Image Aesthetic Assessment Based on Pairwise Comparison – A Unified Approach to Score Regression, Binary Classification, and Personalization

Jun-Tae Lee[†] and Chang-Su Kim[‡]
School of Electrical Engineering, Korea University, Seoul, Korea
jtlee@mcl.korea.ac.kr[†] and changsukim@korea.ac.kr[‡]

Abstract

We propose a unified approach to three tasks of aesthetic score regression, binary aesthetic classification, and personalized aesthetics. First, we develop a comparator to estimate the ratio of aesthetic scores for two images. Then, we construct a pairwise comparison matrix for multiple reference images and an input image, and predict the aesthetic score of the input via the eigenvalue decomposition of the matrix. By varying the reference images, the proposed algorithm can be used for binary aesthetic classification and personalized aesthetics, as well as generic score regression. Experimental results demonstrate that the proposed unified algorithm provides the state-of-the-art performances in all three tasks of image aesthetics.

1. Introduction

As the volume of visual data grows exponentially, the capability of automatically distinguishing high quality images from low quality ones or judging aesthetic values of images becomes increasingly important in image searching, retrieving, and enhancing applications. However, it is challenging due to the subjectiveness and ambiguity of aesthetic criteria. For example, to take high quality images, photographers use several aesthetic rules, including rule of thirds and visual balance [22, 23]. Early assessment techniques [6, 27, 28, 40] adopted various handcrafted features to describe these rules. The rule-based features, however, are not sufficiently effective, and some aesthetic rules might have not been discovered yet. Other approaches leveraged generic image features, such as Fisher vectors [31, 33] and bag-of-visual-words [38], yielding more promising results.

Recently, with the great success of convolutional neural networks (CNNs) in various vision tasks [10, 15, 16, 19, 36], many CNN-based aesthetic assessment techniques have been developed [18, 25, 26, 29, 30, 39]. As human beings evaluate aesthetics based on their experience, these CNN-based techniques learn aesthetic criteria from massive data.

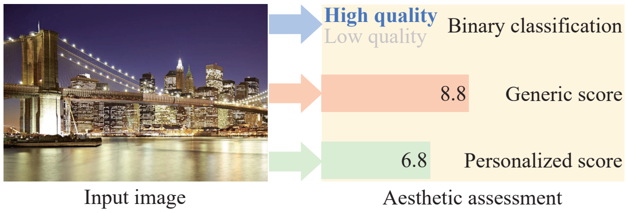


Figure 1. Given an image, the proposed unified algorithm can classify it into either high or low quality class, regress a generic score, and tailor the score to reflect personal preferences.

Although these techniques have made progress in aesthetic assessment, most of them focus on dichotomizing an image into either high or low quality class. However, in some applications, such as image recommendation, image enhancement [20], and personal album curation, it is necessary to estimate a continuous aesthetic score of an image and also tailor the score to meet personal preferences. Relatively little effort has been made for these aesthetic score regression [18] and personalized aesthetics [34], which are more challenging than binary aesthetic classification.

In this paper, we propose a unified approach to the three tasks of aesthetic score regression, binary aesthetic classification, and personalized aesthetics. We first develop an aesthetic comparator, which is a Siamese network, to estimate the ratio of aesthetic scores for two images. Using the comparator, we generate a pairwise comparison matrix for multiple reference images and an input image. Then, via the eigenvalue decomposition of the matrix, we obtain a regressed score of the input image. By modifying the pairwise comparison matrix, the proposed algorithm can achieve all three objectives of score regression, binary classification, and personalization successfully, as illustrated in Figure 1. Experimental results demonstrate that the proposed unified algorithm outperforms the state-of-the-art score regression [18], binary classification [29], and personalization [34] techniques.

To summarize, we make the following contributions:

- We propose the first unified approach to the three tasks of image aesthetic assessment.

- The proposed unified algorithm outperforms the state-of-the-art aesthetic ranker [17], generic score regressor [18], and personalized score regressor [34].

- Especially, the proposed unified algorithm yields a 9.0% higher accuracy that the state-of-the-art algorithm [29] in binary aesthetic classification, which is the most extensively studied task.

2. Related Work

2.1. CNN-Based Aesthetic Assessment

Image aesthetic assessment can be roughly divided into two problems: binary classification and score regression.

Binary classification: It attempts to dichotomize the quality of an image into either high or low class. This binary aesthetic classification has been extensively studied, and there are many CNN-based methods, including [25, 26, 29, 30]. Some methods improve the classification performance by combining global and local information [25, 26, 29]. Lu *et al*. [25] extract aesthetic features using two CNNs which accept an entire image and a randomly cropped patch, respectively. The single patch input, however, may not represent local information faithfully. Moreover, the local CNN does not consider the holistic layout of an image. Thus, Lu *et al.* [26] feed a set of randomly cropped patches into a CNN and aggregate those features. Instead of randomly selecting patches, Ma *et al.* [29] extract more informative patches using an object detector [42] and low-level information, such as saliency and texture. However, as long as an image is divided into small patches, the aesthetics of the global view is not preserved. Also, Mai *et al.* [30] take a whole image as the input to multiple CNNs, from the last layers of which multi-scale local features are extracted. But, near the last layers, most local details are lost, making it difficult to perform local analysis effectively.

Score regression: Compared with binary classification, relatively little effort has been made for aesthetic score regression. This is partly because aesthetic regression is technically more challenging than aesthetic classification. However, score regression is also important in applications. Suppose that a retrieval system should retrieve the top 10% images in terms of aesthetic qualities from a database. In this case, a binary classification algorithm would be of little use. In contrast, with a regression algorithm, it is straightforward to sort the images according to their aesthetic scores.

