Photo Aesthetics Analysis via DCNN Feature Encoding

Hui-Jin Lee, Ki-Sang Hong, Henry Kang, and Seungyong Lee

***Abstract*—We propose an automatic framework for quality assessment of a photograph as well as analysis of its aesthetic attributes. In contrast to the previous methods that rely on manually designed features to account for photo aesthetics, our method automatically extracts such features using a pre-trained deep convolutional neural network (DCNN). To make the DCNN-extracted features more suited to our target tasks of photo quality assessment and aesthetic attribute analysis, we propose a novel feature encoding scheme, SVM-driven sparse restricted Boltzmann machines (SVM-SRBM), which enhances sparseness of features and discrimination between target classes. Experimental results show that our method outperforms the current state-of-the-art methods in automatic photo quality assessment, and gives aesthetic attribute ratings that can be used for photo editing. We demonstrate that our feature encoding scheme can also be applied to general object classification task to achieve performance gains.**

***Index Terms*—photo aesthetics, aesthetic attributes, deep convolutional neural network, feature encoding, restricted Boltzmann machines.**

I. Introduction

With the rapid advances in digital image acquisition and distribution technology, a massive supply of photographs are now readily available to ordinary users. Creating, editing, and sharing photographs have never been easier, which also led people to have increasingly higher expectations on the quality of images that they routinely use. In order to obtain aesthetically pleasing photographs, they often turn to high-quality photo acquisition devices and editing tools. However, generating such a visually appealing photograph also requires understanding and applying a set of complex aesthetic principles in the process of photo acquisition and/or editing. Unfortunately, it takes ordinary users significant training and experience to master such knowledge and skills.

In recent years, the research community has addressed this issue by developing ways to provide automatic photo quality assessment. As shown in Fig. 1, many of them aim to automatically classify images into simple binary categories of *high quality* and *low quality* [1]–[9]. These approaches are common in that they first define a set of image features that they assume to affect the aesthetic quality of photographs, then design some mathematical models to extract them. Such hand-crafting of features, however, is not only difficult but often insufficient to account for the full, complex nature of photo aesthetics and therefore could lead to inaccurate assessments.

In an effort to overcome the limitations of such hand-crafted features, some suggested using generic image features that are typically used for general image recognition problems, and successfully gained enhanced performance for photo quality assessment [10], [11]. Another notable approach to automate the process of selecting and modeling image features involves *deep learning*, in particular, using the deep convolutional neural networks (DCNN) that have been trained on a large-scale image database [12]–[14] to obtain more accurate photo quality assessment than the conventional approaches listed above. These techniques based on automated feature modeling generally perform well in terms of judging whether the given photograph is aesthetically pleasing or not, but not so much in terms of explaining why.

In this paper, we build on and extend the deep-learning based approach, first to further improve on the accuracy of photo quality assessment, and second to not just stop at evaluating the photo quality but *explain* the reasons for such evaluation (Fig. 1), and thereby provide much more useful information and guidance for the user in properly selecting and/or editing target images. Thus, the objective of our system is two-fold: to classify a given photo into a *high* or *low* quality category at a high rate of accuracy, and also to associate the photo with a proper set of descriptive aesthetic attributes, such as *Motion Blur*, *Rule of Thirds*, and so on.

Rather than directly using the features extracted from the pre-trained DCNN, we develop a new encoding scheme to generate a set of customized features tailored to a specific photo classification task. In particular, we employ restricted Boltzmann machines (RBMs) [15] and use the information obtained from the support vector machines (SVMs) [16] for *discrimination* between classes, and the sparsity modeling [17] for *sparseness* of features. We call this scheme *SVM-driven sparse RBM (SVM-SRBM)*. As will be shown via various experimental results, our method outperforms the state-of-the-art approaches in terms of both of our photo analysis tasks: (1) photo quality assessment and (2) analysis of aesthetic attributes, and can provide valuable guidance in many applications including photo editing.

The main contributions of our work are summarized as follows:

- We propose a novel framework to perform automatic photo quality assessment as well as analysis of the

Hui-Jin Lee was with the Department of Electrical Engineering, POSTECH, Namgu, Pohang, Republic of Korea. "This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes [supplementary results]. Contact [leesy@postech.ac.kr] for further questions about this work.

Ki-Sang Hong is with the Department of Electrical Engineering, POSTECH, Namgu, Pohang, Republic of Korea.

Henry Kang is with the Department of Mathematics and Computer Science, University of Missouri, St. Louis, USA.

Seungyong Lee is with the Department of Computer Science and Engineering, POSTECH, Namgu, Pohang, Republic of Korea.

Manuscript received August 29, 2016.

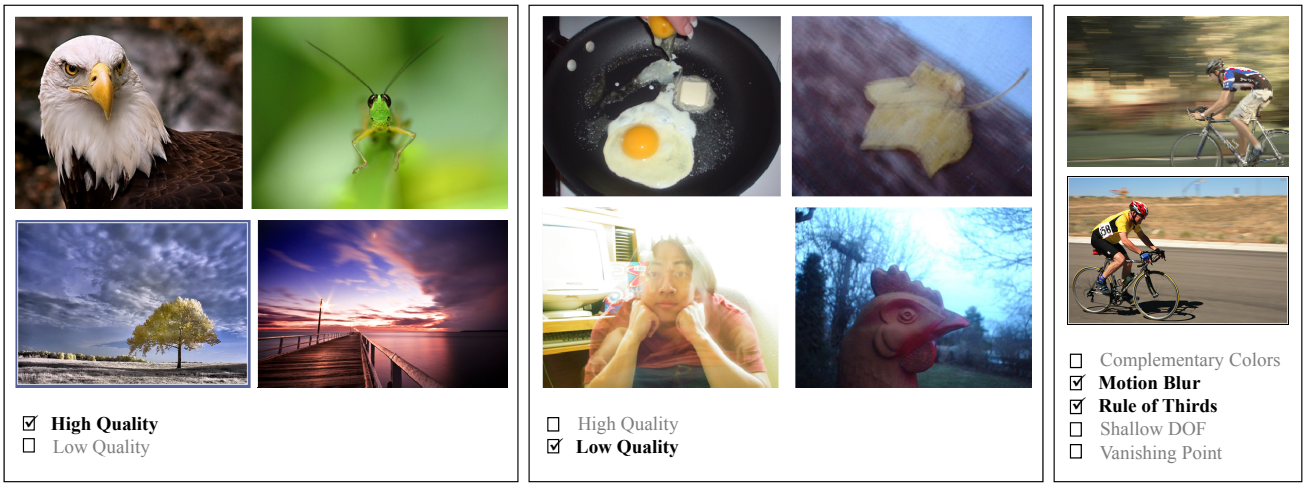


Fig. 1. Example photographs having (Left) "high quality", (Middle) "low quality", and (Right) visual attributes of "motion blur" and "rule of thirds". Our system classifies photographs into one of the two classes (high and low quality) and also analyzes five visual attributes on each photograph.

aesthetic attributes in a unified fashion.

- We develop a novel feature encoding scheme based on RBM, called SVM-SRBM, which helps improve the photo assessment accuracy via enhanced *feature sparsity* and *class discriminability*.
- We show that our system delivers better performance than the state-of-the-art photo quality assessment techniques in classification accuracy.
- We demonstrate that our system provides meaningful analysis on photo aesthetic attributes to assist photo selection/editing.
- We demonstrate that our feature encoding scheme has effectiveness on other classification tasks, such as general object classification.

II. Related Work

A. Photo Quality Assessment

In the task of photo quality assessment, binary classification has often been used, that is, determining whether the given photo is of aesthetically high quality or low quality. Tong et al. [6] used a combination of low-level features such as contrast, blurness, etc. to perform such two-fold classification. However, with the realization of the subjective nature of photo aesthetics, researchers started to use more high-level semantic features such as simplicity, familiarity, balance, etc., resulting in a more accurate classification [2], [5]. Later approaches [3], [7], [8] further improved the classification accuracy by concentrating their analysis on the most important region in the photograph, which typically contains the subject.

Bhattacharya et al. [1] presented an interactive system to perform analysis of photographic quality and thereby provide informed aesthetic suggestions for the task of image recomposition. Aydin et al. [18] developed a photo aesthetic analysis system that generates perceptually calibrated ratings for five aesthetic attributes (sharpness, depth, clarity, tone, and colorfulness), that can be used to provide aesthetic feedback in subsequent photo editing. All of these methods described so far are common in that they select some image features that are believed to influence the photo aesthetics the most, based on common photographic rules or intuition, then model them as some mathematical formulas so they can be extracted and used as the basis for quality assessment.

