IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2019, accepted August 11, 2019, date of publication August 19, 2019, date of current version August 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936289

Property-Specific Aesthetic Assessment With Unsupervised Aesthetic Property Discovery

JUN-TAE LEE[1], (Student Member, IEEE), CHUL LEE[2], (Member, IEEE), AND CHANG-SU KIM[1], (Senior Member, IEEE)

[1]School of Electrical Engineering, Korea University, Seoul 136-701, South Korea
[2]Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Chang-Su Kim (changsukim@korea.ac.kr)

This work was supported in part by the Cross-Ministry Giga Korea Project Grant funded by the Korean Government (MSIT) (Development of 4D Reconstruction and Dynamic Deformable Action Model-Based Hyper-Realistic Service Technology) under Grant GK18P0200, and in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) under Grant NRF-2018R1A2B3003896.

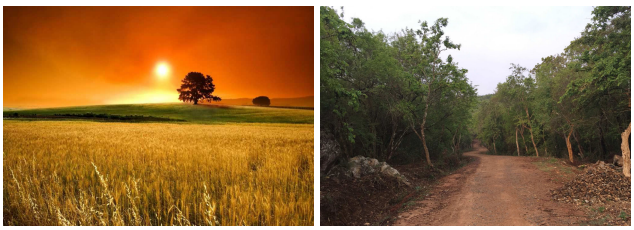
ABSTRACT We propose the property-specific aesthetic assessment (PSAA) algorithm with unsupervised aesthetic property discovery. The proposed PSAA algorithm uses an aesthetic feature extractor, an aesthetic property classifier, and multiple property-specific assessment networks. The aesthetic feature extractor analyzes aesthetics of images to generate features. Using such aesthetic features, we discover diverse aesthetic properties in an unsupervised manner and develop the aesthetic property classifier to predict the aesthetic property of each image. For each discovered aesthetic property, we train a property-specific assessment network. Thus, we can assess the aesthetic quality of an image using the property-specific network that corresponds to its property. Experimental results on a large dataset show that the proposed PSAA algorithm achieves state-of-the-art aesthetic assessment performance. Furthermore, we demonstrate that PSAA is useful for improving aesthetic qualities of images in two applications: contrast enhancement and image cropping.

INDEX TERMS Image aesthetics, aesthetic assessment, image composition, convolutional neural network, unsupervised property discovery, and unsupervised attribute clustering.

I. INTRODUCTION

In art and photography, image aesthetics refers to the principles of beauty conveyed by images. Figure 1 shows examples of aesthetically high- and low-quality images. Whereas the high-quality image is colorful and well-composed, the low-quality one looks pale and ill-composed. The objective of image aesthetic assessment techniques is to computationally distinguish high-quality images from low-quality ones based on aesthetic criteria. The assessment of image aesthetics is important for finding well-taken and appealing photographs. There are lots of potential applications, such as photographic composition [1]–[3], image retrieval [4], [5], and image editing [6], [7]. For example, when retrieving images, image aesthetics can be exploited as one of the ranking factors. Also, image editing systems can produce appealing and polished photographs based on their aesthetic qualities.

The associate editor coordinating the review of this article and approving it for publication was Haimiao Hu.



(a) (b)

FIGURE 1. Examples of images with (a) high-quality and (b) low-quality.

Most aesthetic quality assessment techniques attempt to classify images as either high-quality or low-quality [8]–[14]. However, aesthetic assessment is challenging due to the diversity, subjectiveness, and ambiguity of aesthetic criteria. To describe aesthetic criteria, professional photographers use several rules, *e.g.*, the rule of thirds and visual balance. Thus, early assessment techniques [8]–[10] adopted various handcrafted features, such as the distribution of edges, color histograms, and saliency maps, to describe these

manually-defined aesthetic rules. However, the handcrafted features have limitations for several reasons. First, aesthetic rules were designed based on the experience of professional photographers. Hence, some rules may not have been discovered yet. Second, since aesthetic rules are subjective and ambiguous, the handcrafted features merely approximate these rules. Third, there is no absolute aesthetic criterion. In other words, although a handcrafted feature may represent an aesthetic rule well, it is not applicable to images to which said aesthetic rule is irrelevant. Other approaches tried to leverage generic features, such as Fisher vector [15], [16] and bag-of-visual words [17], which exhibited good performance in image classification tasks. However, these features are also insufficient for assessing aesthetic qualities, since they were designed to describe general image layouts rather than image aesthetics.

Recently, inspired by the success of convolutional neural networks (CNNs) in various computer vision tasks [18]–[22], many CNN-based aesthetic quality assessment techniques have been developed [11]–[14], [23]–[28]. Similarly to how humans evaluate aesthetics based on their experience, CNNs learn aesthetic criteria from massive datasets. Since fine-grained details, as well as holistic image features, are important in aesthetic quality assessment, many CNN-based techniques combine features from both local and global views [11]–[14]. They yield promising results by employing a set of local patches as input [11], [12], [14] or extracting multi-scale features in deep layers [13]. In [26], the layout of an entire image was represented by a graph, and its aesthetic quality was assessed by aggregating aesthetic scores of subgraphs. In [24], [28], the distribution of aesthetic scores for an image was estimated, instead of binary aesthetic classification, to address the inherent subjectivity of aesthetic assessment. To leverage the expert knowledge of photography, CNN features were combined with the hand-crafted or generic features to describe various aesthetic criteria [27]. Furthermore, in many techniques [11]–[14], [23], [25], aesthetic attributes (*e.g.* dynamic range and exposure), scenes, or contents information of images were used to further improve their aesthetic assessment performances. In this work, we refer to this assisting information as *aesthetic properties*.

In assessing the aesthetic quality of an image, aesthetic properties are useful as guidelines for deriving aesthetic criteria. Note that, given an aesthetic property of an image, we can specify aesthetic criteria for the image more easily. For example, suppose that an image is declared to have one of two aesthetic properties, *portrait* or *landscape*. In other words, suppose that an image is annotated either as portrait or landscape. When we assess a portrait image, we focus more on the details in the foreground rather than in the background and on the harmony between foreground and background. In contrast, to assess a landscape image, although it may include people, we consider their details less importantly than in the assessment of a portrait image. Therefore, to exploit such aesthetic property information for quality assessment, previous approaches defined several aesthetic properties manually. However, aesthetic attributes, scenes, and contents are quite diverse and correlated with one another. Hence, it is impractical to define numerous aesthetic properties and their relationship manually. Furthermore, human annotation requires much effort, making it even more difficult. Driven by this issue, an important question arises: *Can we discover diverse aesthetic properties of images without human annotation?*

Motivated by this question, we propose an algorithm called property-specific aesthetic assessment (PSAA) with unsupervised aesthetic property discovery. The proposed PSAA algorithm uses an aesthetic feature extractor and an aesthetic property classifier to perform the property-specific assessment. The feature extractor extracts aesthetic features using multiple deep-layer outputs. Using the aesthetic features, we discover aesthetic properties by employing a CNN-based unsupervised clustering scheme and then train a property-specific network for each discovered property. In the testing phase, for a query image, we determine its aesthetic property and assess its quality using the corresponding property-specific network. Experimental results on a large dataset show that the proposed algorithm outperforms the previous state-of-the-art technique [14].

Main contributions in this work are summarized as follows:

- We propose a novel deep learning-based approach to image aesthetic assessment, which provides the state-of-the-art performance on the largest aesthetic assessment dataset [29].
- We develop the unsupervised aesthetic property discovery scheme to find diverse aesthetic properties effectively, and the aesthetic property classifier to decide the aesthetic properties of query images automatically.
- We design the property-specific assessment network for each discovered property to address the diversity of aesthetic criteria.
- We show that the proposed aesthetic assessment algorithm can be used for two important image processing applications: contrast enhancement and image cropping.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes the proposed PSAA algorithm. Section IV addresses two applications of the proposed algorithm, *i.e.*, aesthetic image enhancement and aesthetic image cropping. Section V discusses experimental results. Finally, Section VI concludes the paper.

