Probabilistic Graphlet Transfer for Photo Cropping

Luming Zhang, Mingli Song, *Member, IEEE*, Qi Zhao, Xiao Liu, Jiajun Bu, *Member, IEEE*, and Chun Chen

***Abstract*— As one of the most basic photo manipulation processes, photo cropping is widely used in the printing, graphic design, and photography industries. In this paper, we introduce graphlets (i.e., small connected subgraphs) to represent a photo's aesthetic features, and propose a probabilistic model to transfer aesthetic features from the training photo onto the cropped photo. In particular, by segmenting each photo into a set of regions, we construct a region adjacency graph (RAG) to represent the global aesthetic feature of each photo. Graphlets are then extracted from the RAGs, and these graphlets capture the local aesthetic features of the photos. Finally, we cast photo cropping as a candidate-searching procedure on the basis of a probabilistic model, and infer the parameters of the cropped photos using Gibbs sampling. The proposed method is fully automatic. Subjective evaluations have shown that it is preferred over a number of existing approaches.**

***Index Terms*— Gibbs sampling, graphlet, probabilistic model, region adjacency graph.**

I. Introduction

PHOTO cropping refers to the removal of an unwanted subject or irrelevant details from a photo, changing its aspect ratio, or the improvement of its overall composition. Conventional photo cropping has been widely used. For example, in the printing industry, a photo is cropped from a panoramic view to enhance its visual aesthetic effects; in telephoto photography, a photo is cropped to enhance the primary subject. However, photo cropping is challenging due to the following three problems. First, the aesthetic features are not well defined, so it is unclear how to preserve the important visual features in the cropped photo. Second, photo assessment is a subjective task, and thus, it is difficult to develop a computational model that automatically measures the quality of each candidate cropped photo. Third, some existing methods require human-computer interaction to obtain an ideal cropped photo.

Photo cropping closely relates to the topic of photo quality assessment. In recent years, several photo cropping and photo quality assessment approaches have been proposed by both perception researchers and computer vision researchers.

Perception researchers utilize visual attention theories and models for evaluating the quality of each region within a photo, and the most visually salient region is recommended as the cropped photo. In particular, researchers on visual attention have developed neuromorphic models that simulate which elements of a visual scene are likely to attract human attention. Given an image, the neuromorphic models compute its saliency map, which topographically encodes the saliency at every location in the visual input by convolving the image with a series of special filters and encoding the response at each pixel location. Then, the image region with the maximum saliency value is recommended as the cropped photo. In [1], Sun *et al.* have proposed a biologically inspired face-sensitive saliency detector to predict visual attention when looking at photos. The difference between the saliency map and the subject mask, *i.e.*, ground truth data obtained from eye-tracking experiments, is used to evaluate the quality of a photo. A top-down personalized photo assessment is then achieved by adjusting the weights of features used in the saliency detection process. You *et al.* [2] have proposed a photo quality assessment approach that is also based on visual attention analysis. The visually salient regions are extracted based on a combination of a bottom-up saliency model and semantic image analysis. Two metrics, peak signal-to-noise ratio and structural similarity, are then computed in the salient regions. Based on the two metrics, a novel photo quality metric is proposed, which adequately exploits the attributes of visual attention information. In [3], Mei *et al.* have extended the visual-attention-based photo quality assessment to a video sequence, and built a comprehensive scheme to model and mine the captured attention of camcorder users. Liu *et al.* [4] have presented a visual attention model to detect the salient regions and prominent lines of each photo. Three measures are defined by the degree of the salient regions, and the prominent lines conform to the basic aesthetic guidelines, such as the rule of the thirds. The three measures are then linearly combined to evaluate the quality of the photo. She *et al.* [5] have proposed the sparse coding [6] of saliency maps to represent each photo. Photos with different semantic context or

Manuscript received March 14, 2012; revised August 4, 2012; accepted September 22, 2012. Date of publication October 9, 2012; date of current version January 10, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 61170142, the National Key Technology R&D Program under Grant 2011BAG05B04, the Zhejiang Province Key S&T Innovation Group Project under Grant 2009R50009, and the Fundamental Research Funds for the Central Universities under Grant 2012FZA5017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph P. Havlicek. *(Corresponding author: M. Song.)*

L. Zhang, M. Song, X. Liu, J. Bu, and C. Chen are with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: brooksong@ieee.org).

Q. Zhao is with the Department of Electrical and Computer Engineering, Sensor-enhanced Social Media Center (SeSaMe), National University of Singapore, 117576, Singapore (e-mail: eleqiz@nus.edu.sg).

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

Digital Object Identifier 10.1109/TIP.2012.2223226

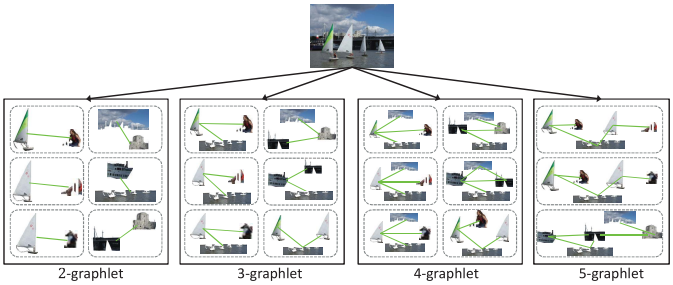


Fig. 1. Graphical illustration of spatial interaction among atomic regions captured by graphlets i-graphlet means a graphlet with i atomic regions and it captures the co-occurrence properties of i atomic regions. When the co-occurrence properties of only two atomic regions are considered, our graphlet-based atomic region's spatial interaction reduces to the omni-range context proposed by Cheng *et al.* [14].

structural information are cropped separately. They firstly used spatial envelop [7] to classify photos into different categories and extracted their saliency maps accordingly. Then for photos in each category, a dictionary is learned by sparse coding of their saliency maps. Given a new photo, the cropped region is selected as the one that can be decoded by the dictionary with the minimum error. Although satisfactory cropping results are empirically observed, visual-attention-based photo cropping methods have three main weaknesses. First, the saliency map used in visual-attention-based photo cropping methods cannot effectively capture the photo aesthetics. As shown in Fig. 1, the spatial interactions of the sky, sailboat, and sea are important visual features that should be preserved in the cropped photo; however, visual-attention-based photo cropping methods fail to capture them. Second, when the photo has little or spurious texture regions, the visual-attention-based model fails to work. Third, there is a semantic gap between the ground truth human data and the existing attention models that are commonly based on low-level features only; in this sense, current attention models have limited predictability of human attention and aesthetic regions.

Computer vision researchers use both low-level and high-level image features[1] for measuring the quality of candidate cropped photos. In [8], Sheikh *et al.* have presented an information fidelity criterion for photo quality assessment by modelling the statistics of natural scenes. Ke *et al.* [9] have designed a group of high-level image features, such as the image simplicity based on the spatial distribution of edges, to imitate people's perception of photo quality. These high-level semantic features are integrated using a probabilistic model for measuring photo quality. Luo *et al.* [10] have proposed a novel photo quality assessment method. Their method first extracts subject regions from a photo, and then formulates a number of high-level semantic features based on the division of the subjects and background. In [11], Yeh *et al.* have proposed a personalized photo ranking system. The system extracts low-level features from professional photos, and then, the weight of each feature is learned based on ListNet [12]. Once the optimal weights are found, photographs can be ranked according to their scores. In addition, to satisfy users' preference, an example-based user interface is developed so that the users can emphasize some features over others by manually adjusting the weights of features. It is noticeable that, the image features used in the above four approaches are designed heuristically, and there is short of evidence that the above features capture the photo aesthetics, such as the spatial interaction of image components in Fig. 1. Besides, Luo *et al.*'s approach relies heavily on a blur detection technique to identify the foreground object's boundary within the frame. This technique works well only with photographs captured by professional single-lens reflex (SLR) cameras that have mechanisms to induce depth-of-field effects and precludes its use with photographs taken using popular point-and-shoot cameras.

Human-computer interaction has been demonstrated to be helpful to further improve the performance of photo cropping by allowing parameters to be tuned for a given photo towards a visually reasonable cropping result. In [15], Bhattacharya *et al.* proposed an interactive framework that improves the visual aesthetics of photos by using spatial recomposition. Users can interactively select a foreground object and the system presents recommendations for where it can be moved in a manner that optimizes a learned aesthetic metric while obeying some semantic constraints. In [16], Santella *et al.* proposed an interactive photo cropping system. The system enables users to look at each photo for a few seconds and records their eye movements accordingly. Then these eye movement data are used to identify the important photo content and further generate cropped photos with any size or aspect ratio. Unfortunately, the human-computer interactive operation of [15] and [16] makes these approaches fail to handle large-scale data sets. Furthermore, the candidate cropped photos are evaluated subjectively so it is difficult to obtain a consistent cropping result for different users.

To avoid the inconvenience brought by those interactive photo cropping methods, Cheng *et al.* [14] have proposed an automatic cropped photo recommendation method by introducing a so-called omni-range context, *i.e.*, the spatial correlation distributions of two arbitrary image patches within an image. To measure the quality of a candidate cropped photo, these omni-context priors are used together with other cues, such as the patch number, to form a posterior probability formulation as a photo quality measure. It is noticeable that the omni-range context only captures the binary spatial interaction of image patches. Higher-order spatial interactions, such as the relative location among the sky, sailboat, and sea described in Fig. 1, fail to be captured. In [17], Nishiyama *et al.* have presented an automatic photo cropping method by training a quality classifier from a large number of photos crawled from the Internet. Their approach first detects multiple subject regions in an image. Each subject region is a bounding rectangle containing the salient part of each subject, such as a treetop and ridge. Then an SVM [18] classifier is trained for each subject region. Finally, the quality of each candidate cropped photo is computed by probabilistically combining the scores of the SVM classifier corresponding to its internal regions.

