Query-Dependent Aesthetic Model With Deep Learning for Photo Quality Assessment

Xinmei Tian, *Member, IEEE*, Zhe Dong, Kuiyuan Yang, and Tao Mei, *Senior Member, IEEE*

***Abstract*—The automatic assessment of photo quality from an aesthetic perspective is a very challenging problem. Most existing research has predominantly focused on the learning of a universal aesthetic model based on hand-crafted visual descriptors. However, this research paradigm can achieve only limited success because 1) such hand-crafted descriptors cannot well preserve abstract aesthetic properties, and 2) such a universal model cannot always capture the full diversity of visual content. To address these challenges, we propose in this paper a novel query-dependent aesthetic model with deep learning for photo quality assessment. In our method, deep aesthetic abstractions are discovered from massive images, whereas the aesthetic assessment model is learned in a query-dependent manner. Our work addresses the first problem by learning mid-level aesthetic feature abstractions via powerful deep convolutional neural networks to automatically capture the underlying aesthetic characteristics of the massive training images. Regarding the second problem, because photographers tend to employ different rules of photography for capturing different images, the aesthetic model should also be query-dependent. Specifically, given an image to be assessed, we first identify which aesthetic model should be applied for this particular image. Then, we build a unique aesthetic model of this type to assess its aesthetic quality. We conducted extensive experiments on two large-scale datasets and demonstrated that the proposed query-dependent model equipped with learned deep aesthetic abstractions significantly and consistently outperforms state-of-the-art hand-crafted feature-based and universal model-based methods.**

***Index Terms*—Deep aesthetic visual abstraction, deep learning, quality assessment.**

I. Introduction

THE objective of photo quality assessment is to automatically determine whether a given image is of "high" or "low" quality from an aesthetic perspective. Such assessments have many attractive applications. For example, they enable the harvesting of high-quality photos from online sources [1]–[3] and can assist home users in managing and editing their digital photos [4]–[8]. As a result, image quality assessment has received increasing attention in recent years [9]–[16].

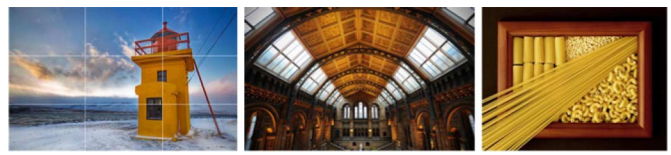


Fig. 1. Three high-quality images following different composition rules: (a) the rule of thirds, (b) a symmetric pattern, and (c) the diagonal rule.

Because of the subjectivity and complexity of humans' aesthetic activities, the automatic aesthetic assessment of images is very challenging. In recent years, many research efforts have been made and various methods have been proposed to solve the photo quality assessment problem [17]–[21]. Most existing research has predominantly focused on the construction of hand-crafted visual descriptors that are related to high-level aesthetic attributes. Those features are designed under the guidance of certain common photography rules and prior knowledge, among which the following are the most commonly applied: rule-of-thirds composition [4]–[19], depth of field (DOF) [17], [9], simplicity [18], colorfulness [17], sharpness [18], [19], exposure [17], [18], and contrast [18]–[19]. When the images of interest are represented in terms of these hand-crafted aesthetic features, a universal aesthetic model can be trained on a dataset consisting of both "high"- and "low"-quality images (labeled by hand) [17]–[20]. The universal aesthetic model can then be applied to assess the quality of any test image.

However, the above research paradigm can achieve only limited success because 1) such hand-crafted features cannot well preserve abstract aesthetic properties and 2) such a universal model cannot always capture the full diversity of image content. Because quality assessment is rather subjective and complex, it is unclear which types of features are correlated with aesthetic value. Unlike the aforementioned hand-crafted features, which are designed heuristically to mimic certain predefined rules, we propose in this paper to mine the underlying aesthetic abstractions automatically using powerful deep convolutional neural networks (DCNNs) [24]. The convolutional network is fed raw pixels and trained end to end, thereby alleviating the shortcomings of hand-engineered features. We believe that compared with hand-crafted features, these automatically trained aesthetic features will produce a representation that can better capture several general underlying aesthetic characteristics from massive training images.

Manuscript received April 29, 2015; revised July 15, 2015 and August 30, 2015; accepted September 13, 2015. Date of publication September 17, 2015; date of current version October 20, 2015. This work was supported by the 973 Project under Contract 2015CB351803, by the NSFC under Contract 61390514 and Contract 61201413, by the Youth Innovation Promotion Association CASunder Grant CX2100060016, by the Fundamental Research Funds for the Central Universities under Grant WK2100060011, Grant WK2100100021, and Grant 61301082, and by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant WJ2100060003. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Guo-Jun Qi.

X. Tian and Z. Dong are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, University of Science and Technology of China, Hefei 230027, China (e-mail: xinmei@ustc.edu.cn; ustcdz@mail.ustc.edu.cn).

K. Yang and T. Mei are with Microsoft Research, Beijing 100190, China (e-mail: kuyang@microsoft.com; tmei@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2479916

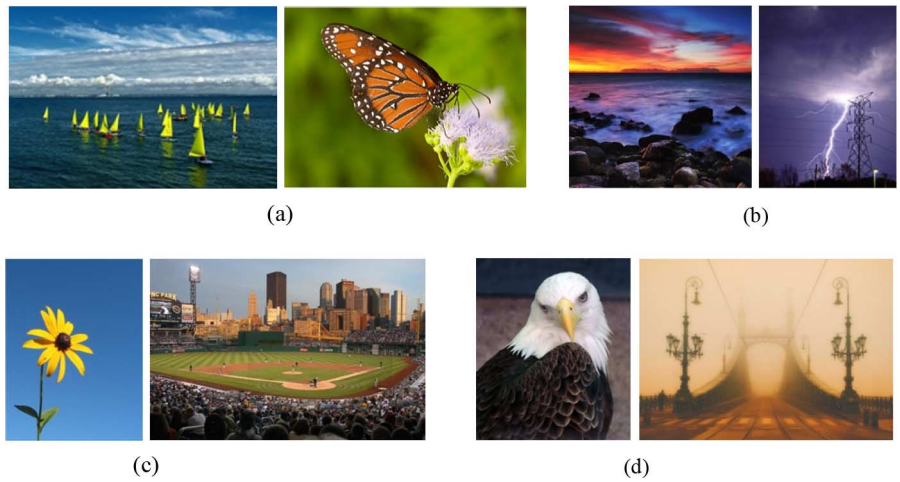


Fig. 2. High-quality images following different rules of photography. There is no single rule that is suitable for all images. (a) Deep (left) or shallow (right) depth of field? (b) Colorful (left) or monotone (right)? (c) Simplicity (left) or complexity (right)? (d) Sharpness (left) or blur (right)?

Regarding the aesthetic model, the underlying assumption in the conventional universal aesthetic model is that all high-quality images share the same photographic/aesthetic rules. However, the optimal photographic rules can vary considerably among different images, and they are affected by many factors, such as the subject of the image (portraiture, landscape, animal, architecture, etc.), the theme (emotive, fun, sad, etc.), the shoot time (day or night), the shoot location (indoor or outdoor), and the type of photograph (photomacrograph, photomicrograph, etc.). There are no fixed rules that can be applied to all images. For example, whether the composition of an image follows the rule of thirds has served as an important feature in previous aesthetic assessment research [4]–[25]. However, beyond the rule of thirds, there are also many other popular rules of photographic composition, such as symmetric patterns (especially for architecture), the diagonal rule, and the S rule. Fig. 1 shows three high-quality images following different composition rules. Fig. 2 provides further examples of images that are all high in quality but follow very different photographic rules.

Because there are no universal photographic rules that are suitable for all images, it is difficult to train a universal aesthetic model that can handle all images well. Photographers tend to employ adaptive photographic rules for capturing different images; therefore, the aesthetic model should also be adaptive to individual images. In this paper, we propose to construct such an adaptive aesthetic model for different testing/query images, i.e., the *query-dependent aesthetic model*. Specifically, given an image to be assessed, we first identify which aesthetic model should be applied to this particular image and then build a unique aesthetic model of this type to assess its aesthetic quality. However, it is very challenging to construct such a query-dependent aesthetic model. As noted above, the optimal photographic rules for a specific image are influenced by many factors, such as the subject, the theme, the shoot time and location, and the type of photograph, as well as many other unknown factors. Modeling all these factors explicitly and exhaustively for the construction of a query-dependent aesthetic model would be an excessively complex task.

We solve this problem by considering that similar images share similar aesthetic models. Under this basic assumption, for a given query image, we first identify other images to which it is visually/semantically similar from the entire training database to construct a query-dependent training set, and the query-specific model is then learned from these images. For example, if the given query image is a photo with a "plant"-related subject, we wish to retrieve other "plant" images from the training database. The aesthetic labels (high quality or low quality) of the returned "plant" images are known. Therefore, we can train a model using those images. The trained model can then be used to determine the aesthetic label of the given "plant" image. When the query image is a photo with an "animal"-related subject, we retrieve other "animal" images from the database to train an aesthetic model that is suitable for "animal" images.