II. RELATED WORK

A. AESTHETIC QUALITY ASSESSMENT

Images that are pleasing to the human eyes are considered to have high aesthetic qualities. Aesthetic quality assessment, hence, is a subjective process. However, various computational algorithms have been proposed to quantify the visual quality of an image based on aesthetic criteria, typically in the form of binary classification. Early assessment techniques adopted handcrafted features to describe a few

aesthetic criteria, including the rule of thirds, color vividness, and visual balance [8]–[10]. However, there are more aesthetic criteria, and it is impractical to design a handcrafted feature for each criterion. Generic features, such as the Fisher vector [15], [16] and bag-of-visual words [17], were also employed in aesthetic assessment tasks. Although generic features are competitive with or even outperform simple handcrafted features, they were designed to represent general image layouts rather than aesthetic characteristics. Hence, these rule-based or generic features are not sufficient for aesthetic assessment.

Recently, CNN-based aesthetic assessment algorithms have been developed, yielding more successful results. Especially, in [11], [12], [14], it was shown that a combination of global and local features can improve assessment performance. Lu *et al.* [11] extracted aesthetic features by training two CNNs, which take an entire image and a randomly cropped patch as global and local input data, respectively. However, the single patch may not faithfully represent local information. Furthermore, the CNN that takes the single patch as input does not consider the holistic layout of the entire image. To overcome this limitation, Lu *et al.* [12] fed a set of randomly cropped patches into a CNN and aggregated the resulting features. Instead of randomly selecting patches, Ma *et al.* [14] extracted more informative patches using an object detector and low-level information, such as saliency and texture. On the other hand, Mai *et al.* [13] used a whole image as the input to multiple CNNs, the last layers of which have different receptive fields to extract multi-scale features. Liu *et al.* [25] extracted local features from semantically salient patches and cascaded the local features in order of human gaze shifts among the patches.

These properties are mostly based on *a priori* knowledge, instead of being learned automatically. Lu *et al.* [11] extracted attribute features using a pre-trained attribute classification network, which categorizes images according to several criteria, *e.g.*, dynamic range, exposure, and depth-of-field. Then, they combined those attribute features with aesthetic features to perform attribute-assisted aesthetic quality assessment. In [12], a similar attribute classification network was designed by fine-tuning an image classification network. In [13], a scene classifier was adopted to label images as human, architecture, landscape, and etc. Kong *et al.* [23] proposed a regression network to exploit ten roughly categorized contents, including humans, animals, and flowers, as well as attributes. Lu *et al.* [25] designed an aesthetic feature by encoding the existence of pre-defined aesthetic properties in an image.

However, only a few aesthetic attributes, scenes, and contents were considered in [11]–[13], [23], [25], [27]. Moreover, although diverse attributes, scenes, and content types are useful for aesthetic quality assessment, it is infeasible to manually mine and annotate all those properties. Therefore, we propose the first algorithm to automatically discover diverse aesthetic properties in an unsupervised manner.

B. UNSUPERVISED VISUAL ATTRIBUTE LEARNING

Visual attributes, such as colors, texture, shapes of objects, and human facial expressions, provide useful mid-level cues in vision tasks, including object description [30], face recognition [31], and object recognition [32]. However, since visual attributes are often ambiguous, manually-defined attributes may be neither reliable nor discriminative in the feature space. Recently, attempts have been made to discover attributes from images [30]–[33]. For example, Berg *et al.* [30] constructed visual attribute vocabularies by mining a large dataset of images and descriptive texts. They trained a visual classifier to measure the so-called visualness of an unseen attribute. Ma *et al.* [31] carried out research on unsupervised relative visual attribute learning. Using an image dataset with class labels, they trained attribute ranking functions, each of which computes class ranks according to a visual attribute, and removed redundant visual attributes iteratively. Singh *et al.* [33] developed an iterative procedure that alternates between clustering and training a discriminative support vector machine (SVM) classifier for each cluster. Each SVM was trained to classify images within the cluster against images not included in any clusters, which were used as negative data. Huang *et al.* [32] performed clustering [33] and CNN training alternately to obtain robust visual attribute clusters and achieve the corresponding CNN feature representation. In this work, we attempt to discover visual attributes for image aesthetics (*i.e.* aesthetic properties) without supervision, even though those properties are ambiguous and diverse.

III. PROPERTY-SPECIFIC AESTHETIC ASSESSMENT

In this section, we propose the PSAA algorithm with unsupervised aesthetic property discovery. We regard aesthetic assessment as a binary classification problem, which classifies an image into either high-quality or low-quality, as in [11]–[14]. Figure 2 shows an overview of PSAA, composed of an aesthetic feature extractor, an aesthetic property classifier, and property-specific networks. The feature extractor generates aesthetic features by pooling and concatenating two features from the baseline network. The property classifier classifies the image into one of the aesthetic properties. Finally, the corresponding property-specific network is used to evaluate the aesthetic quality of the image. Let us describe these three components of PSAA subsequently.

A. AESTHETIC FEATURE EXTRACTOR

The aesthetic feature extractor includes a baseline network for image aesthetic analysis and additional pooling and concatenation layers for aesthetic feature generation, as shown in Figure 2. The baseline network itself is also trained to classify the aesthetic qualities of images into high- or low-quality classes. In other words, it takes an image as input and yields a binary classification result using a soft-max layer with two outputs. In this work, we employ GoogLeNet [34]

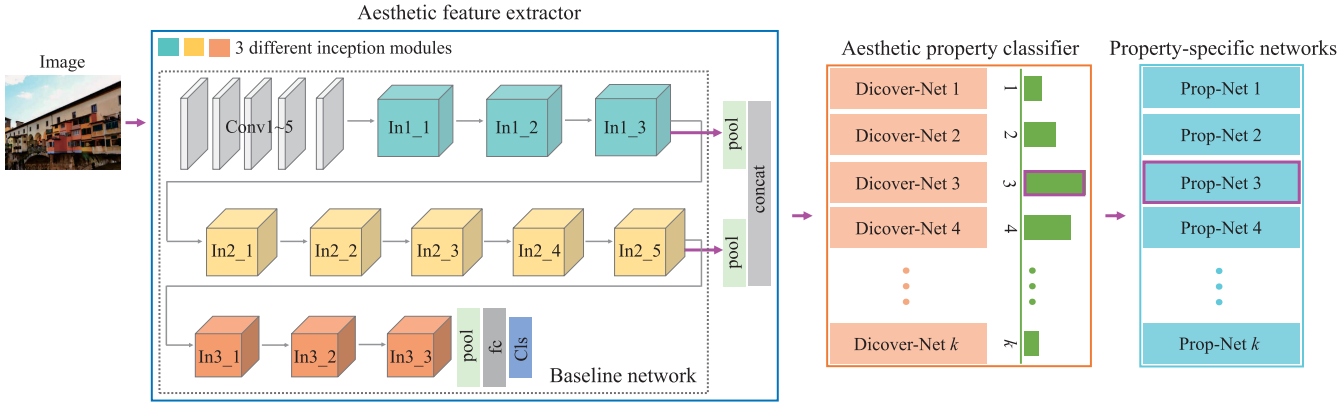


FIGURE 2. An overview of the proposed PSAA algorithm: It is composed of an aesthetic feature extractor, an aesthetic property classifier, and property-specific networks. The baseline network contains five convolution layers (conv1~conv5), eleven inception modules (In1_1~In3_3), an average-pooling layer (pool), a fully connected layer (fc), and a classification layer (Cls). Based on the combination (average-pooling and concatenation) of In1_3 and In2_5, the aesthetic property classifier, which consists of property-discovering networks (Discover-Nets), computes the score for each aesthetic property. Then, the aesthetic quality of the input image is determined by the property-specific network (Prop-Net), corresponding to the maximum score (in this example, Prop-Net 3).

as the backbone of the baseline network. Note that aesthetic qualities are affected by both local details and holistic features [35]. The inception modules of GoogLeNet, which consist of parallel multi-scaled convolution layers, are effective to analyze both local and global aesthetic features.