[1]The low-level features we mention here are those irrelevant to the image semantics, such as the histogram of gradient [13], while the high-level features we mention here convey some semantic cues, *e.g.*, the "simplicity" of the image, whether the sky is clear, *etc*.

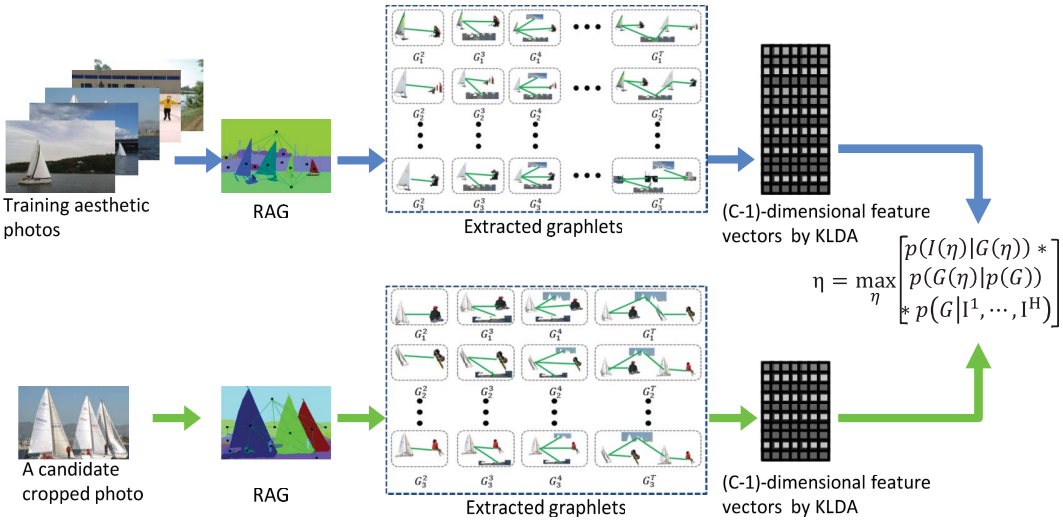


Fig. 2. Pipeline of our approach. First row: the process of extracting graphlets from training aesthetic photos. Second row: the process of evaluating the quality of each candidate cropped photo using a probabilistic model.

Although multiple subjects are considered in [17], their spatial interactions, *e.g.*, whether the sky is below or above the sea, are ignored. To semantically represent the essential features for photo cropping, Dhar *et al.* [19] have proposed a set of high-level attribute-based predictors for evaluating the photo aesthetics. Three types of attribute-based predictors are proposed, *i.e.*, compositional attributes, content attributes, and sky illumination attributes. Experimental results demonstrate that the aesthetic classifier learned from these attributes achieves much better performance compared with those trained solely from low-level image features. The main weakness of [19] is that the attributes are designed manually and are data set dependent. Thus, they cannot be generalized to different data sets.

To solve or at least alleviate the aforementioned problems, we propose graphlets to represent the aesthetic feature of photos and transfer the graphlets from the training aesthetic photos into the cropped photos based on a probabilistic model. As shown in Fig. 2, by segmenting each training photo into a set of atomic regions, we construct a so-called region adjacency graph (RAG) to represent the global aesthetic feature of these atomic regions. To represent the local aesthetic features of the training photos, we extract the graphlets from the RAGs using depth-first-search [20]. Because an RAG can be regarded as a special type of graphlet, the aesthetic similarity of two photos can be formulated as graphlet-to-graphlet matching. To measure the similarity between graphlets, which may have different numbers of vertices, and to obtain a fixed-length feature vector representation for each graphlet, we represent the vertices as well as their spatial interactions using a matrix, compute the kernel between matrices, and further use a Kernel LDA [21] to represent each graphlet by a $(C-1)$-dimensional feature vector, where C is the number of categories of training photos. To evaluate the quality of each candidate cropped photo, we extract the graphlets within the candidate cropped photo and form a posterior probability to measure its quality. Based on the posterior probability formulation, we cast photo cropping as seeking the parameter of a candidate cropped photo with the maximum posterior probability, and Gibbs sampling [22] is applied for parameter inference. Extensive experimental results demonstrate the effectiveness of our approach.

II. Aesthetic Feature Extraction

A. Region Adjacency Graph

A photo usually contains millions of pixels. If we treat each pixel independently, the high computational cost will make photo cropping intractable. Fortunately, pixels are usually highly correlated with their spatially neighboring ones. Thus for each photo, we cluster its pixels into a set of atomic regions, and this photo can be regarded as a set of atomic regions associated with their spatial interactions. A graph is a powerful tool to describe the relationships between objects, and in this work, we propose a region adjacency graph (RAG) to model the global aesthetics of each photo. The RAG construction process is detailed as follows. Given a photo I, we cluster its pixels into a set of atomic regions using an image segmentation algorithm, and an RAG $\mathcal{G}$ is constructed to model I, *i.e.*,

$$\mathcal{G} = (V, E) \tag{1}$$

where V denotes a finite set of vertices, each representing an atomic region; E denotes a set of edges, each connecting a pair of spatially adjacent atomic regions. To make the image segmentation step more stable, we adopt two schemes. First, we use unsupervised fuzzy clustering (UFC) [23] for photo segmentation. One advantage of UFC is that, prior knowledge of the number of segmented atomic regions is not required, and its tolerance bound is flexible to tune. Second, each photo is segmented five times under different tolerance bounds of UFC, *i.e.*, the tolerance bound is tuned from 0.1 to 0.5 with a step of 0.1.

In this work, we use both color and texture information to characterize each atomic region as color and texture are generally complementary to each other in measuring

the region's properties. We detail the feature extraction as follows.

For the color descriptor, we use color moment [24] to represent the three central moments of an atomic region's color distribution in each RGB channel. The three central moments are mean, standard deviation, and skewness. Thus, each atomic region is represented by a 9-dimensional feature vector in RGB channel. For the texture descriptor, we use the well-known histogram of gradient (HOG) [13] to model the texture of each atomic region. The HOG descriptor has the advantage of invariance to local geometric changes, *i.e.*, rotations and photometric transformations. Firstly, we use a finite difference filter, $[-1; 0; +1]$, and its transpose, to compute the gradient of each pixel. Then, each gradient orientation is discretized according to a vector quantization (VQ) codebook, and we obtain a feature map representing both the gradient orientation and intensity of each pixel. This feature map is further divided into 4×4 sub-regions, where the feature map in each sub-region is quantized into an 8-dimensional feature vector. By concatenating the 8-dimensional feature vectors from all the 4×4 sub-regions, we represent each atomic region by a 128-dimensional HOG feature vector in the texture channel.

After extracting the color and texture feature, we have a $9 + 128 = 137$-dimensional feature vector to describe each atomic region, *i.e.*,

$$F(R) = [F_{CM}(R), F_{HOG}(R)] \tag{2}$$

where $F_{CM}(R)$ and $F_{HOG}(R)$ respectively denote the color moment and the HOG feature vector computed from atomic region R.

B. Graphlets as Aesthetic Features

Given a photo, its RAG represents the photo's global aesthetic feature, *i.e.*, all the components within this photo as well as their spatial interactions. To represent the photo's local aesthetic features, it is useful to extract its RAG's graphlets (*i.e.*, connected subgraphs), which capture a subset of components and their spatial interactions. Formally, we define graphlet as a connected subgraph of an RAG. The size of a graphlet is defined as the number of vertices in this grahplet. And we call an i-sized graphlet i-graphlet. As shown in Fig. 3, the 3-graphlet encodes the spatial interactions among the sailboat, waterman, and water, which are important local aesthetic features that should be preserved in the cropped photo.

Given an RAG, a number of its graphlets can be extracted. To measure the similarity between graphlets, a straightforward approach is to concatenate the 137-dimensional feature vector corresponding to the atomic regions from each graphlet into a long feature vector. However, there are two disadvantages of this straightforward strategy. First, different graphlets may have different numbers of vertices, which result in a different dimensional concatenated feature vector. Second, the spatial interactions between atomic regions are totally ignored. As discussed above, the spatial interaction is an essential cue for photo cropping. To solve these two problems, we uniformly represent any sized graphlet by a fixed-length feature vector,

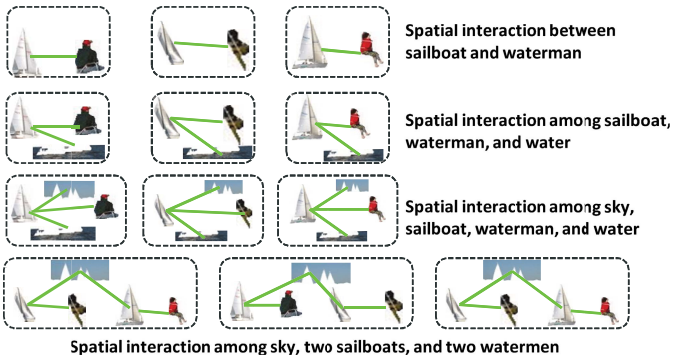


Fig. 3. Local aesthetic features represented by {2, 3, 4}-graphlet.

and the feature vector encodes the spatial interaction of atomic regions. In particular, we first represent each graphlet by a matrix, which captures the color and texture information of each atomic region, as well as the spatial interaction between atomic regions. Because it is infeasible to measure different sized matrices, we derive the kernel between the same sized graphlets; the kernel can be proved to be positive definite. Thus, all kernel-based algorithms in Hilbert space can be adopted. Finally, we adopt Kernel LDA to represent each graphlet by a $(C-1)$-dimensional feature vector, where C denotes the number of categories. Kernel LDA encourages highly discriminative graphlet transfer into the cropped photo. We detail the above steps in the following part of this section.