A limitation of using such hand-crafted features comes from the complex nature of photo aesthetics, which makes it difficult for only a handful of high-level features to fully describe. Marchesotti et al. [10] thus proposed using generic image descriptors such as Bag-of-Visual-Words (BOV) and Fisher Vector (FV) for the aesthetic evaluation of photographs. More recently, deep convolutional neural network (DCNN) was used to automatically extract aesthetic features [13], [14], and achieved improved accuracy of photo quality assessment over the previous techniques that use hand-crafted features or generic image descriptors. These methods however do not provide detailed aesthetic feedback or analysis for the users.

B. Feature Encoding Using RBM

Recently, deep learning based on deep convolutional neural networks (DCNN) has shown remarkable performance in image classification tasks [19], [20]. However, learning DCNN typically requires a large number of annotated image samples for estimating millions of parameters. The classification tasks that are needed in our photo aesthetic analysis come with relatively less amount of training data. Although directly using features extracted from a pre-trained DCNN on a large database achieved enough classification accuracy in the task of photo quality assessment [14], these features might not necessarily be suited for photo aesthetic analysis. We address this issue by way of feature encoding based on a novel Restricted Boltzmann machine (RBM) model designed to customize those features for our photo aesthetics analysis.

RBM [15] is a generative stochastic neural network, and has been applied to many tasks including dimensionality reduction, feature encoding and classification. They can be trained in an either supervised or unsupervised fashion, and efficiently via the gradient-based contrastive divergence algorithm [21]. Supervised RBM training [22]–[25] allows for significant

performance enhancement in classification. These methods aim to achieve performance gain by focusing on either feature sparsity or mathematical feasibility. Unlike other supervised RBMs, our RBM model encodes *sparse* and *discriminative* features in order to improve classification accuracy based on mathematical feasibility provided with the support vector machine (SVM).

III. Construction of Aesthetic Features

In this paper, we focus on two types of classification tasks: 1) classifying a given image into *high* or *low* quality, and 2) identifying which among possible aesthetic attributes, such as *Motion Blur*, *Rule of Thirds*, and *Shallow DOF*, should be associated with the image. Each task requires a proper set of image features to train its own classifier. Fig. 2 illustrates the overall process, which consists of extracting features from pre-trained DCNN (Section III-A) and encoding them for the target task with the proposed method (Section III-B). Section III-C describes how to use those encoded features to train the classifiers which will then be used to predict aesthetic classes of new images.

A. Feature Extraction from Pre-trained DCNN

We employ as a general feature extractor the deep convolutional neural network (DCNN) trained on a 1.2 million subset of the ImageNet dataset [26]. Recently, three DCNN architectures (CNN-F, CNN-M and CNN-S) were proposed [20], considering the trade-off between accuracy and speed. Their architectures are similar to the AlexNet [19], and contain five convolutional layers (C1∼C5) and three fully-connected layers (FC1∼FC3). The output of the last layer is followed by a 1,000-way Softmax that produces a distribution over the given 1,000 image categories. However, some factors are different from the AlexNet, such as the receptive field size and the convolution stride, etc.

In our work, we adopt the pre-trained CNN-S model (Fig. 2) which gives better performance than others. For a given RGB image I_i, we remove FC2 and FC3 layers of the pre-trained CNN-S, and extract 4,096-dimensional activations from the FC1 layer[1]. By normalizing those activations as DCNN features $\mathbf{x}_i$ for I_i, we construct the training set $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$, where $y_i \in \{1, 2, ..., C\}$ is a class label for C classes of the given classification task.

B. Encoding DCNN Features using SVM-SRBM

Features extracted from the pre-trained DCNN (Section III-A) are not necessarily suited to photo aesthetic analysis, since the DCNN was trained on 1,000 general object classes. This issue can be addressed by way of feature encoding. The RBM is one of the efficient feature encoding methods, which allows for both supervised and unsupervised training. In this paper, we propose a newly designed RBM model named SVM-driven sparse RBM (SVM-SRBM) (Fig. 3), and use it to customize DCNN-extracted features for the target classification tasks of photo aesthetics analysis in a supervised fashion. We first review RBM as well as cross-entropy-regularized RBM, then present the details of SVM-SRBM.

1) Restricted Boltzmann Machines: The RBM [15] is a bipartite graphical model that represents a joint distribution over an input vector $\mathbf{x} = [x_1, ..., x_L]^T$ and a hidden vector $\mathbf{h} = [h_1, ..., h_D]^T$, where $\mathbf{h}$ can be viewed in our work as an encoded feature vector. The joint probability of the RBM takes the form

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{x}, \mathbf{h})), \tag{1}$$

where E is an energy function, and $Z = \sum_{\mathbf{x}', \mathbf{h}'} e^{-E(\mathbf{x}', \mathbf{h}')}$ is the partition function. The energy function E of $\mathbf{x}$ and $\mathbf{h}$ is defined as

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{b}^T \mathbf{h} - \mathbf{c}^T \mathbf{x}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{D \times L}$, $\mathbf{b} \in \mathbb{R}^{D \times 1}$, $\mathbf{c} \in \mathbb{R}^{L \times 1}$ are parameters of the RBM. The probability of $\mathbf{x}$ is defined by the sum over all possible hidden vectors:

$$p(\mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{x}, \mathbf{h})). \tag{3}$$

Since the RBM has no intra-layer connections, the input unit activations are mutually independent given the hidden unit activations. That is, the conditional probability of the input vector $\mathbf{x}$, given the hidden vector $\mathbf{h}$, is

$$p(\mathbf{x}|\mathbf{h}) = \prod_{l=1}^{L} p(x_l|\mathbf{h}) = \prod_{l=1}^{L} \sigma(\sum_{d} w_{d,l} h_d + c_l). \tag{4}$$

Likewise, the conditional probability of $\mathbf{h}$ given $\mathbf{x}$ is

$$p(\mathbf{h}|\mathbf{x}) = \prod_{d=1}^{D} p(h_d|\mathbf{x}) = \prod_{d=1}^{D} \sigma(\sum_{l} w_{d,l} x_l + b_d), \tag{5}$$

where $\sigma(r) = (1 + e^{-r})^{-1}$ is the sigmoid function. The RBM is trained by minimizing the negative log-likelihood of training set X with respect to $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$:

$$\underset{\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}}{\arg\min} - \sum_{\mathbf{x} \in X} \log p(\mathbf{x}). \tag{6}$$

Although the gradient is intractable to compute, contrastive divergence (CD) [27] can be used to approximate it. The algorithm performs Gibbs sampling and a gradient descent procedure to update parameters. Thus the RBM is efficiently trained with this CD algorithm, and the parameters are updated using the following update rules:

$$\begin{aligned} \Delta w_{l,d} &\propto \langle x_l h_d \rangle_0 - \langle x_l h_d \rangle_T, \\ \Delta b_d &\propto \langle h_d \rangle_0 - \langle h_d \rangle_T, \\ \Delta c_l &\propto \langle x_l \rangle_0 - \langle x_l \rangle_T, \end{aligned} \tag{7}$$

where $\langle \cdot \rangle$ is defined as the average over the set of training examples, and T is the number of the Gibbs sampling.

[1]We found in our experiments (Section IV) that encoding features of FC1 works better than FC2 for our target tasks. Although FC2 worked well for the original ImageNet classification as the closest layer to the last layer with 1,000-way Softmax, it could be less suitable than FC1 for other applications.

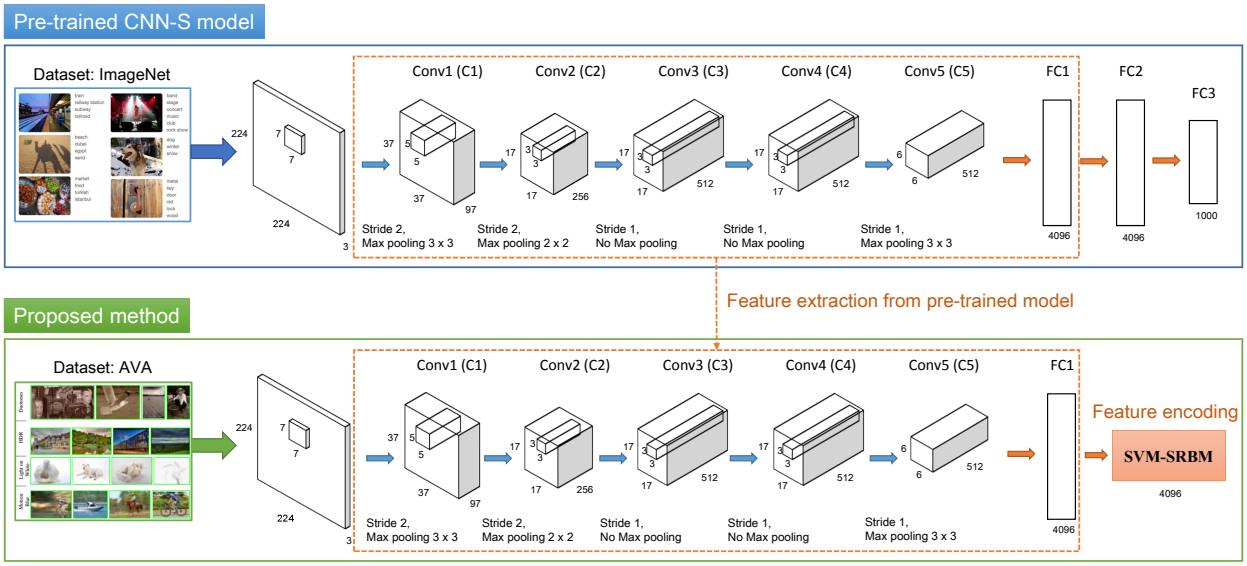


Fig. 2. Overall process for constructing aesthetic features.