Kong *et al*. [18] proposed a CNN to regress the aesthetic score of an image. To train the CNN, they employed a Siamese network with a pairwise ranking loss. They also developed additional networks to extract attribute and content information. Ko *et al*. [17] also proposed a Siamese network, which compares two images and determines the aesthetically better one. However, their algorithm is not a score regressor but a ranker: it does not provide the score of an image and can only rank n images by performing $\binom{n}{2}$ comparisons. Recently, Talebi and Milanfar [39] attempted to estimate the distribution of aesthetic scores for an image to address the subjective nature of aesthetics.

Since aesthetic assessment is inherently a subjective process, it is important to adapt an assessment algorithm to personal preferences. This personalization is challenging, as noted in [5]. Ren *et al*. [34] proposed a regression method which predicts the personalized aesthetic score of an image by adding a user-specific offset to the generic score.

2.2. Pairwise Comparison

It is a fundamental problem to estimate the priorities (or ranks) of multiple entities through pairwise comparison of those entities [1,17,35,37]. For example, in a sports league, teams compete against each other in a pairwise manner, and their ranks are determined according to their numbers of wins. In the classic paper [35], Saaty proposed the scaling method, which can reconstruct absolute priorities up to a scale using only pairwise relative priorities.

In information retrieval, pairwise comparison of training data can be performed to learn a rank function, which measures the relevance of a data item to a query. For instance, Herbrich *et al*. [13] developed an ordinal regression function, called Ranking SVM, to minimize pairwise rank inversion cases. Burges *et al*. [3] proposed RankNet to dichotomize the ordinal relation of a pair of relevance scores into binary classes.

Pairwise comparison is widely used in computer vision as well. Wang *et al*. [41] trained a network for person re-identification, which outputs a high similarity level if two images contain an identical person. Chen *et al.* [4] trained a monocular depth estimation algorithm, by employing different loss functions depending on the ordinal relation between a pair of pixel depths. Recently, Lee and Kim [21] reconstructed relative depths for all pairs of pixels in an image and used them to achieve the state-of-the-art monocular depth estimation performance. Furthermore, pairwise comparison is useful to learn metrics for quantifying perceptual concepts, such as image interestingness [9] and urban appearance [8]. Due to the ambiguity and subjectivity of those concepts, the annotation on individual images is unreliable. Instead, the pairwise comparison (*e.g.* for determining the more interesting one between two images) is relatively easy. For the image interestingness, Fu *et al*. [9] trained a linear regression function by minimizing pairwise errors of regressed interestingness. For the urban appearance, Dubey *et al*. [8] trained a Siamese network by classifying the ordinal relation of two images and regressing their rank difference.

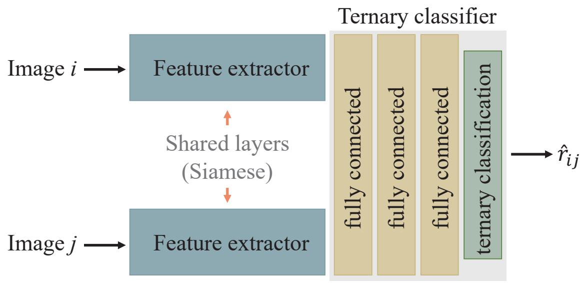


Figure 2. The aesthetic comparator: Given two images, their features are obtained by the coupled extractors, concatenated, propagated to three fully connected layers, and then categorized into one of three classes. Then, the quantized score ratio $\hat{r}_{ij}$ is output.

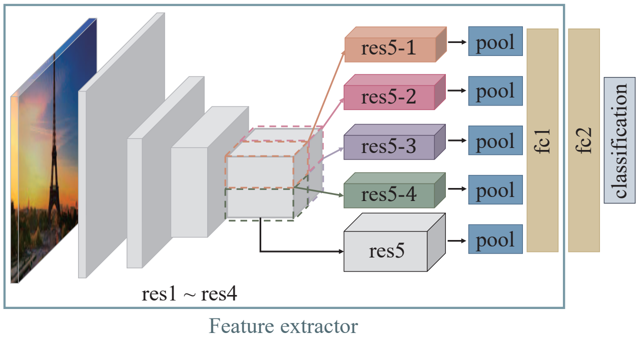


Figure 3. The baseline network contains residual blocks (res1 $\sim$ res4, res5, res5-1 $\sim$ res5-4), pooling layers, fully connected layers (fc1 and fc2), and a classification layer. It is used as the twin feature extractor in Figure 2, after being truncated before fc2.

3. Proposed Algorithm

We propose a unified algorithm to solve the three problems of image aesthetic assessment: score regression, binary classification, and personalized aesthetics. Using an aesthetic comparator, the proposed algorithm forms a pairwise comparison matrix for multiple reference images and an input image. By decomposing the matrix, the proposed algorithm estimates the aesthetic score of the input image. Let us first describe the aesthetic comparator and then explain how to solve each of the three problems by constructing the pairwise comparison matrix differently.

3.1. Aesthetic Comparator

The aesthetic comparator estimates the ratio of aesthetic scores for two images. It is a Siamese network in Figure 2, composed of twin feature extractors and a ternary classifier.

Feature extractors: Let us first describe the baseline network, the truncated version of which is used for the feature extraction. As shown in Figure 3, the baseline network itself is a binary aesthetic classifier to categorize an image into either high or low quality class.