Given a t-sized graphlet, we characterize all its atomic regions as a matrix $M_R \in \mathbb{R}^{t \times 137}$, where each row of M_R denotes a 137-dimensional feature vector representing the color and texture of an atomic region. To represent the spatial interactions of atomic regions in this graphlet, we adopt a $t \times t$ adjacency matrix, *i.e.*,

$$M_S(i, j) = \begin{cases} 1, & \text{if } R_i \text{ and } R_j \text{ are spatially adjacent} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Based on M_R and M_S, we represent t-sized graphlets by a $t \times (137 + t)$ matrix, *i.e.*,

$$M = [M_R, M_S]. \tag{4}$$

To measure the similarity of a pair of t-sized graphlets M_i and M_j, their kernel is defined as follows:

$$k(M_i, M_j) = ||M_i^T M_j||_F^2. \tag{5}$$

Here, we prove that the above kernel function is positive definite. Following [25], a real-valued function $k(x_i, x_j)$ on $\mathcal{X} \times \mathcal{X}$ is positive definite (*resp.* conditional positive definite) if and only if $k(x_i, x_j)$ is symmetric and $\sum_{ij} \gamma_i \gamma_j k(x_i, x_j) \geq 0$, for all $x_1, x_2, \ldots, x_N (x_i \in \mathcal{X})$ and $\gamma_1, \gamma_2, \ldots, \gamma_N (\gamma_i \in \mathbb{R})$ (*resp.* for all $\gamma_1, \gamma_2, \ldots, \gamma_N$ such that $\sum \gamma_i = 0$). Following [26] and [27], each matrix can be deemed as a point on the Grassmann manifold. The positive definiteness follows from the properties of the Frobenius norm. For all $M_1, M_2, \ldots, M_n$

and $\gamma_1, \gamma_2, \ldots, \gamma_n (\gamma_i \in \mathbb{R})$, for any $n \in \mathbb{N}$, we have

$$\sum_{ij} \gamma_i \gamma_j ||M_i^T M_j||_F^2 = \sum_{ij} \gamma_i \gamma_j \mathrm{tr}(M_i M_i^T M_j M_j^T)$$
$$= \mathrm{tr}\Big(\sum_i \gamma_i M_i M_i^T\Big)^2$$
$$= \left|\sum_i \gamma_i M_i M_i^T\right|_F^2 \geq 0. \quad (6)$$

Thus the kernel function $k(M_i, M_j)$ is positive definite.

Due to the large variation in the training photos, conventional photo cropping methods, such as She *et al.*'s [5] and Cheng *et al.*'s [14] approaches, usually employ multiple cropping models. Each cropping model is trained using photos from one category. Given a test photo, they first classified it into a category and then cropped this photo using the cropping model corresponding to this category. This strategy is probably effective for achieving a good cropping result since each cropping model deals only with a subset of photos with small variations. There are two limitations though. First, only the weak global features are used to train the classifier, such as the spatial-envelop [7] and the bag of visual words [28]. When the training photos contain complex structures, the classifiers may fail to accurately predict the class label of a test photo. In this case, the test photo will be cropped under a mismatched cropping model, yielding unsatisfactory cropping results. Second, even if the test photo is classified correctly, in the cropping stage, the local image features, such as the omni-range context in [14], are transferred from the training photos into the cropped photo with identical weights. In practice, however, preferred cropping results will be observed if we assign a larger weight to some "important" local features. For instance, as shown in Fig. 4, there are three training photos from the "sailing" category and one training photo from the "surfing" category. Given a test photo from the "sailboat" category, based on the conventional photo cropping methods, all the image components, such as sky, water and sailboat, will be assigned with identical weights. Thus, all the eight candidate cropped photos will have similar chances of being recommended as the final cropped photo. Suppose we additionally consider the training photo from the "surfing" category and assign a larger weight to some discriminative image components, such as the "sailboat", while assigning a smaller weight to non-discriminative image components, such as the "sky" and "water", the candidate cropped photo from the top row of Fig. 4 will be assigned with higher quality scores, and preferred cropping results will be observed.

To implement the weighting mechanism, we employ a supervised discrimination analysis method to assign a weight to each graphlet, where the weight reflects the discrimination of this graphlet. Besides, to avoid the negative effects brought by the classifier, our approach transfers the weighted graphlets extracted from all training photos into the cropped photo.

As discussed above, each graphlet is represented by a matrix. Thus, it is impossible to use Fisher's LDA [21] to explicitly assign a weight to each graphlet. Instead, we use Kernel LDA to implicitly assign a weight to each graphlet. The weight is a $(C-1)$-dimensional feature vector representing the discrimination of each graphlet, where C denotes the number of categories. We detail the weight-assigning process in the following.

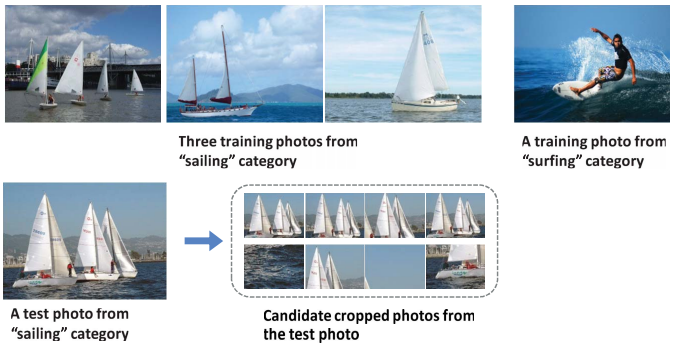


Fig. 4. Example of adding a weighting mechanism to the cropping model.

For the t-sized graphlets, let ϕ be a nonlinear function to map matrix M onto some feature space $\mathcal{F}$. Kernel LDA finds a projection direction w as follows:

$$w = \max_w \frac{w^T S_B^\phi w}{w^T S_W^\phi w} \quad (7)$$

where w denotes the projection matrix; S_B^ϕ and S_W^ϕ denote the between and within class scatter matrices respectively:

$$S_B^\phi = \sum_i N_i (m_i^\phi - m^\phi)(m_i^\phi - m^\phi)^T \quad (8)$$

$$S_W^\phi = \sum_{i=1,2,\ldots,C} \sum_{M \in \mathcal{M}_i} (\phi(M) - m_i^\phi)(\phi(M) - m_i^\phi)^T \quad (9)$$

where $m_i^\phi = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi(M_j^i)$ and $m^\phi = \frac{1}{N} \sum_{i=1}^{N} \phi(M_i)$. Here M_j^i denotes the j-th matrix from the i-th category, and N_i denotes the number of matrices from the i-th category; $\mathcal{M}_i$ denotes all the matrices from the i-th category.

Using the definition of m_i^ϕ we can write:

$$w^T m_i^\phi = \sum_{i=1}^{N} \alpha_i \phi(M_i) m_i^\phi$$
$$= \frac{1}{N_i} \sum_{j=1}^{N} \sum_{k=1}^{N_i} \alpha_j k(M_j, M_k^i) = \alpha^T P_i \quad (10)$$

where $(P_i)_j = \frac{1}{N_i} \sum_{k=1}^{N_i} k(M_j, M_k^i)$; $k(M_i, M_j) = \langle \phi(M_i), \phi(M_j) \rangle = ||M_i^T M_j||_F^2$ is the positive semi-definite kernel function defined in (5).

By using the definition of S_B^ϕ in (8) we can write:

$$w^T S_B^\phi w = \alpha^T P \alpha \quad (11)$$

where $P = \sum_{ij} (P_i - P_j)(P_i - P_j)^T$.

Similarly, we can derive

$$w^T S_W^\phi w = \alpha^T Q \alpha \quad (12)$$

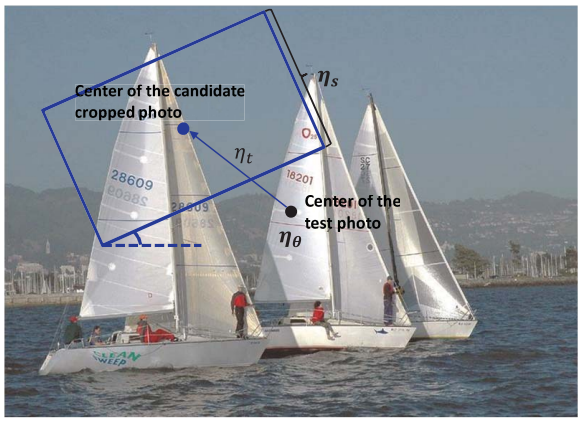


Fig. 5. Illustration of the cropping parameter. Blue rectangle: the candidate-cropped photo.

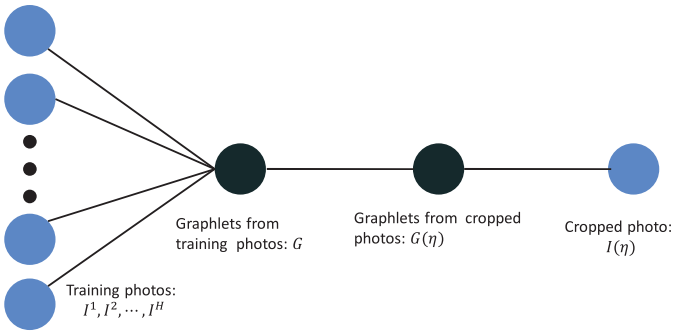


Fig. 6. Undirected graphical model representing the probabilistic model of our photo-cropping process.

where $Q = \sum_{j=1,2,\ldots,C} K_j(I - \mathbf{1}_{N_j})K_j^T$, K_j is an $N \times N_j$ matrix with $(K_j)_{nm} = k(M_n, M_m^j)$, I is an identity matrix and $\mathbf{1}_{N_j}$ is a matrix with all entries $1/N_j$.