The key problem in query-dependent aesthetic model learning is to identify the neighbors of the query image, i.e., the images that are visually/semantically similar to the query image. The most straightforward approach is to use content-based image retrieval. However, at present, methods of content-based image retrieval still suffer from the well-known semantic gap problem. Retrieval based on visual information alone may introduce noise into the retrieval results. However, for most applications, we are concerned with network images that are shared on the Web and are accompanied by rich textual information, e.g., tags, surrounding text, and Exif information. By also invoking this textual information in the retrieval process, a better query-dependent training set can be constructed, resulting in a better query-dependent aesthetic model.

The remainder of this paper is organized as follows. Section II briefly reviews related work. In Section III, the automatic learning of aesthetic features via DCNNs is introduced. In Section IV, the proposed query-dependent image quality assessment method is detailed. Experimental results are presented and analyzed in Section V, followed by the conclusions in Section VI.

II. Related Work

Most existing research on photo quality assessment has focused on the design of aesthetic-related features based on common rules of photography [17]–[26] to mimic human aesthetic perception. For example, Datta *et al.* [17] proposed

a 56-dimensional feature vector to describe several high-level aesthetic attributes, such as light exposure, colorfulness, saturation and hue, and the rule of thirds. Ke *et al.* [18] proposed 7 different kinds of features to capture the simplicity, contrast, brightness, etc. of an image. Marchesotti *et al.* [11] proposed the use of generic image descriptors to assess aesthetic quality. Luo *et al.* [19] first extracted the subject region from a photo and then extracted several features, including clarity, contrast, simplicity, and composition. Lo *et al.* [27] proposed several aesthetic features from the perspective of comparing high- and low-quality image templates derived from a training set. Nishiyama *et al.* [28] assessed the aesthetic quality of photos by evaluating their color harmony and proposed the use of bags of color patterns to characterize color variations in local regions. Lu *et al.* [29] recently proposed the adoption of deep neural networks to classify low- and high-quality images.

With images represented in terms of certain designed aesthetic features, a universal model is typically trained via popular machine learning algorithms (e.g., SVMs, AdaBoost) on a collected training set. The universal model is then applied to various test images. The primary limitation of such a universal aesthetic model is that it assumes that all high-quality images follow the same photographic rules. Recently, several researchers have recognized this problem and have made preliminary efforts to solve it [22]–[31]. Instead of considering all types of images, Li *et al.* [22] focused on one specific genre of photos: consumer photos containing faces. They extracted face-related features and trained a quality evaluation model for images of this particular kind. In [10] and [23], the considered images were manually grouped into different categories based on their subjects ("animal", "plant", "human", etc.), and an aesthetic model was then trained for each category. In [30], an aesthetic model was designed specifically for scenic images. Yin *et al.* [21] restricted their study to scenic images with geo-location information. Specifically, they used the geo-location information to collect images acquired at the same place and also used auxiliary datasets to construct different aesthetic models for scenic images with different contents, e.g., bridges, mountains, or beaches.

Although these works have attempted to address the problems of universal aesthetic models, they suffer from either their limited applicability to specific types of images (face-containing images in [22], scenic images in [21], [30]) or their need for specific auxiliary knowledge (geo-location information in [21]). Furthermore, although these authors acknowledge the limitations of the universal model, they still assume that all images in the same category share a common aesthetic model. As discussed above, the applicability of photographic rules is affected by many factors, and an image's content/subject is only one of them. The aesthetic attributes of images within the same category still vary considerably. It is difficult to use a single aesthetic model to describe them all. Moreover, in these methods, it is necessary to first automatically identify the content to determine which model should be used. Furthermore, it is impractical to build aesthetic models for application to a vast range of categories.

The query-dependent aesthetic assessment method proposed in this paper can effectively address these problems. First, instead of building a universal aesthetic model, our method builds a query-dependent aesthetic model for each query image. For each query image, we construct a query-dependent training set and extract the deep aesthetic features of the images in that set. The query-dependent aesthetic model is then learned from those images. Second, our method requires no additional information beyond the image itself, which makes this method highly generalizable to various applications. In our approach, to assess the quality of a given image, we retrieve other images that are visually similar to it from the training image database using visual features. Therefore, all of the information we require is contained in the visual features of the images. We do not require any auxiliary knowledge, such as the geo-location information required in [21] or the category label information required in [10]. Our approach is applicable to various types of images rather than being limited to a specific genre of images (e.g., face-containing photos as in [22] or scenic images as in [21] and [30]). Third, the proposed method is a general framework that is sufficiently flexible to incorporate auxiliary knowledge when available, for example, the textual information associated with an image, as will be illustrated later.

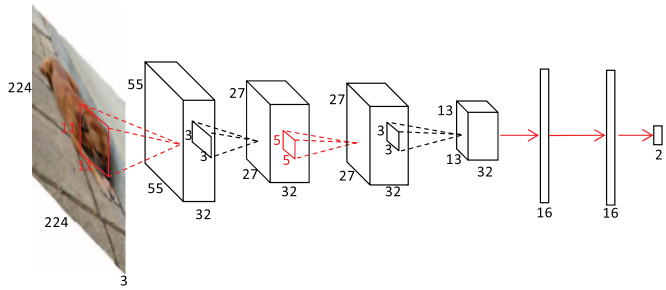


Fig. 3. Architecture of a DCNN for automatic aesthetic feature mining. Layers with weights are marked with red lines. Max-pooling layers without weights are marked with black dashed lines. A local contrast normalization layer, which is not represented in the figure, is applied following the first max-pooling layer.

The query-dependent aesthetic assessment method proposed in this paper can effectively address these problems. First, it constructs a query-dependent aesthetic model for each query image to consider its unique photographic rules. Second, it requires no additional information beyond the image itself, which makes this method highly generalizable to various applications. Third, the proposed method is a general framework that is sufficiently flexible to incorporate auxiliary knowledge when available, for example, the textual information associated with an image, as will be illustrated later.

III. Aesthetic Feature Learning via Deep Convolutional Neural Networks

In this paper, we propose to automatically mine abstract aesthetic features from massive training images using DCNNs. Our network contains five learned layers: two convolutional layers and three fully connected layers. The architecture is schematically illustrated in Fig. 3. The first convolutional layer filters the input image using 32 kernels of $11 \times 11 \times 3$ in size with a stride of 4 pixels. The second convolutional layer takes as input the normalized and pooled output of the first convolutional layer and filters it using 32 kernels of $5 \times 5 \times 32$ in size. The fully connected layers have 16 neurons each. Our objective is to classify the photos into two classes, i.e., "high" quality and

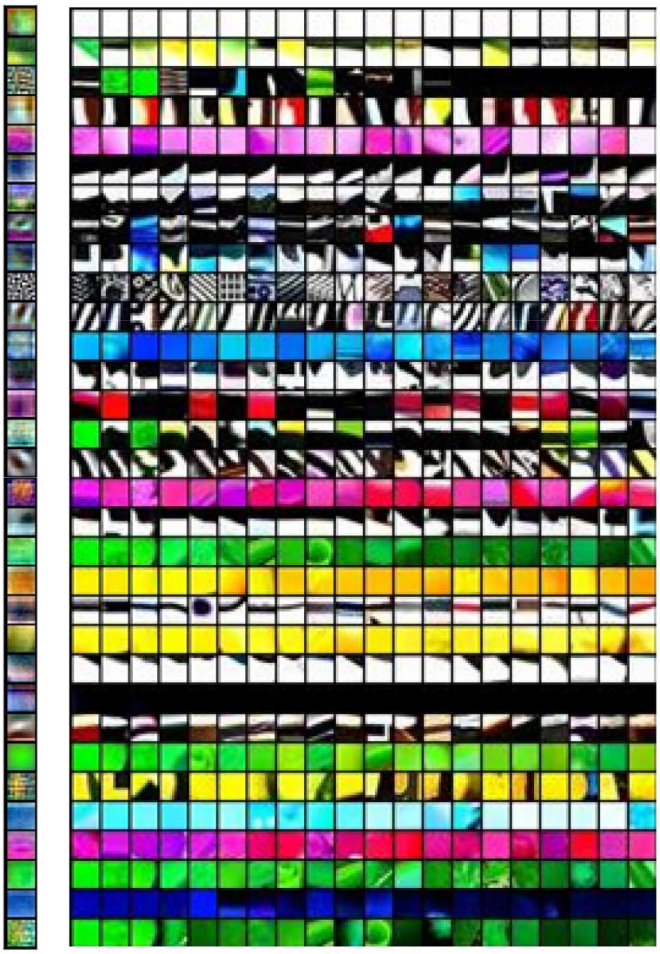


Fig. 4. (a) The 32 convolutional kernels of $11 \times 11 \times 3$ in size learned by the first convolutional layer. (b) Each row visualizes the top 20 local image patches with the highest response to one of the 32 kernels.

Fig. 5. Each row visualizes the top 20 image patches with the highest response to the 32 kernels learned in the second convolutional layer.

"low" quality. Therefore, the last fully connected layer is fed to a 2-way softmax that produces a distribution over the classes. The goal of training is to maximize the probability of the correct class, which is achieved by minimizing the cross-entropy loss for each training example.

