Preference-Aware View Recommendation System for Cameras Based on Bag of Aesthetics-Preserving Features

Hsiao-Hang Su, Tse-Wei Chen, *Member, IEEE*, Chieh-Chi Kao, Winston H. Hsu, *Member, IEEE*, and Shao-Yi Chien*, *Member, IEEE*

***Abstract*—In this paper, the framework for a real-time view recommendation system is proposed. The proposed system comprises two parts: offline aesthetic modeling stage and efficient online aesthetic view finding process. A *preference-aware aesthetic model* is proposed to suggest views according to varied user-favorite photographic styles, where a bottom-up approach is developed to construct an *aesthetic feature library* with *bag-of-aesthetics-preserving features* instead of top-down methods that implement the heuristic guidelines (*rule-specific features*) listed in photography literatures, which is employed in previous works. The proposed model can cover both implicit and explicit aesthetic features and can adapt to users' preferences with a learning process. In the second part, the learned model is employed in a view finder to help the user to locate the most aesthetic view while taking a photograph. The experimental results show that the proposed features in the library (92.06% in accuracy) outperform the state-of-the-art rule-specific features (83.63% in accuracy) significantly in the photo aesthetic quality classification task, and the rule-specific features are also proved to be encompassed by the proposed features. Meanwhile, it is observed from experiments that the features extracted for contrast information are more effective than those for absolute information, which is consistent with the properties of human visual systems. Furthermore, the user studies for the view recommendation task confirm that the suggested views are consistent with users' preferences (81.25% agreements).**

***Index Terms*—View Recommendation, Aesthetic Modeling, Aesthetic View Finding, Photo Aesthetic Quality Classification.**

I. Introduction

With the advances of computing capabilities in digital devices, more and more intelligent functions are included in digital cameras, such as autofocus and face detection. These functions can assist users to take photos of their interests in high quality with appropriate exposure values and focus settings; however, it is still hard for amateur photographers to locate an aesthetic view with a good spatial configuration when taking a picture.

If a wide-angle view can be obtained first with wide-angle lens or panorama functions equipped in many cameras, it is possible to design a view recommendation engine to help users to find the best candidate view that has the best spatial configuration or composition, as illustrated in Fig. 1. The engine can also work on general cameras with the preview data generated during view-finding. Although users can also modify the composition of photographs with post-editing tools, the task is tedious and time consuming; moreover, the resolutions and photo qualities are reduced by the post-editing process. Therefore, there is a great demand to equip these devices with a real-time view recommendation system that guides the users to locate the most aesthetic view when taking photos. A real-time view recommendation system based on the processing of wide-angle views is proposed in this paper as shown in Fig. 1. With the most aesthetic view suggested by the proposed system, the user can adjust the current view for aesthetic photos during photo capturing, which can save the post-editing time and maintain the high quality of photos as well.

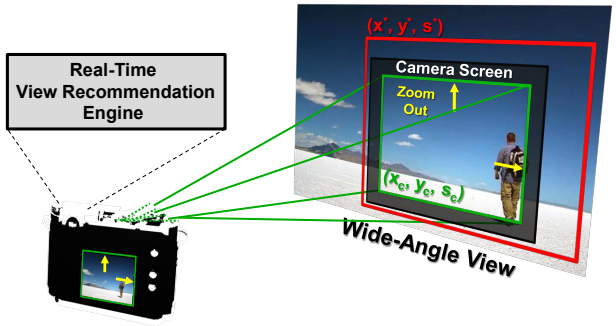
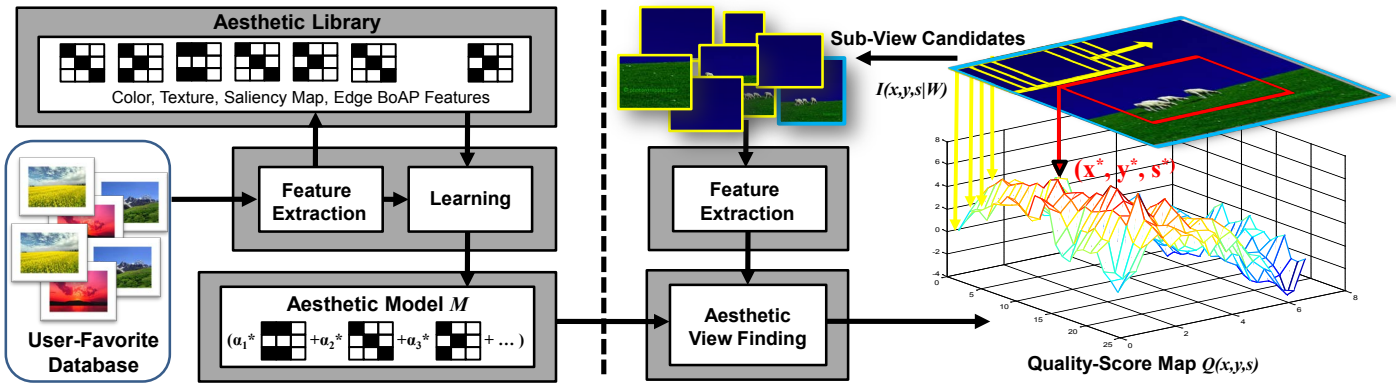


Fig. 1. The proposed real-time view recommendation system by efficiently detecting the most aesthetic regions from a wide-angle view in cameras. Even though there are many possible candidate views that might be captured when taking a picture, with the suggested view (x^*, y^*, s^*) having the best estimated aesthetic score, the user can adjust the position from the current view (x_c, y_c, s_c) for an aesthetic photograph. The user studies demonstrate that the views suggested by the proposed algorithm are consistent with users' preferences. Considering the real-time implementation issues, it can be integrated into cameras or hand-held devices.

The proposed system comprises two stages: offline aesthetic modeling stage and online aesthetic view finding process. The overview of the proposed system is illustrated in Fig. 2. For the aesthetic modeling task at the first stage illustrated in Fig. 2(a), developing a model based on user-preferred photographic styles is an important issue. It is a challenging task because 1) the aesthetic characteristics of photographs are complicated to describe and 2) the photographic styles actually vary for different professional photographers; that is, there are no absolute aesthetic guidelines followed by all professional photographers. Therefore, the great challenge we face is how to comprehensively explore photographic aesthetics, while

Media IC and System Lab, Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, EE2-540, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan sychien@cc.ee.ntu.edu.tw Phone: +886-2-33663668, Fax: +886-2-23681679



(a) Offline aesthetic modeling stage. (b) Online aesthetic view finding process.