The baseline network outputs a soft-max probability vector $y = (y_1, y_2)$, where y_1 and $y_2 = 1 - y_1$ are the probabilities that an input image belongs to the high-quality and low-quality classes, respectively. The input is declared as high-quality, if y_1 is higher than 0.5. For the training, we use the cross-entropy loss, given by

$$L(y, \bar{y}) = -\bar{y}_1 \log y_1 - \bar{y}_2 \log y_2 \tag{1}$$

where $\bar{y} = (\bar{y}_1, \bar{y}_2)$ is the ground-truth one-hot vector.

We extract aesthetic features from the baseline network and use them in the subsequent processes. To take advantage of multi-scale features [36], we concatenate the average-pooled outputs of two inception modules, In1_3 and In2_5, as shown in Figure 2.

B. AESTHETIC PROPERTY CLASSIFIER

1) AESTHETIC PROPERTY

The qualities of images are assessed based on many criteria. To describe and enhance the qualities, professional photographers adopt aesthetic rules based on diverse attributes, including lighting condition, contrast, and photographic composition. Also, when assessing images, different aesthetic attributes are applied according to scene and content types of images. Hence, aesthetic attributes, scenes, and contents, and their relationship are all essential for aesthetic quality assessment, and are called aesthetic properties in this work.

With more aesthetic properties, we can represent more diverse high-quality photographs. However, previous approaches to aesthetic assessment used only several manually-defined attributes [11]–[13], [23]. These attributes, however, are ambiguous and subjective [35], and some attributes may have not been discovered yet. Moreover, as mentioned above, aesthetic attributes are related with scene and content types. Therefore, it is impractical to define all aesthetic properties manually. In this work, we develop an unsupervised algorithm to discover a wide variety of aesthetic properties.

2) UNSUPERVISED AESTHETIC PROPERTY DISCOVERY

The proposed algorithm discovers diverse aesthetic properties without supervision. To achieve this, we develop the aesthetic property classifier in Figure 2, composed of k property-discovering networks (Discover-Nets). Each Discover-Net has three fully connected layers and a soft-max layer. It takes aesthetic features as input, and outputs the soft-max scores for positive and negative classes.

It was observed from our empirical studies that aesthetic properties are more obvious in images with higher aesthetic scores. Hence, we construct the positive set $\mathcal{P}$ of such obviously high-quality images. For example, the AVA dataset [29] provides the aesthetic score of an image between 1 and 10. Images with aesthetic scores higher than 5 are considered as high-quality, otherwise as low-quality in existing techniques [11]–[13], [23], [24]. To exclude border cases, the positive set $\mathcal{P}$ is composed of only the training images with scores higher than 6. On the other hand, the negative set $\mathcal{N}$ is randomly sampled from all training images with scores lower than 5. Note that $\mathcal{N}$ is used to train Discover-Nets. To avoid severe data unbalance during the training, $|\mathcal{N}|$ is set to be about one tenth of $|\mathcal{P}|$.

Using the positive samples in $\mathcal{P}$, we obtain initial aesthetic property clusters using the k-means clustering [37],

$$\mathcal{P} = \cup_{i=1}^{k} \mathcal{C}_i, \quad \text{and } \mathcal{C}_i \cap \mathcal{C}_j = \emptyset \text{ for } i \neq j, \tag{2}$$

where $\mathcal{C}_i$ denotes the ith cluster and k is set to 50 initially. In an ideal case, a cluster should contain images with an identical aesthetic property. However, 50 clusters are not enough to

address diverse aesthetic properties. Hence, starting from the initial clusters, we learn the aesthetic property classifier, update the clusters, and split those clusters iteratively.

Specifically, suppose that there are $k^{(t)}$ aesthetic property clusters at the tth iteration. Then, the aesthetic property classifier consists of $k^{(t)}$ Discover-Nets, one for each cluster. For $1 \leq i \leq k^{(t)}$, we train the ith Discover-Net to discern the positive samples in the ith cluster $\mathcal{C}_i^{(t)}$ from all negative samples in $\mathcal{N}$. It is trained with the cross-entropy loss in (1), as in the baseline network. Next, we update the cluster membership of each positive sample $p \in \mathcal{P}$ as follows. We dichotomize p using each Discover-Net and compute the soft-max score for the positive class, which is regarded as the aesthetic property score. Then, we update the cluster label of p by mapping it to the cluster with the maximum aesthetic property score.

In the split stage, we divide each 'unreliable' cluster into two sub-clusters. When all positive samples in a cluster have the same clearly discernible aesthetic property, the corresponding Discover-Net would classify those samples into the positive class with a high accuracy. Otherwise, the Discover-Net would yield a low accuracy. Hence, to determine whether the ith cluster $\mathcal{C}_i^{(t)}$ is reliable or not, we first partition $\mathcal{C}_i^{(t)}$ into two subsets $\mathcal{C}_{i,1}^{(t)}$ and $\mathcal{C}_{i,2}^{(t)}$ using the k-means clustering with $k = 2$ [37]. Then, we measure the accuracies of the Discover-Net on $\mathcal{C}_i^{(t)}$, $\mathcal{C}_{i,1}^{(t)}$, and $\mathcal{C}_{i,2}^{(t)}$, respectively. For $\mathcal{C}_i^{(t)}$, the accuracy is defined as $N_i^{(t)}/M_i^{(t)}$, where $M_i^{(t)}$ is the number of all samples in the cluster and $N_i^{(t)}$ is the number of the correctly classified samples. Similarly, we compute the accuracies over $\mathcal{C}_{i,1}^{(t)}$ and $\mathcal{C}_{i,2}^{(t)}$. Then, we declare $\mathcal{C}_i^{(t)}$ as unreliable and accept the split, if at least one of the three conditions are satisfied:

- The accuracy on $\mathcal{C}_i^{(t)}$ is lower than a threshold τ_1, which means that the cluster contains images with heterogeneous aesthetic properties.
- The difference between the accuracies on $\mathcal{C}_{i,1}^{(t)}$ and $\mathcal{C}_{i,2}^{(t)}$ is higher than a threshold τ_2, which indicates that the two sub-clusters have different properties.
- The ratio of the size of $\mathcal{C}_i^{(t)}$ to the size of $\mathcal{P} \cup \mathcal{N}$ is larger than a threshold τ_3. This is to limit cluster sizes.

We fix τ_1, τ_2, and τ_3 to 0.3, 0.5, and 0.007, respectively.

The unsupervised aesthetic property discovery is achieved when the number $k^{(t)}$ of clusters converges. The number of finally obtained clusters is 136 for the AVA dataset [29]. Figure 3 shows example images with the discovered aesthetic properties. We see that the images with the same discovered property have similar colors, texture, or contents.

C. PROPERTY-SPECIFIC AESTHETIC ASSESSMENT NETWORKS

Each discovered cluster represents an aesthetic property. As illustrated in Figure 3, it is easier to dichotomize images into either high-quality or low-quality class, when they are grouped according to their properties. Thus, we develop the PSAA algorithm, by constructing a property-specific network (Prop-Net) for each discovered property. Each Prop-Net has the same structure as Discover-Nets. It takes aesthetic features as input and outputs the soft-max scores for the high-quality and low-quality classes.

Contrary to Discover-Nets, Prop-Nets are trained using all training images in the entire score range. First, we determine the aesthetic property of each training image using the aesthetic property classifier: we compute the aesthetic property scores for all properties, and assign the image to the property corresponding to the maximum score. Then, we train each Prop-Net using the assigned training images. For the training, we also use the cross-entropy loss in (1).

D. IMPLEMENTATION DETAILS

In Figure 2, the baseline network within the aesthetic feature extractor is based on GoogLeNet [34]. It consists of five convolution layers (conv1∼conv5), eleven inception modules (In1_1∼In3_3) of three different kinds, a fully connected layer (fc), and a soft-max layer. Since GoogLeNet was trained for the 1,000-way image classification, we modify the output sizes of the fc and soft-max layers for the two-way aesthetic classification. To train the baseline network, we initialize the parameters of all layers, except for the fc layer, with those of the GoogLeNet pre-trained on the ILSVRC-2012 dataset [38]. We initialize the fc layer using the Xavier method [39], which determines the scale of initialization based on the numbers of input and output neurons. We update these parameters using the Adam optimizer [40] with a batch size of 16, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We start with a learning rate $\epsilon = 0.001$ for all layers and shrink it via $\epsilon \leftarrow 0.1\epsilon$ after every four epochs. For each training image, we flip it horizontally with probability 0.5. To use the network in a scale-invariant manner, we resize an input image to 229×229.