Combining (11) and (12), we can rewrite (7) as

$$\alpha = \max_{\alpha} \frac{\alpha^T P \alpha}{\alpha^T Q \alpha}. \tag{13}$$

Thus, the problem of calculating α can be solved by finding the leading eigenvector of $Q^{-1}P$. It is noticeable that, the above Kernel LDA training process is carried out T times, where T denotes the maximum size of graphlets. And a set of parameters $\{\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(T)}\}$ is obtained in the training stage. Given a matrix obtained from a new t-sized matrix M, its weight is calculated by an implicit projection, *i.e.*,

$$w \cdot \phi(M) = \sum_{i=1}^{N} \alpha_i^{(t)} k(M_i, M). \tag{14}$$

III. Probabilistic Model for Photo Cropping

Given a photo, its weighted graphlets capture both the local and global aesthetics of training photos (graphlets are connected subgraphs of an RAG; thus, RAG, which captures the global aesthetic features can be deemed as a special type of graphlet), with the weight indicating the importance of each graphlet. To effectively integrate these weighted graphlets for photo cropping, we propose a probabilistic model to measure the quality of each candidate cropped photo.

A. Probabilistic Model

Given a test photo I, we define its cropped photo as $I(\eta)$. $\eta = (\eta_s, \eta_\theta, \eta_t)$ is a 5-dimensional cropping parameter. As illustrated in Fig. 5, η_s is a 2-dimensional variable denoting the XY coordinate scale of the cropped photo. $\eta_\theta \in [0, 2\pi]$ is a 1-dimensional variable denoting the rotation angle of the cropped photo. η_t is a 2-dimensional variable denoting the translation from the center of the test photo to that of the cropped photo.

Given a set of training photos $I^1, I^2, \ldots, I^H$ and a test photo I, the cropped photo $I(\eta)$ should maximally preserves the training aesthetic features, *i.e.*, weighted graphlets. Let G denote all the training weighted graphlets and $G(\eta)$ denote all the weighted graphlets from the cropped photo. The training photos and the cropped photo are highly correlated through their respective weighted graphlets. In particular, there are strong correlations between the following three pairs of variables: 1) $I^1, I^2, \ldots, I^H$ and G; 2) G and $G(\eta)$; and 3) $G(\eta)$ and $I(\eta)$. Thus, we propose a probabilistic graphical model [29], [30] to make use of this prior knowledge, as shown in Fig. 6. The undirected graphical model illustrates the process of photo cropping, where $I^1, I^2, \ldots, I^H$ denotes the state of each training photo and $I(\eta)$ the state of the cropped photo; G denotes the state of all training weighted graphlets and $G(\eta)$ the state of weighted graphlets from $I(\eta)$. Our probabilistic model contains two types of nodes: observable nodes (colored blue) and hidden nodes (colored gray). Edges are used to describe the relationships between nodes. These two types of nodes form four layers. The first layer corresponds to all the training photos $I^1, I^2, \ldots, I^H$. The second layer denotes all the training weighted graphlets G. The third layer represents all the weighted graphlets from the cropped photo $G(\eta)$, and the fourth layer denotes the cropped photo $I(\eta)$. The relationship between the first layer and the second layer is formulated as $p(G|I^1, I^2, \ldots, I^H)$. The relationship between the second and the third layer is $p(G(\eta)|G)$ and the relationship between the third and the fourth layer is $p(I(\eta)|G(\eta))$.

The photo cropping model can be regarded as a process that maximally transfers the extracted weighted graphlets from the training photos to the cropped photo. This process can be formulated into the following maximum a posterior (MAP) framework:

$$\begin{aligned}\eta &= \max_{\eta} p(I(\eta)|I^1, I^2, \ldots, I^H) \\ &= \max_{\eta} p(I(\eta)|G(\eta)) * p(G(\eta)|G) * p(G|I^1, I^2, \ldots, I^H).\end{aligned} \tag{15}$$

For ease of expression, we rearrange the three probabilities in (15) as:

$$\begin{aligned}&p(I(\eta)|G(\eta)) \\ &\quad = p(I(\eta)|G^1(\eta), G^2(\eta), \ldots, G^T(\eta)) \\ &\quad = \frac{p(G^1(\eta), G^2(\eta), \ldots, G^T(\eta)|I(\eta))p(I(\eta))}{p(G^1(\eta), G^2(\eta), \ldots, G^T(\eta))}\end{aligned}$$

$$\propto p(G^1(\eta), G^2(\eta), \ldots, G^T(\eta)|I(\eta))p(I(\eta))$$
$$= \prod_{i=1}^{T} p(G^i(\eta)|I(\eta))p(I(\eta))$$
$$= \prod_{i=1}^{T}\prod_{j=1}^{Y_i} p(G_j^i(\eta)|I(\eta))p(I(\eta)) \quad (16)$$

$p(G(\eta)|G)$

$$= p(G^1(\eta), G^2(\eta), \ldots G^T(\eta)|G^1, G^2, \ldots G^T)$$
$$\propto \prod_{i=1}^{T} p(G^i(\eta)|G^1, G^2, \ldots G^T)$$
$$= \prod_{i=1}^{T}\prod_{j=1}^{Y_i(\eta)} p(G_j^i(\eta)|G^1, G^2, \ldots, G^T) \quad (17)$$

$p(G|I^1, I^2, \ldots, I^H)$

$$= p(G^1, G^2, \ldots, G^T|I^1, I^2, \ldots, I^H)$$
$$= \prod_{i=1}^{T} p(G^i|I^1, I^2, \ldots, I^H)$$
$$= \prod_{i=1}^{T}\prod_{j=1}^{Y_i} p(G_j^i|I^1, I^2, \ldots, I^H) \quad (18)$$

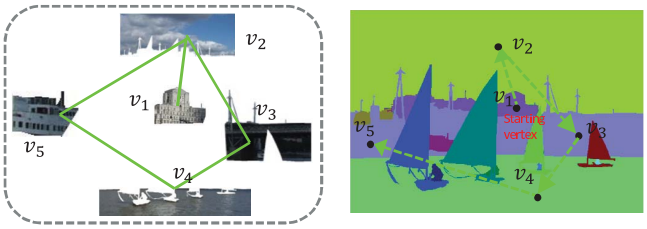
where G^i denotes all the training weighted i-graphlets (i-graphlet represents a graphlet with i vertices) and represents the aesthetic features described by the i-sized training graphlets, G_j^i denotes the j-th weighted graphlet from all the training weighted i-graphlets and is the basic element representing the photo aesthetics, $G^i(\eta)$ denotes all the weighted i-graphlets from the cropped photo and represents the aesthetic features captured by the i-sized graphlets from the cropped photo, $G_j^i(\eta)$ denotes the j-th weighted graphlet from all the weighted i-graphlets in the cropped photo, Y_i denotes the number of i-sized graphlets obtained from the training photos, and $Y_i(\eta)$ is the number of i-sized graphlets obtained from the cropped photo.

To calculate the three probabilities $p(I(\eta)|G(\eta))$, $p(G(\eta)|G)$ and $p(G|I^1, I^2, \ldots, I^H)$, we define several probabilities as follows.

$p(G_j^i|I)$ is the probability of extracting weighted graphlet G_j^i from photo I. As shown in Fig. 7, the procedure of graphlet extraction can be deemed as traversing the vertices on an RAG. We first choose a starting vertex in an RAG with probability $p(Y)\frac{1}{Y}$, where Y is the number of atomic regions in photo I and $P(Y)$ is the probability of Y atomic regions in image I. We then visit the spatially adjacent vertices one by one, and the probability of visiting a spatially adjacent vertex is decided by the degree of the current vertex, *i.e.*, $\frac{1}{\sum_d p_d(R_l)d(R_l)}$ where $p_d(R_l)$ denotes the probability of the degree of the current atomic region R_l. The visiting process stops when the maximum size of the graphlet is reached. Based on the above graphlet extraction procedure, we define $p(G_j^i|I)$ as follows:

$$p(G_j^i|I) \propto p(Y)\frac{1}{Y}\prod_{l=1}^{i-1}\frac{1}{\sum_d p_d(R_l)d(R_l)} \quad (19)$$

where $p_d(R_l)$ and $P(Y)$ are defined as Gaussian kernels, *i.e.*, $p_d(R_l) \propto \exp(\frac{-||R-\bar{R}||^2}{\sigma_d^2})$; and $p(Y) \propto \exp(-\frac{||Y-\bar{Y}||^2}{\sigma_Y^2})$. Here $\bar{R}$ and $\bar{Y}$ respectively denotes the Gaussian centers of $p_d(R_l)$ and $p(Y)$; σ_d and σ_Y respectively denotes the Gaussian covariance of $p_d(R_l)$ and $p(Y)$. The four parameters $\bar{R}$, $\bar{Y}$, σ_d and σ_Y are set by the empirical values from the training photos.



(a) (b)

Fig. 7. (a) Graphical illustration of graphlet. (b) Extraction from an RAG. Dashed line arrows: traverse procedure.

Let $p(G_j^i|I^1, I^2, \ldots, I^H)$ be the probability of weighted graphlet G_j^i coming from all training photos $I^1, I^2, \ldots, I^H$; it is defined as:

$$p(G_j^i|I^1, I^2, \ldots, I^H) = 1 - \prod_{k=1}^{H}\left(1 - p(G_j^i|I^k)\right). \quad (20)$$

Let $p(I(\eta))$ denote the probability of a photo I cropped using the parameter η; it is defined as:

$$p(I(\eta)) \propto \exp\left(\frac{||\eta - \bar{\eta}||^2}{\sigma_\eta^2}\right). \quad (21)$$

Lastly, let $p(G_j^i(\eta)|G^1, G^2, \ldots, G^T)$ be the probability of graphlet $G_j^i(\eta)$ existing in $G^1, G^2, \ldots, G^T$; it is defined as:

$$p(G_j^i(\eta)|G^1, G^2, \ldots, G^T) \propto \exp\left(-\frac{\sum_{G\in G^1, G^2, \ldots, G^T}||G_j^i - G||}{|G^1, G^2, \ldots, G^T|}\right). \quad (22)$$

B. Parameter Inference

We can see that the posterior probability in (15) is complicated and has no explicit analytical solution. Therefore, to derive the optimal cropping parameter, we adopt the commonly used Gibbs sampling [22]. One advantage of Gibbs sampling is that one only considers univariate conditional distributions, *i.e.*, the distribution when all the random variables but one are assigned fixed values. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms.