Fig. 2. Overview of the proposed system. (a) illustrates the offline aesthetic modeling stage. Given the aesthetic library with the proposed BoAP features, a preference-aware aesthetic model is learned through the offline learning process. (b) illustrates the online view finding process. The learned model M serves as an aesthetic view finder to scan through all sub-view candidates. The estimated score per sub-view is entered in the quality-score map Q at its corresponding scale-space position. The red triangle point (x^*, y^*, s^*) indicates the suggested view with the highest aesthetic quality score.

adaptively imitating the styles from different photographers or different photo galleries. It is not easy to build a generalized aesthetic model suitable for every user, as has been attempted in most existing works [1]–[9]. These works adopt a rule-based approach, which is a top-down process, to implement simple heuristic guidelines individually, such as "Rule of Thirds" and "Visual Weight Balance" illustrated in Fig. 3. These individual features are called *rule-specific features* in this paper. However, there should be some underlying rules that professional photographs follow in common, while they are too sophisticated to be formulated and listed in photography literatures.

In contrast to previous works, a bottom-up aesthetic modeling methodology is proposed to comprehensively analyze and mimic the essence of aesthetics. The idea of the proposed approach is that although photographic styles vary, there still exists an *aesthetic feature library*, which covers most possible implicit rules adhered to by different professional photographers. Based on the idea of constructing a feature library, a new image representation, with *bag-of-aesthetics-preserving* (BoAP) features, is proposed. The proposed BoAP features in the aesthetic feature library are extracted from different feature spaces to model color, texture, saliency, and edge information jointly as shown in Fig. 2(a).

Given the constructed aesthetic feature library with BoAP features, a *preference-aware aesthetic model* consisting of weighted discriminative features is built by learning with a database composed of user-preferred photographs. This model is then employed in an online aesthetic view finder at the second stage as Fig. 2(b) illustrates. Based on the mature and robust real-time software [10] and hardware [11], [12] implementation for object (face) detection, the proposed system is designed with similar operations to share the same hardware architecture of the built-in object (face) detection function. Therefore, the cost and power consumption overhead to embed this engine in digital cameras can be reduced.

The contributions of this work can be listed as follows: 1) a new bag-of-aesthetics-preserving (BoAP) image representation is developed to cover most possible implicit aesthetic rules, which are too complicated to be listed in photographic literatures; 2) the proposed BoAP features can model both the absolute and relative relations among image patches (i.e., absolute features and contrast features); 3) an adaptive learning framework is presented for different photographic styles of different professional photographers; 4) an efficient solution to the aesthetic view finding problem is proposed by formulating it as an *object detection problem* [10].

This paper is organized as follows. A brief introduction to the related works and their limitations is given in Sec. II. Next, the details of the the aesthetic modeling procedure, the first stage of the proposed system as Fig. 2(a) illustrates, are introduced in Sec. III and IV. In Sec. III, an aesthetic feature library with BoAP features is developed; in Sec. IV, the construction procedure for the preference-aware aesthetic model is introduced. Sec. V then presents the process of the online aesthetic view finding illustrated in Fig. 2(b) and its real-time performance. The objective and subjective experimental results and analyses are discussed in Sec. VI. Finally, the conclusion and future work are presented in Sec. VII.

II. Related Work

A. Aesthetic Modeling

Most existing works [1], [2], [4], [6]–[9], [13], [14] adopt a rule-based feature extraction approach, which is a top-down process, to implement heuristic photographic guidelines individually. These features are called rule-specific features in this paper. Some examples of the rule-based extraction approaches are given as follows. One of the famous composition rules, the "Rule of thirds" shown in Fig. 3(a)), is modeled based on encoding the position information of salient subject(s) obtained from a saliency map with respect to the four stress points or with respect to the background regions. The other rule called "Visual Weight Balance," as shown in Fig. 3(b), is modeled by calculating the ratios of different regions [2]. In addition to composition rules, color and texture information are used to model the rules, including color harmony, colorfulness, and emotions by [6]. They propose to

(a) Rule of Thirds. (b) Visual Weight Balance.

Fig. 3. Two famous aesthetic rules in the photography literature. (a) In this photograph, the primary subject's center of mass (indicated by red cross-hair) is placed at the position near one of the four stress points (indicated by yellow cross-hairs) in order to satisfy the "Rule of Thirds." (b) According to the rule of "Visual Weight Balance," the visual weights of different regions (indicated by red and green dashed lines) satisfy the Golden Ratio (approximately equal to 1.61803). We will show that such ad-hoc rules can be systematically included in our view recommendation by the proposed learning framework and bag-of-aesthetics-preserving features.

model these aesthetic principles by extracting features for the foreground and background separately.

While the above works extract features at the scope of the whole image and lose the within-image information, Cheng et al. [3] propose to decompose the image into several patches first and model the joint distributions for patch pairs and the spatial distribution for each single patch. The maximum number of patches they model concurrently is two. The image representation approach proposed by [3] models the absolute values, rather than the contrast values of the color and texture distributions. However, [15], [16] show that humans are more sensitive to contrast. In addition, all of co-occurring elements in a photograph have correlations; therefore, they should be modeled altogether rather than merely as pairs of elements.

To summarize, there are three major limitations of the aforementioned works shown as follows. 1) Composition rules are encoded using the position information for a single salient subject and background without the color and texture information. The other rules are modeled by considering the global image information in most papers. 2) Previous computational approaches focus on implementing simple heuristic guidelines individually. Therefore, these ad hoc rule-specific features with low-dimensionality may lose some aesthetic characteristics. That is, they may not provide complete models for professional photos. 3) The methods they propose model the absolute feature distribution while the contrast information is not taken into account.

To overcome the aforementioned limitations, a bottom-up aesthetic modeling approach, called bag-of-aesthetics-preserving image representation, is proposed to model the absolute and relative relations among image patches.

B. View Finding Algorithm

Several view recommendation algorithms have been proposed by other papers. Some works recommend views based on salient areas or visual attention models from a wide view or a continuous view sequence [9], [14]. In contrast to targeting the detection of salient areas, Cheng et al. [3] propose a view finding system based on an omni-range context modeling method. The images in the large-scale professional photo database are first segmented into basic patches. Second, large-scale training photos are pre-segregated into sub-topics, such as sunsets, fields, beaches, etc. Then they use Gaussian mixture models (GMM) to model the aesthetics. As a wide-view image is fed into their system, they use a sampling based optimization method for optimal view finding. This system can function as an intelligent professional view guider based on real-time view quality assessment.

Another work presented by Chang et al. [13] is a stochastic optimization search algorithm aimed at looking for good views within a panoramic scene and choosing suitable reference images from masterpiece photographs. Given any initial location in the panoramic scene, their algorithm is able to suggest a better view that would often yield professional-like photo composition. However, there are some limitations of this algorithm: It solves the problem by finding the best view that would look very similar to the exemplar chosen in the selected famous photographer's photo gallery. That is, the photographic style of that specific selected photographer is not learned by their approach. That is, when a wide view image is fed into this system, their algorithm could not handle the cases that the selected photographer did not take pictures with the same surroundings. Moreover, in their experimental results, their system could not run in real-time.