Discover-Nets and Prop-Nets have the same architecture, *i.e.*, three fc layers and a soft-max layer with two output neurons. The three fc layers have 1024, 512, and 2 neurons, respectively. To train Discover-Nets and Prop-Nets, we initialize the fc layers with the Xavier method starting with a learning rate $\epsilon = 0.0001$ and shrink it via $\epsilon \leftarrow 0.1\epsilon$ after ten epochs. We set the other hyper-parameters in the same way as we do for training the baseline network.

IV. APPLICATIONS: AESTHETIC IMAGE ENHANCEMENT

In addition to the development of an effective image aesthetic assessment algorithm, we demonstrate how it can be applied to practical image processing applications. Specifically, we show that the proposed algorithm can be used for (1) aesthetic contrast enhancement and (2) aesthetic region cropping. Both enhancement schemes use an aesthetic activation map, which represents the aesthetic importance of each pixel.

A. AESTHETIC ACTIVATION MAP

Let us first describe how to generate an aesthetic activation map. There are various tools for CNNs that can identify important pixels, *i.e.* [41]–[43]. The aesthetic activation map represents the level of importance of each pixel in terms

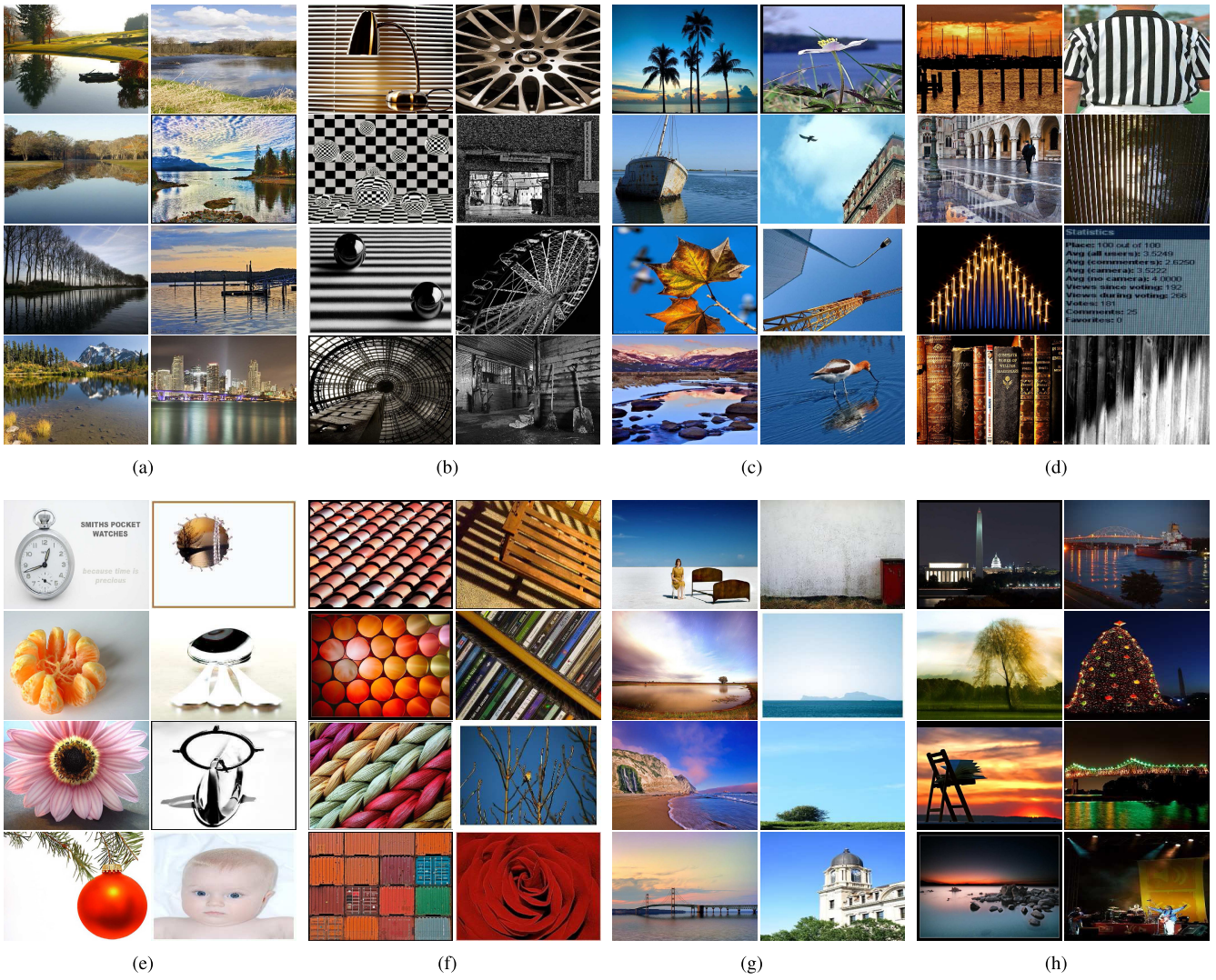


FIGURE 3. Example images that are declared to have the first eight aesthetic properties, discovered by applying the proposed algorithm to the AVA dataset [29]. For each property, we show high-quality images in the left column and low-quality images in the right one: (a) landscapes with symmetry composition, (b) patterns in gray-scale images, (c) blue background, (d) vertical composition, (e) oval-shaped foreground with flat background, (f) patterns in colorful scenes, (g) horizontal composition, and (h) dim scenes.

of image aesthetics. An aesthetically important pixel should have a high value in the aesthetic activation map. To this end, as similarly done in [41], we use output feature maps of more than one layers in the baseline network. More specifically, we use the output feature maps of two inception modules, In1_3 and In2_5, which have 288 channels of size 35×35 and 768 channels of size 17×17, respectively.

Let F_c denote the cth channel of the feature map of In1_3. Given an image, the proposed PSAA algorithm yields the soft-max probability vector $p = (p, q)$. By back-propagating the probability p for the high-quality class to F_c, we obtain the gradient map $G_c = \frac{\partial p}{\partial F_c}$. Then, we define the significance level w_c of the cth channel as

$$w_c = \max_{1 \le x,y \le 35} |G_c(x, y)|. \tag{3}$$

Next, we obtain the aesthetic activation map $A_{\text{In1\_3}}$ for In1_3, by superposing the feature maps using the significance levels as weights

$$A_{\text{In1\_3}}(x, y) = \max \left\{ \sum_c w_c F_c(x, y),\ 0 \right\}. \tag{4}$$

Note that, in (4), we only consider the positive influence to the high-quality class. We also obtain the aesthetic activation map $A_{\text{In2\_5}}$ for In2_5 similarly.

Finally, we obtain the final aesthetic activation map A by aggregating $A_{\text{In1\_3}}$ and $A_{\text{In2\_5}}$ as

$$A = A_{\text{In1\_3}} + \widetilde{A}_{\text{In2\_5}} \tag{5}$$

where $\widetilde{A}_{\text{In2\_5}}$ is the resized version of $A_{\text{In2\_5}}$ to that of $A_{\text{In1\_3}}$.

Figure 4 shows examples of aesthetic activation maps. It is observable that the aesthetic activation maps show higher values at the composition elements such as symmetric, horizontal, and diagonal lines [3] as well as salient objects.