For use as training images, we downloaded approximately 19,000 images from DPChallenge.com, a popular social photo sharing website that allows users to share, comment on and score photos online. Please refer to Section V-A for details. We trained the model using the stochastic gradient descent technique with a mini-batch size of 128 examples, with a dropout rate of 0.5 added to two fully connected layers. The 32 convolutional kernels of $11 \times 11 \times 3$ in size learned by the first convolutional layer are presented in Fig. 4(a), and the top 20 local image patches with the highest response to each of these 32 kernels are given in Fig. 4(b). These kernels can be grouped into two major classes, one consisting of frequency and orientation kernels and one consisting of color-related kernels. The frequency and orientation kernels detect patterns of variation in the images. They are potentially related to the general aesthetic rules of sharpness (high frequency), contrast, local structures, and so on. The color-related kernels detect color-harmonious image patches. It has been proven that color harmony is a key factor among the various contributions to the perceived quality of a photo [32], and many studies of hand-crafted aesthetic feature extraction have considered the issue of color harmony [28], [19].

The kernels learned in the first convolutional layer detect predominantly local features. The kernels in the second convolutional layer view a larger contextual window and thus can derive aesthetic features at higher levels of abstraction. The top 20 image patches with the highest response to each of the 32 kernels learned in the second convolutional layer are presented in Fig. 5. We can observe that these patches exhibit more complex compositional structures and color patterns. With the three fully connected layers, more abstract and global aesthetic features will be derived. We adopt the 16-dimensional features of the last hidden layer as the final deep aesthetic feature representation of the image.

Compared with the model adopted in [29], the architecture of our model is much simpler and has fewer parameters, especially in the fully connected layers. [29] mainly follows the architecture of ImageNet [24]. However, unlike ImageNet, which is a highly complex 1000-class classification task, photo quality assessment is only a binary classification task. Moreover, in [29]

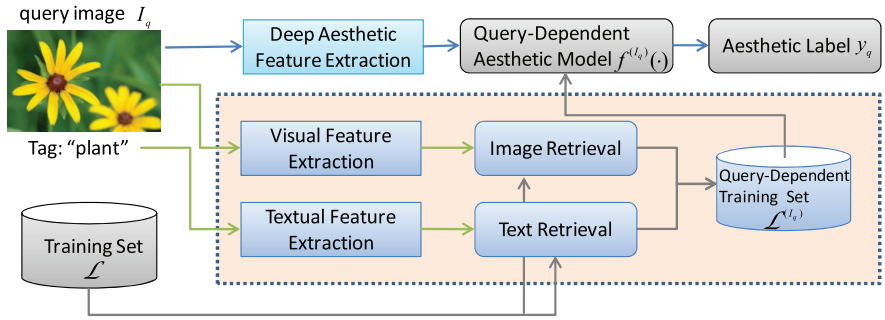


Fig. 6. The framework of the proposed query-dependent aesthetic assessment system. For a given query image, we first extract its visual and textual features for the retrieval of similar images from the entire training set. By combining the results of image and text retrieval, a query-dependent training set is constructed. Using images representing the extracted deep aesthetic features, the query-dependent model is learned from the query-dependent training set. The aesthetic label y_q is predicted by the learned query-dependent aesthetic model. The proposed query-dependent aesthetic model is comparable to the manner in which humans learn to rate image aesthetics from experience gained by viewing similar images.

all images with varying aesthetic quality levels were used while in this paper we discarded the ambiguous images in the middle of the quality range, leading to a much easier task. Therefore, there is no need for as many neurons in the fully connected layers in this case. In fact, we tested various numbers of neurons in the fully connected layers and observed that redundant and repeated neurons were generated. Thus, we gradually reduced the number of neurons in the fully connected layers and found that 16 is a suitable choice.

Our model also has fewer convolutional layers than the model in [29]. To investigate whether this two convolutional layers CNN model is strong enough to infer high level subjective aesthetic meanings, we tried two alternative experiments to involve more global information. In the first experiment, we increased the number of convolutional layers from 2 to 5. In the second experiment, we trained a multi-scale DCNN model which is similar to [33]. However, neither experiment brings performance increase. We have tried our best to analyse the reasons. We think the most possible reason is two-fold. On one hand, the size of kernels in the two convolutional layers are 11×11 and 5×5, respectively. Therefore, the size of the receptive filed after two convolutional layers is 55×55. As shown in Fig. 5, the patch in this size contains considerable part of the whole image and can reflect partial aesthetic quality. On the other hand, the outputs of the second convolutional layer are fed into the fully connected layers, in which they are combined together for subsequent learning. Therefore, global information can be partially involved in this step. Besides, aesthetic quality assessment is highly abstract and very challenging. It needs us devote more efforts. We will continue investigating this interesting problem in the future.

IV. Query-Dependent Aesthetic Model

In this section, we first introduce the conventional universal aesthetic model and then present the proposed query-dependent aesthetic assessment method, including image-retrieval-based, text-retrieval-based and multi-view query-dependent aesthetic models.

A. Universal Aesthetic Model

Given a query image I_q, the objective of aesthetic quality assessment is to predict its aesthetic quality y_q based on a training set $\mathcal{L} = \{(I_1, y_1), (I_2, y_2), \cdots, (I_L, y_L)\}$, where I_i is the i-th training image in $\mathcal{L}$ and y_i is the aesthetic quality of I_i. Currently, most studies of aesthetic quality analysis treat it as a binary classification problem, i.e., the aesthetic quality is represented by $y \in \{-1, 1\}$, with $y = 1$ denoting high quality and $y = -1$ denoting low quality.

To predict the aesthetic label y_q of query image I_q based on the training set $\mathcal{L}$, we need to estimate the posterior probability $p(y|I_q)$. The label y_q should be the one with the maximum a posteriori probability given the query image I_q and the training set $\mathcal{L}$

$$y_q = \text{argmax}_{y \in \{1,-1\}} p(y|I_q, \mathcal{L}). \tag{1}$$

Currently, y is typically estimated using a model denoted by $f(\cdot)$, i.e., $y = f(\cdot)$, via sophisticated machine learning methods, e.g., SVMs [17], Bayesian classifiers [18], or AdaBoost [19]. The model is trained on $\mathcal{L}$ by minimizing the training errors on this set.

In universal aesthetic quality assessment, it is assumed that all images share a common aesthetic model; therefore, the universal model $f(\cdot)$ is trained on a set $\mathcal{L}$ that is independent of the query image I_q.

B. Query-Dependent Aesthetic Model

The universal model suffers from the limitation that it cannot be effectively applied to a broad variety of query images. As discussed in Section I, the aesthetic model $f(\cdot)$ should be dependent on the query image. In other words, instead of being independent of I_q, a query-dependent model $f^{(I_q)}(\cdot)$ should be learned from a suitable training set $\mathcal{L}^{(I_q)}$ that reflects the unique aesthetic characteristics of I_q. Thus, the key problem in query-dependent aesthetic model learning is to properly construct the query-dependent training set $\mathcal{L}^{(I_q)}$. In this paper, we propose to construct $\mathcal{L}^{(I_q)}$ by exploring the neighbors of the query image in a joint visual and textual space.

The framework of the proposed query-dependent aesthetic assessment method is illustrated in Fig. 6. For a given query image I_q, we first extract its visual and textual features for the retrieval of similar images from the entire training set $\mathcal{L}$. By combining the results of image and text retrieval, a query-dependent training set $\mathcal{L}^{(I_q)}$ is constructed. Then, the query-dependent aesthetic model $f^{(I_q)}(\cdot)$ is learned from $\mathcal{L}^{(I_q)}$. The aesthetic label y_q is finally predicted by the learned query-dependent model.

The proposed framework is very flexible. In the case that only visual information is available, we can construct $\mathcal{L}^{(I_q)}$ via image retrieval alone. In this case, we refer to the procedure as image-retrieval-based query-dependent aesthetic learning. When only text retrieval is used for the construction of $\mathcal{L}^{(I_q)}$, the method is known as text-retrieval-based query-dependent aesthetic learning. When $\mathcal{L}^{(I_q)}$ is constructed from both visual and textual information, we refer to this approach as multi-view query-dependent aesthetic learning. Moreover, it is also quite simple to incorporate other information to derive $\mathcal{L}^{(I_q)}$ more precisely.

For query-dependent aesthetic learning, we need to construct a query-dependent training set and train a query-dependent model for each query image; therefore, both efficiency and effectiveness are important. The image- and text-retrieval-based query-dependent aesthetic learning methods presented below are designed to ensure both high efficiency and high accuracy.

1) Image-Retrieval-Based Query-Dependent Aesthetic Learning: In image retrieval, the essential problem is to measure the visual similarity between the query image and each of the images in the database precisely and efficiently. The retrieval accuracy depends on the visual features considered, and the retrieval efficiency depends on the index strategy. For the visual features, we adopt two popular methods of visual representation. One is the widely used bag-of-words (BOW) visual representation, and the other is the newly emerged representation based on generic features learned via large CNNs.

Image Retrieval Based on BOW Features: We first discuss image retrieval using the BOW visual representation. For each image, the DoG detector is used for interest point detection and the SIFT approach is used to describe local features [34]. A codebook is trained with local features from all training images via K-means clustering. Each local feature is quantized into its corresponding visual word; thus, each image is represented as a sequence of visual words.