In contrast with the works presented above, a view finding algorithm that uses a real-time boosting-based detection process [10] is proposed based on the bag-of-aesthetics-preserving features.

III. Aesthetic Feature Library Construction

The aesthetic feature library in Fig. 2(a) is proposed to encompass all the possible implicit rules that might be adhered to by different professional photographs.

The proposed library is constructed by a bottom-up approach, which includes two steps: 1) partitioning an image into multi-size image patches and 2) extracting the bag-of-aesthetics-preserving (BoAP) features by systematically manipulating the patches in different feature spaces. The extracted BoAP features constitute the library.

In contrast to the set of Haar-like features utilized in an object (face) detection task [10], which is a large set of exhaustive features extracted from all the combinations of x and y coordinates per sub-window in an image (e.g. 45,396 features are extracted for a 24×24 window), the proposed BoAP features consider the spatial layout of the entire view for the aesthetic composition, as shown in Fig. 4. That is, for the photographic composition, the features should be extracted based on basic patches at the scope of the entire view, while the Haar-like features for object detection are extracted at the scope of sub-window containing target objects (faces) to merely capture the subtle intensity differences between facial features or features among objects. In addition, the Haar-like feature extraction is applied on a grayscale image, which is sufficient for object (face) detection. In contrast, the proposed BoAP features are extracted from different feature spaces to model color, texture, saliency, and edge information jointly.

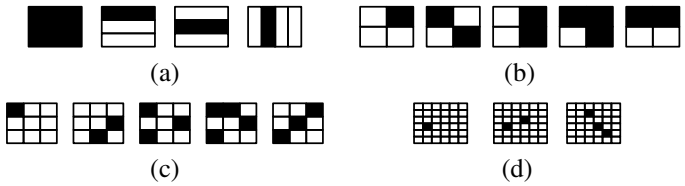


Fig. 4. Some examples in four types of geometric compositions: (a) Global line composition. (b) 2×2 composition. (c) 3×3 composition. (d) 6×6 composition. The black patches can be viewed as the pre-defined foreground regions while the white patches can be seen as the background regions.

A. Multi-Resolution Image Decomposition

In order to mine the relations among multiple subjects in aesthetic landscape photographs, the first step of aesthetic modeling is to partition the images into several basic patches prior to extracting BoAP features.

Both [3] and [17] propose to extract the features for multiple subjects in photographs for aesthetic modeling. Nishiyama et al. [17] utilize the k-means clustering algorithm as the decomposition method. In [17], images are partitioned into 13 pre-defined segments. Another decomposition method utilized by Cheng et al. [3] is a graph-based segmentation algorithm [18]. In [3], images are segmented into 50 to 70 homogeneous patches in average.

In contrast, a multi-resolution grid-based decomposition method, which partitions an image into predefined basic patches of different grid sizes, is proposed to model views in a coarse-to-fine manner. There are four simple types of geometric compositions defined in this paper: global line composition, 2×2 composition, 3×3 composition, and 6×6 composition, and some examples are illustrated in Fig. 4. These black and white patches are further utilized to extract BoAP features in Sec. III-B. In contrast to the linear-time decomposition methods utilized in [17] (k-means) and [3] (graph-based), the proposed approach divides images into grids of different sizes in constant time with very low timing overhead. For Fig. 4(a), the images are not divided or are divided into several horizontal/vertical lines at the global scope. For Fig. 4(b)–(d), denoted as n×n composition, the images are partitioned into n×n spatial grids, i.e., n×n basic image patches. The concept of this image decomposing step is to utilize these coarse-to-fine geometric compositions to further mine the correlation among all the decomposed patches. Hence, structural arrangement in the views can be captured at multiple resolutions. The proposed method has the following advantages: it considers the spatial layout of views at multiple resolutions; patches can be quickly generated in constant time.

B. Bag-of-Aesthetics-Preserving Feature Extraction

In this section, a detailed introduction to the process of extracting BoAP features after an image is transformed into numerous multi-size patches is given. The extracting process is comprised of two steps: 1) the first step is to describe multi-size image patches using feature vectors and 2) the second step is to generate different sets of BoAP features by applying a few simple patch-wise operations on the feature vectors in different feature spaces.

Fig. 5. Illustration of patterns generated by **permutation** $\binom{4}{2}$ in 2×2 composition. The image is decomposed into four patches P, where $P = A, B, C, D$. Each patch is represented by feature vectors $f_k(P)$, where k denotes the kinds of feature images.

1) Feature vectors of image patches: Before extracting BoAP features, each image patch is needed to be represented by feature vectors in different feature spaces. Inspired by the relation between the feature space and its corresponding modeled rules built by prior works (e.g., a saliency map corresponds to rules about composition, contrast, and clarity; HSV or other color spaces correspond to colorfulness and color harmony), we propose to elaborate and analyze photographic aesthetics from different aspects by applying BoAP feature extraction to different feature spaces separately [19]. There are four types of information modeled in the proposed approach: color, texture, saliency, and edge information. To begin with, an image I is transformed into a number of *feature images* I_k, where k denotes the kinds of feature images extracted from the aforementioned information. Based on the feature images I_k, each decomposed patch P in an image I is represented by feature vectors $f_k(P)$. The feature images and the corresponding patch feature vectors for each information type are introduced as follows.

For color information, two kinds of color spaces, RGB and HSV, are utilized. R, G, B, H, S, and V feature images are extracted and analyzed separately [19]. A local binary pattern (LBP) [20] feature image is extracted to model the texture information. Each location is assigned a decimal value by encoding the neighboring eight pixels with respect to the pixel value at that location. In order to model the importance of each pixel and the locations of salient regions in an image, a saliency map [21] is also employed as a feature image.

For feature images I_k, where k = R, G, B, H, S, V, LBP, and saliency map, the first two moments, mean (μ) and variance (σ^2) descriptors, are utilized to describe the statistic properties of the decomposed patches within an image. The mean and variance within the patch P can be calculated from the following equations:

$$\mu_k(P) = \frac{1}{N}\sum_{i=1}^{N} I_k(i), \tag{1}$$

$$\sigma_k^2(P) = \frac{1}{N-1}\sum_{i=1}^{N} (I_k(i) - \mu_k(P))^2, \tag{2}$$

TABLE I
ABSOLUTE AND CONTRAST FEATURES OF TOP-LEFT PATTERN IN **FIG. 5**.