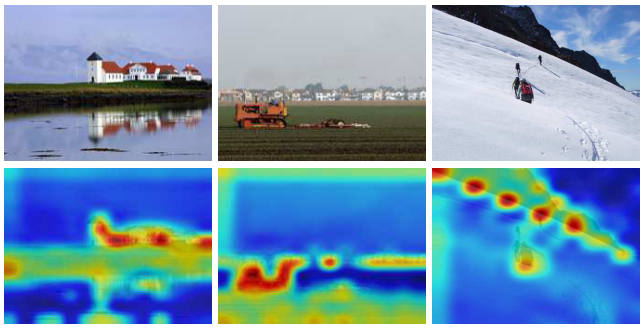


FIGURE 4. Examples of aesthetic activation maps: (top) input images and (bottom) their aesthetic activation maps. In these examples, high activation values are on salient contents and regions, which determine photographic composition rules [3]: symmetric, horizontal, and diagonal.

B. AESTHETIC CONTRAST ENHANCEMENT

Many contrast enhancement techniques provide user-adjustable parameters to control the level of enhancement. Optimal parameters are generally found on a trial-and-error basis to maximize output qualities. In this work, we use aesthetic activation maps to determine optimal parameters automatically and adaptively. To demonstrate the effectiveness of the proposed algorithm in the parameter decision, we test two enhancement algorithms: gamma correction (GC) [44] with a parameter γ and histogram-based contrast enhancement (HCE) [45] with two parameters (μ, β). However, note that the proposed algorithm can be applied to any contrast enhancement techniques with user-adjustable parameters. Let us briefly review GC and HCE to describe the roles of the parameters.

GC is one of the simplest contrast enhancement algorithm, which converts intensity l of an image using the transformation function

$$T(l) = l_{\max}\left(\frac{l}{l_{\max}}\right)^{\gamma} \tag{6}$$

where $l_{\max}$ denotes the maximum intensity value. In this work, $l_{\max} = 255$.

HCE is a histogram modification technique based on the logarithm function. The histogram of pixel intensities in an image $h = [h_0, h_1, \ldots, h_{255}]^T$ is modified to $m = [m_0, m_1, \ldots, m_{255}]^T$. The modified histogram m_k for intensity k is obtained as

$$m_k = \frac{\log(h_k h_{\max} 10^{-\mu} + 1)}{\log(h_{\max}^2 10^{-\mu} + 1)} \tag{7}$$

where $h_{\max}$ denotes the maximum element in h and μ is the parameter to control the level of contrast enhancement. As μ increases, the input histogram is less strongly modified and HCE becomes more similar to the original histogram equalization, which may yield over-enhancement artifacts. On the other hand, as μ decreases, the input image is less strongly enhanced. In [45], in addition to contrast enhancement, power consumption is also controlled by the parameter β. As β gets larger, the overall brightness of an enhanced image is dimmed to reduce the power consumption more aggressively.

For HCE-based aesthetic contrast enhancement, we apply HCE by varying μ in $\{2.0, 5.0, 5.5, 6.5\}$ and β between 0.8 and 2.2 with step size 0.2 [45]. Let $A_{\mu,\beta}$ denote the activation map in (5) with a parameter pair (μ, β). Then, for an enhanced image with (μ, β), we measure its aesthetic score by

$$S^{\mathrm{a}}_{\mu,\beta} = \frac{1}{N}\sum_{x,y} A_{\mu,\beta}(x, y) \tag{8}$$

where $N = 35 \times 35$ is the size of $A_{\mu,\beta}$. We assume that an image with pixels that affect the high class probability p more strongly has a higher aesthetic score. Thus, we determine the optimal pair of (μ^*, β^*) to yield the maximum score,

$$(\mu^*, \beta^*) = \arg\max_{(\mu,\beta)} S^{\mathrm{a}}_{\mu,\beta}. \tag{9}$$

Similarly, for GC-based enhancement, we apply GC to an image by varying γ between 0.5 and 1.5 with step size 0.1, and select the optimal γ^* to maximize the aesthetic score.

C. AESTHETIC IMAGE CROPPING

Aesthetic cropping attempts to retain the most appealing view in a photograph, while excluding less important regions. We use 368 windows with different sizes and aspect ratios. Then, we obtain cropping candidates by sliding windows within the image. In this work, 5,632 candidates are considered in total.

Unlike contrast enhancement, if we consider aesthetic quality only, a cropping result may lose major subjects or include the entire input image. To avoid both cases, we attempt to preserve the contextual information of an original image I_0, and also trim off insignificant region as much as possible. To this end, for each cropping candidate I_n, we firstly extract its contextual feature f_n, which is the output of the last pooling layer of ResNet-50 [46] trained on the ImageNet-2012 classification dataset [38]. Then, we measure the contextual similarity between I_n and I_0 by computing the context-preservation score

$$S^{\mathrm{p}}_n = -\chi^2(f_n, f_0) \tag{10}$$

where f_0 is the contextual feature of I_0 and $\chi^2(\cdot, \cdot)$ denotes the chi-square distance.

Secondly, we measure the size difference between the candidate and input by normalizing the size of the cropping candidate as

$$S^{\mathrm{z}}_n = -\frac{|I_n|}{|I_0|}, \tag{11}$$

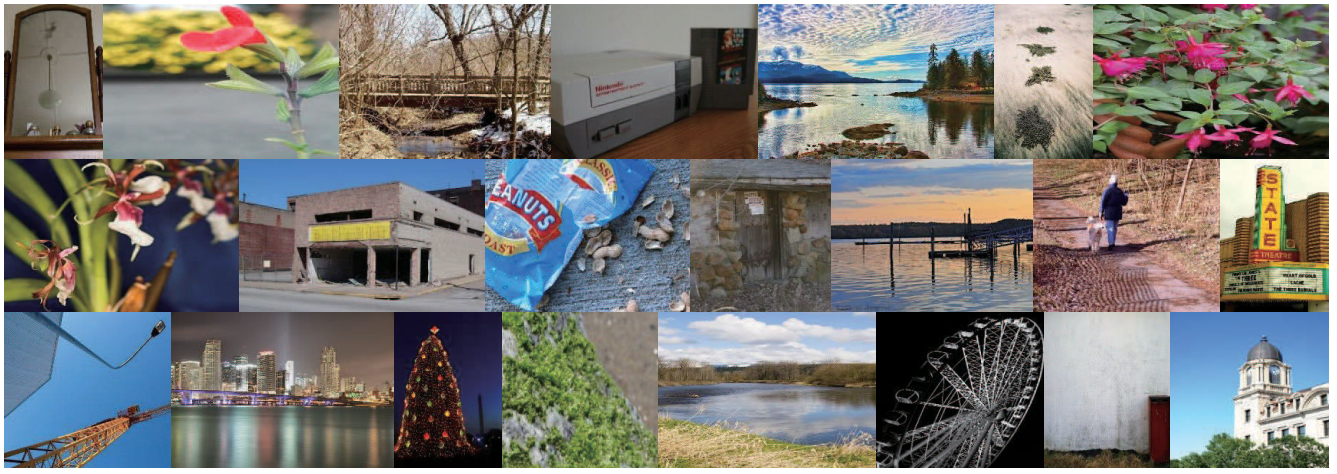
where $|\cdot|$ denotes the size of an image.

To find the optimal cropping region, we measure the overall cropping score S^{c}_n of each candidate I_n by

$$S^{\mathrm{c}}_n = S^{\mathrm{a}}_n + \lambda_1 S^{\mathrm{p}}_n + \lambda_2 S^{\mathrm{z}}_n \tag{12}$$

where S^{a}_n is the aesthetic score of I_n, computed in the same way as (8). Finally, we choose the candidate with the maximum cropping score as the optimal cropping region I_{n^*},

(a) Predicted as high-quality images



(b) Predicted as low-quality images

FIGURE 5. Examples of test images in the AVA dataset, which are categorized into the high- and low-quality classes. Notice that the landscape images in Figure 3(a) are also included in this figure. Even for humans, their aesthetic classification is easier in Figure 3(a) than in this figure.

where

$$n^* = \arg\max_n S_n^c. \tag{13}$$

In (12), λ_1 and λ_2 are empirically set to 0.05 and 0.59, respectively, to maximize the subjective qualities of cropping results.