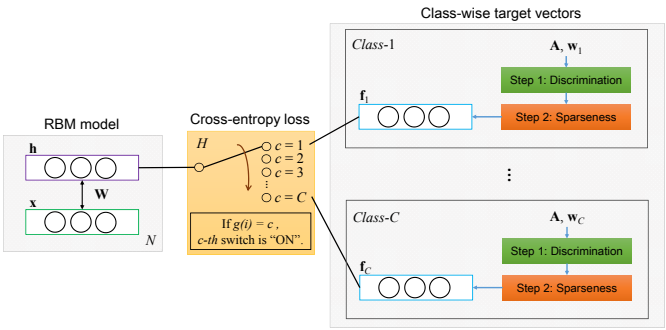


Fig. 3. Structure of a SVM-SRBM.

2) Cross-entropy-regularized RBM: RBMs form the latent structure $\mathbf{h}$ of input data by minimizing only the negative log-likelihood of training set X (Eq. (6)). However, given the nature of an encoding task, it may not be a desirable approach. The cross-entropy-regularized RBM [17] is designed to adjust feature activations of the hidden vector. For this, the cross-entropy loss regularization is defined to minimize the difference between the hidden vector $\mathbf{h}$ and a target vector $\mathbf{f}$. The optimization problem with this regularization is as follows:

$$\underset{\{\mathbf{W},\mathbf{b},\mathbf{c}\}}{\arg\min} -\sum_{i=1}^{N}\log p(\mathbf{x}_i) - \lambda\sum_{d=1}^{D} H(f_{i,d}, h_{i,d}), \tag{8}$$

where $f_{i,d}$ and $h_{i,d}$ are the d-th target and hidden unit activations of i-th data, respectively, and $H(f_{i,d}, h_{i,d}) = -f_{i,d}\log h_{i,d} - (1-f_{i,d})\log(1-h_{i,d})$ is the cross-entropy loss between these activations. The update rule of the RBM in Eq. (7) is modified as follows:

$$\Delta w_{l,d} \propto \langle x_l z_d\rangle_0 - \langle x_l h_d\rangle_T, \tag{9}$$

where $z_d = (1-\alpha)h_d + \alpha f_d$ is the weighted sum of the hidden and target activations. When $\alpha = 0$ or $h_d = f_d$, this update rule is the same as the original update rule of the contrastive divergence.

3) SVM-driven Sparse RBM: Our goal is to encode DCNN features into suitable features for a target classification task, such as classifying high/low qualities (binary classes) or aesthetic attributes (several classes). We aim to learn the RBM so that it enables encoded features to share the class-wise properties obtained in a supervised fashion (Fig. 3). We use the cross-entropy-regularized RBM method [17] that promotes target properties through regularization. To control hidden vectors with the class-wise conditions, we rewrite Eq. (8) as follows:

$$\underset{\{\mathbf{W},\mathbf{b},\mathbf{c}\}}{\arg\min} -\sum_{i=1}^{N}\log p(\mathbf{x}_i) - \lambda\sum_{d=1}^{D} H(f_{g(i),d}, h_{i,d}), \tag{10}$$

where $g(i) \in \{1, 2, ..., C\}$ is the class label of the i-th training data, and our model is manipulated with target vectors for C classes. It is essential to properly define target vectors for given classes, as it directly affects the quality of feature encoding. To design the class-wise target vectors, we consider two kinds of properties that are important in a classification task: *discrimination* and *sparseness*.

a) Discrimination: Designing features that encourage proper discrimination (separation) of target classes is crucial in classification [28], [29]. In order to model target vectors with good discrimination characteristics, we use the support vector machine (SVM) which is a discriminative classifier. We assume that a target vector can be modeled as a linear combination of sparse bases representing parts of a given training set. Given a basis matrix $\mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_K]^T \in \mathbb{R}^{D\times K}$ consisting of K basis vectors, the class-wise target vectors $\mathbf{f}_c$ are modeled by finding the set of coefficients $\mathbf{\Phi_c} = [\phi_{c,1}, \phi_{c,2}, ..., \phi_{c,K}]^T$ that have good discrimination characteristics for each class c. We thus pose the objective

function as follows:

$$\arg\max_{\Phi_c} \mathbf{w}_c^T(\mathbf{A}\mathbf{\Phi_c}) + z_c - \frac{1}{C-1}\sum_{c' \in C \setminus c} \mathbf{w}_{c'}^T(\mathbf{A}\mathbf{\Phi_c}) + z_{c'}, \quad (11)$$

where $\mathbf{w}_c \in \mathbb{R}^{D\times 1}$ and z_c are parameters of the SVM classifier learned to properly classify training data of the c-th class. For each target vector $\mathbf{f}_c = \mathbf{A}\mathbf{\Phi_c}$, the first term is the classification score for the classifier of the corresponding class, and the second term is the average classification score for classifiers of other classes. From this process, we obtain $\mathbf{\Phi_c}$ that makes target vector $\mathbf{f}_c$ fit for the classifier of the corresponding class but distant from classifiers of other classes. This objective function can be easily optimized by the gradient ascent procedure with respect to the coefficient vector $\mathbf{\Phi_c}$.

b) Sparseness: Sparse representations have a number of theoretical and practical advantages. They are particularly beneficial for classifiers, because they make classification easier in higher dimensional spaces [30]. We thus further transform the target vector $\mathbf{f}_c$ to have the *sparseness* property by applying the sparsity modeling [17]. Each unit activation $f_{c,d}$ in a target vector $\mathbf{f}_c$ is transformed as:

$$f_{c,d} = (R(f_{c,d}, \mathbf{f}_c))^{(1/\mu)-1}, \quad (12)$$

where $R(f_{c,d}, \mathbf{f}_c)$ assigns a value from 0 to 1 based on the rank of $f_{c,d}$ in $\mathbf{f}_c$. That is, it gives a value of 0 to the smallest value, and 1 to the largest. The target mean μ ($0 < \mu < 1$) creates the power-law relationship. If $\mu < 0.5$, the distribution of values obtained from $R(f_{c,d}, \mathbf{f}_c)$ will be positively skewed. It means only a few values in a target vector $\mathbf{f}_c$ are high while most remain low.

c) Iteration: Our SVM-SRBM model (Eq. (10)) and target vectors $\mathbf{f}$ (Eq. (11) and Eq. (12)) are updated several times, alternating with each other. When updating the target vectors $\mathbf{f}$, computing the basis matrix $\mathbf{A}$ and SVM parameters $\{\mathbf{w}_c, z_c\}$ in Eq. (11) requires the encoded feature (hidden) vectors $\mathbf{h}_i$ of all training data $\mathbf{x}_i$. At first, these encoded vectors $\mathbf{h}_i$ are obtained from a randomly initialized SVM-SRBM model (using random $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ values). Given $\mathbf{h}_i$, we learn the SVM parameters $\{\mathbf{w}_c, z_c\}$ so that $\mathbf{h}_i$ would find proper classes, and also construct the basis matrix $\mathbf{A}$ using non-negative matrix factorization (NMF)[2] [31]. Once the target vectors are constructed using Eq. (11) and Eq. (12), we update SVM-SRBM model with these target vectors using Eq. (10), then obtain new $\mathbf{h}_i$ from the updated SVM-SRBM model.

d) Class-wise multiple target vectors: So far, we have assumed a single target vector for each class. However, most classes have intra-class variations, such as viewpoint, scale, and light condition changes. To incorporate such intra-class variations, we model M target vectors for each class, so that each class can have M clusters of representative features. We cluster, for each class, the encoded vectors $\mathbf{h}_i$ of all training data $\mathbf{x}_i$ into M groups using K-means clustering, and thereby obtain a total of $G = C \times M$ groups for all C classes. Accordingly, in Eq. (10), the label $g(i)$ of a corresponding target vector for each training data $\mathbf{x}_i$ is extended to one of G group labels.