Features Type	Operation
absolute	$\frac{1}{2}(f_k(A)+f_k(D))$
contrast$_1$	$f_k(A)-f_k(D)$
contrast$_2$	$f_k(A+D)-f_k(B+C)$

$$f_k(P) = (\mu_k(P), \sigma_k^2(P)), \tag{3}$$

where N is the pixel number in the patch P and $\forall i \in P$; k is the feature image type. Therefore, for each basic patch P in a specific feature image I_k, it is represented by a two-dimensional vector $f_k(P)$.

For edge information, each basic patch P in an image is represented by a histogram of oriented gradient (HOG) [22]. HOG features are obtained by accumulating gradient values over pixels in one patch into a histogram of D gradient direction. There are three kinds of bin number D (i.e., defined by orientations) extracted in the experiment, and the value of D is set to 2 (horizontal/vertical), 4, and 8. That is, for k = HOG, each patch P is represented and described by a $D-$dimensional vector $f_k(P)$, where D =2, 4, and 8.

2) Generation of BoAP Features by patch-wise operations: Up to now, lots of feature vectors representing multi-size patches have been extracted in each analyzed feature space. For each geometric composition type illustrated in Fig. 4, some simple and fast *patch-wise operations* are defined to generate thousands of BoAP features including **absolute** and **contrast** information. By applying the defined patch-wise operations on $f_k(P)$, a set of BoAP features is extracted for one type of feature image I_k (denoted as BoAP$_k$ features). First, a **permutation** operation is utilized to generate lots of patterns by rearranging a number of patches in an organized order. That is, there are a number of $\binom{n\times n}{m}$ patterns to be generated for one type of n×n composition ($n \times n$ stands for the total patch number, m for the black patch number, and $n \times n - m$ for the white patch number in Fig. 4). The m black patches can be viewed as the predefined foreground regions, and the $n \times n - m$ white patches can be seen as the background regions. Taking the 2×2 composition type for example, the permutation operator first generates $\binom{4}{1}$+$\binom{4}{2}$+$\binom{4}{3}$ patterns. Fig. 5 illustrates the feature extraction procedure of one pattern from $\binom{4}{2}$. In this example, an image is partitioned into four patches. Each patch is represented by feature vectors $f_k(P)$, where $P = A, B, C, D$ in this case. Based on applying patch-wise operations on feature vectors of patches, the **absolute** and **contrast** features are extracted per pattern as shown in Table I, where the top-left pattern in Fig. 5 is taken as an example. The **absolute** features model the absolute value distribution of m black patches while the **contrast**$_1$ features model the within-contrast among multiple black patches and **contrast**$_2$ model the relative contrast between multiple black patches and white patches. The contrast is shown more consistent with human visual perceptions [15], [16]. Therefore, different from the two-patches modeling methods [3], [17], the proposed approach not merely models one or two visual elements concurrently but multiple patches at different image resolutions simultaneously.

All sets of BoAP features extracted through above procedure from different spaces (i.e. color, texture, saliency, and edge) constitute the aesthetic feature library. For k = R, G, B, H, S, V, LBP, and saliency map, a set of BoAP$_k$ features is of 3072 dimensions. For k = HOG, a set of BoAP$_k$ features with $939\times D$+561 dimensions is extracted per image, where D= 2, 4 and 8.

The advantages of the proposed BoAP image representation are: 1) images can be viewed in multiple resolutions by the proposed decomposition method; 2) images can be described from different aesthetic aspects, including color, texture, saliency, and edge, by applying the proposed patch-wise operations on all feature spaces; 3) contrast information, which humans are more sensitive to, is taken into consideration (i.e., Table III(b) demonstrates that contrast features are more effective than absolute ones); 4) the computation process is highly efficient via patch-wise operations.

IV. AESTHETIC MODEL LEARNING

With thousands of the BoAP features in the library, a method for efficiently and adaptively creating a model that best represents a specific user-favorite photographic style becomes an important issue. The offline learning flow chart is illustrated in Fig. 2(a).

If the prepared photo database for the learning process is a collection of high quality photographs taken from different photographers gathered on the Internet, the learned model is a generalized one that will assess the aesthetic quality of photos from a public point of view. If the database is a photo gallery created by a well-recognized or user-favorite photographer, the model that uses this gallery for learning will judge the quality of a photo based on the rules normally followed by the corresponding popular photographer. Moreover, a collection of aesthetic photos of scenery taken by the users themselves can be utilized to construct a personalized aesthetic model, which comprises a collection of subjective photo quality judging criteria based on users' preferences. Therefore, users can load different training photo databases to adaptively create a generalized/personalized aesthetic model based on their tastes (i.e., preference-aware model).

There have been many learning metrics used for constructing aesthetic models (e.g., SVM, GMM, Bayes, Adaboost, etc.) in photo quality assessment problems. Because different photographers have their own photographic styles and follow different aesthetic principles, the learning metric should allow models to adaptively learn based on the provided training photos. Because of this, similar to its use in tasks such as detection and recognition [10], [23], [24], Adaboost is utilized for discriminating feature selection in this paper to explore and analyze the relation between the extracted BoAP features and photo database provided by a user. Every feature in the library has its own judging criterion and discriminating power in assessing the qualities of different types of photos; therefore, different features may be chosen to constitute the learned model based on the different sources for the photo collections. Moreover, if performance in photo quality classification tasks is considered, Adaboost is also the most suitable choice [6].

In the algorithm, each of the BoAP features in the library serves as a hypothesis (denoted as h) and corresponds to a weak classifier. The selection process chooses the most discriminative feature $h_n(\cdot)$ with the minimum classification error in differentiating high quality and low quality photos at each iteration, and assigns a weight α_n to it at the same time based on its discriminating power. Finally, all N selected features $h_n(\cdot)$ are integrated and boosted together to form a strong aesthetic model M using the following equation:

$$M(\Phi) = \sum_{n=1}^{N} \alpha_n \cdot h_n(\Phi), \tag{4}$$

where Φ represents a photo in the training database; M is the learned preference-aware model, a weighted combination of all of the selected discriminative features $h_n(\cdot)$ from aesthetic feature library. The final learned model M in (4) is the formulized user-favorite photographic style. That is, photograph represented by Φ can be assessed by a set of selected aesthetic features the user-favorite photos adhere to. The value of $M(\Phi)$ (i.e., confidence value, which is related to the margin) can be interpreted as the aesthetic quality score of photograph. Model M is analyzed in Sec. VI-B by measuring the performance in a photo aesthetic quality classification problem. It is further utilized in the proposed view finding task in Sec. VI-C. The details of the proposed view finding algorithm are introduced in Sec. V.