V. EXPERIMENTAL RESULTS

A. DATASET

We evaluate the performance of the proposed PSAA algorithm on the AVA dataset [29]. To the best of our knowledge, AVA is the largest publicly available aesthetic assessment dataset. It contains about 250,000 images, and the aesthetic quality of each image was rated by about 200 human annotators. The ratings range from 1 to 10, with 10 indicating the highest quality. For a fair comparison, we use the same partition of training and testing data as the conventional algorithms do [11]–[13], [23], [29]: 235,599 images for training and 19,930 images for testing. Also, we follow the same procedure as the conventional algorithms to assign a binary aesthetic label to each image; images with mean ratings smaller than 5 are labeled as low-quality, otherwise high-quality.

B. AESTHETIC QUALITY ASSESSMENT

We first evaluate the aesthetic assessment performance of the proposed PSAA algorithm qualitatively. Figure 5 shows some test images, which are classified into the high- and low-quality classes by PSAA. Notice that the eight landscape images in Figure 3(a) are also included in Figure 5. When they are mixed with other images of different aesthetic properties, it is more difficult to determine their quality classes. Even for humans, the classification is easier in Figure 3(a) than in Figure 5. This is why we perform the unsupervised aesthetic property discovery and the property-specific assessment.

The performance of the proposed algorithm is assessed quantitatively, by measuring the accuracy score

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{14}$$

TABLE 1. Comparison of the aesthetic assessment accuracy of the proposed PSAA algorithm with those of conventional algorithms. The best result is boldfaced.

	Accuracy (%)
AVA [29]	67.0
RDCNN [11]	74.4
DMA-Net-ImgFu [12]	75.4
MNA-Net-Scene [13]	77.4
Regress-Net [23]	77.3
NIMA [24]	81.5
GPF-CNN [28]	81.8
Feature Fusion [27]	82.0
A-Lamp [14]	82.5
SDAL [25]	83.1
ASPP FCN-GC [26]	83.6
Proposed PSAA	**84.3**

where N_{total} and N_{correct} are the numbers of total and correctly classified images, respectively.

1) COMPARISON WITH CONVENTIONAL ALGORITHMS

In an extensive survey of aesthetic assessment techniques in [35], it was shown that CNN-based algorithms outperform the others. Hence, we compare the proposed PSAA algorithm only with the recent algorithms in [11]–[14], [23]–[29], most of which are CNN-based. The AVA algorithm [29] is based on handcrafted and generic features. The other algorithms use CNNs. In addition to CNN features, external information, such as learning-based aesthetic attribute classification in [11], [12], scene categorization in [13], attribute and content classification in [23], salient object detection in [14], aesthetic property classification [25], and hand-crafted and generic feature in [27], is employed. In [11], [12], [14], [25], both local and global characteristics are analyzed. Also, the aesthetic score of an image is computed as the mean of the aesthetic score distribution in [24], [28], or by averaging the scores in local regions in [26]. Table 1 compares the accuracy scores. We see that, whereas the traditional AVA algorithm provides the lowest accuracy, the proposed PSAA algorithm yields the highest accuracy. PSAA outperforms ASPP FCN-GC [26], which is the best-performing conventional algorithm, by a gap of 0.7%. Let us analyze subsequently how this sate-of-the-art performance is achieved.

2) EFFECTIVENESS OF PROPERTY-SPECIFIC ASSESSMENT

We verify the efficacy of the property-specific assessment strategy of the proposed PSAA algorithm. Notice that the baseline network itself can be used as a binary quality classifier, but it does not exploit any aesthetic properties. Hence, we compare PSAA with the baseline network to measure the effectiveness of the property-specific assessment. To confirm its general effectiveness, in addition to GoogLeNet, we employ two additional backbone networks: VGG-16 [47] and ResNet-50 [46]. As well as GoogLeNet, both VGG-16 and ResNet-50 have been widely employed as backbones in various CNN-based image processing and computer vision techniques [21], [24], [36].

TABLE 2. Comparison of the accuracy scores of the baseline networks and PSAAs using three different backbones on the AVA dataset.

Backbone	Accuracy (%)	
	Baseline network	PSAA
GoogLeNet	73.1	84.3
VGG-16	72.7	84.1
ResNet-50	74.2	84.5

TABLE 3. The accuracy scores and the numbers, k, of aesthetic properties of PSAA according to the split thresholds τ_1, τ_2, τ_3.

τ_1	0.2	0.3	0.3	0.4	0.5
τ_2	0.5	0.5	0.5	0.4	0.3
τ_3	0.009	0.008	0.007	0.006	0.004
k	110	125	136	153	169
Accuracy (%)	82.7	83.1	84.3	83.8	83.7

Similar to the GoogLeNet-based baseline network in Sections III, the additional baseline networks are constructed based on VGG-16 and ResNet-50. For the VGG-based network, we adopt its conv1∼conv5 blocks and first two fc layers. Then, we modify the structure of the last fc layer and the soft-max layer for binary aesthetic quality classification. In training, the conv1∼conv5 blocks and the first two fc layers are initialized with those of the pre-trained VGG-16 on ILSVRC-2012 [38], and the modified fc layer is trained from scratch. Similarly, for the ResNet-based network, we use the residual blocks res1∼res5, and redesign its fc and soft-max layers for binary classification. To train it, we initialize the res1∼res5 blocks with those of the pre-trained ResNet-50 using ILSVRC-2012, and train the other layers from scratch. As done in Figure 2, we extract multi-scale aesthetic features by average-pooling and concatenating conv4 and conv5 for VGG-16 and res4 and res5 for ResNet-50.

Table 2 lists the accuracy scores of the baseline networks and PSAAs using the three backbones on the AVA dataset. Notice that all three PSAAs significantly outperform their corresponding baseline networks. More specifically, PSAAs using GoogLeNet, VGG-16, and ResNet-50 backbones improve the accuracies by 11.2%, 11.4%, and 10.3%, respectively. These large performance gaps between the baseline networks and PSAAs confirm the efficacy of the aesthetic property discovery and the property-specific assessment. Also, note that the proposed algorithm is independent of the architecture of a backbone network and thus can employ any CNN-based feature extractor as a backbone.

Furthermore, notice from Tables 1 and 2 that, regardless of the choice of the backbone, the proposed PSAA algorithm outperforms the recent CNN-based algorithms [13], [14], [23]–[28].

TABLE 4. Comparison of contrast enhancement algorithms using objective quality metrics: Lee *et al.*'s algorithm [45], Regress-Net-based algorithm [23], and the proposed algorithm. A boldfaced number denotes the best result for each test.

	Input	HCE-based			GC-based	
		Lee *et al.* [45]	Regress-Net [23]	Proposed	Regress-Net [23]	Proposed
DE (↑)	7.11	6.48	6.58	**6.60**	6.58	**6.69**
AMBE (↓)	-	26.68	26.11	**17.74**	21.04	**19.63**
EME (↑)	36.38	**46.78**	40.15	41.58	32.52	**37.15**
PixDist (↑)	28.67	26.79	28.22	**30.09**	**28.21**	28.00

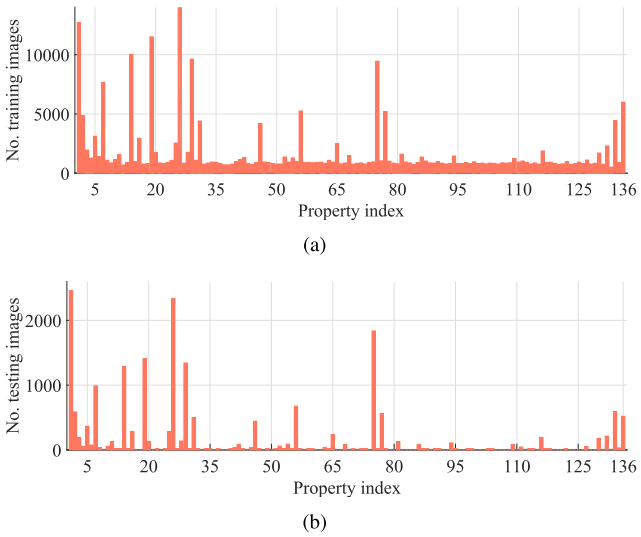


FIGURE 6. The distribution of training images (a) and that of testing images (b) in the AVA dataset according to the property index.