[2]Since NMF produces non-negative sparse bases from given data, the obtained bases can be used as the part-based representations.

Fig. 4 illustrates which properties the modeled target vectors capture for the respective classes. The left of Fig. 4 shows training photos whose encoded features have the closest distances to the three target vectors[3] modeled for the respective classes. In *High Quality* class for photo quality assessment, the target vectors capture three types of properties: geometric patterns, nature with high dynamic range, and well-focused persons. In *Motion Blur* class for aesthetic attribute analysis, the target vectors capture three types of blur: a blur by a circle motion, a blur by a moving object and a blur by a fast movement. To test the robustness of our method, we conducted the same experiment described above on all the test photos. As shown on the right of Fig. 4, we observed that most of the test photos closest to a given target vector were well matched with those of the training photos. It means feature encoding learned from the training set was effectively transferred to the test set, and features of test photos were encoded similarly to the matching training photos. More examples of the representative images for different target vectors can be found in the supplementary material. Table I illustrates enhanced classification performance via using multiple target vectors.

To summarize the overall procedure to develop our SVM-SRBM including the concept of class-wise multiple target vectors: (a) Parameters $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ of the SVM-SRBM are randomly initialized; (b) The encoded features $\mathbf{h}_i$ of all training data $\mathbf{x}_i$ are obtained from the SVM-SRBM; (c) All $\mathbf{h}_i$ are clustered into G groups; (d) K sparse bases and classifiers of G groups are constructed using all $\mathbf{h}_i$; (e) G target vectors are modeled using Eq. (11) and Eq. (12); (f) The SVM-SRBM is trained using Eq. (10). The steps (b) $\sim$ (f) are repeated until the termination condition (e.g., the number of iterations) is met.

C. Classifier Training with Encoded Features

Given a 4,096-dimensional DCNN feature $\mathbf{x}_i$ extracted from an image I_i, we encode $\mathbf{x}_i$ into a feature vector $\mathbf{h}_i$ of the same dimension using the SVM-SRBM learned as described in Section III-B3. We use this encoded feature $\mathbf{h}_i$ as an aesthetic feature of I_i, and construct the training set $H = \{(\mathbf{h}_1, y_1), (\mathbf{h}_2, y_2), ..., (\mathbf{h}_N, y_N)\}$, where $y \in \{1, 2, ..., C\}$ are class labels. For photo quality classification (*high* or *low*, $C = 2$), we train a single SVM classifier on H, which separates a given training data into proper two classes. For multi-class classification such as C aesthetic attributes, we train a 1-vs-rest SVM classifier for each attribute, and thereby obtain total C classifiers. To train the classifiers, we use the L2-regularized L1-loss SVC model in the LIBLINEAR [16].

IV. Experiments

We conducted experiments to evaluate the performance of our SVM-SRBM in photo quality assessment (Section IV-A) and aesthetic attribute analysis (Section IV-B). We also analyzed the correlation between photo quality and aesthetic attributes (Section IV-C), and the effectiveness of

[3]In Section IV, using *three* target vectors for each class achieves best classification performance both in photo quality assessment and aesthetic attribute analysis.

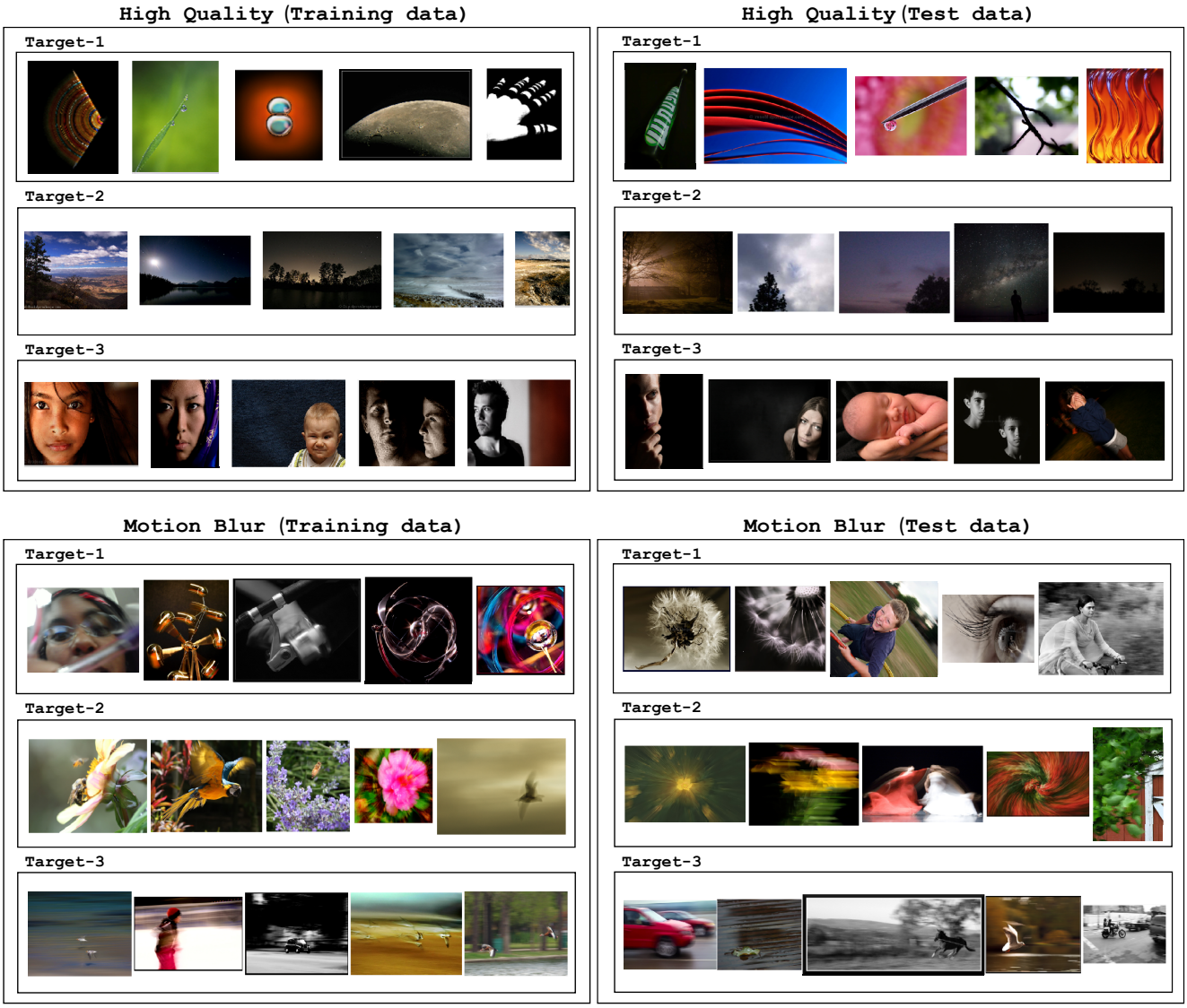


Fig. 4. Training and test photos with the closest distances to the three target vectors modeled for the respective classes.

our SVM-SRBM on a different classification task, general object classification (Section IV-D). We used the pre-trained DCNN provided in the MATLAB toolbox for Convolutional Neural Networks (CNNs), called MatConvNet [32]. We set the parameters of SVM-SRBM as follows: the number of bases $K = 50$ and the target mean $\mu = 0.1$. Our SVM-SRBM model and target vectors were updated through five iterations. Additional discussions of our experiments were presented in Section IV-E, and more example images can be found in the supplementary materials.

A. Photo Quality Assessment

The goal of our first classification task, *photo quality assessment*, is to determine whether the aesthetic quality of a given photo is *high* or *low*. We used the AVA dataset for this experiment and compared our classification accuracy with other related methods.

1) AVA Dataset: Aesthetic Visual Analysis (AVA) [11] is a large-scale dataset constructed for photo aesthetic analysis, and contains more than 250 thousand images from DPChallenge.com. Each image is tagged with a distribution of scores voted by different viewers, which ranges from one to ten. We computed the average score using the distribution and used it as the ground truth aesthetic quality scores. We followed the experimental settings in the previous works [9], [14] that had used the same dataset for image quality assessment. Using the quality scores, top 10% and bottom 10% of the photos were marked as high and low quality classes, respectively, and the remaining ambiguous photos were excluded in the experiments. For each class, half of the photos were randomly selected for the training, and the other half were used for the test. To address the question on random splitting of training and test data in AVA, we tested our method 5 times with randomly split data, and obtained average accuracy.