Based on the learned model M, high and low aesthetic quality photos can be discriminated by the following equation:

$$H(\Phi) = \text{sign}(M(\Phi)), \tag{5}$$

where $H(\Phi) = 1$ indicates that the testing photo represented by the proposed BoAP features possesses the high aesthetic quality while $H(\Phi) = -1$ means the testing photo is of low aesthetic quality. Experimental results of the binary classification conducted in Sec. VI-B show that the proposed BoAP features outperform the state-of-the-art rule-specific features utilized in previous works.

Note that the symbol Φ in the view finding process (cf. Sec. V and Sec. VI-C) denotes a sub-view candidate $I(x, y, s|W)$ trimmed from a wide-angle view W, while in the photo aesthetic quality classification it denotes the entire testing photo (cf. Sec. VI-B).

V. Online Aesthetic View Finding

With the learned model M in (4), the efficient online aesthetic view finding algorithm [10] is developed in this paper. The procedure is illustrated in Fig. 2(b).

By modeling the moving of camera views with translating along four directions, right, left, up, and down, in the wide-angle view W and approximating zooming with scaling the view W, the problem of view finding is skillfully formulated as the object detection problem [10]. Therefore, given a wide-angle view W as the system input, the goal of the proposed system is to find the optimal scale-space position (x^*, y^*, s^*) of the most aesthetic view $I(x^*, y^*, s^*|W)$ in real-time. Based on the real-time online object detection process from Viola and Jones [10], the learned model M serves as an aesthetic view finder to evaluate lots of sub-view candidates $I(x, y, s|W)$ trimmed from W.

To begin with, the wide-angle view W is first transformed into a number of feature images. Similar to the implementation in [10], integral images [10] and integral histograms [25] are utilized as the intermediate representations for feature images to improve the feature extraction efficiency. Based on these intermediate representations for W, BoAP features of each sub-view candidate $I(x, y, s|W)$ can be computed rapidly. Integral histograms are utilized for k =HOG while integral images are utilized for other feature images except for k =HOG.

A map called *quality-score map* $Q(x, y, s)$ is defined in the proposed algorithm to indicate aesthetic qualities of all sub-view candidates. Each sub-view candidate is represented by the BoAP features (denoted as $\Phi_I(x, y, s|W)$). The learned model M makes a judgment to its aesthetic quality based on a few discriminative dimensions selected from the evaluated sub-view's BoAP features $\Phi_I(x, y, s|W)$ and outputs a corresponding aesthetic quality score. The estimated score is then entered in the quality-score map $Q(x, y, s)$ at its corresponding scale-space position as illustrated in Fig. 2(b). The map generation is formulated as follows:

$$Q(x, y, s) = M(\Phi_I(x, y, s|W)), \tag{6}$$

The idea of utilizing the confidence map (i.e. called quality-score map Q in this work) to measure the aesthetic degree of all evaluated views is motivated by the work proposed by Grabner et al. [26], where the confidence map is utilized for the purpose of real-time object tracking via on-line boosting algorithm [27].

The complete quality-score map Q is derived after the scanning process terminates. The scale-space position of the best view $I(x^*, y^*, s^*|W)$ possessing the highest aesthetic score in the quality-score map Q is determined by:

$$\{x^*, y^*, s^*\} = \arg\max_{x,y,s} Q(x, y, s), \tag{7}$$

Based on the position of the current view (x_c, y_c, s_c) and the suggested view (x^*, y^*, s^*), the displacement $(\Delta x, \Delta y)$, which is calculated by $(x^*, y^*) - (x_c, y_c)$, guides the user by panning up, down, left, or right, while s^* instructs the user how to adjust the zoom lens (zoom in/out) to take an aesthetic picture in real-time (cf. Fig. 1). The user studies conducted in Sec. VI-C demonstrate that the views suggested by the proposed algorithm are consistent with users' preferences.

The system was performed on a PC with Intel Quad-Core CPU Q9400 (2.66GHz) and 2GB memory. In our current implementation with C++, the proposed system can process a 960 by 720 pixel image in about 0.072 seconds (5 kinds of candidate view's size and a step size of 20 pixels). Actually, the VLSI hardware implementation is in our ongoing process to further accelerate the system.

VI. Experimental Results

To evaluate our system comprehensively from both objective and subjective aspects, the conducted experiments cover two parts: photo aesthetic quality classification (objective) and user study (subjective).

TABLE II
PHOTO AESTHETIC QUALITY CLASSIFICATION ACCURACY OF BoAP$_k$ FEATURES WITH DIFFERENT COMBINATIONS OF k AND THE RULE-SPECIFIC FEATURES.

Color						Texture	Saliency	Edge		
R	G	B	H	S	V	LBP	Saliency	HOG (D=2)	HOG (D=4)	HOG (D=8)
82.73%	81.67%	83.14%	80.77%	79.46%	82.65%	85.02%	75.86%			
R+G+B (RGB Color Space)			H+S+V (HSV Color Space)					75.86%	88.71%	90.18%
87.15%			90.02%							
Combined All: Color (HSV) + Texture(LBP) + Saliency + Edge (HOG:D=8)										
92.06%										
State-of-the-Art Rule-Specific Features ([1], [5], [6])										
83.63%										

A. Database Collection

Most of the previous works evaluate their photo aesthetic quality assessment systems using their own private photo collections. The ideal dataset for our purpose is required to contain classified high/low aesthetic quality scenic photos based on the various styles of different photographers. Nevertheless, there is still no publicly available dataset suitable for our application. In order to compare our work with state-of-the-art method, we utilize the same public available dataset [1] including 6000 highest-rated and 6000 lowest-rated photographs collected from photograph contest website *DPChallenge* (also used in [5] and [6]). Because we focus on mining and characterizing the underlying aesthetic rules for scenic photos, only photos from the scenic category of that database are used in the experiments. In the following experiments, our dataset is split into training and testing set. The training set has 340 high quality photos and 680 low quality photos, while the others (up to 3000 photos) constitute the testing set.

Because this dataset is a mixed photo collection containing all styles of scenic photos assessed by a massive number of observers, it is utilized in the experiments to prove that the proposed system not only builds a personalized aesthetic model but also has the ability to build a generalized model based on massive tastes and learn the implicit rules adhered to by all high quality photos.

Moreover, because high quality photos may have some underlying aesthetic properties in common, in contrast to the previous work by Cheng et al. [3], a prior knowledge of the topic represented by a photo is not utilized. The database utilized in this paper contains all kinds of landscape classes without performing scene topic classification a priori.

B. Photo Aesthetic Quality Classification

To evaluate the representation effectiveness of the proposed BoAP features, the learned model H in (5) serves as an aesthetic quality classifier to perform photo quality classification, which the most common way to approve the effectiveness of the features adopted by related works. In this section, an experiment of photo aesthetic quality classification is conducted. Photos in the database (including training and testing photos; cf. Sec. VI-A) are represented by Φ in (5) and are further classified into two classes by the quality classifier H.