3) NUMBER OF DISCOVERED AESTHETIC PROPERTIES

In Section III-B, we use three split thresholds (τ_1, τ_2, τ_3) to control the number k of aesthetic properties. Let us analyze the impacts of these thresholds on the assessment performance. Table 3 lists the accuracy scores and the numbers of clusters of PSAA for several combinations of the thresholds on the AVA dataset. Too large a k degrades the assessment performance, since each cluster does not contain sufficient samples for training Prop-Nets. On the other hand, too small a k cannot represent diverse aesthetic properties faithfully. Hence, we choose the split thresholds $(0.3, 0.5, 0.007)$ to achieve the highest accuracy, striking a balance between reliable training and faithful representation. For this setting, the number of discovered aesthetic properties is $k = 136$.

The proposed aesthetic property classifier can be regarded as a clustering scheme. To analyze its contribution to the overall assessment performance, we replace the aesthetic property classifier with the k-means clustering with $k = 136$ [37], where only the positive set $\mathcal{P}$ is used. Then, we map all training data into their nearest neighbor (NN) clusters and train Prop-Nets. In the testing phase, for a query image, the NN cluster is found and the corresponding Prop-Net is employed for the aesthetic assessment. This k-means clustering approach yields the accuracy score of 80.8%, which is lower than the proposed PSAA algorithm by 3.5%.

TABLE 5. Comparison of the conventional and proposed image cropping algorithm on the human crop dataset [53] in terms of the MaxOverlap scores. The best result and the second best are boldfaced and underlined, respectively.

	Average MaxOverlap
Yan *et al.* [54]	0.64
Fang *et al.* [53]	0.70
Kao *et al.* [55]	0.75
Li *et al.* [56]	0.82
Wang *et al.* [57]	<u>0.83</u>
Proposed PSAA	**0.84**

This confirms that the proposed aesthetic property classifier is more effective in grouping aesthetic features than the k-means clustering.

Figure 6 shows the number of images that are assigned to each aesthetic property for the AVA training and testing datasets, respectively. We discover diverse aesthetic properties, but images are not uniformly distributed in terms of the properties. In other words, although aesthetic properties are quite diverse, some properties are more frequently adopted than the others. However, even to the least adopted property for the AVA training dataset, 704 training images are assigned. They are then used to train the corresponding Prop-Net. On the other hand, there are much fewer test images than training images in the AVA dataset. Thus, for the AVA testing dataset, 52 properties are not used, and the aesthetic assessment is performed by the remaining 84 Prop-Nets.

C. AESTHETIC CONTRAST ENHANCEMENT

Next, the proposed PSAA algorithm is used to select optimal parameters for contrast enhancement. For comparison, we also test Regress-Net [23]. Regress-Net yields a continuous aesthetic score between 0 and 1, whereas PSAA performs the binary classification. Therefore, instead of the aesthetic score in (8), we use the output of Regress-Net directly to decide user-controllable parameters (μ, β) for HCE and γ for GC. For HCE-based aesthetic contrast enhancement, we also compare the proposed algorithm with the original HCE [45], where the parameters are fixed as $(\mu, \beta) = (5, 1)$.

We first compare the performances objectively on the 93 low-contrast images in [48] using four quality metrics: discrete entropy (DE) [49], absolute mean brightness error (AMBE) [50], measure of enhancement (EME) [50], and PixDist [51]. DE measures the amount of information

FIGURE 7. HCE-based contrast enhancement: from top to bottom, input images, enhancement results of Lee *et al.*'s algorithm [45], Regress-Net [23], and the proposed algorithm.

FIGURE 8. Aesthetic gamma correction: from top to bottom, input images, enhancement results using Regress-Net [23] and the proposed algorithm.

in an image. A high DE means that the image conveys more information. AMBE measures the absolute difference between input and output mean brightness levels. It yields a lower value when an algorithm well preserves the mean brightness of the input. EME approximates the contrast in an image by computing a score based on the block-wise minimum and maximum intensity levels. PixDist also represents a contrast level by computing the average mutual intensity

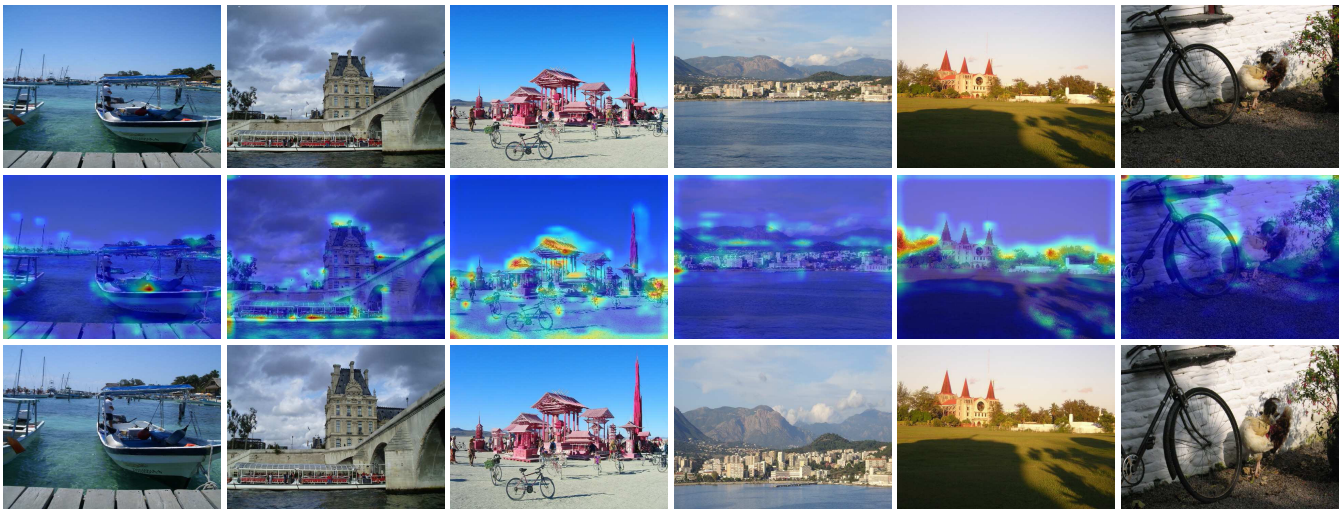


FIGURE 9. Examples of aesthetic image cropping: (top) original images, (middle) aesthetic activation maps, and (bottom) cropping results.

FIGURE 10. More examples of aesthetic image cropping: (top) original images, (middle) aesthetic activation maps, and (bottom) cropping results.

differences for all pixel pairs in an image. Therefore, higher DE, EME, and PixDist and a lower AMBE imply higher image quality.

Table 4 lists the average performances of HCE-based and GC-based enhancement algorithms on the test images. Overall, HCE-based algorithms outperform simple GC-based algorithms. Let us first compare the HCE-based algorithms. For DE, the proposed algorithm conveys more information than the other algorithms. Note that, because of the information processing inequality [52], no algorithm can yield a higher DE than an input image. In terms of AMBE and PixDist, the proposed algorithm achieves the best performances. This is because the proposed algorithm selects images with well-preserved mean brightness and with uniformly distributed intensity histograms. With fixed parameters, Lee *et al.*'s algorithm yields the highest EME. However, it is less adaptive to image characteristics as illustrated in Figure 7. In both GC-based and HCE-based tests, the proposed algorithm performs better than Regress-Net in terms of all metrics, except for PixDist in the GC-based test. This indicates that the proposed algorithm more accurately assesses image aesthetics, of which image contrast is one of the most important factors.

Figures 7 and 8 show enhancement results of HCE- and GE-based algorithms, respectively. In Figure 7, while Lee *et al.*'s algorithm and Regress-Net provide under- or over-enhanced results, the proposed algorithm yields better results overall. In Figure 8, the proposed algorithm provides more reliable results than Regress-Net.