2) Classification Accuracy: To check the effectiveness of our encoding method, we first compared several settings of our method with the *baseline* that uses only the features extracted from the pre-trained DCNN model (CNN-S) without any encoding (Table I). In the baseline, "FC1" and "FC2" denote image features extracted from the FC1 and FC2 layers of the CNN-S, respectively. "FC1 + SVM-SRBM" means features from the FC1 layer are encoded with our SVM-SRBM. Also, "T1", "T3", "T5" and "T10" denote the number of target vectors for each class. The result shows that feature encoding with our SVM-SRBM achieves improved accuracy in photo aesthetics assessment over directly using the DCNN features (FC1 and FC2). Using multiple target vectors to exploit intra-class variations in each class gave a better result than using a single target vector, where three target vectors performed best.

Table I also compares our classification accuracy with other feature encoding methods. Our SVM-SRBM was developed on RBM model, thus we tested other RBM-based methods for comparison purposes: Standard RBM [15], Cross-entropy-regularized RBM [17], and BBP RBM [23]. As we did with ours, we applied these encoding methods to DCNN features extracted from the FC1 layer. We also tested with a CNN weight transfer approach (T-CNNW) [33] in which the pre-trained parameters of C1∼C5 and FC1 layers were fixed, and only the FC2 layer was re-trained for the target task with back propagation[4]. The test results show that our SVM-SRBM performed best among all of these methods. In particular, the fact that our model was directly learned from parameters of the SVM classifier, we believe, gives an edge over others, including T-CNNW that focuses on minimizing the classification error of the Softmax classifier.

Table II also shows performance comparisons with other existing photo quality assessment methods [2]–[5], [10], [12]–[14], [34][5]. Note that DCNN-based methods [12]–[14], [34] deliver higher accuracy than others. Dong et al. [14] achieved an additional performance gain by using a spatial pyramid (SP) to combine features from different regions. The last two columns of Table II show that our method (FC1 + SVM-SRBM-T3) outperformed DCNN-based methods [12]–[14], [34], and also that we achieved the best performance by employing a spatial pyramid (Ours + SP). Additionally, we tested our method on a whole set of successfully downloaded AVA images[6] (instead of just top 10% and bottom 10% of the photos), and obtained 81.02% accuracy (cf. 83.44% for top 10% and bottom 10%).

[4]Parameters of the FC2 layer were updated with an initial learning rate of 0.1 which is lowered by $1/10$ after three epochs, and with a momentum of 0.9. The numbers of training epochs and the batch size were set to 15 and 30, respectively. The error rate of the training data was converged within a few epochs (5∼6).

[5]The accuracy values in Table II were quoted from the corresponding papers. Experimental setting in three methods [10], [12], [34] was different from our setting. According to Tian et al. [13], the performance of Lu et al. [12] on the same setting as ours (top and bottom 10% rated images) was roughly 74.54%.

[6]The current website disallows acquiring the entire set of AVA images. This issue was also mentioned in [9].

B. Photo Aesthetic Attribute Analysis

The goal of our second classification task is to identify descriptive aesthetic attributes for a given photograph. While theoretically there is no limit in the number of attributes our model can handle, we chose five aesthetic attributes for experimental purposes: *Complementary Colors*, *Motion Blur*, *Rule of Thirds*, *Shallow DOF*, and *Vanishing Point*.

1) 5-Style Dataset: In AVA dataset, 14 photographic style labels are provided for some images. Many of these styles are related to the camera parameters, such as shutter speed, exposure, and ISO level. AVA dataset provides lists of training and test images for a style classifier, where the style annotations are single-labeled for training images and multi-labeled for test images. For analysis of aesthetic attributes, we selected five styles from this dataset, and constructed a new dataset named 5-Style dataset. The selection of the five styles is based on the aesthetic attributes analyzed in the previous works: *Complementary Colors* [3], [9], [35], *Motion Blur* [2], [5], [9], [18], *Rule of Thirds* [1]–[3], [36]–[38], *Shallow DOF* [2], [3], [9], [18], except for *Vanishing Point* which we selected based on the observation in AVA dataset that photos having this property are highly likely to be classified as high quality (Fig. 7). The numbers of images associated with these styles are as follows: Complementary Colors (949), Motion Blur (609), Rule of Thirds (1,031), Shallow DOF (710), and Vanishing Point (674).

2) Classification Performance: The test images for this experiment have multi-labeled class annotations, and we measured the average precision for each class using the precision-recall curve. We then used the mean average precision (MAP) of the five classes to evaluate the classification accuracy. In Table I, we compared the accuracy of classification based on our SVM-SRBM features with the baseline DCNN features (FC1 and FC2). As shown in the table, we witnessed even bigger performance gain over the baselines compared to the case of photo quality assessment, which suggests that the pre-trained DCNN features are increasingly ill-equipped for the target classification task as the number of classes goes up. Also, Table I shows that our method outperformed all other encoding methods in this task as well, and achieved the best result when combined with the spatial pyramid method (Ours + SP).

C. Relation between Photo Quality and Aesthetic Attributes

1) AVA & 5-Style Dataset: Using the classifiers $\{\mathbf{w}_c, z_c\}$, $c = 1, ..., 5$, that have been learned for the five aesthetic attributes, we can compute a classification score $s = \mathbf{w}_c^T \mathbf{h}_i + z_c$ of each attribute c for a given photo $\mathbf{x}_i$. The score s is normalized by the sigmoid function $\sigma(s) = (1 + e^{-s})^{-1}$, and regarded as the aesthetic attribute score. Fig. 5 shows the normalized scores of the five aesthetic attributes for example photos. Fig. 6 shows the results of Top-5 and Bottom-5 ranked photos based on the computed scores for each aesthetic attribute. The ranked results were remarkably consistent with the query aesthetic attributes.

Using our aesthetic attribute scores computed for AVA images, we analyzed how much contribution each aesthetic

TABLE I
CLASSIFICATION ACCURACY COMPARISON FOR VARIOUS SETTINGS (ON AVA AND 5-STYLES DATASETS).

Methods	Feat-dim.	AVA, Accuracy(%)	5-Styles, MAP(%)
Baseline, FC1 (w/o encoding)	4,096	78.24	62.28
Baseline, FC2 (w/o encoding)	4,096	81.44	68.99
FC1 + SVM-SRBM-T1	4,096	82.29	74.40
FC1 + SVM-SRBM-T3	4,096	**83.44**	**75.53**
FC1 + SVM-SRBM-T5	4,096	83.27	74.91
FC1 + SVM-SRBM-T10	4,096	83.15	75.27
FC1 + Standard RBM [15]	4,096	78.08	71.34
FC1 + BBP RBM [23]	4,096	79.72	70.26
FC1 + Cross-entropy-regularized RBM [17]	4,096	82.20	74.20
FC1 + T-CNNW [33]	4,096	81.11	70.07
FC1 + SVM-SRBM-T3 + SP	4,096	**87.98**	**77.82**

TABLE II
CLASSIFICATION ACCURACY COMPARISON WITH EXISTING METHODS (ON AVA DATASET).