We implement the baseline network using the first five residual blocks (res1 $\sim$ res5) of ResNet-50 [12]. The last block (res5) describes global features of an image holistically, while taking less account of local characteristics of smaller regions. In aesthetic assessment, local features are as important as global ones. To extract local features, the conventional techniques [25, 26] use locally cropped patches as input to their networks. However, when processing visual information, brains handle local views in deeper steps, by analyzing already deeply processed information from the previous processing [11]. Hence, we extract local aesthetic features from a deep layer. Specifically, we add four local residual blocks res5-k, $1 \leq k \leq 4$, in parallel with res5. In Figure 3, each res5-k analyzes a quadrant of the output of res4. To aggregate both global and local features, the output responses of res5 and res5-1, . . . , res5-4 are average-pooled and concatenated. Subsequently, we use two fully connected layers. Finally, the classification layer yields a softmax probability vector for the two classes. To train the network, we use the cross-entropy loss.

We truncate the baseline network before the fc2 layer and use it to initialize the twin feature extractors in the Siamese network in Figure 2.

Ternary classifier: It is difficult (even for a human being) to estimate a continuous ratio between aesthetic scores of two images. We hence quantize the ratio into one of three classes: the first image is aesthetically ‘superior,’ ‘similar,’ or ‘inferior’ to the second one. In other words, we design the ternary classifier in Figure 2, which takes two feature vectors and yields one of the three class labels. The classifier consists of fully connected layers and a softmax layer. Finer quantization, such as 5-ary or 7-ary classifier, is also possible, but the ternary classifier is the most effective for the proposed aesthetic assessment, as will be verified in Section 4 and the supplemental document.

To obtain ground-truth classes, we quantize the aesthetic score ratios of pairs of images in a training dataset. Let s_i and s_j denote the ground-truth aesthetic scores of images i and j, respectively. Also, let the score ratio be $r_{ij} = \frac{s_i}{s_j}$. Note that the distribution of score ratios is reciprocally symmetric with respect to 1. In other words, for each score ratio r_{ij}, its reciprocal $r_{ij}^{-1} = \frac{s_j}{s_i}$ is also a score ratio. Therefore, we quantize a continuous ratio r_{ij} into

$$\hat{r}_{ij} = \begin{cases} \gamma & \text{if } \theta \leq r_{ij}, & (i \text{ is superior to } j) \\ 1 & \text{if } \theta^{-1} \leq r_{ij} < \theta, & (i \text{ is similar to } j) \\ \gamma^{-1} & \text{if } r_{ij} < \theta^{-1}, & (i \text{ is inferior to } j) \end{cases} \quad (1)$$

where $\gamma > 1$ is the reconstruction level for the superior case, and θ is the decision level.

We determine these levels γ and θ, by modifying the Lloyd algorithm [24] to satisfy the reciprocal constraints in

(1). We first compute the reconstruction level by

$$\gamma = \frac{\int_{\theta}^{\infty} rp(r)dr}{\int_{\theta}^{\infty} p(r)dr} \tag{2}$$

where $p(r)$ is the probability distribution of score ratios in a training dataset. Second, θ is set to be the midpoint $\frac{1+\gamma}{2}$ to satisfy the nearest neighbor criterion. These two steps are iterated until the convergence.

The entire aesthetic comparator is trained in an end-to-end manner. In other words, the twin feature extractors are fine-tuned and the ternary classifier is trained from scratch. We train the aesthetic comparator with the cross-entropy loss, given by $L_{\mathrm{c}}(\mathbf{p}, \bar{\mathbf{p}}) = -\sum_{k=0}^{2} \bar{p}_k \log p_k$, where $\mathbf{p} = (p_0, p_1, p_2)$ represents the estimated probabilities that an image pair belongs to the three classes and $\bar{\mathbf{p}} = (\bar{p}_0, \bar{p}_1, \bar{p}_2)$ is the ground-truth.

3.2. Aesthetic Score Regression

The comparator analyzes two images comparatively to yield their score ratio. In this section, by extending the Saaty's scaling method for priorities [35], we propose an aesthetic score regressor that processes pairwise comparison results among multiple reference images and an input image to predict the aesthetic score of the input. Then, we describe how to select the reference images and extract their features in advance to perform the regression efficiently.

Score regression: To predict the score of an image, we use R reference images in a training dataset, whose scores are known. Using the known scores, we first construct the pairwise comparison matrix $\mathbf{A}_{\mathrm{ref}}$ of size $R \times R$ for the reference images,

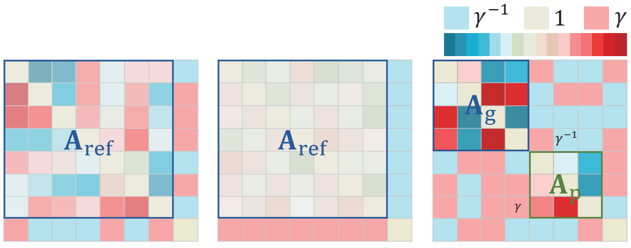
$$\mathbf{A}_{\mathrm{ref}} = \begin{bmatrix} a_1/a_1 & a_1/a_2 & \cdots & a_1/a_R \\ a_2/a_1 & a_2/a_2 & \cdots & a_2/a_R \\ \vdots & \vdots & \vdots & \vdots \\ a_R/a_1 & a_R/a_2 & \cdots & a_R/a_R \end{bmatrix} \tag{3}$$

where a_i denotes the aesthetic score of ith reference image. Thus, each element $a_{ij} \triangleq a_i/a_j$ in $\mathbf{A}_{\mathrm{ref}}$ is an aesthetic score ratio. $\mathbf{A}_{\mathrm{ref}}$ is a reciprocal matrix, since $a_{ij} = \frac{1}{a_{ji}}$.