Based on the concept of Gibbs sampling, we start by selecting an initial value of $\eta_t^{(1)}$ and $\eta_\theta^{(1)}$ based on the distribution given as follows:

$$p(\eta_t) \propto \exp\left(-\frac{1}{\sigma_t^2}||\eta_t - \bar{\eta}_t||^2\right) \quad (23)$$

$$p(\eta_\theta) = \frac{1}{2\pi}\eta_\theta \quad (24)$$

where η_t denotes the translation from the center of the test photo to that of the cropped photo and $\bar{\eta}_t$ is the Gaussian center; $\eta_\theta \in [0, 2\pi]$ is the rotation angle of the cropped photo.

We give a graphical illustration of (23) and (24) in Fig. 8. The term $||\eta_t - \bar{\eta}_t||^2$ in (23) reflects that the closer the distance between the center of the cropped photo and $\bar{\eta}_t$, the more probability of this center will be accepted. The term $\frac{1}{2\pi}\eta_\theta$ reflects that the cropped photo can be evenly rotated to any angle in the ranges between 0 and 2π.

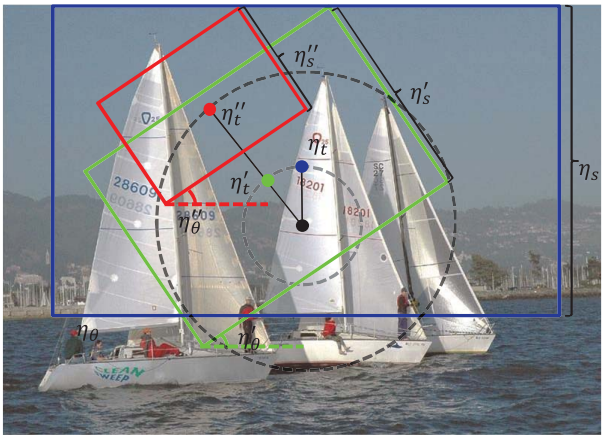


Fig. 8. Graphical illustration of (23)–(25). Green, blue, and red points: centers of three candidate cropped photos. Green, blue, and red rectangles: maximum size of the candidate-cropped photos.

Based on $\eta_t^{(1)}$ and $\eta_\theta^{(1)}$, we then generate a new value of $\eta_s^{(1)}$ based on the conditional probability as follows:

$$p(\eta_s|\eta_t,\eta_\theta) \propto \exp\Big(-\frac{||\eta_s-\bar{\eta}_s||^2}{\sigma_S^2} * \frac{1}{||\eta_t||} * \frac{1}{\eta_\theta \mathrm{mod}\pi/2}\Big) \quad (25)$$

where η_s denotes the XY coordinate scale of the cropped photo. As shown in Fig. 8, the term $\frac{1}{||\eta_t||}$ reflects that, the further the distance between the center of the cropped photo and that of the test photo, the higher is the probability of obtaining a smaller sized cropped photo. The term $\frac{1}{\eta_\theta \mathrm{mod}\pi/2}$ reflects that the closer the rotation angle of the cropped photo η_θ between $\{0, \pi/2, \pi, 3/2\pi, 2\pi\}$, the more likely of obtaining a larger sized cropped photo.

Based on the three probabilities above, we update η iteratively until the convergence criteria is met, *i.e.*, the posterior probabilistic in (15), which computes based on η, becomes stable. Given that the cropping parameter space is only 5-dimensional, the convergence of the sampling procedure is quite fast. Finally, the cropping parameter that yields the highest value in (15) represents the optimal cropping parameter, and the corresponding cropped photo will be our solution.

C. Probabilistic Graphlet Model for Photo Cropping

We present the procedure of the proposed probabilistic model for photo cropping in Algorithm I. Firstly, we use unsupervised fuzzy clustering to decompose each photo into a set of atomic regions, and extract all graphlets with size $t \in \{1, 2, \ldots, T\}$. For each atomic region, we extract a 137-dimensional feature vector to represent the color and texture information. Secondly, for each graphlet, we represent it by a $t \times (t + 137)$ matrix, and use Kernel LDA to transfer the matrix into a $(C-1)$-dimensional feature vector that captures the discriminative aesthetic features. Thirdly, we compute the optimal parameter of the cropped photo based on Gibbs sampling, and output the cropped photo based on the optimal cropping parameter.

Algorithm 1 Probabilistic Graphlet Model for Photo Cropping

input: a set of labeled training photos $I^1, I^2, \ldots, I^H$; a test photo I and the maximum graphlet size T.
output: a cropped photo $I(\eta)$
begin:

1) Apply unsupervised fuzzy clustering to segment each photo; extract the $\{1, 2, \ldots, T\}$-sized graphlets from the training photos; for each segmented atomic region, extract the 137-dimensional feature vector from (2).
2) Compute the matrix of each graphlet from (4); use Kernel LDA to transfer each graphlet into a $(C-1)$-dimensional feature vector according to (14).
3) Use Gibbs sampling to select an optimal cropping parameter η based on (15); output the cropped photo $I(\eta)$.

end

IV. Experimental Results and Analysis

In this section, we justify the effectiveness of the proposed probabilistic model for photo cropping. The first set of experiments show the effectiveness of our discriminative graphlets in capturing the aesthetic feature of training photos. The second set of experiments evaluate the proposed method in comparison with representative photo cropping methods. The third set of experiments discuss the influence of the maximum size of graphlets T and the number of training photos H on the output cropped photo. Additional comparisons of cropping results with interactive photo cropping methods are also given to further validate the proposed photo cropping model.

A. Data Collection and Preprocessing

As far as we know, there are still no public available standard data sets for evaluating the performance of photo cropping. Thus we firstly use text queries to crawl massive photos from two online photo sharing websites: PhotoSig[2] and Flicker.[3] The total training data set contains more than 12,000 photos, *i.e.*, 6,000 highly ranked photos and 6,000 low ranked photos. It is noticeable that, for those cropping models that evaluate the photo quality using a classifier, such as sensation-based photo cropping proposed by Nishiyama *et al.* [17], both the highly ranked and low ranked photos are used to train the classifier, *i.e.*, the highly ranked photos are used as positive samples and the low ranked photos as negative samples. For other cropping methods, such as omni-range context based cropping proposed by Cheng *et al.* [14] and our approach, the cropping model only transfers the aesthetic features from the training photos into the cropped photo, and thus, we only use the highly ranked photos for training.

As discussed above, some cropping methods need a prepressing step to group the training photos into several

[2]http://www.photosig.com.
[3]http://www.flicker.com.

categories, such as the omni-range context-based cropping proposed by Cheng *et al.* and our approach. However, since the categories of our crawled photos are not assigned, we conduct a rough photo classification to assign each training photo with a class label. In particular, we use the well-known spatial envelop [7] as the descriptor of each training photo. The spatial envelop is a set of perceptual dimensions (naturalness, openness, roughness, ruggedness and expansion) that are related to the shape of space. We extract the spatial envelop from the scene data set published by Feifei *et al.* [31] and train a 13-class SVM to predict the class label of each training photo. As described in Feifei *et al.*'s publication, the 13 classes are respectively highway, inside of cities, tall buildings, streets, suburb residence, forest, coast, mountain, open country, bedroom, kitchen, living room, and office.

To evaluate the performance of the proposed approach, we notice that previous photo cropping experiments employ either 4:3 aspect ratio photos or panoramic photos for testing. For example, the test photos used in Liu *et al.*'s and Bhattacharya *et al.*'s experiments are with an aspect ratio of 4:3; while Cheng *et al.*'s experiment uses all panoramic photos. In our experiment, we construct two groups of test photos. The first group contains 314 badly composed photos. All these photos are with an aspect ratio of 4:3. We intend obtaining a well-composed photo by cropping a sub-region from the original photo. The second group contains 313 panoramic photos crawled from the Internet, and most of these photos are well composed. We intend maximally preserving the aesthetic features from the panoramic photo into the cropped normal 4:3 aspect ratio photo. Due to space limitation, only 13 sets of comparative cropping results obtained from the first group of test photos are presented and evaluated in this paper. In addition, we present 22 sets of comparative cropping results obtained from the second group of test photos in the supplementary video.

It is worth emphasizing the following three points. First, cropping panoramic photos is more challenging compared with cropping those normal aspect ratio photos. This is because when cropping panoramic photos, the cropping parameter searching space is much larger, *i.e.*, much more candidate cropped photos will be generated from a panoramic photo. Second, we notice that the photos (both 4:3 aspect ratio and panoramic) we collect from the Internet are always well aligned horizontally.This is because unaligned photos severely affect the photo aesthetics and users seldom upload them to a photo-sharing website. But in practice, users may obtain a large number of unaligned photos, especially when taking photos using a cell-phone camera. Therefore, it is meaningful to test the effectiveness of the rotation variable η_θ of the cropping parameter. Conventional photo cropping experiments, such as Cheng *et al.*'s, usually ignore the rotation variable. In our experiment, we use Photoshop to rotate the test photos and then use the rotated photos as the input photo for cropping. Third, towards a pair comparison of our approach with the previous cropping methods, we restrict the aspect ratio of the cropped photo output from all the compared cropping methods to 4:3.