Methods	[3]	[4]	[2]	[5]	[10]	[12]	[34]	[13]	DCNNAesth [14]	DCNNAesth+SP [14]	Ours	Ours+SP
Accuracy(%)	61.49	68.13	68.67	71.06	68.55	71.20	75.41	80.38	78.92	83.52	83.44	**87.98**

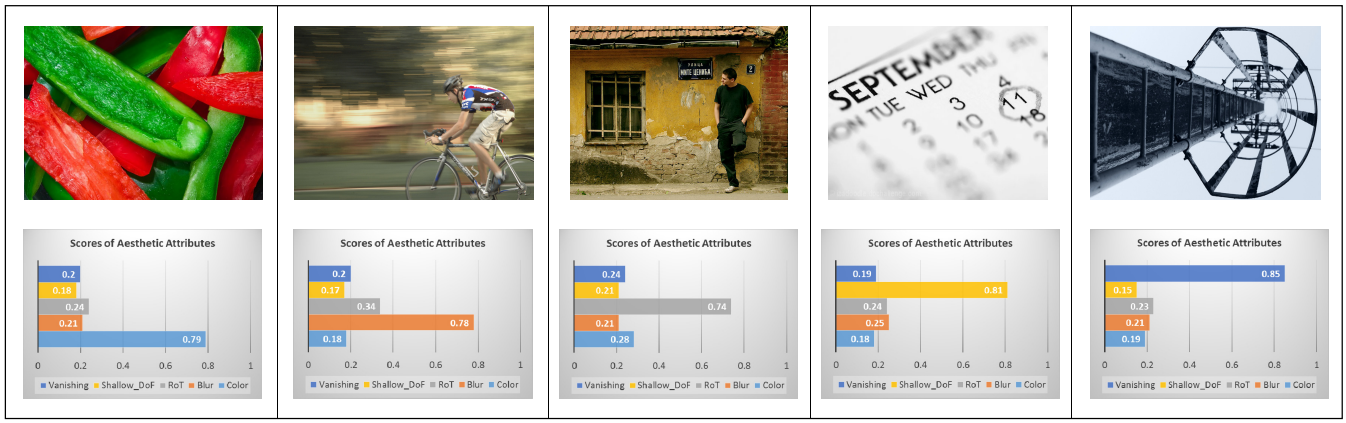


Fig. 5. Normalized scores of the 5-Style aesthetic attributes for example photographs.

attribute makes to the overall photo quality. For this analysis, we selected Top-100 photos with the highest scores for an aesthetic attribute or a combination of aesthetic attributes. Based on the binary quality label (high or low) on AVA dataset, we calculated the ratio of high and low quality photos for each attribute or attribute combination (Fig. 7).

As shown in Fig. 7, *Vanishing Point* is one of the most important aesthetic attributes, and the combination of *Motion Blur* and *Shallow DOF* gives the least pleasing results. *Rule of Thirds*, as expected, is an important attribute for high quality photos, but the test reveals that combining it with *Shallow DOF* improves the photo quality even more. While *Motion Blur* alone does not seem to contribute much for photo quality, Fig. 7 shows that when it is combined with other properties, such as *Vanishing Point*, it could make a positive impact. While *Shallow DOF* is generally considered a desirable feature of a good photograph, *Shallow DOF* in AVA dataset often occurs with other attributes (e.g., *Motion Blur*), and as a result, pictures with high *Shallow DOF* could be classified either way (good or bad quality) due to the influences of other attributes.

These results suggest a variety of possible photo editing strategies to make a given photograph more attractive. One obvious strategy would be to edit the given picture to incorporate or strengthen an aesthetic attribute that would improve the aesthetic quality (Fig. 8). As an experiment, we selected three attributes that our analysis reported to generate high quality photos, *Vanishing Point*, *Rule of Third*, and *Rule of Thirds + Shallow DOF*. For a given photograph, we applied photo editing to change each of these attributes, and compared the quality scores of two versions with *weak* and *strong* attributes. As for *Vanishing Point*, we generated two photos with weak and strong vanishing structures by applying cropping and scaling to an image. For *Shallow DOF*, we introduced the effect by applying blur to the background region using Adobe PhotoShop. Finally, *Rule of Thirds* property was simply changed by cropping an image. In all cases, the versions that have strong desired attributes marked higher aesthetic photo quality scores.

2) AADB Dataset: Aesthetics and Attributes Database (AADB) [43] has a more balanced distribution of professional

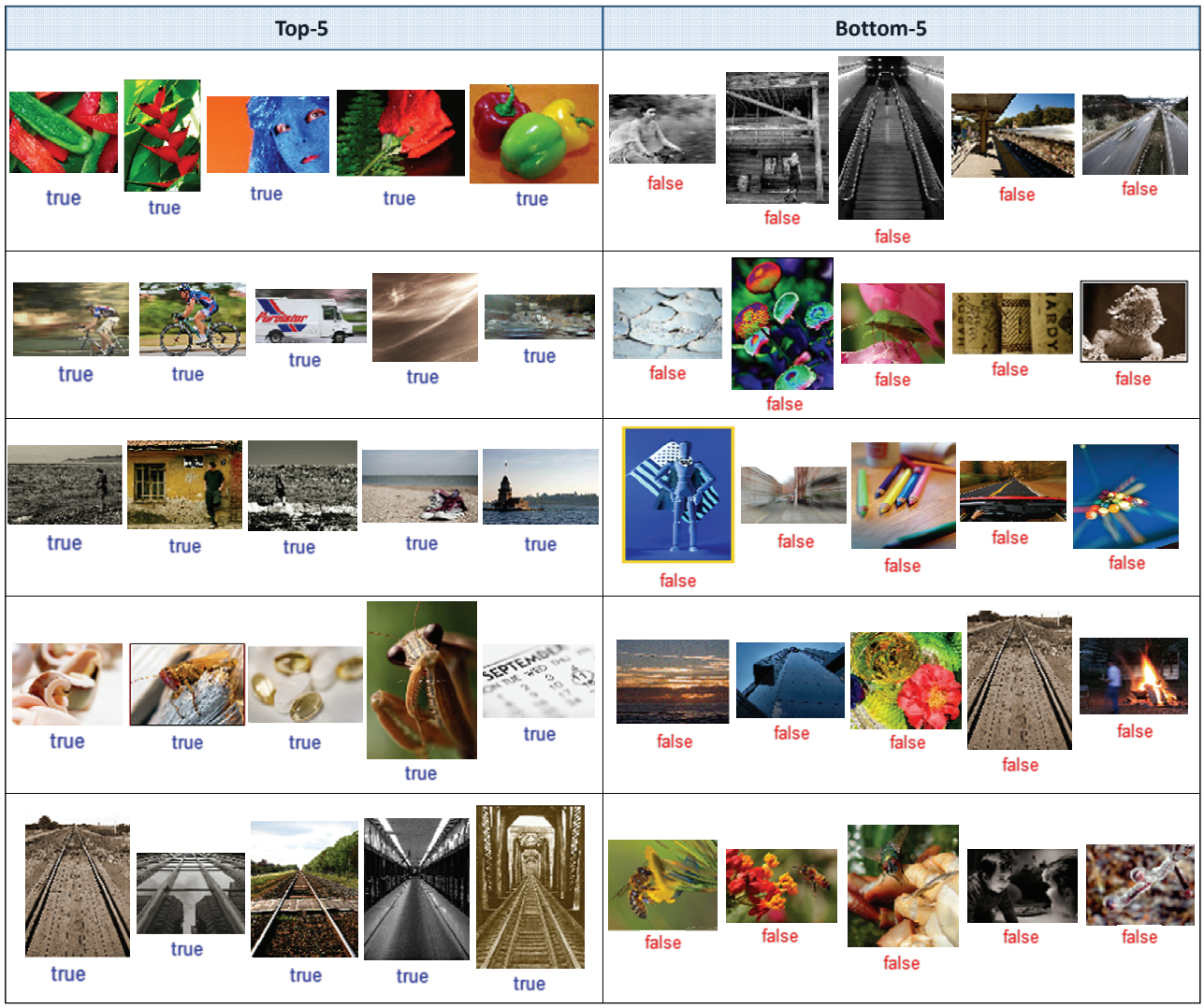


Fig. 6. Top-5 and Bottom-5 ranked images for each aesthetic attribute (on 5-Styles dataset). Each row corresponds to an aesthetic attribute in the order (from the top) of Complementary Colors, Motion Blur, Rule of Thirds, Shallow DOF, and Vanishing Point. Each image is marked "true" if it is labeled in the dataset to have the corresponding aesthetic attribute, or "false" otherwise, showing our aesthetic attribute scores are consistent with the ground truth labels in the dataset.

TABLE III
CLASSIFICATION ACCURACY COMPARISON WITH BASELINE METHODS (ON AADB DATASET).

Methods	Quality, AP(%)	11-Attributes, MAP(%)
Baseline, CNN-S (FC1)	66.90	50.60
Baseline, CNN-S (FC2)	69.27	54.44
CNN-S (FC1) + SVM-SRBM	**77.03**	**61.84**

and consumer photos collected from Flickr, compared to AVA dataset which mostly consists of professional images. AADB dataset contains a total of 10,000 images, that are split into training (8,500), validation (500), and testing (1,000) images. Each image is annotated with quality and eleven attribute scores that are averaged by five raters. The eleven attributes are *interesting content*, *object emphasis*, *good lighting*, *color harmony*, *vivid color*, *shallow depth of field*, *motion blur*, *rule of thirds*, *balancing element*, *repetition*, and *symmetry*. They cover traditional photographic principals of color, lighting, focus and composition.

Since AADB dataset was constructed for rating and ranking images in terms of aesthetics, it did not provide any binary classification labels. For images of each class, we re-assigned binary labels by thresholding scores of the images with a mean score of each class. In Table III, we compared classification accuracy with the baseline methods for photo quality (Quality) and eleven attributes (11-Attributes). The comparison result shows that our method clearly improves the classification accuracy over baselines.