1) Results with Different Feature Images: To explore the effectiveness of the proposed BoAP features in different aspects,

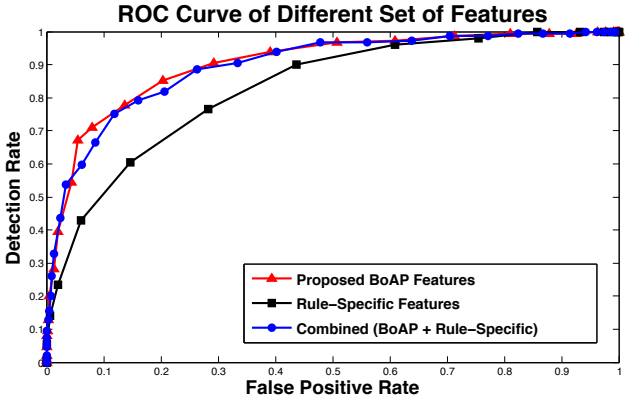


Fig. 6. Comparison in the classification performance of the rule-specific features, the proposed BoAP features, and the combined (i.e. rule-specific + BoAP) features. The result shows that the proposed BoAP features outperform the state-of-the-art rule-specific features significantly. The performance of the combined features, including rule-specific features into the proposed BoAP feature library, does not improve. It shows that the proposed BoAP features encompass the professional photographic guidelines (i.e. encoded in the rule-specific features).

BoAP$_k$ features are extracted on single and combined feature images (I_k). For each set of BoAP$_k$ features, a corresponding aesthetic model is trained to evaluate its accuracy in binary classification with the dataset introduced in Sec. VI-A. The experimental results show that the model with HSV color space features (90.02% accuracy), obtained from summing over BoAP$_k$ features, where k =H, S, and V (3072×3 dimensions), is more effective in describing photographic aesthetics than that with RGB color space features (87.15% accuracy), which are obtained from summing over BoAP$_k$ features, where k =R, G, and B (3072×3 dimensions). The final combined BoAP features are obtained by adding all features from color (HSV), texture (LBP), saliency, and edge (HOG with bin number=8) together (23,433 dimensions). Based on the final combined BoAP features, the corresponding built model can achieve the highest accuracy 92.06%, which beats the performance achieved by the rule-specific features (83.63% accuracy). We implement the rule-specific features combining the works in [1], [5], [6]. Classification accuracy of BoAP$_k$ features with different combinations of k and the state-of-the-art rule-specific features [1], [5], [6] are listed in Table II. The final built aesthetic model with combined BoAP features from all spaces (92.06% accuracy) is further utilized in the following experiments.

2) Evaluation with ROC Curve: To further compare the classification performance with the rule-specific features and

TABLE III
CHARACTERISTICS OF THE FINAL BUILT MODEL WITH 110 SELECTED FEATURES.

(a) Percentage of features from different spaces.

Feature Image	Percentage
Color(HSV)	35.09%
Texture(LBP)	5.26%
Saliency Map	15.79%
Edge(HOG:D=8)	43.86%

(b) Percentage of absolute and contrast features.

Feature Type	Percentage
absolute	24.69%
contrast	75.31%

the proposed BoAP features, the receiver operating characteristic (ROC) curves are plotted by adjusting the threshold of the learned aesthetic model with Adaboost from $-\infty$ to ∞. Adjusting the threshold to ∞ will yield a detection rate of 0.0 and a false positive rate of 0.0. Adjusting the threshold to $-\infty$, however, increases both the detection rate and false positive rate. From Fig. 6, the proposed BoAP features achieve better performance than the rule-specific features. As Fig. 6 demonstrates, if these two types of features are combined by adding the rule-specific features into the proposed BoAP feature library, the performance is almost the same as that with the proposed features. Therefore, a conclusion can be drawn from experiments that the proposed features encompass guidelines listed in photographic manuals (i.e. encoded in the rule-specific features). That is, the rule-specific features extracted by rule-based method are well covered by ours.

3) Effective BoAP Features: Characteristics of the final built model with combined BoAP features from all spaces are analyzed in this section. The highest accuracy (92.06%) with minimum testing error of this model occurs at the iteration number = 110 (i.e., the number of weak classifiers). That is, 110 discriminative BoAP features selected from the library constitute the final built model. Table III(a) shows the percentage of selected features from different feature spaces. It reveals that color and edge (HOG with bin number=8) information are more effective than others. Table III(b) demonstrates that among the 110 selected features, **contrast** features are more effective than **absolute** features. This result is consistent with the fact that humans are more sensitive to contrast [15], [16]. The top five selected features among 110 features are listed in Table IV in order. From these selected features, it is observed that contrast (the $1^{st}, 2^{nd}, 4^{th}, 5^{th}$ row in Table III(b)) and absolute (the 3^{rd} row) features are both important in describing aesthetic characteristics while contrast features are more important. Besides, the relation between multiple patches and background (the $1^{st}, 2^{nd}, 5^{th}$ row) and the relation among multiple patches (the $3^{rd}, 4^{th}$ row) are both needed to be taken into consideration.

C. User Studies for View Recommendation

Because aesthetic assessments are highly subjective, this part is needed to prove the presented system with the proposed features (which have been approved in the classification tasks) can work successfully by considering the subjectivity of aesthetics. The final aesthetic model M with the combined all BoAP features (92.06% accuracy) functions as a view finder in this experiment.

16 wide-angle views have been collected on the Internet, and then the proposed view finding algorithm is applied on them. For each wide-angle view, the view with the highest aesthetic score in quality-score map is found as the best view and the view with the lowest aesthetic score is found as the worst view in our evaluation. Because the model built in Sec. VI-B is learned through the training dataset (cf. Sec. VI-A) with massive human tastes, it is a generalized aesthetic model. We anticipate that the views suggested by the proposed algorithm are consistent with the user studies.

There are 24 subjects (testers) involved in the experiment. The training database in our experiments was actually acquired by crawling from a photo contest website, DPChallenge. Each good photo of training data is rated by at least a hundred users in its community. The users may contain both amateurs and professional (or experienced) photographers. Since the built aesthetic model depends on the training database, the testers in the conducted user study are composed of both types of photographers to approve our system for the consistency in training and testing. For each image pair (the best view and the worst view) found in 16 wide-angle views for testing, they make a decision about which one has the best aesthetic characteristics, and then the results are compared with the corresponding quality scores of each image pair evaluated by the proposed system.