B. Aesthetic Features Represented by Graphlets

In this experiment, we evaluate the effectiveness of our discriminative graphlets in capturing the aesthetic features. In particular, we experiment on the Stanford event data set [32] associated with the annotation provided by Lotus Hill Institute (LHI) [33]. This data set contains 10 sports event categories collected from the Internet. In each category, we use half the photos for training and leave the rest for testing. We set the maximum size of graphlet T to 5. In each category, we calculate the discrimination of a graphlet by adding a normalization factor $\frac{1}{N}$ to (14):

$$g(M) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \alpha k(M_i, M) \tag{26}$$

where M denotes the matrix obtained from the graphlet and N the number of training graphlets. In Fig. 9, we present the top four discriminative graphlets from each training photo (one photo from each category is given). To compare our approach with conventional visual-attention-based photo cropping, we further compute the saliency map based on the well-known algorithm proposed by Itti *et al.* [34]. In conclusion, our approach shows the following advantages. First, graphlets capture the spatial interactions among image components, which is essential for photo cropping. As shown in the "rowing" and "ice-skate" categories, the girls form a line and this spatial interaction captures the aesthetics of training photos and should be preserved in the cropped photo; while using conventional visual-attention-based models, these essential features are ignored. Second, visual-attention-model-based photo cropping methods select the most salient region as the cropped photo, yet sometimes salient regions are not consistent with the image regions that should be preserved in the cropped photo. For instance, in the "rowing" category, the visual attention model selects the trees as the most important cues that should be preserved in the cropped photo. However, compared with the waterman and the sailboat which are more relevant cues to the "rowing" category, trees are less representative cues for photo cropping. Third, in most categories, the background captures important cues for cropping, such as the white snowfield in the "snowboarding" category and the red track in the "hurdles" category, yet the visual-attention-based model generally does not consider background information.

To further demonstrate the advantages of our approach over the features used for photo cropping or photo quality evaluation proposed by computer vision researchers, we compare our discriminative graphlet with three features proposed by Luo *et al.* [10], Ke *et al.* [9] and Yeh *et al.* [11], and the saliency model proposed by Itti *et al.* [34] is also employed for comparison. In particular, we experiment on the data set collected by Yeh *et al.*, which contains 6000 highly aesthetic as well as 6000 low aesthetic photos collected from DPChallenge.[4] In Table I, we detail the five compared features for photo cropping or photo quality evaluation. To compare the effectiveness of the five features, we use each feature to predict whether a test photo is highly aesthetic or low aesthetic. We use the same split of training and test

[4]http://www.dpchallenge.com.

TABLE I
DETAILS OF THE FIVE COMPARED FEATURES

Luo *et al.*	Composition+clarity+simplicity[5] +color distribution+lighting
Ke *et al.*	Spatial distribution of edges+color distribution+hue count+blur+contrast level+brightness
Yeh *et al.*	Simplicity[6] +texture+contrast+intensity average+region blur
Itti *et al.*	Saliency map based on color, intensity and orientation of local patch
Ours	Graphlet based on color and texture of segmented region

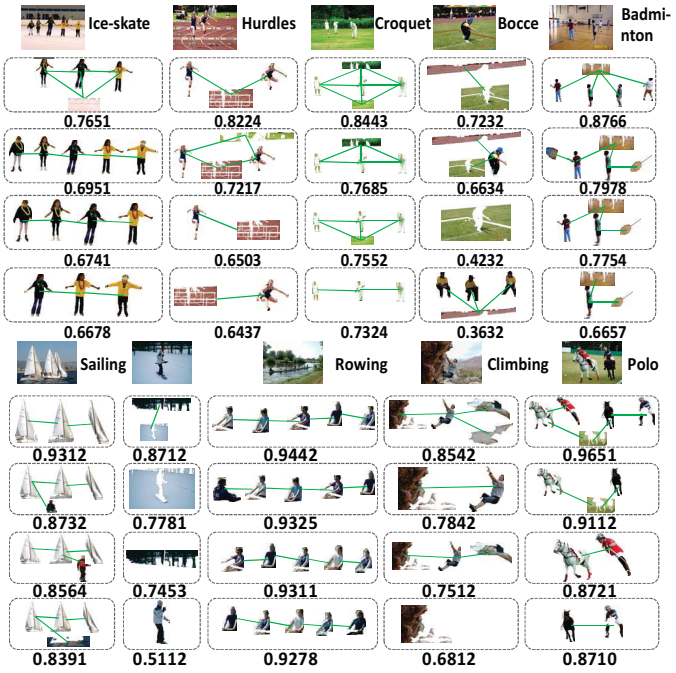


Fig. 9. Aesthetics captured by four top-ranked discriminative graphlets (i.e., the discrimination is marked below each graphlet) and the saliency map proposed by Itti *et al.* [34].

sets as in the program provided by Yeh *et al.*, and then train a binary SVM classifier based on the five features. Note that the discriminative graphlets in our approach cannot be used for classification directly because different photos may contain different numbers of graphlets. To address this problem, inspired by the graph kernel [35] that measures the similarity of two graphs by comparing all their respective subgraphs, we construct a kernel to measure the aesthetic similarity of two photos I and I', *i.e.*,

$$k(I, I') = \frac{1}{N_I * N_{I'}} \sum\nolimits_{G \in I, G' \in I'} k(F(G), F(G')) \quad (27)$$

where N_I and $N_{I'}$ respectively denotes the number of graphlets in photo I and I'; $F(G)$ and $F(G')$ are the $(C-1)$-dimensional feature vectors corresponding to graphlet G and G' respectively. In addition, to classify the saliency map generated using the algorithm by Itti *et al.*, we resize it to a 22×32 matrix and stack this matrix to a 704-dimensional feature vector.

[5]Color distribution of the background.

[6]Size of ROI segments associated with the simplicity feature proposed by Luo *et al.* [10].

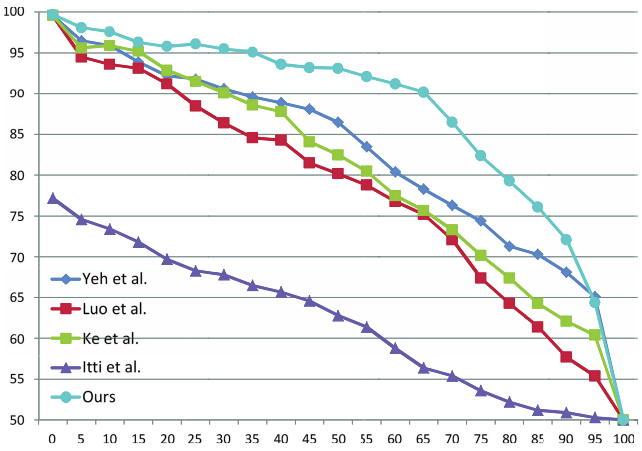


Fig. 10. Precision–recall curve of the five compared features.

As shown in Fig. 10, our graphlets outperform the four compared features significantly. The reasons are given as follows. First, the simplicity feature in Luo *et al.*'s and Yeh *et al.*'s approaches are based on the assumption that photos are taken by SLR cameras where the foreground and background can easily be discriminated. However, the data set collected by Yeh *et al.* contains a large number of photos taken by point-and-shoot cameras. Second, there is short of evidence that the concatenated global features can effectively capture the photo aesthetics, since each global feature is defined intuitively. Third, the worst performance is achieved by the saliency model from Itti *et al.* This is because the saliency map only tells the conspicuity of each pixel and it fails to capture important aesthetic features of a photo, such as color or texture information. This is consistent with the observation that the saliency map is seldom used for photo aesthetics evaluation or photo cropping alone.

C. Relations to Well-Known Aesthetic Rules

The proposed graphlet closely relates to three prominent aesthetics rules. We conclude them as follows.

1) Diagonal dominance, a well-known aesthetic guideline illustrated in Grill and Scanlon's book [36], discovered that viewers prefer the visually salient objects distributed along the diagonal line in a photo. This property can be appropriately captured by graphlets and the associated Kernel LDA [21]-based weighting scheme, *i.e.*, assigning a large weight to graphets if they locate closely to the diagonal line. In Fig. 11, as shown in the two left photos in the first row, the houses locating closely to the diagonal line are assigned with large weights and well preserved in the cropped photo. Besides, in the second

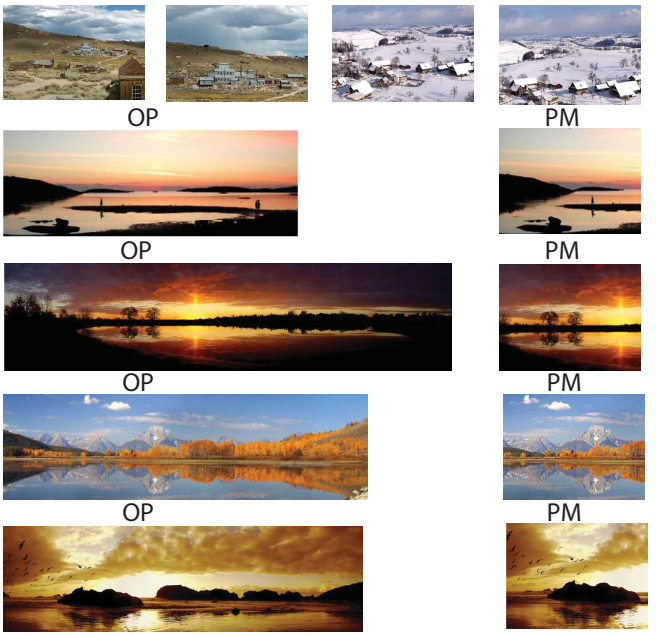


Fig. 11. Diagonal dominance, visual balance, and color harmony preserved in the cropped photo. OP: original photo. PM: cropped photo produced by our approach.