Using the aesthetic comparator in Section 3.1, we estimate the quantized score ratios between reference and input images. Let $\mathbf{b} = [b_1, b_2, \ldots, b_R]^{\mathrm{T}}$ be the resultant vector, where $b_i \in \{\gamma^{-1}, 1, \gamma\}$ is the score ratio between the ith reference image and the input image. Then, we form the pairwise comparison matrix $\mathbf{A}$ for the reference and input images, given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\mathrm{ref}} & \mathbf{b} \\ \check{\mathbf{b}}^{\mathrm{T}} & 1 \end{bmatrix} \tag{4}$$

where $\check{\mathbf{b}} = [b_1^{-1}, b_2^{-1}, \ldots, b_R^{-1}]^{\mathrm{T}}$ denotes the element-wise inverse of $\mathbf{b}$. Figure 4(a) is an example of the pairwise comparison matrix $\mathbf{A}$ for the score regression.


γ^{-1}
1
γ
$\mathbf{A}_{\mathrm{ref}}$
$\mathbf{A}_{\mathrm{ref}}$
$\mathbf{A}_{\mathrm{g}}$
γ^{-1}
$\mathbf{A}_{\mathrm{p}}$
γ

(a) (b) (c)

Figure 4. Examples of pairwise comparison matrices for (a) generic score regression, (b) binary classification, and (c) personalized score regression. The ratios within green or blue boxes are computed using known scores.

Note that $\mathbf{A}$ is also a reciprocal matrix, and its all elements are positive. Therefore, the priority vector $\mathbf{u}$ of aesthetic scores of the reference and input images can be obtained by solving the eigenvalue problem [35],

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{u}, \tag{5}$$

where λ denotes an eigenvalue. In the ideal case that the aesthetic score ratios in $\mathbf{A}$ are error-free and consistent, this is a trivial problem since $\mathrm{rank}(\mathbf{A}) = 1$. In such a case, the only non-zero eigenvalue is $\lambda_{\max} = R + 1$, and the corresponding eigenvector is equal to any column in $\mathbf{A}$. However, in practice, the score ratios in $\mathbf{b}$ may contain classification and quantization errors. As a result, the score ratios in $\mathbf{A}$ may be inconsistent. Even in this noisy case, all score ratios in $\mathbf{A}$ are positive. Therefore, by the Perron-Frobenius theorem [14], the eigenvalue decomposition of $\mathbf{A}$ yields a positive maximum eigenvalue $\lambda_{\max}$, whose modulus exceeds all the other eigenvalues. The corresponding eigenvector (principal eigenvector) has nonnegative entries. It can be used as a scaled aesthetic score vector, since it is the column vector for the best rank-1 approximation of $\mathbf{A}$ in terms of the Frobenius norm [2].

Let $\mathbf{u} = [\mathbf{u}_{\mathrm{ref}}^{\mathrm{T}}, u]^{\mathrm{T}}$ denote the principal eigenvector, where $\mathbf{u}_{\mathrm{ref}}$ is the priority vector for the R reference images and u is the priority of the input image. Then, we obtain the score vector $\mathbf{s} = [\mathbf{s}_{\mathrm{ref}}^{\mathrm{T}}, s]^{\mathrm{T}}$ by scaling $\mathbf{u}$,

$$\mathbf{s} = \kappa \mathbf{u} \tag{6}$$

where κ is a scale factor. Note that the ground-truth scores of the reference images are available. Let $\bar{\mathbf{s}}_{\mathrm{ref}}$ be the ground-truth vector. The optimal coefficient κ^* is determined to minimize the squared error $\|\bar{\mathbf{s}}_{\mathrm{ref}} - \mathbf{s}_{\mathrm{ref}}\|^2 = \|\bar{\mathbf{s}}_{\mathrm{ref}} - \kappa \mathbf{u}_{\mathrm{ref}}\|^2$, which is given by

$$\kappa^* = \frac{\mathbf{u}_{\mathrm{ref}}^{\mathrm{T}} \bar{\mathbf{s}}_{\mathrm{ref}}}{\mathbf{u}_{\mathrm{ref}}^{\mathrm{T}} \mathbf{u}_{\mathrm{ref}}}. \tag{7}$$

Last, we compute the aesthetic score of the input image by

$$s = \kappa^* u. \tag{8}$$

Table 1. Testing times per image for the three assessment tasks.

Task	R	Testing time (sec)
Score regression	110	1.4×10^{-2}
Binary classification	30	7.3×10^{-3}
Personalized aesthetics	110	1.4×10^{-2}
Personalized aesthetics	200	2.9×10^{-2}

Reference image selection: For the score regression, we use R reference images to compose the pairwise comparison matrix $\mathbf{A}_{\text{ref}}$ in (3). The performance of the proposed score regression method depends on the variety of reference images, as well as on the accuracy of pairwise comparison between reference and test images. Hence, we select reliable reference images as follows. First, we select R_{init} reference images from the training images, where $R_{\text{init}} = 200$. We attempt to make the scores of the reference images uniformly distributed, by dividing the entire score range into 10 equal partitions and randomly sampling $0.1R_{\text{init}}$ training images from each partition. Next, for each reference image, we compare it with the validation images using the aesthetic comparator, and measure the accuracy of the pairwise comparison. We use it as the reliability of the reference image. Then, at each step, we remove the five most unreliable images. As a result, for example, $R = 110$ reference images are selected for the AVA dataset [32].

Testing time: In testing, the proposed algorithm compares an input image with each of R reference images using the aesthetic comparator. For efficient computing, we extract the CNN features of those reference images in advance. In other words, during the test, the feature extraction of the reference images is not necessary. Thus, when $R = 110$, the score regression of an image takes 0.6×10^{-2} sec for the input feature extraction, 0.3×10^{-2} sec for the shallow ternary classifier, and 5.4×10^{-3} sec for the eigenvalue decomposition using a PC with a GTX 1080 ti GPU. Therefore, as listed in Table 1, the proposed algorithm takes merely 1.4×10^{-2} sec in total to regress the score of an image.