Results, part of which shown in Fig. 7, demonstrate that the best and worst views assessed by the proposed system are consistent with testers' preferences, achieving 81.25% agreements. In Fig. 7, the views with highest aesthetic scores are indicated by red rectangles and the views with lowest aesthetic scores are indicated by green rectangles. Meanwhile, it is observed that the proposed algorithm with the generalized aesthetic model doesn't model ad-hoc professional guidelines explicitly, but the horizontal line obeys the rule of "Visual Weight Balance." Besides, the subject region is placed at a position near one of the four stress points, which follows the "Rule of thirds." However, in some image pairs, the evaluations of the proposed system are not consistent with most testers' preferences. Two examples are illustrated in Fig. 8(a)-(b) (i.e., 15 out of 24 testers voted on the green rectangle as the high aesthetic quality view in Fig. 8(a); 21 out of 24 testers in Fig. 8(b)).

Although the preferences of testers conflict with the assessments of the proposed system, we can see that the placements of the horizontal lines in the views with highest evaluated quality scores satisfy the rule "Visual Weight Balance." These inconsistent estimations might be due to that the subjectivity of aesthetic evaluation. Therefore, putting aside the highly subjective evaluation results from the testers, the two suggested views still meet the expectations.

The user study conducted here provides another way to prove that the proposed system not only builds a personalized aesthetic model but also has the ability to build a generalized model based on massive tastes and learn the implicit rules

TABLE IV
THE TOP FIVE SELECTED FEATURES AND THEIR MEANINGS IN THE FINAL BUILT MODEL.

Feature Image	Feature Type	Modeled Relation	Composition
Edge(HOG:D=8)	contrast	patches vs. background	3×3
Edge(HOG:D=8)	contrast	horizontal line vs. background	Global line
Edge(HOG:D=8)	absolute	among patches	3×3
Color(Saturation)	contrast	among patches	3×3
Color(Hue)	contrast	patches vs. background	3×3

adhered to by the high quality photos.

VII. CONCLUSION AND FUTURE WORK

A preference-aware view recommendation system is proposed in this paper. The proposed system comprises two stages: offline aesthetic modeling stage and online aesthetic view finding process. For the stage of aesthetic modeling, a new image representation with bag-of-aesthetics-preserving (BoAP) features is developed, and the preference-aware model is built by a learning process with database composed of user-preferred photographs. The experimental results show that the proposed features (92.06% in accuracy) outperform the state-of-the-art rule-specific features (83.63% in accuracy) in the photo aesthetic quality classification task. In addition, these experiments further demonstrate that the rule-specific features are covered by the proposed features. Meanwhile, it is observed from experiments that the features extracted for contrast information are more effective than those for absolute information, which is consistent with the fact that humans are more sensitive to contrast in the visual perception systems. For the second stage, the robust real-time object detection procedure is skillfully reformulated to the proposed view finding process with the quality-score map. Based on the suggested view with the highest quality score, users can adjust the camera to take aesthetic pictures according to their favored photographic styles, which is approved in the user studies.

In the future, the aesthetic photo ranking problem will be addressed. The RankBoost algorithm will be adopted to generate a list of relative ranking for a training or testing data. Moreover, some experiments will be conducted with the proposed BoAP features on not only scenic photos but also other general photographs, such as portrait images with human faces or objects, to analyze the proposed features' limitations and further improve the feature design.

REFERENCES

[1] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," *IEEE CVPR*, vol. 1, pp. 419–426, 2006.
[2] S. Bhattacahrya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *ACM MM*, 2010, pp. 271–280.
[3] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *ACM MM*, 2010, pp. 291–300.
[4] W. Jiang, A. Loui, and C. Cerosaletti, "Automatic aesthetic value assessment in photographic images," in *IEEE ICME*, 2010, pp. 920 – 925.
[5] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system," in *ACM MM*, 2010, pp. 211–220.
[6] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008, pp. 386–399.
[7] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.
[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006, pp. 7–13.
[9] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *ACM MM*, 2009, pp. 541–544.
[10] P. Viola and M. Jones, "Robust real-time object detection," in *Int. J. Comput. Vision*, 2001.
[11] K. Khattab, J. Dubois, and J. Miteran, "Cascade boosting-based object detection from high-level description to hardware implementation," *EURASIP J. Embedded Syst.*, vol. 2009, pp. 2:1–2:12, January 2009.
[12] T. Theocharides, G. Link, N. Vijaykrishnan, M. Irwin, and W. Wolf, "Embedded hardware face detection," in *International Conference on VLSI Design*, 2004.
[13] Y.-Y. Chang and H.-T. Chen, "Finding good composition in panoramic scenes," in *IEEE ICCV*, 29 2009.
[14] S. Banerjee and B. Evans, "In-camera automation of photographic composition rules," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1807 –1820, 2007.
[15] B. T. Barrett, I. E. Pacey, A. Bradley, L. N. Thibos, and P. Morrill, "Nonveridical visual perception in human amblyopia," *Investigative Ophthalmology Vis. Sci.*, 2003.
[16] A. McNamara, "Visual perception in realistic image synthesis," *Comput. Graph. Forum*, vol. 20, no. 4, pp. 211–224, 2001.
[17] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *ACM MM*, 2009, pp. 669–672.
[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, pp. 167–181, September 2004.
[19] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE J. of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 236 –252, 2009.
[20] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.
[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2007, pp. 545–552.
[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, vol. 1, 2005, pp. 886 –893 vol. 1.
[23] T. Darrell and K. Wohn, "Pyramid based depth from focus," in *IEEE CVPR*, Jun. 1988, pp. 504 –509.
[24] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416 –431, 2006.
[25] F. Porikli, "Integral histogram: a fast way to extract histograms in cartesian spaces," in *IEEE CVPR*, vol. 1, 2005, pp. 829 – 836 vol. 1.
[26] H. Grabner and H. Bischof, "On-line boosting and vision," in *IEEE CVPR*, vol. 1, 2006, pp. 260 – 267.
[27] N. Oza, "Online bagging and boosting," in *IEEE SMC*, vol. 3, 2005, pp. 2340 – 2345 Vol. 3.