To check the relationship between photo quality and this expanded set of aesthetic attributes, as we did with AVA dataset, we selected Top-30 photos with the highest scores for an aesthetic attribute. Based on the binary quality label (high or low) on AADB dataset, we calculated the ratio of high and low quality photos for each attribute (Fig. 9). In this dataset, *shallow depth of field* is one of the most important aesthetic attributes, and *Motion Blur* gives the least pleasing

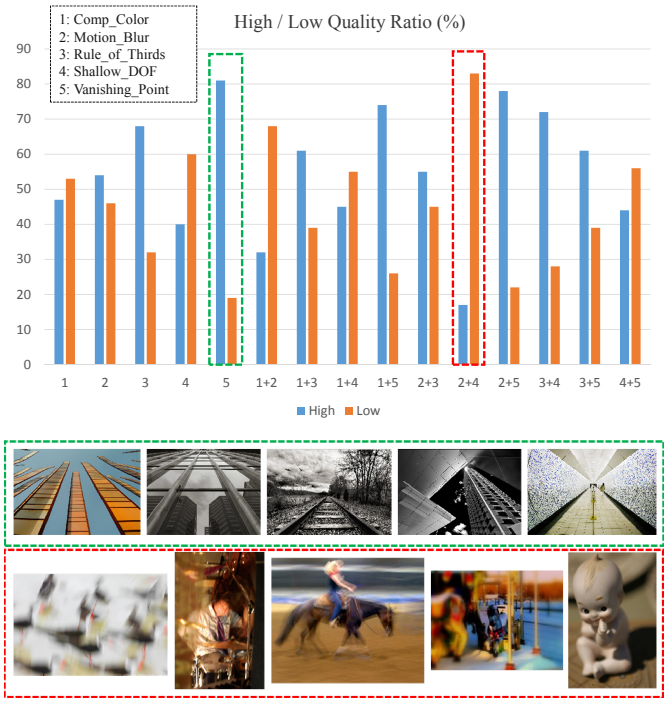


Fig. 7. Ratio of high and low quality photos for each aesthetic attribute or a combination of attributes (on AVA dataset). Also shown below are samples of images that contributed to the statistics in the green/red dotted boxes in the graph.

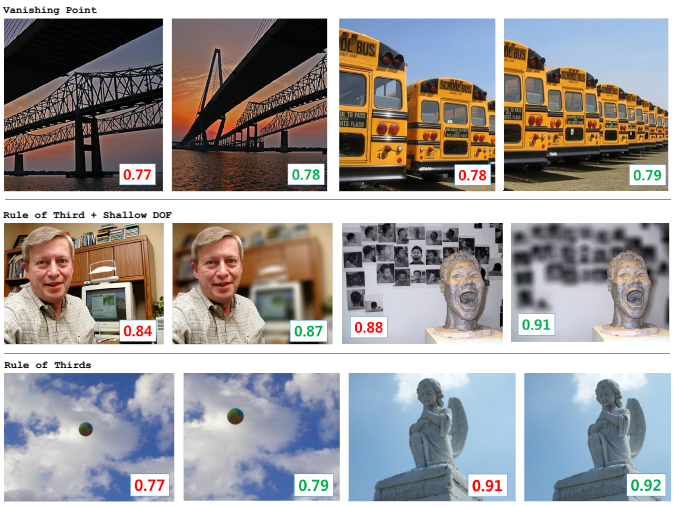


Fig. 8. Example images that are edited in various ways to make them visually more appealing. Green scores indicate better quality after editing. The scores are obtained from our quality assessment scheme.

results.

D. General Object Classification

We applied our SVM-SRBM model to general object classification task, and identified that our method can achieve similar performance gain on this task via enhanced *feature sparsity* and *class discriminability*. We used the a-Pascal dataset for this experiment and compared our classification accuracy with other related methods.

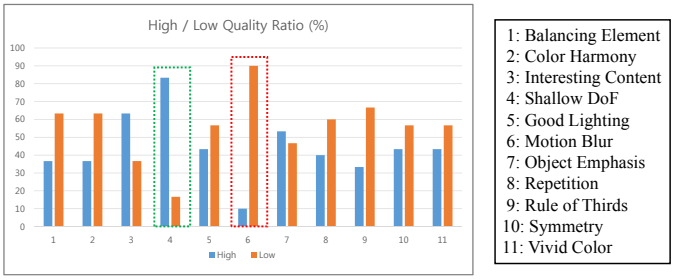


Fig. 9. Ratio of high and low quality photos for each aesthetic attributes (on AADB dataset)

Fig. 10. Failure cases with misclassified labels.

1) a-Pascal Dataset: a-Pascal [39] is datasest created for classification of 20 visual object classes in a variety of natural poses, viewpoints, and orientations. The object classes are people, bird, cat, cow, dog, horse, sheep aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, and tv/monitor. This dataset is split into training and testing sets, each containing over 6400 images.

2) Classification Performance: To evaluate the classification accuracy, we measured both "overall" and "mean per class" accuracies. The "overall" accuracy is the ratio of correctly classified test images over total test images, and the "mean per class" accuracy is the mean of class-wise classification accuracies. In Table IV, we compared the classification accuracy of our method with other related methods. Our baselines are DCNN-features extracted from pre-trained models [19], [20] on the large database (ImageNet), and "CNN-S (FC1) + SVM-SRBM" means DCNN-features extracted from the FC1 layer of the CNN-S model are encoded with our SVM-SRBM. As shown in the table, using DCNN-based features achieved significant performance gains over low-level and DBN-based features, and using features encoded by our SVM-SRBM improved the classification accuracy over the baselines and other related method.

E. Discussion

1) Computation Time: We implemented our system using MATLAB on a PC with Intel Core(TM) i7 4770 @ 3.40 GHz, 32GB RAM and NVIDIA GeForce GTX 650 Ti. We also used CUDA to extract features from the pre-trained DCNN. The computation time for constructing target vectors, SVM parameters, and sparse bases, took about 34.6s, and the training time of our SVM-SRBM model was about 54.4s. For a given 224×224 color test image, it took 90ms to extract features from the pre-trained DCNN, and 8ms for feature encoding with our SVM-SRBM.

2) More Deep Architecture: Our SVM-SRBM model can encode DCNN-features extracted from any deep architectures.

TABLE IV
CLASSIFICATION ACCURACY COMPARISON WITH EXISTING METHODS (ON A-PASCAL DATASET).

Low-level Features	
Methods	**Overall (Mean Per-class) (%)**
Farhadi et al. [39]	59.4 (37.7)
Wang et al. [40]	62.2 (46.2)
DBN-based Features	
Methods	**Overall (Mean Per-class) (%)**
Chung et al. [41]	59.4 (37.7)
Mittelman et al. [42]	63.2 (46.1)
DCNN-based Features	
Methods	**Overall (Mean Per-class) (%)**
AlexNet [19]	75.8 (64.7)
CNN-S [20]	77.9 (67.3)
CNN-S (FC1) + T-CNNW [33]	77.1 (65.5)
CNN-S (FC1) + SVM-SRBM	$\mathbf{82.1(71.2)}$

For an experiment, we applied our SVM-SRBM to more deep architecture, the vgg-verydeep-19 (with 19 layers) [44]. For 5-Style dataset (Section IV-B), the pre-trained vgg-verydeep-19 model reached 71.48% MAP, which is better than CNN-S model. By encoding features extracted from the vgg-verydeep-19 with our SVM-SRBM, we achieved the improved performance with 77.14% MAP, compared to 75.53% using CNN-S model. Similarly, for a-PASCAL dataset (Section IV-D) the pre-trained vgg-verydeep-19 model reached 81.8% overall classification accuracy, which is much better than other deep architectures (with 7 layers). By encoding features extracted from the vgg-verydeep-19 with our SVM-SRBM, we could achieve the best performance on this dataset with 84.10% overall classification accuracy.

3) Using Pre-trained DCNN: Instead of using a pre-trained DCNN, we could have trained a DCNN ourselves. The benefit of using a pre-trained DCNN comes from the fact that the size of the ImageNet dataset (15 million images) for image classification is significantly larger than that of the AVA dataset (250 thousand images) for aesthetics analysis, and thus the features learned from ImageNet would have a stronger representation power. In fact, Lu et al. used AVA dataset for self-training [12] or fine-tuning [34] their DCNNs, while Dong et al. [14] used features directly extracted from a DCNN that has been pre-trained with ImageNet dataset. Even though Dong et al. simply used a traditional DCNN, Table II shows that higher aesthetics analysis accuracy could be achieved than both self-trained and fine-tuned DCNNs of Lu et al. Since it is a huge task to build a dataset as large as ImageNet, we focused our efforts on developing and training a new encoding scheme which generates a set of customized features tailored to a specific photo classification task.