Visual words should serve to describe images in manner similar to compact and descriptive text words. Thus, we can build a highly efficient content-based image retrieval system in which images are indexed and retrieved via Inverted File Indexing [35], one of the most popular information retrieval strategies. In our image retrieval system, each image I is represented as a histogram

$$\mathbf{t}_I = [t_1, t_2, \cdots, t_H]^T \tag{2}$$

where H is the codebook size, $t_i = n_i\omega_i$, and n_i is the frequency of the i-th visual word in I. The ω_i are weighting constants defined as the inverse document frequencies (IDFs), $\omega_i = \ln\frac{N}{N_i}$, where N is the total number of images in the database and N_i is the number of images in the database that contain the i-th visual word.

Finally, the similarity between the query image I_q and a database image I_d is calculated as follows:

$$Sim_{visual-BOW}(I_q, I_d) = 1 - \frac{1}{2}\|\frac{\mathbf{t}_q}{\|\mathbf{t}_q\|_1} - \frac{\mathbf{t}_d}{\|\mathbf{t}_d\|_1}\|_1 \tag{3}$$

where $\mathbf{t}_q$ and $\mathbf{t}_d$ are the histograms of I_q and I_d, respectively.

Image Retrieval Based on CNN Features: Recent research indicates that the generic features extracted by large CNNs trained on the diverse ImageNet database are very powerful [24]. Therefore, we also adopt these CNN-based visual features as an image representation for measuring the visual similarity between images.

We extract the CNN-based visual features using the open-source deep learning framework Caffe.[1] The architecture of the CNN model is similar to that presented in [24]. It consists of five convolutional layers and three fully connected layers. The neural network is trained on part of the ImageNet image database, which contains millions of images in 1000 categories. For an image I, we input it into the learned CNN, and the 4096-dimensional output of the final fully connected layer is adopted as its visual representation $\mathbf{x}_I = [x_1, x_2, \cdots, x_{4096}]^T$.

The similarity between the query image I_q and a database image I_d is calculated as follows:

$$Sim_{visual-CNN}(I_q, I_d) = 1 - \frac{1}{2}\|\frac{\mathbf{x}_q}{\|\mathbf{x}_q\|_1} - \frac{\mathbf{x}_d}{\|\mathbf{x}_d\|_1}\|_1. \tag{4}$$

Query-Dependent Training Set Construction via Image Retrieval: When no textual information is available, we can construct the query-dependent training set $\mathcal{L}^{(I_q)}$ for I_q based solely on the image retrieval results as follows:

$$\mathcal{L}^{(I_q)} = \{I_d, I_d \in \mathcal{L} \text{ and } I_d \in \mathcal{N}(I_q)\} \tag{5}$$

where $\mathcal{N}(I_q)$ denotes the set of neighboring images of I_q in the BOW-based or CNN-based visual feature space, which can be derived using (3) or (4), respectively. We denote these image-retrieval-based query-dependent aesthetic learning methods as QDep_IR$_{\text{BOW}}$ and QDep_IR$_{\text{CNN}}$ for short.

2) Text-Retrieval-Based Query-Dependent Aesthetic Learning: The image retrieval system can efficiently return images that are visually similar to the query image. However, because of the well-known semantic gap problem, some irrelevant images may be returned. Currently, many images are shared on the Web that are associated with rich textual information. Therefore, when an image's textual information is available, we can leverage mature text retrieval systems to conduct text-based retrieval for the evaluation of images that are textually similar to the query image.

As discussed in Section II, several previous works have addressed the shortcomings of universal aesthetic learning by categorizing images into different groups based on their tags and training independent aesthetic models for each category. This type of method is a special case of our proposed text-retrieval-based query-dependent aesthetic learning method in which simple Boolean text retrieval is performed using the image's tag information. This type of method is termed QDep_Tag in this paper.

QDep_Tag constructs the query-dependent training set $\mathcal{L}^{(I_q)}$ based on the following rule:

$$\mathcal{L}^{(I_q)} = \{I_d, I_d \in \mathcal{L} \text{ and } Tag\{I_q\} \cap Tag\{I_d\} \neq \emptyset\} \tag{6}$$

where $Tag\{I\}$ denotes the tags associated with image I. It is equivalent to conducting Boolean text retrieval, in which the similarity between the query image I_q and a database image I_d is defined as

$$Sim_{text}(I_q, I_d) = \begin{cases} 1 & \text{if } Tag\{I_q\} \cap Tag\{I_d\} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

[1]"Caffe: An open source convolutional architecture for fast feature embedding," [Online]. Available: http://caffe.berkeleyvision.org/

TABLE I
PERFORMANCE COMPARISON OF VARIOUS AESTHETIC FEATURES AND CLASSIFICATION MODELS ON THE AVA DATASET. THE RESULTS SHOW THAT 1) THE PROPOSED DEEP AESTHETIC FEATURES SIGNIFICANTLY OUTPERFORM ALL OTHER HAND-CRAFTED FEATURES REGARDLESS OF WHETHER THE UNIVERSAL MODEL OR A QUERY-DEPENDENT MODEL IS APPLIED; 2) THE PROPOSED QUERY-DEPENDENT AESTHETIC MODELS $\text{QDep\_IR}_{\text{BOW}}$ AND $\text{QDep\_IR}_{\text{CNN}}$ CONSISTENTLY OUTPERFORM THE UNIVERSAL MODEL FOR ALL TYPES OF AESTHETIC FEATURES; 3) $\text{QDep\_IR}_{\text{CNN}}$ DEMONSTRATES BETTER PERFORMANCE THAN DOES $\text{QDep\_IR}_{\text{BOW}}$; AND 4) $\text{QDep\_IR}_{\text{CNN}}$ WITH THE DEEP AESTHETIC FEATURES ACHIEVES THE HIGHEST PERFORMANCE WITH AN 80.38% CLASSIFICATION ACCURACY, WHEREAS THE UNIVERSAL MODEL WITH HAND-CRAFTED FEATURES CAN ACHIEVE A CLASSIFICATION ACCURACY OF ONLY APPROXIMATELY 70%

	Accuracy (%)		
Aesthetic Features	Universal model	$\text{QDep\_IR}_{\text{BOW}}$	$\text{QDep\_IR}_{\text{CNN}}$
Luo [19]	61.49	66.19	*76.11*
Efficiency [27]	68.13	69.27	*77.76*
Datta [17]	68.67	69.50	*75.86*
Ke [18]	71.06	76.05	*78.60*
Image Descriptors [11]	68.55	69.22	*75.49*
RAPID [29]	74.54	75.88	*79.35*
Deep Aesthetic Features	**75.89**	***77.15***	***80.38***

Although there are many mature text retrieval systems available and many additional textual features beyond the tags can be used in our proposed framework, we simply use the Boolean retrieval procedure described above using tag information to ensure a fair comparison with the QDep_Tag method presented in [10]–[23].

3) Multi-View Query-Dependent Aesthetic Learning: The textual and visual features of an image describe the image from different perspectives and are complementary to each other. To obtain a better query-dependent training set $\mathcal{L}^{(I_q)}$, it is natural to combine these two types of information. Many studies of multi-view learning have been reported [36], [37]. Here, we consider only the simplest linear combination

$$Sim(I_q, I_d) = \alpha Sim_{text}(I_q, I_d) + (1-\alpha) Sim_{visual}(I_q, I_d). \quad (8)$$

The combination coefficient $\alpha \in [0, 1]$ is a trade-off factor between the two components. For query image I_q, we retrieve the K most similar images from the entire training set $\mathcal{L}$ according to (8) to construct $\mathcal{L}^{(I_q)}$, and we then derive its multi-view query-dependent aesthetic model $f^{(I_q)}(\cdot)$ by training its on $\mathcal{L}^{(I_q)}$.

The multi-view query-dependent aesthetic learning procedure reduces to QDep_Tag when $\alpha = 1$ and reduces to QDep_IR when $\alpha = 0$. In this paper, we simply set $\alpha = 0.5$.

V. EXPERIMENTS

A. Experimental Setting

1) Datasets: There are several publicly available datasets for aesthetic image quality assessment. In general, these images can be downloaded from social photo-sharing websites, such as photo.net and DPChallenge.com. These platforms allow users to share, comment on and score photos online. In this paper, we report experiments conducted on two popular datasets, AVA [23] and CUHKPQ [10].

AVA is a large-scale dataset for aesthetic visual analysis [23]. It contains 255,530 images collected from the website DPChallenge.com. The provider of AVA does not release the images but rather their web links. We successfully downloaded 193,077 images; the links to the remaining images could not be accessed. Each image has a distribution of quality scores (from 1 to 10, where 10 is the highest) contributed by photographers from the website. 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 [18], the top 10% and bottom 10% of the photos were designated as high- and low-quality images, respectively, and the ambiguous images in the middle of the quality range were discarded.

CUHKPQ consists of 17,690 images collected from professional photography websites and contributed by amateur photographers [10]. Each image has been labeled as being of either "high" or "low" aesthetic quality by 10 viewers. In addition, the images in this dataset are divided into seven categories, with each image being assigned one of the following seven tags: "animal", "architecture", "human", "landscape", "night", "plant", and "static".

For each dataset, we randomly selected half of the images for training and the remaining images for testing. The deep neural network for deep aesthetic feature extraction was trained using the images in the training set.