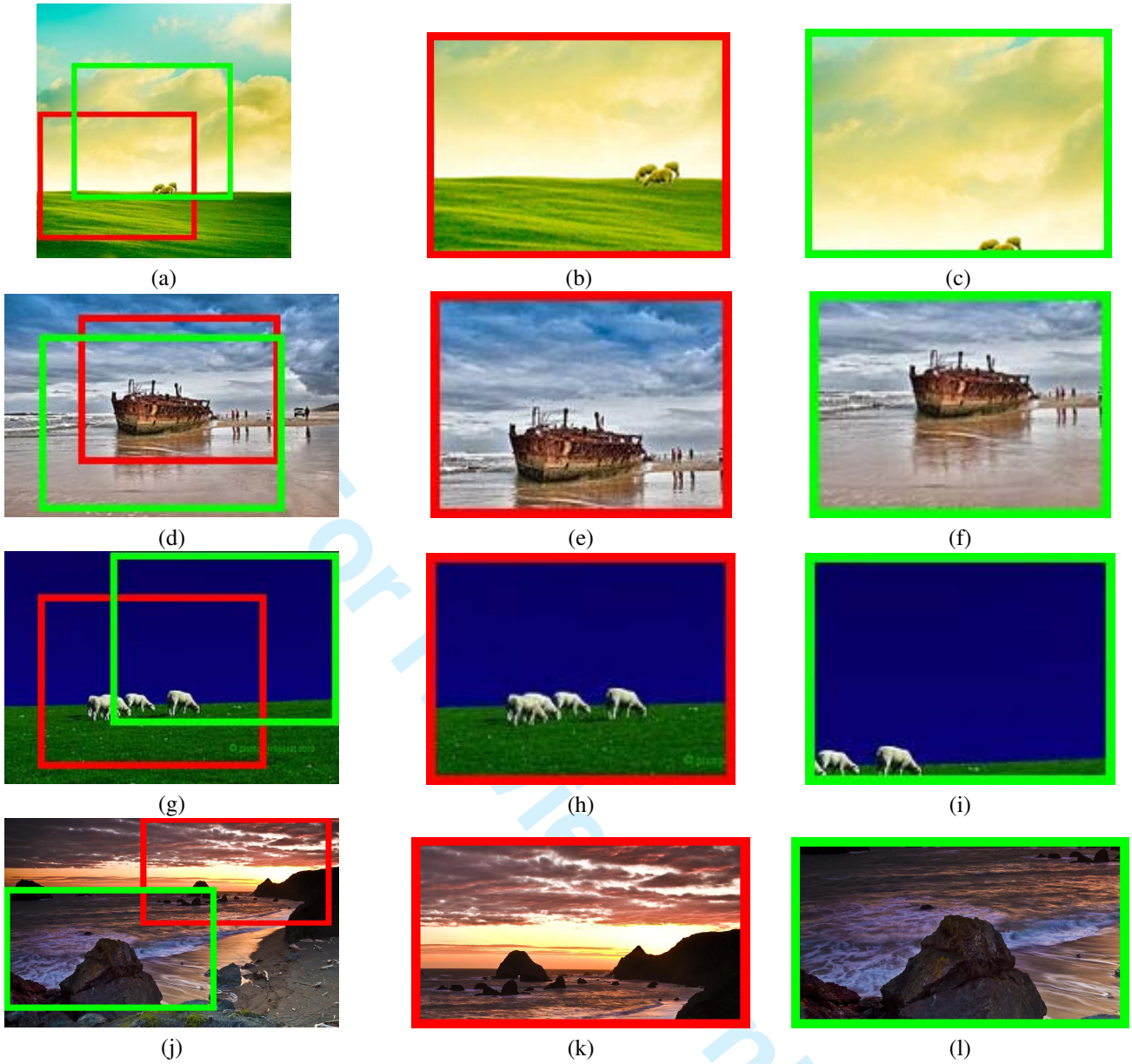


Fig. 7. (a)-(l) are the results of the proposed view recommendation system: Photos in the first column are the wide-angle views for testing; photos in the second and third columns are the views with the highest (indicated by red rectangles) and lowest (indicated by green rectangles) aesthetic quality scores.

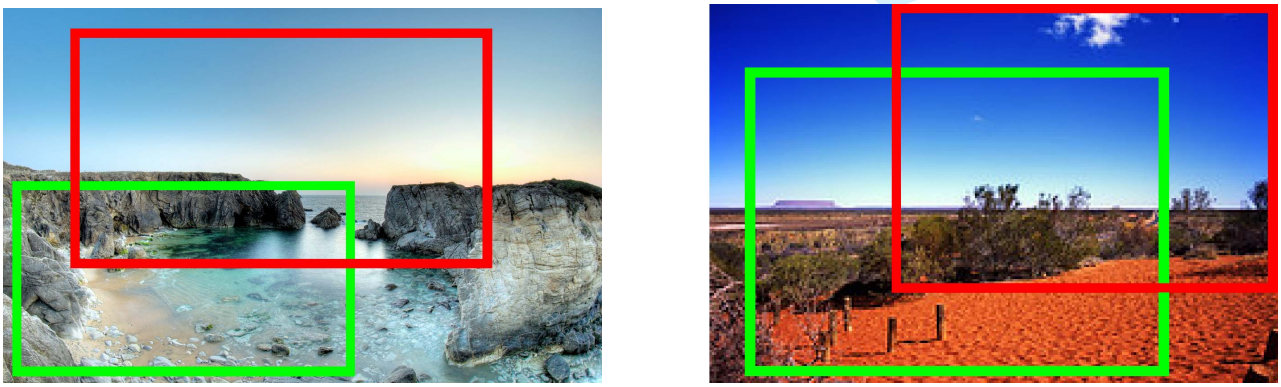


Fig. 8. Illustration of the two image pairs estimated inconsistently. The red and green bounded views are obtained as in Fig. 7. The views indicated by red rectangles are evaluated as high quality ones by the proposed algorithm but rated lower by most testers.

Hsiao-Hang Su received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2010. She is currently working toward the M.S. degree in the Media IC and System Laboratory, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan. Her research interests include multimedia analysis, machine learning, and associated VLSI architectures.

Tse-Wei Chen (S'07-M'11) received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2006, and the Ph.D. degree from the Graduate Institute of Electronics Engineering (GIEE), NTU, in 2010.

In 2010, he became a Foreign Joint Researcher with the Graduate School of Information Science, Nagoya University, Nagoya, Japan. His research interests include computer vision, pattern recognition, machine learning, and associated VLSI architectures.

Chieh-Chi Kao received his B.S. degree in Electrical Engineering, from the National Taiwan University (NTU), Taipei, Taiwan, in 2009. He is currently a research student in the Media IC and System Laboratory, Graduate Institute of Electronics Engineering, NTU. His research interests are in computer vision, machine learning, multimedia analysis, and related VLSI architecture.

Winston H. Hsu received the Ph.D. degree from the Department of Electrical Engineering, Columbia University, New York, NY. He is an Associate Professor in the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, since February 2007. Prior to this, he was in the multimedia software industry for years. He is also with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. His research interests include multimedia content analysis, image/video indexing and retrieval, machine learning and mining over largescale databases.

Shao-Yi Chien (S'99–M'04) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, in 1999 and 2003, respectively. During 2003 to 2004, he was a research staff in Quanta Research Institute, Tao Yuan Shien, Taiwan. In 2004, he joined the Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, as an Assistant Professor. Since 2008, he has been an Associate Professor. His research interests include video segmentation algorithm, intelligent video coding technology, image processing, computer graphics, and associated VLSI architectures.