4) Limitations: Fig. 10 shows failure cases of our system. The classification fails when the image does not contain the major features represented by the target vectors for the class that the image actually belongs to. In addition, features extracted from DCNN may have captured some semantic scene elements, such as car, horse, pasture, etc. While these semantic elements could provide useful information for our classification tasks (e.g., car - Motion Blur), they can also mislead when the relevant aesthetic attributes are not strong enough in the image.

V. CONCLUSION

We have presented a novel method to perform both photo quality assessment and analysis of aesthetic attributes in a unified fashion. Using a pre-trained DCNN, our method automatically extracts relevant high-level features that account for the aesthetics of a given photograph. In particular, we developed a novel SVM-SRBM feature encoding method to customize the DCNN-extracted features to our two classification tasks and thereby improve the classification accuracy. We have demonstrated via experimental results that our method outperforms the state-of-the-art techniques in both classification tasks, and could provide valuable guidance in an application of image editing. In addition to aesthetic analysis, we have applied our method to general object classification task, and demonstrated that our method can achieve similar performance gain. Interesting future research direction is visualization of image features modeled by our photo aesthetic analysis system [45].

ACKNOWLEDGMENTS

This work was supported by Institute for Information and Communications Technology Promotion (IITP) Grant (R0126-16-1078) and the National Research Foundation of Korea (NRF) Grant (NRF-2014R1A2A1A11052779) both funded by the Korea government (MSIP).

REFERENCES

[1] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. ACM Int. Conf. Multimedia, Florence, IT*, 2010, pp. 271–280.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. 9th ECCV*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 288–301.

[3] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. 10th ECCV*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 386–399.

[4] K. Lo, K. Liu, and C. Chen, "Assessment of photo aesthetics with efficiency," in *Proc. 21st Int Conf Pattern Recog., Tsukuba, Japan*, 2012, pp. 2186–2189.

[5] 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.*, 2006, pp. 419–426.

[6] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang, "Classification of digital photos taken by photographers or home users," in *Proc. 5th PCM*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, vol. 3331, pp. 198–205.

[7] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proc. 16th IEEE Int. Conf. Image Process.*, Nov 2009, pp. 997–1000.

[8] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[9] Z. Dong and X. Tian, "Effective and efficient photo quality assessment," in *Proc. IEEE Int. Conf. SMC*, Oct 2014, pp. 2859–2864.

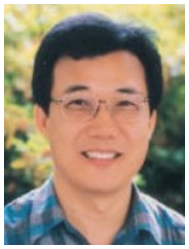
[10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.* IEEE, 2011, pp. 1784–1791.

[11] 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.* IEEE, 2012, pp. 2408–2415.

[12] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, 2014, pp. 457–466.

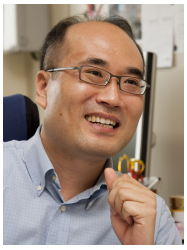
[13] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment." *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.
[14] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *Proc. 21st Int. Conf. MMM, Sydney, NSW, Australia*, 2015, pp. 524–535.
[15] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986.
[16] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *JMLR*, vol. 9, pp. 1871–1874, 2008.
[17] H. Goh, N. Thome, and M. Cord, "Biasing restricted boltzmann machines to manipulate latent selectivity and sparsity," in *Proc. NIPS*, 2010.
[18] T. Aydin, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 31–42, 2015.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS 25*. Curran Associates, Inc., 2012, pp. 1097–1105.
[20] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC, Nottingham, UK*, 2014.
[21] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proc. 10th AISTATS*, 2005, pp. 33–40.
[22] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proc. 25th ICML*. New York, NY, USA: ACM, 2008, pp. 536–543.
[23] R. Mittelman, H. Lee, B. Kuipers, and S. Savarese, "Weakly supervised learning of mid-level features with beta-bernoulli process restricted boltzmann machines." in *Proc. IEEE int. Conf. Comput. Vis. Pattern Recog.* IEEE, 2013, pp. 476–483.
[24] H. Goh, N. Thome, M. Cord, and J.-H. Lim, "Unsupervised and supervised visual codes with restricted boltzmann machines." in *Proc. ECCV*, ser. Lecture Notes in Computer Science, vol. 7576. Springer, 2012, pp. 298–311.
[25] ——, "Top-down regularization of deep belief networks," in *Proc. NIPS 26*. Curran Associates, Inc., 2013, pp. 1878–1886.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE int. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
[27] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[28] H. Yu, J. Yang, W. Wang, and J. Han, "Discovering compact and highly discriminative features or combinations of drug activities using support vector machines," in *Proc. IEEE conf. CSB*. IEEE, 2003, pp. 220–228.
[29] Y. Guo, G. Zhao, and M. Pietikinen, "Discriminative features for texture description." *Pattern Recog.*, vol. 45, no. 10, pp. 3834–3843, 2012.
[30] M. Ranzato, Y. Boureau, and Y. L. Cun, "Sparse feature learning for deep belief networks," in *Proc. NIPS 20*. Curran Associates, Inc., 2008, pp. 1185–1192.
[31] P. O. Hoyer and P. Dayan, "Non-negative matrix factorization with sparseness constraints," *JMLR*, vol. 5, pp. 1457–1469, 2004.
[32] A. Vedaldi and K. Lenc, "Matconvnet - convolutional neural networks for MATLAB," in *Proc. 25th ACM int. conf. Multimedia*, 2015, pp. 689–692.
[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE int. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1717–1724.
[34] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, December 2015.
[35] N. Sawant and N. J. Mitra, "Color harmonization for videos," in *Proc. ICVGIP*, 2008, pp. 576–582.
[36] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition." *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.
[37] 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 Transactions on Multimedia*, vol. 14, no. 3-2, pp. 833–843, 2012.
[38] 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, 2013.
[39] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE int. Conf. Comput. Vis. Pattern Recog.*, 2009.
[40] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proc. 11th ECCV*, ser. ECCV'10. Springer-Verlag, 2010, pp. 155–168.
[41] J. Chung, D. Lee, Y. Seo, and C. D. Yoo, "Deep attribute networks," *CoRR*, vol. abs/1211.2881, 2012.
[42] *Weakly Supervised Learning of Mid-Level Features with Beta-Bernoulli Process Restricted Boltzmann Machines*, 2013.
[43] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. European Conference on Computer Vision (ECCV)*, 2016.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[45] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013.

Hui-Jin Lee received the B.S. degree in Electronic Engineering from Chonbuk National University, S. Korea in 2010, and is currently pursuing the Ph.D degree from the Department of Electrical Engineering, POSTECH of S. Korea. Her current research interests include feature encoding and image classification.



Ki-Sang Hong received the B.S. degree in Electronic Engineering from Seoul National University, S. Korea in 1977, the M.S. degree in Electrical and Electronic Engineering in 1979, and the PhD degree in 1984 from KAIST, S. Korea. From 1984 to 1986, he was a researcher with Korea Atomic Energy Research Institute, and in 1986 he joined POSTECH, Korea, where he is currently a professor with the Division of Electrical and Computer Engineering. From 1988 to 1989 he was a visiting professor with Robotics Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include computer vision, augmented reality, pattern recognition, and color image processing.

Henry Kang is an Associate Professor of Computer Science at the University of Missouri – St. Louis, USA. He received his B.S. degree in computer science from Yonsei University in 1994, his M.S. and Ph.D. in computer science from the Korea Advanced Institute of Science and Technology (KAIST) in 1996 and 2002, respectively. His current research interests include computer graphics and visualization, image/video processing, and computer vision.



Seungyong Lee is a professor of computer science and engineering at the Pohang University of Science and Technology (POSTECH), Korea. He received the BS degree in computer science and statistics from Seoul National University in 1988 and the MS and PhD degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST) in 1990 and 1995, respectively. From 1995 to 1996, he worked at the City College of New York as a postdoctoral researcher. Since 1996, he has been a faculty member of POSTECH, where he leads the Computer Graphics Group. During his sabbatical years, he worked at MPI Informatik (2003-2004) and the Creative Technologies Lab at Adobe Systems (2010-2011). His technologies on image deblurring and photo upright adjustment have been transferred to Adobe Creative Cloud and Adobe Photoshop Lightroom. His current research interests include image and video processing, deep learning, and 3D scene reconstruction.