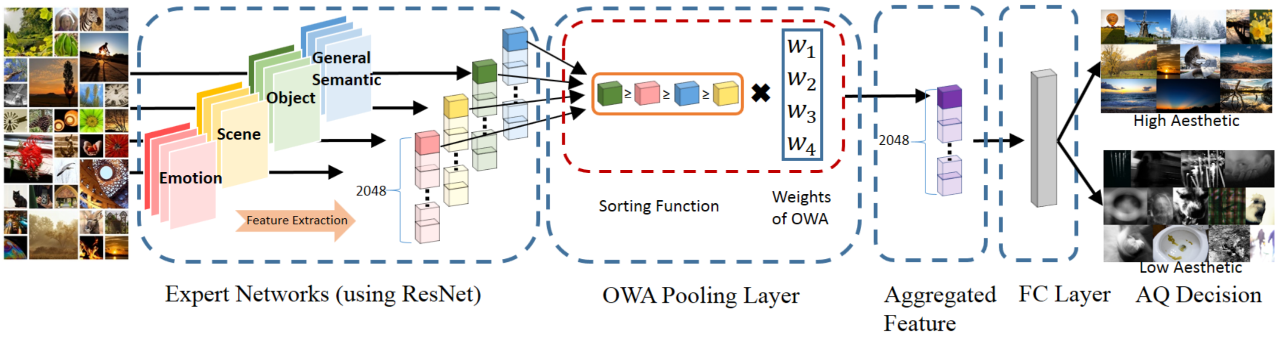


Fig. 1. Deep semantic aggregation of expert networks fuses the high-level concepts by an OWA pooling layer.

dropout (that avoids over-fitting), Krizhevsky et al. [15] successfully train a 5-layer CNN on a huge image collection, ILSVRC [16]. Since then, several powerful deep CNN models have been proposed to further boost the image classification performance. The VGG network [17] uses a small-sized 3×3 convolution kernel (instead of diverse-sized ones) in a deeper CNN model and double the number of channels when strides are performed. The GoogLeNet [18] designs a repeatable structure called Inception, and cascades this structure to form a deeper network. Both networks have about 15 layers and perform more favorably than AlexNet [15].

However, as the network becomes deeper, the gradient-vanishing problem [19] occurs and makes the training demanding. To avoid this difficulty, a better initialization [20] and batch-normalization [21] are widely used in the latest network training. Another problem of a deeper network (e.g., excessing 150 layers) is the slow convergence. More recently, the residual network (ResNet) [10] solves this problem by adding regularly the "short cuts" acrossing layers. ResNet can be understood as a combination of many shallow-to-deep networks. Analogous to the dropout training that results in a combination of various networks with different "widths," the ResNet combines many networks of different "depths" and achieves more favorable performance. Although deeper, ResNet's model is smaller than VGG's because the averaged pooling is applied to replace full-connected layers. The network model trained on the ILSVRC dataset [16] can serve as a good feature extractor for images. In this paper, we adopt the ResNet model and adapt it to the tasks of scene recognition, semantic classification, and emotion understanding of images, and then combine them for the AQ evaluation.

Deep CNNs for AQ Assessment. With the success of deep CNNs, some researches [4, 5] apply deep CNN for AQ Assessment. Lu et al. [4] introduce a double-column CNN architecture that takes holistic images and image patches as global and local features, respectively. In addition to aesthetic values, the style attributes are added to help determine the AQ. The approach in [5] generates multiple patches from a single images, and uses two network layers to aggregate the multiple patches. Better results are obtained.

In this study, we seek to use existing networks capable of recognizing objects for AQ evaluation. Through fine-tuning the networks to handle various high-level concepts such as scene, general semantic, and emotion, we conduct an AQ estimation network that fuses these attributes for the prediction. A unified architecture can then be used for different purposes and achieves more favorable performance on AQ assessment.