3.3. Binary Aesthetic Classification

In binary classification, an image is declared as high quality if its aesthetic score is higher than the median level (*e.g.* 5 in the AVA dataset), and low quality otherwise.

Therefore, for binary classification, we compare an input image to reference images with middle scores. More specifically, we construct the set of reference images, by selecting the training images whose scores are closest to the median level. This is more desirable than using the reference images with a uniform score distribution. Then, as in (4), we form the pairwise comparison matrix $\mathbf{A}$, but all elements in the sub-matrix $\mathbf{A}_{\text{ref}}$ are close to 1, as illustrated in Figure 4(b). The remaining steps are identical to the score regression. If the resultant score s is higher than the median level s_{med}, the image is declared to be of high quality. Otherwise, it is of low quality.

3.4. Personalized Image Aesthetics

Aesthetic assessment is a subjective process. Although people may have a collective consensus about the aesthetic qualities of images, their preferences differ in general. However, it is not practical to train a personalized aesthetic model from scratch. It takes too much effort for a user to provide a sufficient number of annotated examples. Thus, we propose a personalized aesthetic score regression algorithm, requiring only a few user-annotated images. To this end, the personalized regression algorithm exploits the generic preferences of people, by extending the generic regression algorithm in Section 3.2.

We employ R_{g} generic reference images in a training dataset, whose scores are assessed by hundreds of annotators and then averaged [18], and R_{p} personal reference images, scored by a single user. For practical use, we set $R_{\text{g}} \geq R_{\text{p}}$. Then, similarly to (4), we construct the overall comparison matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\text{g}} & \mathbf{A}_{\text{gp}} & \mathbf{b}_{\text{g}} \\ \check{\mathbf{A}}_{\text{gp}}^{\text{T}} & \mathbf{A}_{\text{p}} & \mathbf{b}_{\text{p}} \\ \check{\mathbf{b}}_{\text{g}}^{\text{T}} & \check{\mathbf{b}}_{\text{p}}^{\text{T}} & 1 \end{bmatrix} \tag{9}$$

where $\mathbf{A}_{\text{g}}$ and $\mathbf{A}_{\text{p}}$ are the comparison matrices for the generic and personal reference images, respectively. $\mathbf{A}_{\text{gp}}$ records the score ratio between each pair of generic and personal reference images. Also, $\mathbf{b}_{\text{g}}$ and $\mathbf{b}_{\text{p}}$, respectively, record the relative scores of the generic and personal reference images with respect to an input image. As illustrated in Figure 4(c), $\mathbf{A}_{\text{gp}}$, $\mathbf{b}_{\text{g}}$, and $\mathbf{b}_{\text{p}}$ are computed by the aesthetic comparator in Section 3.1.

Through the eigenvalue decomposition of $\mathbf{A}$ in (9), we obtain the principal eigenvector $\mathbf{u} = [\mathbf{u}_{\text{g}}^{\text{T}}, \mathbf{u}_{\text{p}}^{\text{T}}, u]^{\text{T}}$, where $\mathbf{u}_{\text{g}}$, $\mathbf{u}_{\text{p}}$, and u represent the aesthetic priorities of the generic reference images, the personal reference images, and the input image, respectively. Then, as in (7) and (8), the input priority u is scaled to the personalized aesthetic score by

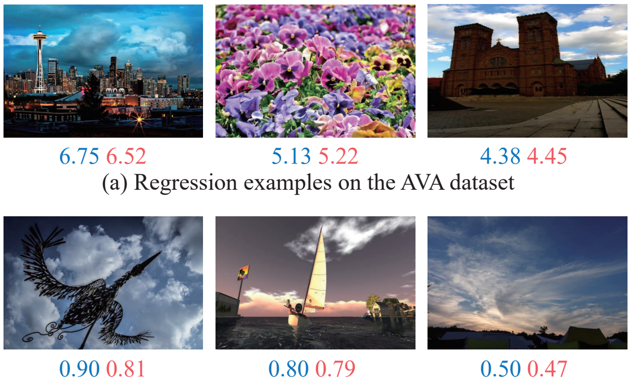
$$s = \frac{\mathbf{u}_{\text{g}}^{\text{T}} \bar{\mathbf{s}}_{\text{g}} + \mathbf{u}_{\text{p}}^{\text{T}} \bar{\mathbf{s}}_{\text{p}}}{\mathbf{u}_{\text{g}}^{\text{T}} \mathbf{u}_{\text{g}} + \mathbf{u}_{\text{p}}^{\text{T}} \mathbf{u}_{\text{p}}} u \tag{10}$$

where $\bar{\mathbf{s}}_{\text{g}}$ and $\bar{\mathbf{s}}_{\text{p}}$ are the ground truth score vectors of the generic and personal reference images, respectively.

4. Experimental Results

4.1. Datasets

We assess the proposed algorithms on three dataset: AVA [32] for binary classification and generic regression, AADB [18] for generic regression, and FLICKER-AES [34] for personalized regression.



(a) Regression examples on the AVA dataset

(b) Regression examples on the AADB dataset

Figure 5. Results of the proposed score regressor: ground-truth and regressed scores are reported in blue and red, respectively.