	OP	DAPC	OCBC	PM	PPR	SBPC	SCSM	Score
OP	--	10	22	10	15	19	10	86
DAPC	19	--	16	13	14	17	16	95
OCBC	7	13	--	11	12	14	11	68
PM	19	16	18	--	18	18	17	106
PPR	14	15	17	11	--	17	15	89
SBPC	10	12	15	11	12	--	10	70
SCSM	19	13	18	12	14	19	--	95

Preference matrix from the cropping results in the first row (filled by 29 volunteers)

	OP	DAPC	OCBC	PM	PPR	SBPC	SCSM	Score
OP	--	19	13	8	21	14	17	92
DAPC	14	--	8	6	18	10	7	63
OCBC	20	25	--	10	11	25	15	106
PM	25	27	23	--	17	25	18	135
PPR	12	15	22	16	--	12	19	96
SBPC	19	23	8	8	21	--	7	86
SCSM	16	26	18	15	14	26	--	115

Preference matrix from the cropping results in the second row (filled by 33 volunteers)

	OP	DAPC	OCBC	PM	PPR	SBPC	SCSM	Score
OP	--	16	20	12	20	21	14	103
DAPC	17	--	5	5	14	11	18	70
OCBC	13	28	--	15	22	23	20	121
PM	21	28	18	--	28	28	27	150
PPR	13	19	11	5	--	13	14	75
SBPC	12	22	10	5	20	--	12	81
SCSM	19	15	13	6	19	21	--	93

Preference matrix from the cropping results in the third row (filled by 33 volunteers)

	OP	DAPC	OCBC	PM	PPR	SBPC	SCSM	Score
OP	--	14	18	15	13	19	21	100
DAPC	17	--	15	14	13	19	15	93
OCBC	13	16	--	15	14	14	16	88
PM	16	17	16	--	26	25	24	124
PPR	18	18	17	5	--	17	17	92
SBPC	12	12	17	6	14	--	12	73
SCSM	10	16	15	7	14	19	--	81

Preference matrix from the cropping results in the fourth row (filled by 31 volunteers)

Fig. 12. Comparison of our approach with well-known cropping methods as well as the preference matrix filled by volunteers in Zhejiang University.

row of Fig. 11, the graphlet constructed from the boat, the benchland, and the sky locate closely to the diagonal line are nicely preserved in the cropped photo also.

2) Visual balance, another popular aesthetic guideline from Grill and Scanlon's book [36], claims that viewers prefer the visually salient objects distributed evenly around the photo center. Empirical results of our approach show that, for those well-composed natural scene images, graphlets with balanced structures, such as circle structure (an atomic region edge-connected by a collection of surrounding ones) and linear structure, are usually assigned with large weights. In Fig. 11, as shown in the two right photos in the first row, these circle-structurally distributed houses are well preserved in the cropped photo. As shown in the second and the third rows of Fig. 11, the linearly arranged sky, land and lake are also aesthetically preserved in the cropped photo.
3) Color harmony, a widely used aesthetic rule illustrated by Daniel *et al.* [37], measures the distribution of a set of colors in terms of human perceived visual harmony. Although colors in some atomic regions may be disharmonically distributed, colors in graphlets generated from these atomic regions are probably distributed harmonically. Our method can generate a number of graphlet-level color harmonic patterns, and the graphlet weighting scheme dynamically adjusts their importance. The importance level influences the cropping result based on the probabilistic model. As shown in the last three rows in Fig. 11, the harmonically distributed colors are kept in the cropped photo.

D. Comparative Evaluations of Photo Cropping

The proposed weighted graphlets can not only capture the photo aesthetics, but it can also be incorporated into a probabilistic model for photo cropping. Given a set of training photos, we extract graphlets to represent their aesthetic features and our probabilistic model enforces these aesthetic features to maximally transfer into the cropped photo.

Fig. 12 compares the proposed approach (PM) against several representative approaches, including sparse coding of saliency maps (SCSM [5]), sensation based photo cropping (SBPC [17]), omni-range context based cropping (OCBC [14]), personalized photo ranking (PPR [11]) and describable attribute for photo cropping (DAPC [19]). Sparse coding of saliency maps selects the cropped region that can be decoded by the dictionary learned from training saliency maps with the minimum error. Sensation-based photo cropping selects the cropped region with the maximum quality score, which is computed by probabilistically integrating the SVM scores corresponding to the detected subjects in a photo. Omni-range context-based cropping integrates the prior of spatial distribution of two arbitrary image patches into a probabilistic model to score each candidate cropped photo, and the candidate cropped photo with the maximum score is selected as the cropped photo.

Because those photo quality evaluation methods, such as personalized photo ranking proposed by Yeh *et al.* [11] and the describable attribute for photo cropping proposed by Dhar *et al.* [19], only output a score representing the quality of each photo, it is impossible to compare our approach with them directly because our approach outputs the cropped region of each photo. Fortunately, it is easy and straightforward to transform each of those photo quality evaluation methods into a photo cropping method. Typically, a photo cropping method

contains three steps: i) Candidate cropped photos sampling: Employing size-changeable and rotatable windows to slide on the original photo with a fixed XY-coordinate step. The region inside the sliding window is deemed as the candidate cropped photo. ii) Candidate cropped photo scoring: Evaluating the quality of each candidate cropped photo based on photo quality evaluation methods. iii) Cropped photo selection: The highest scored candidate cropped photo is deemed as the most qualified and will be selected. The first and the last steps are common with most of the photo cropping algorithms, while the second step is a technically challenging procedure. Hence, the key contribution of a photo cropping methods is usually a novel photo quality evaluation method, *e.g.*, Cheng *et al.*'s omni-range spatial context. Thus, it is fair to compare a photo quality evaluation method with a cropping method by transforming the photo evaluation method into a photo cropping method. Particularly, equip a photo quality evaluation method with a standard first and last cropping step.

The experimental settings of the two photo quality evaluation methods are given as follows. For personalized photo ranking, we extract low-level aesthetic features from the photo ranking system proposed by Yeh *et al.* These low-level aesthetic features are used to train a classifier for measuring the quality of each candidate cropped photo. For a describable attribute for photo cropping, we use the public code from Li *et al.* [38] to extract the attributes from each photo. These attributes are combined with the low-level features proposed by Yeh *et al.* to train a classifier to evaluate the quality of each candidate cropped photo.

In order to make the evaluation comprehensive, we adopt a typical subjective evaluation method. A paired comparison-based user study is carried out to evaluate the effectiveness of the proposed photo cropping method. This strategy was also used in [14] for evaluating the quality of a cropped photo. It is worth emphasizing that both rating and ranking are not suitable here as it would be an unnatural task for observers. Paired comparison is to present each subject with a pair of cropped photos from two different approaches. Participants are then required to indicate a preference, for one of the two cropped photos. Evaluation results are stored in the preference matrix. For example, considering the first preference matrix from Fig. 12, the entry in column SBPC and row DAPC has a value of 17, indicating that 17 subjects prefer the cropped photo produced from DAPC than that produced by SBPC. Additionally, to evaluate whether aesthetics of these suboptimally-composed original photos (OPs) are enhanced after cropping, original photos are also included for the paired comparison.

In this paper, the paired comparison was conducted by a group of volunteers who made paired comparison that fills the preference matrix. Most of the volunteers were from the computer science department of Zhejiang University, and were experienced in digital photography. We designed active Web pages that included the evaluation criteria and the resulting photos needed to be compared. The evaluation criteria suggested volunteers to click a resulting photo (out of two) that was more user-satisfied, according to their understanding of the criteria. Each set of resulting photos was evaluated by at least 29 volunteers and one preference matrix was set up for each set of resulting photos. In Fig. 12, we present the four preference matrices corresponding to the four sets of resulting photos produced by the compared cropping methods. We also show the overall scores for all evaluations, where the overall score is the sum of the scores in each row. The evaluation results clearly confirm the effectiveness of the proposed method for transferring the aesthetic features from the training photos into the cropped photo against a number of state-of-the-art photo cropping methods.

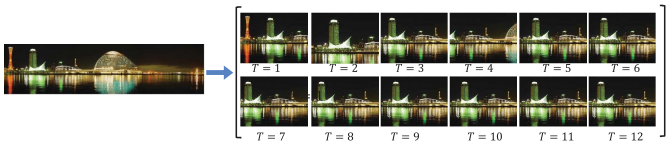


Fig. 13. Performance of the proposed approach under different maximum sizes of a graphlet.

The time consumption analysis of the proposed method is as follows. All experiments were carried out on a personal computer with Intel E8500 and 4 GB RAM. Our approach was implemented on a Matlab platform. Different from those compared methods that evaluate a large number of candidate cropped photos, the convergence of Gibbs sampling in our approach is fast. Given a test photo with a width of 1024 pixels, it usually takes around one minute to obtain a cropped photo, including photo segmentation, graphlet extraction, and Gibbs-sampling-based parameter inference. For the compared methods, by sequentially sampling, we usually obtain more than 1000 candidate cropped photos for evaluation, and it usually takes more than five minutes to select a qualified photo from those candidate cropped photos.

E. Parameter Analysis

In this experiment, we study how free parameters affect the performance of the proposed approach and how to set parameters to achieve a good cropping result. Particularly, we have two free parameters to be tuned, *i.e.*, the maximum size of a graphlet T, and the number of training photos H.

To analyze the effects of the maximum size of graphlets on photo cropping, we set up an experiment by varying T continuously. In Fig. 13, we present the cropped photos corresponding to T ranging from 1 to 12. We do not experiment with T larger than 12 because it becomes computationally intractable. As shown, for the sequence of cropped photos, the cropped photos become more aesthetic from $T = 1$ to $T = 5$. When T is larger than 5, the cropped photos become stable. This may be because few aesthetic features are captured by graphlets with a size larger than 5.