3. AESTHETICS QUALITY ASSESSMENT USING HIGH-LEVEL SEMANTIC FEATURES

In this section, we introduce the expert networks trained for the associated high-level attributes. We use ResNet-152 pre-trained on ImageNet [10] for object classification as our initial model, and fine-tune it on the datasets of different high-level attributes to build the individual expert network.

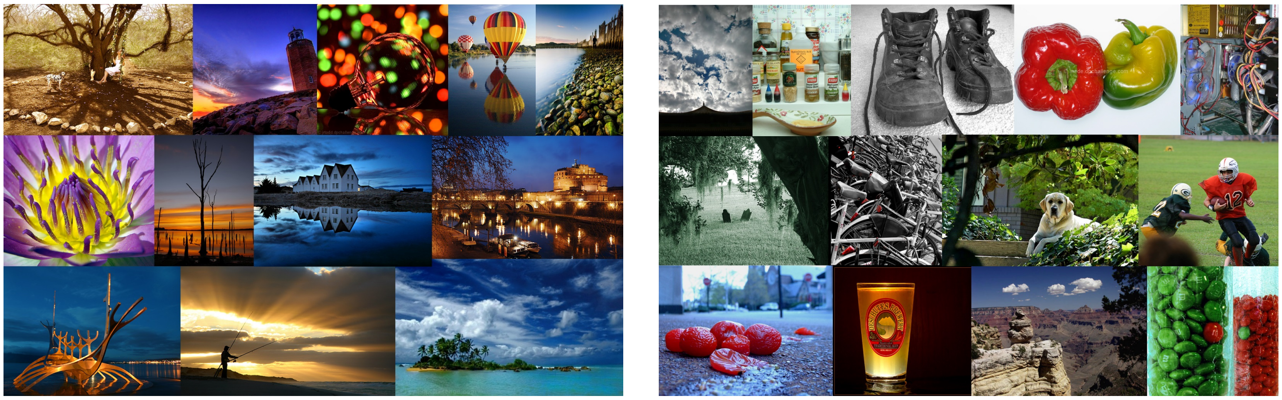
3.1. Scenes of images

Unlike object recognition, scene recognition aims to identify place-related concepts, such as natural landscapes and artificial buildings. The scenes provide essential background information that affects the AQ [22, 8].

To adapt the ResNet-152 to a scene-aware network, we exploit the Places dataset [22]. It is a large-scale dataset containing 205 scene categories and 2.5 millions of images with a category label for each image. The categories in Places are abundant and diverse, ranging from outdoor scenes like airfield, street and underwater to indoor scenes such as gallery, kitchen and pub etc. We use the Places dataset to fine-tune a ResNet-152 model for scene understanding, and the resulted model is referred to as the scene expert.

3.2. General semantic annotation of images

In addition to the AQ ratings, more plentiful annotations describing the semantic meanings of the images are also given in the AVA dataset [6]. There are 66 textual tags not only including the type of image but also some specific theme like sports and travel etc. Due to the diversity, we call it general semantic



(a) Images with high aesthetic quality

(b) Images with low aesthetic quality

Fig. 2. Example images of (a) high and (b) low aesthetic quality on AVA dataset

annotation in this paper. There are about 200k images having at least one tag, and 150k images possessing 2 tags. We treat each image-tag pair as one instance and thus there are about 400k instances. We use these general-semantic instances to fine-tune a ResNet-152 model using the ImageNet initialized parameters, resulting in the semantic expert in this paper.

3.3. Objects in images

Hosla et al. [23] observe that the existence of objects is a fairly effective feature for predicting the popularity of images. Having an expert for object recognition is thus helpful to the AQ assessment. In this paper, we use directly the public ResNet-152 model trained on ImageNet as the object expert.

To verify the influence of the object categories on the AQ assessment, we use our object expert to classify the images in AVA, and compute the correlation coefficients between the occurrence of objects and AQ. Some examples of the correlations between aesthetics and objects are listed below, where positive correlation coefficient indicates that images have higher probability to be of high AQ when the objects present, and vice versa.