AVA [32]: AVA is a large publicly available aesthetic assessment dataset, containing about 250,000 images. We use the same partition of training data and testing data as the conventional algorithms [18, 25, 26, 30, 32] do: 235,599 images for training and 19,930 images for testing. Then, as a validation set, we randomly select 2,000 images from the training images. The aesthetic quality of each image was rated by about 200 human annotators. Ratings range from one to ten, with ten indicating the highest quality. The mean rating of an image is set to be its continuous score. An image is labeled as high quality when its mean rating is higher than 5, and low quality otherwise.

AADB [18]: The aesthetics and attribute database (AADB) is for scoring and ranking images in terms of aesthetics. It contains 10,000 images in total, which are split into 8,500 images for training, 500 images for validation, and 1,000 images for testing. Each image was annotated with an aesthetic score and confidence scores for eleven attributes, averaged by five annotators. Aesthetic scores range from 0 to 1 with 1 denoting the highest quality, and confidence scores from -1 to 1, where 1 indicates that the corresponding attribute is manifested to the maximum.

FLICKER-AES [34]: Raw aesthetic scores range from 1 to 5, representing the lowest to the highest aesthetic levels. Each image was rated by about five workers and its ground truth score was set to be the mean of their scores. 210 workers participated in the annotation of FLICKER-AES, which was split into 35,263 images for training and 4,737 images for testing. For personalized applications, the workers of the training images were different from those of the test images. Specifically, the training images were rated by 173 workers, and the test images by the other 37 workers. As for the test images, each worker rated about 137 images.

4.2. Aesthetic Score Regression

We assess the performances of the proposed aesthetic score regressor on the AVA and AADB datasets. As shown in Figure 5, regressed scores are close to the ground-truth scores in most cases.

Table 2. Comparison of the proposed regression algorithm with Reg-Net and PAC-Net on the AVA and AADB datasets. The best results are boldfaced.

	AVA dataset		AADB dataset	
Methods	$\rho(\uparrow)$	MASD($\downarrow$)	$\rho(\uparrow)$	MASD($\downarrow$)
Reg-Net [18]	0.558	0.0582	0.678	0.1268
PAC-Net [17]	0.871	-	0.837	-
Proposed	**0.918**	**0.0229**	**0.879**	**0.1141**

To quantify the score regression performance, we adopt the Spearman's coefficient [7, 18] and the mean of absolute score differences (MASD). The Spearman's coefficient is the correlation coefficient between the ranks, obtained from ground-truth scores and regressed scores, respectively. More specifically, the Spearman's coefficient ρ is given by

$$\rho = 1 - \frac{6\sum_i (r_i - \hat{r}_i)^2}{N^3 - N} \tag{11}$$

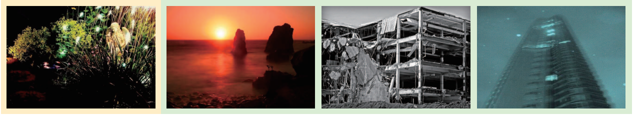
where N is the number of test images, and r_i and $\hat{r}_i$ are the ground-truth and predicted ranks of the ith test image. The Spearman's coefficient measures the degree of the monotonic relationship between two rank vectors. Hence, it does not assess the quality of a regressed score directly. MASD measures the differences between ground-truth and regressed scores directly and averages them, which is defined as

$$\text{MASD} = \frac{1}{N}\sum_i |s_i - \hat{s}_i| \tag{12}$$

where s_i and $\hat{s}_i$ are the ground-truth and regressed scores of the ith image, normalized to the range $[0, 1]$.

Comparative evaluation: We compare the proposed regression algorithm with the conventional regression [18] and ranking [17] algorithms. Similarly to the proposed algorithm, given an image, the regression network Reg-Net [18] yields its score. In contrast, the ranking algorithm PAC-Net [17] does not provide a score. Note that it is straightforward to obtain the ranks of N images from their scores. Any sorting algorithm can be used. On the contrary, it is hard to estimate the aesthetic scores of N images, annotated by humans, from the ranks only.

Table 2 compares the results. The Spearman's coefficients of the conventional algorithms are from the respective papers [17, 18], and the MASDs of Reg-Net are computed using their source codes. As mentioned above, the ranking algorithm PAC-Net does not yield scores, so its MASD cannot be measured. We see that the proposed algorithm performs better than Reg-Net and PAC-Net on both datasets. In terms of ρ, the proposed algorithm outperforms PAC-Net by 0.047 and 0.042 on the AVA and AADB datasets, respectively. Also, for MASD, the proposed algorithm out-



(a) Reference (b) Superior (c) Similar (d) Inferior

Figure 6. Ternary classification results of the proposed aesthetic comparator on the AVA dataset: images in (b), (c) and (d) are declared to be superior, similar, and inferior to the reference image in (a), respectively. The ground-truth scores of (a)∼(d) are 5.02, 6.37, 5.05, and 3.13.

Table 3. The overall aesthetic score regression performances, when different classifiers are used in the aesthetic comparator.

	AVA dataset		AADB dataset	
Comparator	$\rho(\uparrow)$	MASD($\downarrow$)	$\rho(\uparrow)$	MASD($\downarrow$)
3-ary classifier	0.918	0.0229	0.879	0.1141
5-ary classifier	0.791	0.0555	0.867	0.1713
7-ary classifier	0.779	0.0528	0.867	0.1783

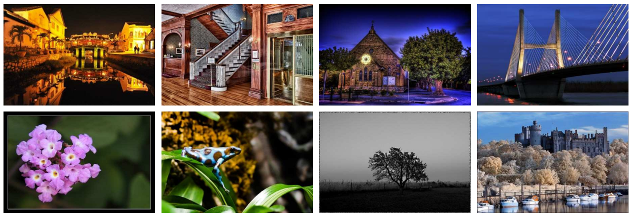
performs Reg-Net by 0.0353 and 0.0127 on the AVA and AADB datasets, respectively.