2) Image Representation: As discussed in Section IV-B1, the widely used BOW visual representation and the recently developed representation based on generic features learned via large CNNs were adopted for image representation (IR). For the BOW representation, the DoG detector is used for interest point detection and the SIFT [34] approach is used for local feature description. The codebook is learned via hierarchical K-means with 6 levels and 10 centers at each level [35]. For the CNN feature representation, the 4096-dimensional output of the final fully connected layer is adopted as the visual representation. For the representation of textual features for text retrieval, we directly used the tag information associated with each image in the datasets.

3) Method Comparison: To validate the effectiveness of the aesthetic features learned with deep neural networks, we compared the results with those of several state-of-the-art hand-crafted features proposed in recent years [17]–[27] as well as those of [29].

To thoroughly study and compare universal aesthetic learning and various query-dependent aesthetic learning methods, we implemented four methods and compared them with each other.

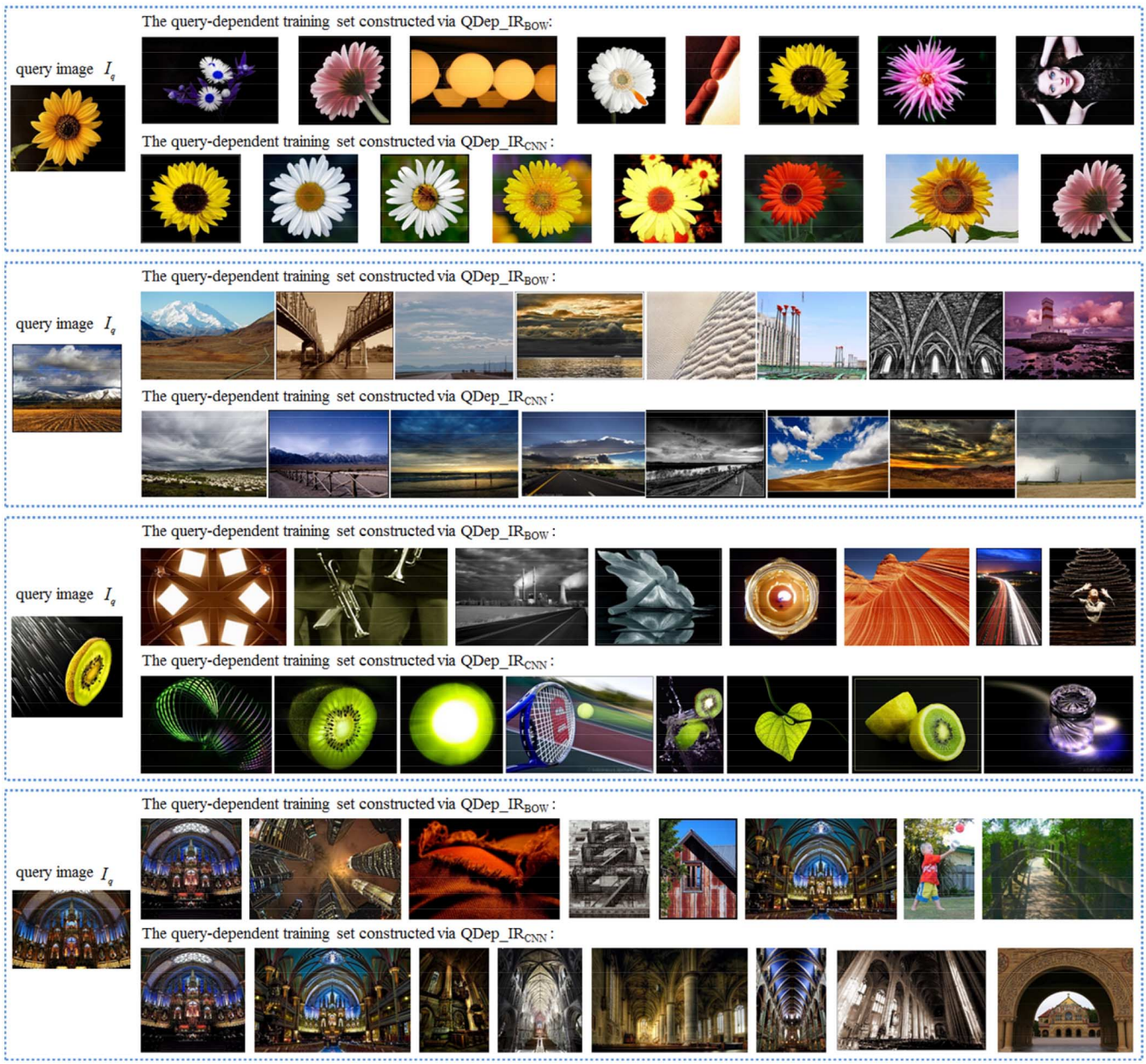


Fig. 7. The query-dependent training sets constructed via the query-dependent aesthetic models $\mathrm{QDep\_IR_{BOW}}$ and $\mathrm{QDep\_IR_{CNN}}$. (The top eight images for each are shown.)

Universal: In this method, a universal aesthetic model is directly trained on the training set and then the model is applied to all test images. Most previous aesthetic analysis research has relied on this type of method. In this case, all query images share the same aesthetic model.

QDep_Tag: Each image in CUHKPQ is associated with tag information. Therefore, we can train an aesthetic model for each tag, as in [10], [23]. Specifically, we separate the dataset into 7 categories according to the tags. Then, a model is trained for each tag. Regarding the query-dependent aspect, query images with the same tag share the same aesthetic model.

QDep_IR: The objective of this model is to adaptively build an individual aesthetic model for each query image through image retrieval. For each query image I_q in the testing set, we first find the top N images from the training set that are most visually similar to the query image via our image retrieval system. Then, the N returned images are used as the query-dependent training set to build the query-dependent aesthetic model for I_q. Finally, the aesthetic label of I_q is predicted using this query-dependent model.

QDep_Tag&IR: The query-dependent training set is constructed via multi-view query-dependent aesthetic learning with $\alpha = 0.5$. The other settings are similar to those in QDep_IR.

For all methods, the SVM [38] classifier, which has been widely used in previous aesthetic quality assessment studies [17]–[20], was adopted for aesthetic model training. All parameters were selected via 5-fold cross-validation on the training set. For QDep_IR, the size of the query-dependent training set was empirically set to $N = 50$. The average classification accuracy for each method is reported.

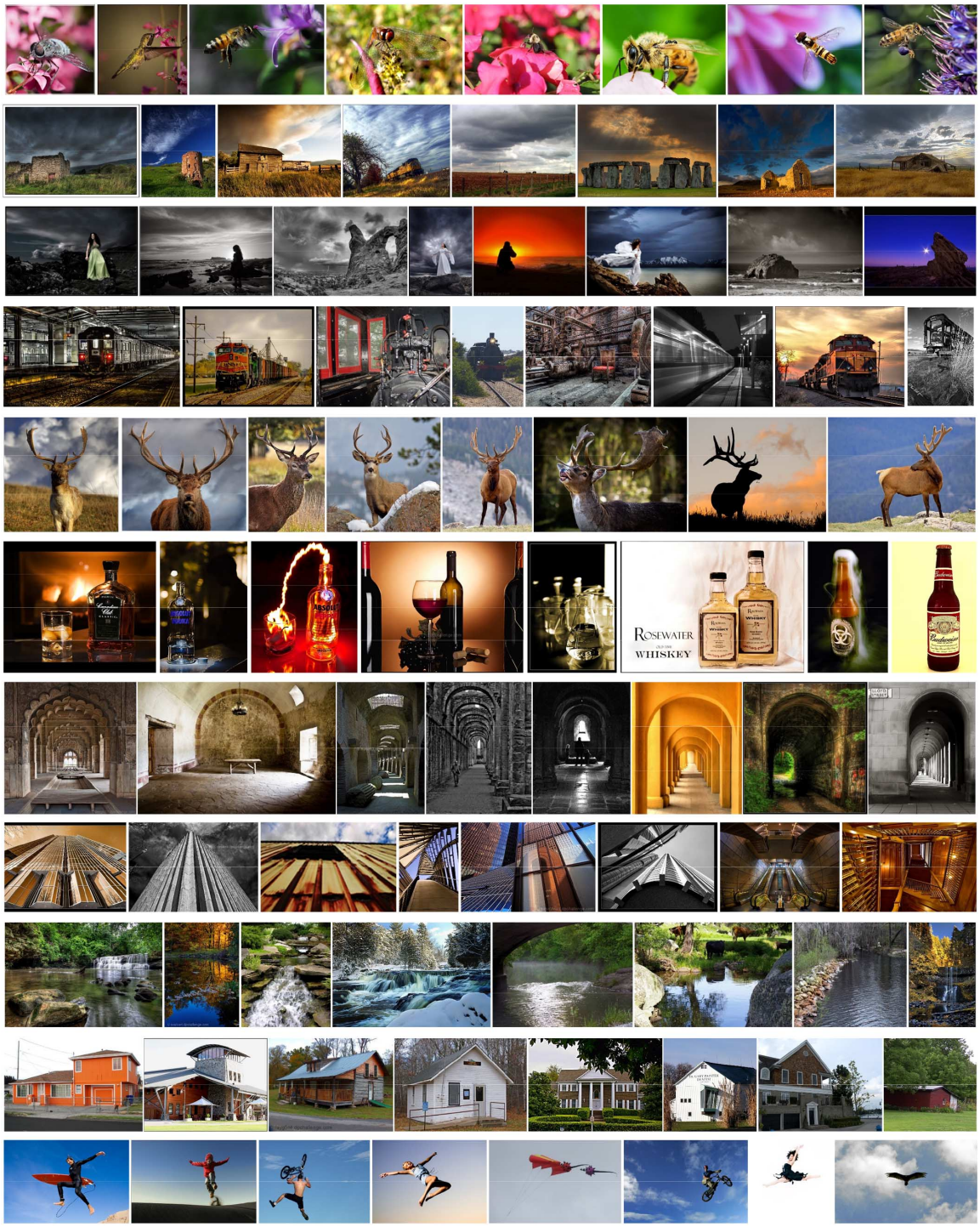


Fig. 8. More examples of query-dependent training sets constructed via $\mathrm{QDep\_IR_{CNN}}$. In each row, the first image is the query image and the remainder are the top seven images returned via image retrieval using CNN features.