- Positive correlation: seashore, lake, alp, valley, pier;
- Negative correlation: packet, digital clock, candle, pot;

We can observe that natural landscapes usually bring human the feeling of beauty, while objects of artificial supplies have negative relation with AQ. This confirms that the object in an image influences the AQ rating of human.

3.4. Emotions of images

A photographer can convey the emotion that he/she wants to express within the photos, and often constructs photos to elicit a specific response by the viewer. Emotion prediction of the image can be seen as clues of the photographers [9]. In Peng et al. [24], a new dataset called Emotion6 that consists six basic emotions (anger, disgust, fear, joy, sadness and surprise) and neutral is presented. Each emotion category has 330 images, and there are totally 1,980 images in this dataset.

The amount of images in Emotion6 dataset is not plentiful, and thus we adopt data augmentation to enlarge the dataset. Each image is randomly cropped, mirrored and rotated among a small range, which does not change people's feeling about the image but provides extra images with reasonable labels. We use approximately 360k image-data pairs in the training phase and obtain the emotion expert.

4. DEEP SEMANTIC AGGREGATION

The four expert networks extract high-level features of the image, which are then combined for AQ prediction. Here, we use the 2048-dimensional vector from "pool5" layer of the ResNet-152 as the feature representation for each expert network. As an expert network can predict the AQ rating individually as well, we report the AQ assessment results by using each expert network at first. For an expert network constructed in Section 3, we fine-tune it by using the training data of AVA dataset and present the testing performance (as shown in Table 1). As can be seen, the object, emotion, scene, and semantic experts achieve the accuracies around 76% $\sim$ 77%, which already outperforms the other competitive methods shown in Table 1. It reveals that these high-level concepts are related to AQ and ResNet is a powerful model for AQ assessment.

Since these high-level features would be complementary to each other, we then introduce the integration of them for further performance enhancement. An intuitive integration is directly concatenating features to obtain a high-dimensional representation [4, 25]. Instead of concatenation, [5] proposed two feature aggregation approaches. In this paper, we design a generalized aggregation method considering both ordinal information and statistical functions in [5]. We utilize ordered-weighted averaging (OWA) operator which is flexible and can automatically learn the aggregation rule.

Table 1. Accuracy (%) of aesthetic quality assessment using different compared methods on the AVA dataset

Method	AVA Dataset
Our Scene-only CNN	76.9
Our Semantic-only CNN	77.1
Our Object-only CNN	77.2
Our Emotion-only CNN	76.0
Our OWA CNN	**78.6**
Murry et al. [6]	68.0
SCNN [4]	71.2
AVG-SCNN [4]	69.9
DCNN [4]	73.3
RDCNN [4]	74.5
AlexNet [5]	72.3
DMA-Net-ImgFu$_{stat}$ [5]	75.4
DMA-Net-ImgFu$_{fc}$ [5]	75.4

4.1. Ordered Weighted Averaging (OWA)

Yager [11] proposes a aggregation operator, order-weighted-average (OWA) operator. OWA is a mapping $\mathbb{R}^n \to \mathbb{R}$ that sums the ordered elements with an associated weights W. The OWA operator is defined as

$$\text{OWA}(O, W) = \sum_{i=1}^{n} w_i o_i, \tag{1}$$

where $w_i \in [0, 1]$ and $\forall i = 1, ..., n$ and $\sum_{i=1}^{n} w_i = 1$. Let $O = \{o_i\}$ be the ordered values set of $x_i, \forall i = 1, ..., n$, where $o_1 \geq o_2 \geq ... \geq o_n$ are sorted in descending order.

OWA is a parameterized class of mean type aggregation operators e.g., max, min, mean and median. Some special cases are showed in the following

- if W=$[1, 0, ..., 0]$, OWA(O,W) = $max(O)$
- if W=$[0, ..., 0, 1]$, OWA(O,W) = $min(O)$
- if W=$[1/n, ..., 1/n, 1/n]$, OWA(O,W) = $mean(O)$

From the above examples, the statistics layer and sorting layer in [5] could be regarded as special cases of OWA layer.