Although PAC-Net is comparable to the proposed algorithm in the ρ performance, but it is not practical. It requires the pairwise comparison between all possible pairs in the test dataset. On the AVA dataset, the number of such pairs is $\binom{19930}{2} \cong 1.99 \times 10^6$, and it takes about 71 hours for testing. In contrast, the proposed algorithm computes the score of each image and obtains the ranks of all images by sorting the scores. The proposed algorithm takes 1.4×10^{-2} sec for computing each score and thus requires about 5 minutes only for obtaining the rank vector of 19,930 images.

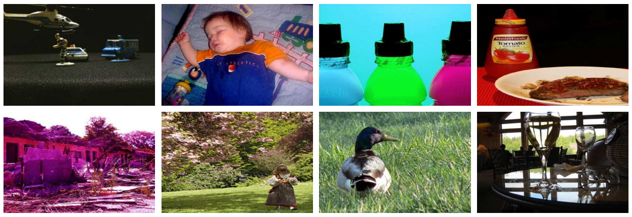
Finer quantization in aesthetic comparator: We analyze the quantization effects of score ratios in the aesthetic comparator. More specifically, we design 5-ary and 7-ary classifiers, as well as the ternary classifier in Figure 2. Table 3 shows the overall aesthetic score regression performances, when these alternative classifiers are employed instead of the ternary-classifier. We see that the proposed ternary classifier provides the best performances in terms of ρ and MASD on both datasets. This is because, although the ternary classifier performs the coarsest quantization, it is the most reliable and yields the highest classification accuracy. Figure 6 shows comparison examples of the ternary classifier.

4.3. Binary Aesthetic Classification

Binary classification is the most extensively researched topic in image aesthetic assessment [17,18,25,26,29,30,39]. We evaluate the proposed binary classification algorithm on the AVA dataset. Figure 7 shows how the proposed algorithm classifies images into the high or low quality classes. It uses the 30 reference images whose scores are the closest to the median score among the training images. This number of reference images, $R = 30$, is sufficient for the binary classification, even though it is smaller than that ($= 110$) for the score regression. Thus, as listed in Table 1, the proposed algorithm takes 7.3×10^{-3} sec only to classify an image.



(a) High quality class



(b) Low quality class

Figure 7. Binary classification results: images in (a) are declared by the proposed algorithm as high quality, and images in (b) as low quality.

Table 4. Comparison of the accuracy scores of binary classification on the AVA dataset. The best result and the second best result are boldfaced and underlined, respectively.

Methods	Accuracy (%)
AVA [32]	67.0
RDCNN [25]	74.4
DMA-Net-ImgFu [26]	75.4
Reg-Net [18]	77.3
MNA-Net-Scene [30]	77.4
PAC-Net [17]	82.2
A-Lamp [29]	<u>82.5</u>
Baseline network	78.7
Proposed	**91.5**

We measure the accuracy score

$$\text{Accuracy} = \frac{N_{\text{c}}}{N} \tag{13}$$

where N_{c} is the number of correctly classified images and N is the number of total test images.

Comparative evaluation: Table 4 compares the proposed binary aesthetic classification algorithm with the recent algorithms in [18, 25, 26, 29, 30, 32] on the AVA dataset. Based on handcrafted and generic features, the AVA algorithm [32] yields the lowest accuracy. The other conventional algorithms are based on CNNs. Most of them exploit external information such as attribute classification [25,26], scene categorization [30], attribute and content classification [18], and salient object detection [29], whereas the proposed algorithm uses *no* such information.

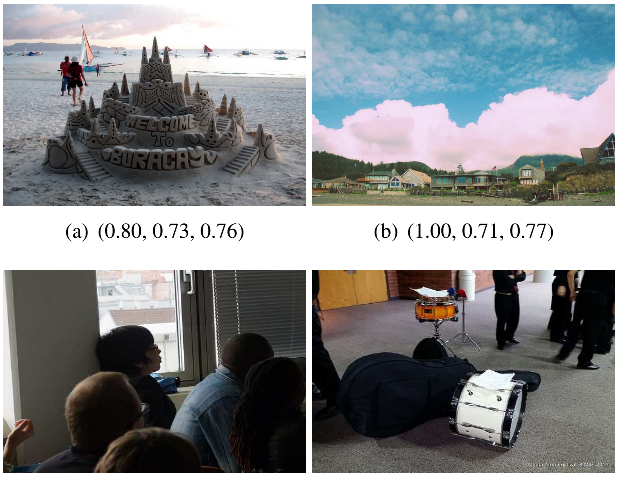

(a) (0.80, 0.73, 0.76)
(b) (1.00, 0.71, 0.77)
(c) (0.40, 0.42, 0.40)
(d) (0.40, 0.49, 0.45)

Figure 8. Examples of the personalized score regression for a test worker. For each image, (the worker's annotated score, regressed generic score, regressed personalized score) are reported, where all scores are normalized to [0, 1].

In Table 4, we also include the performance of the proposed baseline network. Even the baseline network yields a comparable accuracy to the conventional CNN-based algorithms. Furthermore, the proposed algorithm based on pairwise comparison improves the performance of the baseline network by 12.8%. Consequently, notice that the proposed algorithm outperforms the previous state-of-the-art method, A-Lamp [29] by a significant gap of 9.0%.