B. Experimental Results on AVA

The experimental results obtained on the AVA dataset are summarized in Table I. Let us first compare the proposed deep aesthetic features with several hand-crafted features [17]–[27] and [29]. The comparison indicates that the proposed deep aesthetic features significantly outperform all other aesthetic features on this dataset, regardless of whether the universal model or a query-dependent model is applied. This demonstrates that the aesthetic features learned via our DCNNs are more discriminative for photo quality assessment. Compared with the deep learning model adopted in [29], the fully connected layers in our model contain much fewer neurons. This is because in this study, we discarded the ambiguous images in the middle of the quality range, thereby simplifying the binary classification problem.

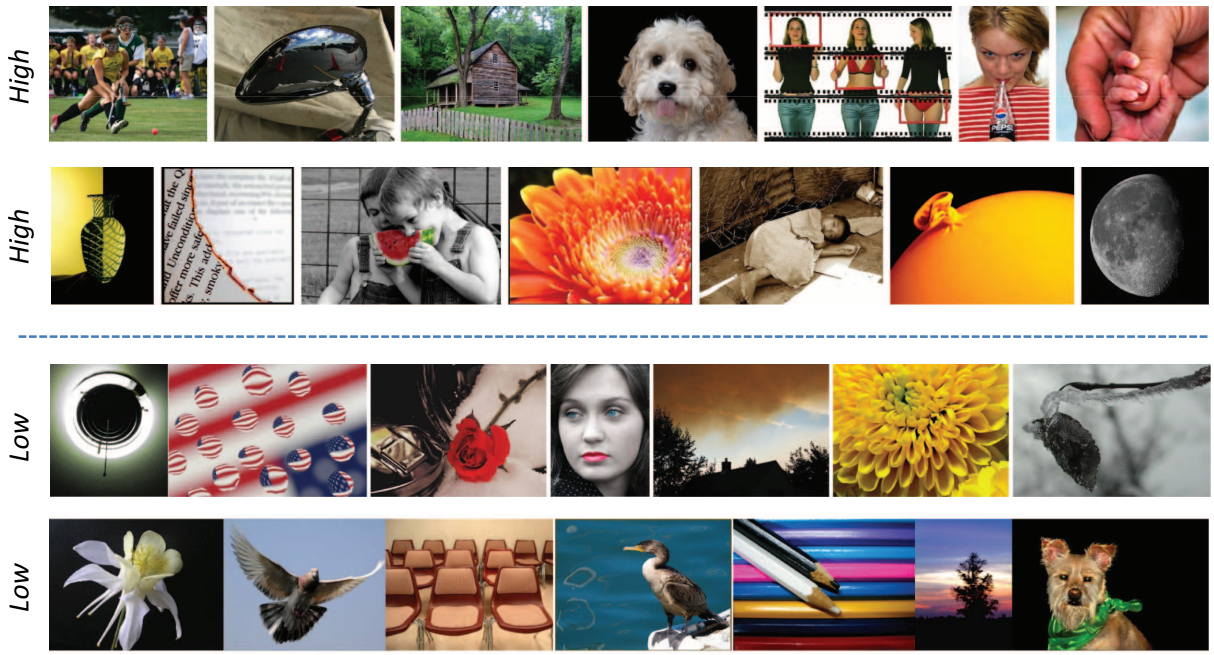


Fig. 9. Example images that are misclassified by our model. The label indicates the ground-truth aesthetic quality.

Now, we turn to a comparison of different aesthetic models. Table I shows that the proposed query-dependent aesthetic models $\text{QDep\_IR}_{\text{BOW}}$ and $\text{QDep\_IR}_{\text{CNN}}$ consistently outperform the universal model over all considered aesthetic features. The query-dependent model $\text{QDep\_IR}_{\text{CNN}}$ with the deep aesthetic representation yields the best performance (80.38%), whereas the Universal model with hand-crafted feature can achieve a classification accuracy of only approximately 70%.

Of the two query-dependent models, $\text{QDep\_IR}_{\text{CNN}}$ demonstrates better performance than does $\text{QDep\_IR}_{\text{BOW}}$. It is obvious that the adaptive training set used significantly affects the quality assessment performance. Fig. 7 shows the query-dependent training sets constructed by $\text{QDep\_IR}_{\text{BOW}}$ and $\text{QDep\_IR}_{\text{CNN}}$ for two query images. We can see that the query-dependent training sets returned by $\text{QDep\_IR}_{\text{CNN}}$ consist of images that are more similar to the query image. Therefore, $\text{QDep\_IR}_{\text{CNN}}$ achieves better performance. To further demonstrate and investigate the query-dependent training sets returned by $\text{QDep\_IR}_{\text{CNN}}$, more retrieval results are provided in Fig. 8. From this figure, we can see that the query image and the images returned by the retrieval system are visually similar, with similar shooting subjects, similar shooting backgrounds/environments, similar lighting conditions, similar shooting angles, and so on. Because many images in the AVA dataset do not have tag information, we could not apply QDep_Tag. We believe that if precise tag information were available for this dataset, the performance could be further improved by using QDep_IR&Tag.

Although our method achieves the best performance, about 20% images are still misclassified. Fig. 9 shows some examples of the misclassified images. The label indicates the ground-truth aesthetic quality. The reasons that may potentially cause the misclassification are complex. For some images, their ground truth labels may be incorrect. Though we have discarded the ambiguous images in the middle of the quality range, it is inevitable that some noise still exists. For some images, their content is very rare and we cannot retrieve sufficient images which are closely similar with it to construct a good query-dependent training set. For some images, their aesthetic quality labels may also be affected by other factors beyond the photographic rules, for example, the rare shooting object, the emotion and culture behind them, etc. These factors are too abstract to be captured by our model.

The influence of the query-dependent training set size N is illustrated in Fig. 10. In this figure, N varies from 10 to 250. We can see that the query-dependent aesthetic model can achieve good performance even at very small query-dependent training set sizes (≤ 50). In most cases, it exhibits stable performance when N is larger than 50. This indicates that the aesthetic model is truly dependent on the query image and can be learned from images located in a small neighborhood surrounding the query image.

C. Experimental Results on CUHKPQ

The experimental results obtained on the CUHKPQ dataset are summarized in Table II. Let us first compare the different aesthetic features, i.e., the proposed deep aesthetic features and the other baseline aesthetic features [17]–[29]. From Table II, we can observe that the proposed deep aesthetic features significantly outperform all other aesthetic features, regardless of whether the universal model or a query-dependent model is adopted.

Now, we turn to a comparison of the different query-dependent aesthetic models. This table yields the following observations. First, all query-dependent aesthetic models consistently outperform the universal model across all considered aesthetic features. Second, QDep_Tag&IR ($\text{QDep\_Tag\&IR}_{\text{BOW}}$ and $\text{QDep\_Tag\&IR}_{\text{CNN}}$) outperforms QDep_Tag and QDep_IR

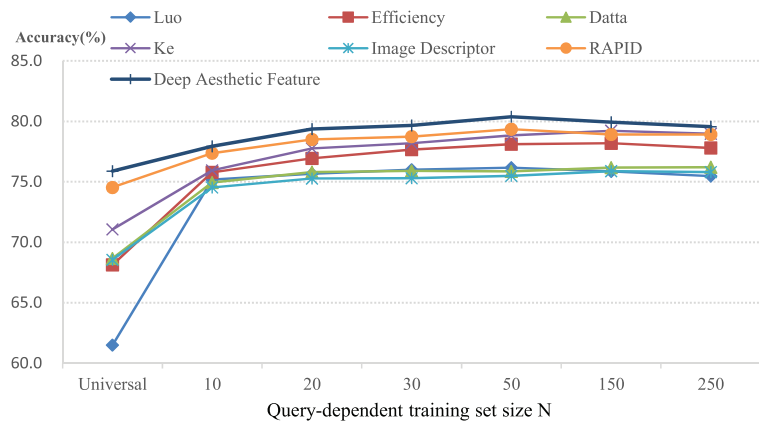


Fig. 10. Classification accuracy of the universal model and query-dependent models with varying query-dependent training set size N (AVA).