4.2. Details of learning

Given the inputs vectors V_m^n with $m = 2048$ dimensions from $n = 4$ expert networks, we first implement a sorting function $\phi(V_m^n) = (o_m^1, o_m^2, ..., o_m^n)$ where $o_m^1 \geq o_m^2 \geq ... \geq o_m^n$. n feature vectors are combined with the dot products of o_m^n and w_n as illustrated in Figure 1. There are two constrains of OWA that should be considered for updating w_i in the training stage. The first constraint is $\sum_{i=1}^{n} w_i = 1$, and w_i should be normalized by a scale. The scale of w_i could be absorbed in the later fully connected layer, so we do not have to normalize w_i in the OWA layer. Another constrain is $w_i \geq 0$, and we use projected gradient descent to handle this bound-constrained optimization. When the gradient of w_i tries to make $w_i < 0$, the w_i will be projected onto the feasible set (≥ 0).

Table 2. Accuracy (%) of aesthetic quality assessment using different aggregation methods on the AVA dataset

Aggregation Method	AVA Dataset
Our OWA CNN	78.6
Our Max CNN	78.4
Our Min CNN	78.1
Our Mean CNN	78.1

5. EXPERIMENTAL RESULTS

This section details the performance of the proposed deep neural network. The impact of the proposed learning methodology and features are clarified and compared on Aesthetic Visual Analysis (AVA) dataset [6]. The AVA dataset contains a total of 250k images, and each image has about 210 votes in average. Example images are illustrated in Figure 2. We follow the same experimental setting in [6, 4, 5], and evaluate our approach on the same 230k images and 20k images for training and testing respectively.

First, we evaluate the performance of different methods on AVA dataset. As shown in Table 1, the performance using only object, emotion, scene, and semantic experts is better than the state-of-the-art RDCNN [4] and DMA-Net-ImgFu$_{stat}$ [5]. This result reveals that the high-level concepts with ResNet are useful for AQ assessment. When applying only single feature for accessing the AQ, object expert network performs better than others, and the performance of rest experts in descending orders are general semantic, scene and emotion. The proposed OWA combines the four types of expert networks, and achieves the highest performance than the CNN using single expert and previous works [4, 5, 6].

In this experiment, we discover the performance of aggregation approaches. Table 2 shows the results of OWA and the other related special cases of aggregations, e.g., max, min and mean. The OWA aggregation performs better than the other aggregation methods. In addition to performance improvement, the OWA operator is more flexible, since the parameters of OWA could be automatically learned from training stage.

6. CONCLUSION

In this paper, we build the expert networks of high-level semantic concepts using the state-of-the-art deep CNN, and fuse them by the introduced OWA pooling layer. Our experiments show that single expert attribute outperforms the state-of-the-art approaches, and the OWA aggregation could further boost the performance for AQ assessment.

7. ACKNOWLEDGEMENT

This work was supported in part under the grants MOST 104-2221-E-001-023-MY2.

8. REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006.

[2] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006.

[3] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *ICCV*, 2011.

[4] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *ACM MM*, 2014.

[5] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *ICCV*, 2015.

[6] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012.

[7] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *ACM MM*, 2010.

[8] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Assessment of photo aesthetics with efficiency," in *ICPR*. IEEE, 2012.

[9] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[11] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE TSMC*, vol. 18, no. 1, pp. 183–190, 1988.

[12] M.-C. Yeh and Y.-C. Cheng, "Relative features for photo quality assessment," in *ICIP*, 2012.

[13] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008.

[14] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *CVPR*, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[19] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE TNN*, vol. 5, no. 2, pp. 157–166, 1994.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[23] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?," in *ACM WWW*, 2014.

[24] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *CVPR*, 2015.

[25] X. Zhang, H. Zhang, Y. D. Zhang, Y. Yang, M. Wang, H. Luan, J. T. Li, and T.-S. Chua, "Deep fusion of multiple semantic cues for complex event recognition," *IEEE TIP*, 2016.