For the performance of the proposed approach with different numbers of training photos, we evaluate the performance of our approach for different values of H by sampling 10% training photos to 100% training photos from our data set, with a step of 10%. In each sampling, the proportion of training photos in each category is the same as that in the entire data set. We present the performance of our approach for different numbers of training photos in Fig. 14. As illustrated, more

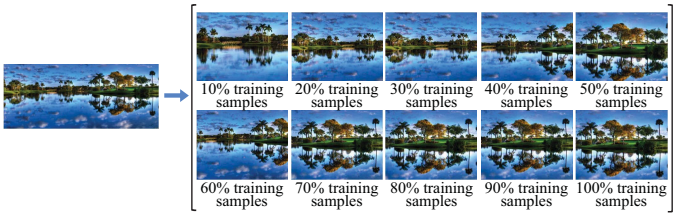


Fig. 14. Performance of the proposed approach under different numbers of training photos.

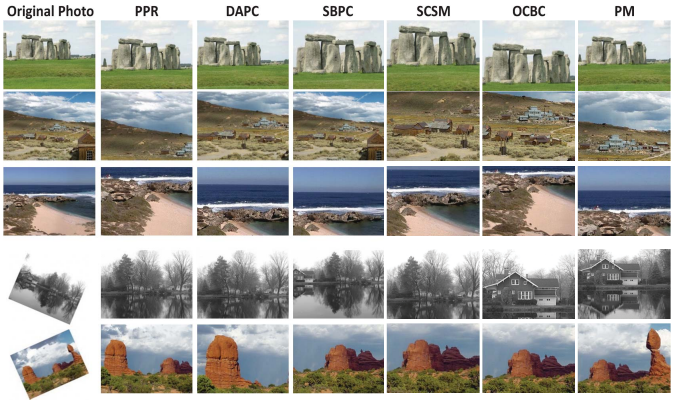


Fig. 15. More examples on the comparison of our approach with several well-known cropping methods.

training samples lead to more structural areas preserved in the cropped photo, though more time and space costs are required with an increasing number of training samples.

F. Further Examples for the Proposed Method

In this subsection, we present more cropping results of our approach with the five compared cropping methods described in Section 4.2. As shown in Fig. 15, we make the following observations. First, our approach achieves a good balance between the foreground objects and the background objects, as shown in the first three rows. Second, our approach prefers to preserve more structured objects in the cropped photo, such as the residential quarter in the second row and the villa in the fourth row. Third, all the compared cropping methods perform well with rotated input photos. As shown in the last two rows, the output cropped photos are well aligned horizontally.

Beyond the five fully-automatic cropping methods (*i.e.*, no human interaction is needed in the cropping process) described in Section 4.2, we further compare our approach with two human interactive cropping methods: gaze-based photo cropping (GBPC) proposed by Santella *et al.* [16] and interactive photo quality enhancement (IPQE) proposed by Bhattacharya *et al.* [15]. Briefly, gaze-based photo cropping enables users to look at each photo for a few seconds, while the system records their eye movements, which are used to identify important photo contents. Interactive photo quality enhancement lets users interactively select a foreground object and the system feedback to users for where the foreground object can be optimally located. Both the methods need human interaction, and different cropped photos may be produced by different users. Towards a fair comparison, we use four experimental

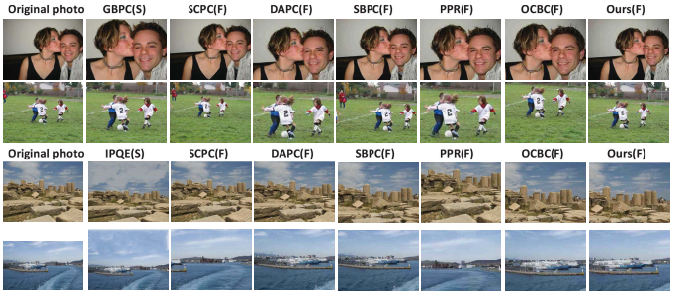


Fig. 16. Comparison of our approach with two representative human-interactive photo-cropping methods. F: fully automatic approach. S: semiautomatic approach.

photos from the publication of [15] and [16], and a comparison of the two human interactive methods as well as the five fully-automatic methods is given in the first four rows of Fig. 16. As illustrated, the cropping results produced by our approach are competitive to the two human interactive methods.

V. Conclusion

Photo cropping is a widely used technique in the printing, graphic design, and photography industries. In this paper, we propose graphlets to capture the photo aesthetics and further develop a probabilistic model to maximally transfer the graphlets from the training photos to the cropped photo. In particular, by segmenting each photo into a set of regions, we construct a so-called region adjacency graph (RAG) to represent the spatial relations of atomic regions. Next, we extract graphlets from the RAGs, and these graphlets capture the aesthetics from the training photos. Finally, we cast photo cropping as a candidate cropped photos searching procedure based on a probabilistic model and infer the cropping parameter using Gibbs sampling. The proposed method is fully-automatic. Thorough empirical studies demonstrate the effectiveness of our approach in comparison with a group of popular photo cropping and photo quality evaluation methods.

In the future, we plan to study the influence of different image segmentation schemes on the cropping results. Besides, we want to employ more participants in the pair comparison-based user study.

References

[1] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. 17th ACM Int. Multimedia Conf.*, 2009, pp. 541–544.

[2] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. 17th ACM Int. Multimedia Conf.*, 2009, pp. 561–564.

[3] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li, "Modeling and mining of users' capture intention for home videos," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 66–77, Jan. 2007.

[4] L. G. Liu, R. J. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.

[5] J. She, D. Wang, and M. Song, "Automatic image cropping using sparse coding," in *Proc. 1st Asian Conf. Pattern Recognit.*, Nov. 2007, pp. 490–494.

[6] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Neural Inf. Process. Syst. Conf.*, 2006, pp. 1–4.

[7] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[8] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[9] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 419–426.
[10] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Comput. Vis. Conf.*, 2008, pp. 386–399.
[11] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system," in *Proc. Int. Conf. ACM Multimedia*, 2010, pp. 211–220.
[12] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 129–136.
[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2005, pp. 886–893.
[14] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. Int. ACM Multimedia Conf.*, 2010, pp. 291–300.
[15] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. Int. ACM Multimedia Conf.*, 2010, pp. 271–280.
[16] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze based interaction for semi-automatic photo cropping," in *Proc. CHI Conf.*, 2006, pp. 771–780.
[17] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 669–672.
[18] C.-C. Chang and C.-J. Lin. (2012). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/faq.html
[19] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2011, pp. 1657–1664.
[20] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA: MIT Press, 2001.
[21] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[22] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boston, MA: Chapman & Hall, 1996.
[23] X. Xiong and K. L. Chan, "Toward an unsupervised optimal fuzzy clustering algorithm for image database organization," in *Proc. 15th Int. Conf. Pattern Recognit.*, 2000, pp. 897–900.
[24] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. Storage Retr. Image Video Databases Conf.*, 1995, pp. 381–392.
[25] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures," in *Proc. Conf. Workshop Item*, 2005, pp. 136–143.
[26] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.
[27] X. Wang, Z. Li, and D. Tao, "Subspaces indexing model on Grassmann manifold for image search," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2627–2635, Sep. 2011.
[28] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 221–228.
[29] S. Mingli, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 341–357, Jan. 2011.
[30] M. Song, D. Tao, C. Chen, X. Li, and C. W. Chen, "Color to gray: Visual cue preservation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1537–1552, Sep. 2010.
[31] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2005, pp. 524–531.
[32] L.-J. Li and L. Fei-Fei, "What, where, and who? Classifying event by scene and object recognition," in *Proc. 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[33] B. Yao, X. Yang, and S.-C. Zhu, "Introduction to a large scale general purpose ground truth dataset: Methodology, annotation tool, and benchmarks," *Energy Min. Meth. Comput. Vis. Pattern Recognit.*, pp. 169–183, Aug. 2007.
[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[35] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *Proc. IEEE Comput. Vis. Pattern Recognit. Conf.*, 2007, pp. 1–8.
[36] T. Grill and M. Scanlon, *Photographic Composition*. New York: Amphoto Books, 1990.
[37] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," in *Proc. ACM SIGGRAPH Conf.*, 2006, pp. 624–630.
[38] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. Neural Inf. Process. Syst. Conf.*, 2010, pp. 1–9.

Luming Zhang is currently pursuing the Ph.D. degree in computer science with Zhejiang University, Hangzhou, China.

His current research interests include visual perception analysis, image enhancement, and pattern recognition.

Mingli Song received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2006.

He is currently an Associate Professor with the College of Computer Science, Zhejiang University. He has authored and co-authored more than 60 papers in journals and conferences, including IEEE T-PAMI, T-IP, T-MM, PR, CVPR, ECCV, ACM MM. His current research interests include visual surveillance, visual perception analysis, image enhancement, and face modeling.

Qi Zhao received the Ph.D. degree in computer science from the University of California, Santa Cruz.

She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. She is involved in research on human perceptive mechanisms in computer vision and image processing.

Xiao Liu is currently pursuing the Ph.D. degree in computer science with Zhejiang University, Hangzhou, China.

His current research interests include visual surveillance, image statistics, and pattern recognition.

Jiajun Bu is currently a Professor with the College of Computer Science, Zhejiang University, Hangzhou, China. His current research interests include computer vision, computer graphics, and embedded technology.

Chun Chen is currently a Professor with the College of Computer Science, Zhejiang University, Hangzhou, China. His current research interests include information retrieval, computer vision, and embedded systems.