TABLE II
EXPERIMENTAL RESULTS OBTAINED ON THE CUHKPQ DATASET. OUR PROPOSED DEEP AESTHETIC FEATURES SIGNIFICANTLY OUTPERFORM ALL OTHER HAND-CRAFTED AESTHETIC FEATURES. THE QUERY-DEPENDENT MODELS CONSISTENTLY OUTPERFORM UNIVERSAL MODEL

	Accuracy (%)					
Aesthetic Features	Universal	QDep_Tag	QDep_IR$_{BOW}$	QDep_Tag&IR$_{BOW}$	QDep_IR$_{CNN}$	QDep_Tag&IR$_{CNN}$
Luo [19]	76.91	79.50	78.35	79.94	*88.86*	88.55
Efficiency [27]	81.76	84.47	82.66	85.26	90.31	*90.81*
Datta [17]	83.48	85.89	83.56	86.27	90.28	*91.00*
Ke [18]	81.70	82.98	82.08	84.24	90.12	*90.55*
Image Descriptors [11]	79.53	83.10	81.73	85.17	89.45	*90.18*
RAPID [29]	83.80	84.15	84.81	86.13	90.15	*90.29*
Deep Aesthetic Features	**86.20**	**86.35**	**86.38**	**87.02**	**91.59**	***91.94***

(QDep_IR$_{BOW}$ and QDep_IR$_{CNN}$), which demonstrates that a combination of visual and textual information can yield a better query-dependent training set. Third, query-dependent models with CNN-based IR significantly outperform query-dependent models with BOW-based IR because the CNN representation can describe the visual content of an image more comprehensively and thus retrieve more similar query-dependent training images. One example is shown in Fig. 11. This figure presents the query-dependent training sets constructed using different IR methods. A better retrieval result gives rise to higher accuracy in the aesthetic quality classification results. Fourth, the query-dependent model QDep_Tag&IR$_{CNN}$ with the deep aesthetic representation always achieves the best performance.

We further investigated the performance of the Universal model and the query-dependent model QDep_Tag&IR$_{CNN}$ on each of the 7 tags, as shown in Fig. 12. The proposed deep aesthetic features were used in both models. This figure shows that QDep_Tag&IR$_{CNN}$ significantly and consistently outperforms the universal model across all 7 categories. The results further demonstrate the effectiveness and robustness of our proposed aesthetic features learned via deep neural networks.

D. Comparison Between CNN Features and Learned Deep Aesthetic Features

In our query-dependent aesthetic assessment system, there are two types of features learned from DCNNs. One is the CNN features extracted using the open-source deep learning framework Caffe trained on ImageNet, as described in Section IV-B. The other is the deep aesthetic features extracted using the DCNN model we specifically trained for aesthetic feature mining, as described in Section III. We use the CNN features for a search for similar images to construct the query-dependent training set, and we use the deep aesthetic features for query-dependent aesthetic model (SVM) training. During the similar-image search, our goal is to retrieve images that have similar visual content to the query image but not images that are of similar aesthetic quality to the query image. Our assumption is that each image has its own adaptive quality assessment model and that this model can be learned from images that are visually similar to it. For example, when we assess the quality of a "landscape" image, we would like to use a model learned from other "landscape" images in the database. Therefore, we search for other "landscape" images using CNN features, which are powerful for use in similar-image searches, and then extract their deep aesthetic features to train the query-dependent model via the SVM classifier. In a case in which we wish to assess the quality of a "human" image, we should search for other "human" images from the database using CNN features. In summary, CNN features are used to search for visually similar images to construct the query-dependent training set, whereas learned aesthetic features are used to train the query-dependent aesthetic model because features of the latter type are more discriminative with regard to the subjective image quality.

To further verify this approach, we conducted the following experiments. Either of these two types of features could be used in the retrieval step and in the SVM classifier training step. Therefore, there are four possible combinations: 1) CNN features for retrieval and learned deep aesthetic features (Deep-

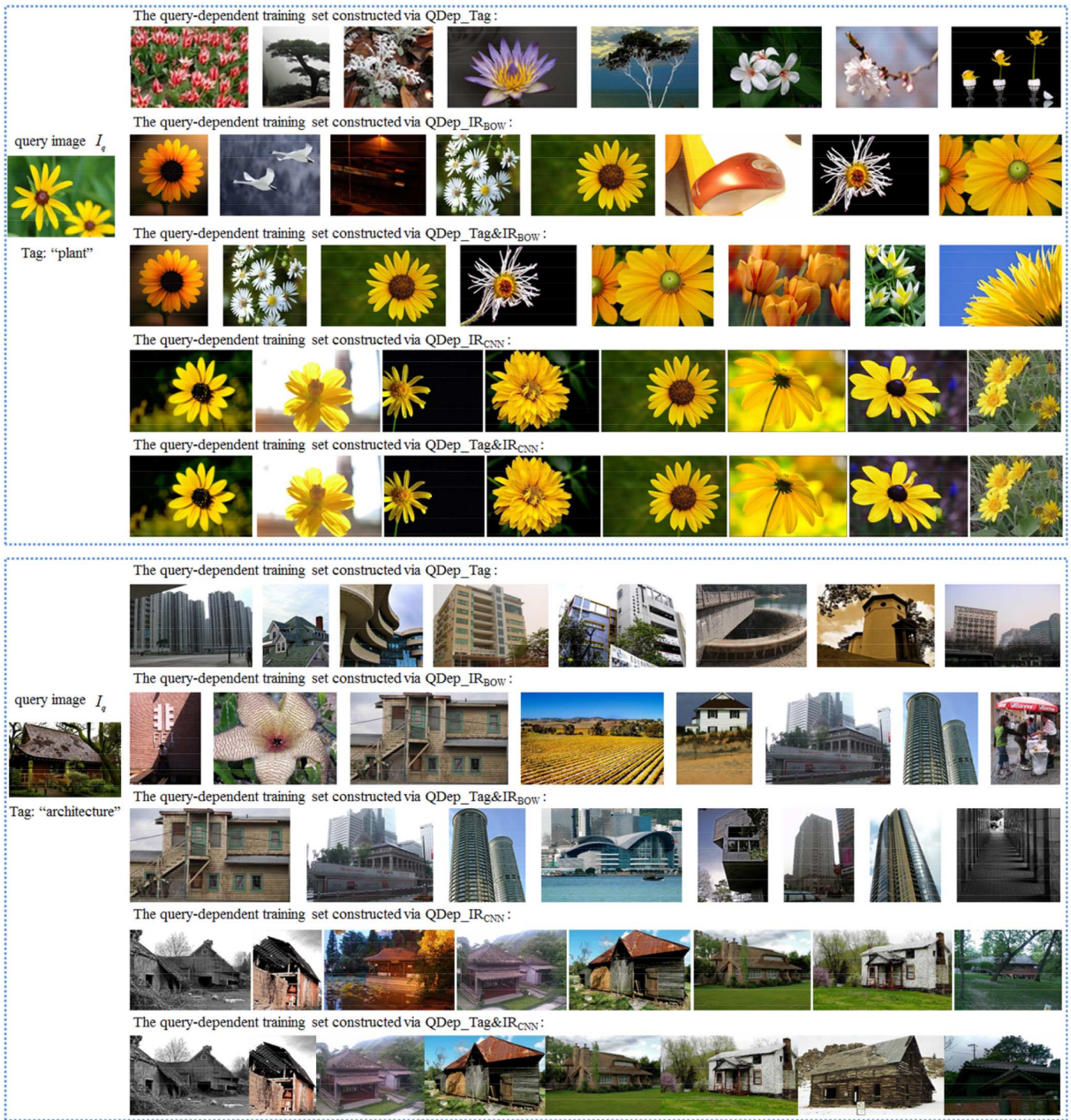


Fig. 11. Query-dependent training sets constructed via the query-dependent aesthetic models QDep_Tag, QDep_IR$_{\mathrm{BOW}}$, QDep_Tag&IR$_{\mathrm{BOW}}$, QDep_IR$_{\mathrm{CNN}}$, and QDep_Tag&IR$_{\mathrm{CNN}}$ (the top eight images for each are shown.).

Aesth) for SVM training, IR$_{\mathrm{CNN}}$+SVM$_{\mathrm{DeepAesth}}$; 2) CNN features for retrieval and CNN features for SVM training, IR$_{\mathrm{CNN}}$+SVM$_{\mathrm{CNN}}$; 3) DeepAesth features for retrieval and CNN features for SVM training, IR$_{\mathrm{DeepAesth}}$+SVM$_{\mathrm{CNN}}$; and 4) DeepAesth features for retrieval and DeepAesth features for SVM training, IR$_{\mathrm{DeepAesth}}$ + SVM$_{\mathrm{DeepAesth}}$. For the experiments presented in Section V-B and V-C, the IR$_{\mathrm{CNN}}$+SVM$_{\mathrm{DeepAesth}}$ combination was adopted. In the experiments reported here, we tested the performances of the other three cases and compared them with that of IR$_{\mathrm{CNN}}$ + SVM$_{\mathrm{DeepAesth}}$. We conducted the experiments on the AVA dataset, and the results are summarized in Table III.