D. AESTHETIC IMAGE CROPPING

We evaluate the performance of the proposed aesthetic image cropping algorithm on the 500 ill-composed test images in the human crop dataset [53], in which each image has ten ground-truth cropping results annotated by experienced photographers. As in [53], the performance of an image cropping result is measured by the maximum-overlap ratio (MaxOverlap) between the cropping result R and the ground-truth set

$G = \{G_1, \ldots, G_{10}\}$, given by

$$\text{MaxOverlap}(R, G) = \max_k \text{IoU}(R, G_k) \tag{15}$$

where $\text{IoU}(R, G_k)$ is the intersection over union ratio between R and G_k computed by

$$\text{IoU}(R, G_k) = \frac{|R \cap G_k|}{|R \cup G_k|}. \tag{16}$$

We compare the proposed algorithm with five conventional cropping algorithms: Yan *et al.* [54], Fang *et al.* [53], Kao *et al.* [55], Li *et al.* [56], and Wang *et al.* [57]. Whereas Yan *et al.* and Fang *et al.* use handcrafted features and generic image descriptors to measure the changes in contents and composition after cropping, the others use feature maps obtained by CNNs.

Table 5 compares the average MaxOverlap scores of the proposed algorithm with those of the conventional algorithms on the test images. The proposed algorithm significantly outperforms the Yan *et al.*'s algorithm and the Fang *et al.*'s algorithm. This confirms that aesthetic features of the proposed algorithm are more effective than handcrafted features and generic image descriptors for the purpose of cropping. Also, the proposed algorithm performs better than the other aesthetic-based algorithms [55]–[57]. This shows quantitatively that the proposed algorithm can crop out aesthetically important regions more effectively. Figures 9 and 10 present some cropping results. We see that the proposed algorithm successfully determines aesthetically attractive regions and preserves major contents of original images.

VI. CONCLUSIONS

We developed a property-specific image aesthetic assessment algorithm, called PSAA, that consists of the aesthetic feature extractor, the aesthetic property classifier, and the property-specific assessment networks. The aesthetic feature extractor generates aesthetic features of an input image by analyzing image aesthetics via a CNN. The aesthetic property classifier then predicts the aesthetic property of the input image. Then, the property-specific network, corresponding to the predicted property, categorizes the image into either high-quality or low-quality class. Experimental results showed that the proposed PSAA algorithm outperforms conventional state-of-the-art techniques. Moreover, it was demonstrated that PSAA can be employed in the two applications of contrast enhancement and image cropping.

REFERENCES

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

[2] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 833–843, Jun. 2012.

[3] J.-T. Lee, H.-U. Kim, C. Lee, and C.-S. Kim, "Photographic composition classification and dominant geometric element detection for outdoor scenes," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 91–105, Aug. 2018.

[4] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016.

[5] X. He, "Laplacian regularized D-optimal design for active learning and its application to image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 254–263, Jan. 2010.

[6] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003.

[7] S.-M. Hu, F.-L. Zhang, M. Wang, R. R. Martin, and J. Wang, "PatchNet: A patch-based image representation for interactive library-driven image editing," *ACM Trans. Graph.*, vol. 32, no. 6, p. 196, Nov. 2013.

[8] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 386–399.

[9] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1657–1664.

[10] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2011, pp. 2206–2213.

[11] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. ACM Multimedia*, Nov. 2014, pp. 457–466.

[12] 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. (ICCV)*, Dec. 2015, pp. 990–998.

[13] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 497–506.

[14] S. Ma, J. Liu, and W. C. Chang, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4535–4544.

[15] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[16] 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.*, Dec. 2011, pp. 1784–1791.

[17] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proc. ACM Multimedia*, Nov. 2011, pp. 1213–1216.

[18] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[21] J.-T. Lee, H.-U. Kim, C. Lee, and C.-S. Kim, "Semantic line detection and its applications," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3249–3257.

[22] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on Fourier domain analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 330–339.

[23] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 662–679.

[24] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[25] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, "Deep active learning with contaminated tags for image aesthetics assessment," *IEEE Trans. Image Process.*, to be published.

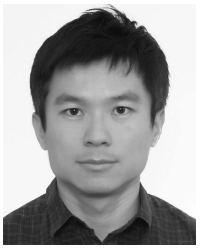
[26] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Modeling image composition for visual aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2019, pp. 1–3.

[27] M. Kucer, A. C. Loui, and D. W. Messinger, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5100–5112, Oct. 2018.

[28] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, to be published.

[29] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2408–2415.
[30] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy Web data," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 669–676.
[31] S. Ma, S. Sclaroff, and N. Ikizler-Cinbis, "Unsupervised learning of discriminative relative visual attributes," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 61–70.
[32] C. Huang, C. C. Loy, and X. Tang, "Unsupervised learning of discriminative attributes and visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5175–5184.
[33] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 73–86.
[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[35] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017.
[36] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2874–2883.
[37] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *J. Roy. Stat. Soc.*, vol. 28, no. 1, pp. 100–108, Oct. 1979.
[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, May 2010, pp. 249–256.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–15.
[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2017, pp. 618–626.
[42] C. Gan, N. Wang, Y. Yang, D. Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Comput. Vis. Pattern Recognit.*, vol. 2015, pp. 2568–2577.
[43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 818–833.
[44] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[45] C. Lee, C. Lee, Y.-Y. Lee, and C.-S. Kim, "Power-constrained contrast enhancement for emissive displays based on histogram equalization," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 80–93, Jan. 2012.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.
[48] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, Dec. 2013.
[49] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[50] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.
[51] Z. Chen, B. R. Abidi, D. L. Page, and M. A. Abidi, "Gray-level grouping (GLG): An automatic method for optimized image contrast enhancement—Part I: The basic method," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2290–2302, Aug. 2006.
[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[53] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proc. ACM Multimedia*, Nov. 2014, pp. 1105–1108.
[54] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 971–978.
[55] Y. Kao, R. He, and K. Huang, "Automatic image cropping with aesthetic map and gradient energy map," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1982–1986.
[56] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for image cropping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8193–8201.
[57] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.

JUN-TAE LEE (S'14) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2013, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include computer vision and machine learning, especially in the problems of photographic composition and semantic line detection.



CHUL LEE (S'06–M'13) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2003, 2008, and 2013, respectively.

He was with Biospace Inc., Seoul, from 2002 to 2006, where he was involved in the development of medical equipment. From 2013 to 2014, he was a Postdoctoral Scholar with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA. From 2014 to 2015, he was a Research Scientist with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. From 2015 to 2019, he was an Assistant Professor with the Department of Computer Engineering, Pukyong National University, Busan, South Korea. In March 2019, he joined the Department of Multimedia Engineering, Dongguk University, Seoul, where he is currently an Assistant Professor. His current research interests include image processing and computational imaging with an emphasis on restoration and high dynamic range imaging.

He received the Best Paper Award from the *Journal of Visual Communication and Image Representation*, in 2014. He is currently an Editorial Board Member of the *Journal of Visual Communication and Image Representation*.

CHANG-SU KIM (S'95–M'01–SM'05) received the Ph.D. degree (Hons.) in electrical engineering from Seoul National University, in 2000. From 2000 to 2001, he was a Visiting Scholar with the Signal and Image Processing Institute, University of Southern California, Los Angeles, CA, USA. From 2001 to 2003, he Coordinated the 3D Data Compression Group, National Research Laboratory for 3D Visual Information Processing, SNU. From 2003 and 2005, he was an Assistant Professor with the Department of Information Engineering, Chinese University of Hong Kong. In September 2005, he joined the School of Electrical Engineering, Korea University, where he is currently a Professor. He has published more than 270 technical papers in international journals and conferences. His research interests include image processing and computer vision. He is a member of the Multimedia Systems and Application Technical Committee (MSATC) of the IEEE Circuits and Systems Society. In 2009, he received the IEEK/IEEE Joint Award for Young IT Engineer of the Year. In 2014, he received the Best Paper Award from the *Journal of Visual Communication and Image Representation (JVCI)*. He served as an Editorial Board Member of JVCI and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is also a Senior Area Editor of JVCI and an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He was an APSIPA Distinguished Lecturer for term, from 2017 to 2018.