4.4. Personalized Image Aesthetics

Figure 8 shows examples of the proposed personalized score regression. In this test, 100 generic reference images are used to form $\mathbf{A}_\text{g}$, and 10 personal reference images, annotated by a test worker, are employed to construct $\mathbf{A}_\text{p}$ in (9). The personalized regression predicts the worker's aesthetic preferences more accurately than the generic regression. For example, the generic regression determines that Figure 8(a) is aesthetically superior to Figure 8(b). On the contrary, the personalized regression declares that Figure 8(b) is better, coinciding with the worker's preferences.

Comparative evaluation: We evaluate the proposed personalized score regression algorithm on the FLICKER-AES dataset. We randomly select R_g generic reference images from the training set, where $R_\text{g} = 100$. For each test worker, we randomly sample R_p personal reference images scored by the worker. Then, the remaining images, scored by the same worker, are used to evaluate the personalized regression performance. We compare the proposed algorithm with the conventional algorithm, PAM [34], which computes a user-specific offset and adds it to the generic aesthetic score. As done in [34], we test two cases of $R_\text{p} = 10$ and $R_\text{p} = 100$. This is why we set R_g to 100. In other words, we select the smallest R_g under the condition $R_\text{g} \geq R_\text{p}$. In terms of testing time, in Table 1, the proposed algorithm takes 1.4×10^{-2} sec and 2.9×10^{-2} sec per image at $R_p = 10$ and $R_p = 100$, respectively.

Table 5. Comparison of the Spearman's coefficients (ρ) on the FLICKER-AES dataset. Here, $+\alpha$ means that the coefficient is increased by α through the personalization, as compared with the generic regression.

		Personalized	
	Generic	$R_\text{p} = 10$	$R_\text{p} = 100$
PAM [34]	0.514	+0.006	+0.039
Proposed	0.668	+0.040	+0.044

In Table 5, when only generic reference images are used, the proposed algorithm achieves the Spearman's coefficient $\rho = 0.668$. The generic model of PAM yields $\rho = 0.514$. Then, we measure the improvement due to personal reference images. When $R_\text{p} = 10$, the proposed algorithm increases ρ by 0.040 while PAM does by 0.006 only. Note that the increase 0.040 is even bigger than the increase ($= 0.039$) of PAM at $R_\text{p} = 100$. This indicates that the proposed algorithm achieves the personalization more effectively using less personal references. Thus, the proposed algorithm reduces the burden of user annotations for personalization meaningfully.

5. Conclusions

We proposed a unified approach to the three tasks of aesthetic score regression, binary aesthetic classification, and personalized aesthetics. We developed the aesthetic comparator, composed of twin feature extractors and a ternary classifier. Using the aesthetic comparator, we constructed a pairwise comparison matrix for reference and input images. Using the principal eigenvector of the matrix, we regressed the score of the input. It was shown that the proposed algorithm can be used for binary classification and personalization, as well as score regression, by varying the pairwise comparison matrix. The proposed unified algorithm outperforms the state-of-the-art generic score regressor [18], binary aesthetic classifier [29], and personalized score regressor [34]. Especially, for binary classification, the proposed algorithm surpasses the state-of-the-art technique [29] by a notable gap of 9.0%.

Acknowledgement

This work was supported partly by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2018R1A2B3003896), and partly by the Agency for Defense Development (ADD) and Defense Acquisition Program Administration (DAPA) of Korea (UC160016FD).

References

[1] Tammo HA Bijmolt and Michel Wedel. The effects of alternative methods of collecting similarity data for multidimensional scaling. *Int. J. Res. Mark.*, 12(4):363–371, Nov. 1995. 2

[2] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of Data Science. 2015. 4

[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML*, 2005. 2

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 2

[5] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag.*, 34(4):80–106, Jul. 2017. 2

[6] Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011. 1

[7] Persi Diaconis and Ronald L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268, Apr. 1977. 6

[8] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *ECCV*, 2016. 2

[9] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. Interestingness prediction by robust learning to rank. In *ECCV*, 2014. 2

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[13] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *ICANN*, 1999. 2

[14] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge, 2 edition, 2012. 4

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[17] Keunsoo Ko, Jun-Tae Lee, and Chang-Su Kim. PAC-Net: Pairwise aesthetic comparison network for image aesthetic assessment. In *ICIP*, 2018. 2, 6, 7

[18] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 1, 2, 5, 6, 7, 8

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[20] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2-D histograms. *IEEE Trans. Image Process.*, 22:5372–5384, Dec. 2013. 1

[21] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *CVPR*, 2019. 2

[22] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Semantic line detection and its applications. In *ICCV*, 2017. 1

[23] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *J. Vis. Commun. Image Represent.*, 55:91–105, Aug. 2018. 1

[24] Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, Mar. 1982. 3

[25] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In *ACM Multimedia*, 2014. 1, 2, 3, 6, 7

[26] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, 2015. 1, 2, 3, 6, 7

[27] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *ICCV*, 2011. 1

[28] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008. 1

[29] Shuang Ma, Jing Liu, and Wen Chen Chang. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, 2017. 1, 2, 7, 8

[30] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016. 1, 2, 6, 7

[31] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011. 1

[32] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 5, 6, 7

[33] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 1

[34] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *ICCV*, 2017. 1, 2, 5, 6, 8

[35] Thomas L Saaty. A scaling method for priorities in hierarchical structures. *J. Math. Psychol.*, 15(3):234–281, Jun. 1977. 2, 4

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep fisher networks for large-scale image classification. In *NIPS*, 2013. 1

[37] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological Review*, 112(4):881, Oct. 2005. 2

[38] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H. Hsu, and Shao-Yi Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *ACM Multimedia*, 2011. 1

[39] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, Aug. 2018. 1, 2, 7

[40] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Trans. Multimedia*, 15(8):1930–1943, Dec. 2013. 1

[41] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 2

[42] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, 2016. 2