Table III yields the following observations. CNN features are more suitable than DeepAesth features for image retrieval. Both IR$_{\mathrm{DeepAesth}}$+SVM$_{\mathrm{CNN}}$ and IR$_{\mathrm{DeepAesth}}$+SVM$_{\mathrm{DeepAesth}}$ underperform with respect to IR$_{\mathrm{CNN}}$ + SVM$_{\mathrm{CNN}}$ and IR$_{\mathrm{CNN}}$ + SVM$_{\mathrm{DeepAesth}}$. The reason is that when DeepAesth features are used for image retrieval, the returned images are similar to the query image in aesthetic quality but do not have contents similar to that of the query image. Comparing IR$_{\mathrm{CNN}}$+SVM$_{\mathrm{DeepAesth}}$ and IR$_{\mathrm{CNN}}$+SVM$_{\mathrm{CNN}}$, we can see that the former significantly outperforms the latter, which demonstrates that our learned deep aesthetic features are much more discriminative with regard to the subjective image quality.

E. Complexity Analysis

Unlike the universal method, our proposed query-dependent aesthetic learning method requires the training of a query-dependent aesthetic model for each query image. The additional computational cost introduced in this online training process

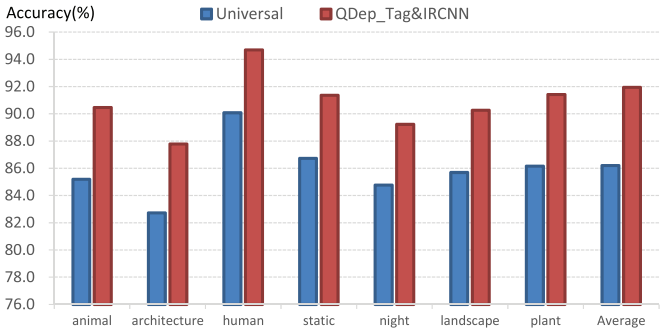


Fig. 12. Classification accuracies of the Universal model and the QDep_Tag&IR_{CNN} model on the seven categories of the CUHKPQ dataset. The proposed deep aesthetic features were used in both models. The results show that QDep_Tag&IR_{CNN} consistently outperforms the Universal model across all seven categories.

TABLE III
COMPARISON BETWEEN CNN FEATURES AND LEARNED DEEP AESTHETIC FEATURES

	Accuracy (%)
$IR_{CNN} + SVM_{DeepAesth}$	**80.38**
$IR_{CNN} + SVM_{CNN}$	77.68
$IR_{DeepAesth} + SVM_{CNN}$	75.67
$IR_{DeepAesth} + SVM_{DeepAesth}$	75.73

includes the image retrieval task and the SVM training of the query-dependent aesthetic model training.

The image retrieval system was implemented using KD-Tree via OpenCV. To ensure high retrieval performance, 2048 parallel kd-trees are constructed. Our image retrieval system is very efficient. For a given query image, it requires less than 0.01 s to return the retrieval result from the database.

For the SVM training of the query-dependent aesthetic model, as we observed in the experiments presented above, the proposed method demonstrates good, stable performance at a small N (e.g., 50). Therefore, we tested the average SVM training time cost at $N = 50$, and it was found to be only 0.0008 s. The time cost of SVM testing is negligible.

In sum, the total time cost of the online query-dependent aesthetic learning process is less than 0.01 s, which is sufficiently small for real-time applications. The entire system was implemented using C++. The time cost reported above was obtained on a PC with a 3.4 GHz Intel Core CPU and 4 GB of memory.

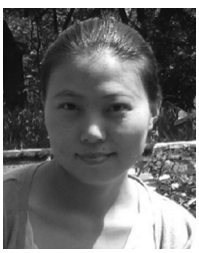
VI. CONCLUSION

In this paper, we propose the automatic mining of abstract aesthetic features using deep convolutional neural networks and also propose a general framework for query-dependent aesthetic quality assessment to solve the problems inherent to the current universal aesthetic learning methodology. In our proposed method, a query-dependent aesthetic model is built for every given query image to describe its unique aesthetic attributes. The query-dependent model is learned from the neighbors of the query image in both visual and textual space. By leveraging mature image and text retrieval systems, high efficacy and efficiency of the query-dependent aesthetic method are ensured. Extensive experiments on two popular datasets demonstrate that our proposed deep aesthetic features outperform the state-of-the-art hand-crafted aesthetic features and that our proposed query-dependent scheme significantly and consistently outperforms the conventional universal scheme.

REFERENCES

[1] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li, "The role of attractiveness in web image search," in *Proc. ACM MM*, 2011, pp. 63–72.
[2] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver, "The role of tags and image aesthetics in social image search," in *Proc. 1st SIGMM Workshop Social Media*, 2009, pp. 65–72.
[3] N. Murray, L. Marchesotti, F. Perronnin, and F. Meylan, "Learning to rank images using semantic and aesthetic labels," in *Proc. BMVC*, 2012, pp. 1–10.
[4] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. Int. Conf. Multimedia*, 2010, pp. 271–280.
[5] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 833–843, Jun. 2012.
[6] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li, "Oscar: On-site composition and aesthetics feedback through exemplars for photographers," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 353–383, 2012.
[7] Y. Jin, Q. Wu, and L. Liu, "Aesthetic photo composition by optimal crop-and-warp," *Comput. Graph.*, vol. 36, no. 8, pp. 955–965, 2012.
[8] F.-L. Zhang, M. Wang, and S.-M. Hu, "Aesthetic image enhancement by dependence-aware object recomposition," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1480–1490, Nov. 2013.
[9] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1657–1664.
[10] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2206–2213.
[11] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.
[12] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proc. ECCV*, 2010, pp. 1–14.
[13] C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: A photo quality assessment and photo selection system," in *Proc. Int. Conf. Multimedia*, 2010, pp. 827–830.
[14] W.-T. Chu, Y.-K. Chen, and K.-T. Chen, "Size does matter: How image size affects aesthetic perception?," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 53–62.
[15] O. Wu, W. Hu, and J. Gao, "Learning to predict the perceived visual quality of photos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 225–232.
[16] Z. Dong and X. Tian, "Multi-level photo quality assessment with multi-view features," *Neurocomputing*, vol. 168, pp. 308–319, 2015.
[17] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. ECCV*, 2006, pp. 288–301.
[18] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 419–426.
[19] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. ECCV*, 2008, pp. 386–399.
[20] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 997–1000.
[21] W. Yin, T. Mei, and C. W. Chen, "Assessing photo quality with geo-context and crowdsourced photos," in *Proc. IEEE Conf. Vis. Commun. Image Process.*, Nov. 2012, pp. 1–6.
[22] C. Li, A. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3221–3224.
[23] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2408–2415.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, vol. 1, no. 2, p. 4.

[25] Y. Wu, C. Bauckhage, and C. Thurau, "The good, the bad, and the ugly: Predicting aesthetic image labels," in *Proc. IEEE Int. Conf. Pattern Recog.*, Aug. 2010, pp. 1586–1589.
[26] S. Bhattacharya, R. Sukthankar, and M. Shah, "A holistic approach to aesthetic enhancement of photographs," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 7, no. 1, p. 21, 2011.
[27] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Assessment of photo aesthetics with efficiency," in *Proc. 21st Int. Conf. Pattern Recog.*, Nov. 2012, pp. 2186–2189.
[28] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 33–40.
[29] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. Int. Conf. Multimedia*, 2014, pp. 457–466.
[30] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proc. ACM MM*, 2011, pp. 1213–1216.
[31] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.
[32] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 624–630, 2006.
[33] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labelling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[35] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2161–2168, IEEE.
[36] G.-J. Qi, C. C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.
[37] G.-J. Qi, M.-H. Tsai, S.-F. Tsai, L. Cao, and T. S. Huang, "Web-scale multimedia information networks," *Proc. IEEE*, vol. 100, no. 9, pp. 2688–2704, Sep. 2012.
[38] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.



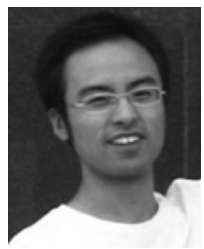
Xinmei Tian (M'13) received the B.E. degree and Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

She is currently an Associate Professor with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, University of Science and Technology of China. Her current research interests include multimedia information retrieval and machine learning.

Prof. Tian was the recipient of the Excellent Doctoral Dissertation of Chinese Academy of Sciences Award in 2012 and the Nomination of National Excellent Doctoral Dissertation Award in 2013.

Zhe Dong received the B.E. degree from the Xidian University, Xi'an, China, in 2013, and is currently working towards the M.Sc. degree at the University of Science and Technology of China, Hefei, China.

His research interests include image quality assessment, multimedia search, and machine learning.



Kuiyuan Yang received the B.E. and Ph.D. degrees in automation from the University of Science and Technology of China, Hefei, China, in 2007 and 2012, respectively.

He is currently a Research Staff Member with the Web Search and Mining Group, Microsoft Research, Beijing, China. His current research interests include computer vision, multimedia, and deep learning.

Tao Mei (M'06–SM'11) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is a Lead Researcher with Microsoft Research, Beijing, China. He has authored or coauthored over 100 papers in journals and conferences, and holds eight U.S. granted patents. His current research interests include multimedia information retrieval and computer vision.

Dr. Mei is an Associate Editor of *Neurocomputing* and the *Journal of Multimedia*. He was the recipient of several Best Paper Awards, including the Best Paper Awards at ACM Multimedia in 2007 and 2009, and the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award 2013.