Fusion of Multichannel Local and Global Structural Cues for Photo Aesthetics Evaluation

Luming Zhang, Yue Gao, *Member, IEEE*, Roger Zimmermann, *Senior Member, IEEE*,
Qi Tian, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Photo aesthetic quality evaluation is a fundamental yet under addressed task in computer vision and image processing fields. Conventional approaches are frustrated by the following two drawbacks. First, both the local and global spatial arrangements of image regions play an important role in photo aesthetics. However, existing rules, e.g., visual balance, heuristically define which spatial distribution among the salient regions of a photo is aesthetically pleasing. Second, it is difficult to adjust visual cues from multiple channels automatically in photo aesthetics assessment. To solve these problems, we propose a new photo aesthetics evaluation framework, focusing on learning the image descriptors that characterize local and global structural aesthetics from multiple visual channels. In particular, to describe the spatial structure of the image local regions, we construct graphlets small-sized connected graphs by connecting spatially adjacent atomic regions. Since spatially adjacent graphlets distribute closely in their feature space, we project them onto a manifold and subsequently propose an embedding algorithm. The embedding algorithm encodes the photo global spatial layout into graphlets. Simultaneously, the importance of graphlets from multiple visual channels are dynamically adjusted. Finally, these post-embedding graphlets are integrated for photo aesthetics evaluation using a probabilistic model. Experimental results show that: 1) the visualized graphlets explicitly capture the aesthetically arranged atomic regions; 2) the proposed approach generalizes and improves four prominent aesthetic rules; and 3) our approach significantly outperforms state-of-the-art algorithms in photo aesthetics prediction.**

***Index Terms*—Multi-channel, structural cues, aesthetic evaluation, probabilistic model.**

I. Introduction

PHOTO aesthetics evaluation is a widely used technique in image retrieval [32], [34], graphic design [29], [30], and *etc.* For example, a successful photo management system should rank photos based on the human perception of photo aesthetics, so that users can conveniently select their favorite pictures into albums. Moreover, an effective photo aesthetics evaluation algorithm can help photographers to crop an aesthetically pleasing sub-region from an original poorly framed photo. However, photo aesthetics evaluation is still a challenging task due to the following two problems.

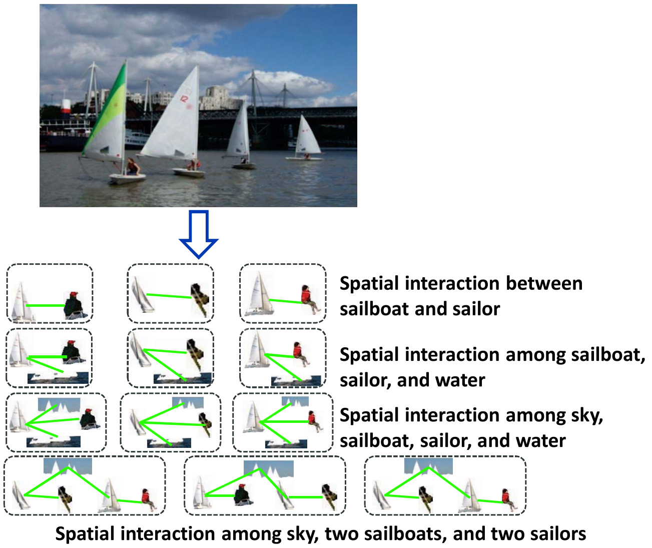

Spatial interaction between sailboat and sailor
Spatial interaction among sailboat, sailor, and water
Spatial interaction among sky, sailboat, sailor, and water
Spatial interaction among sky, two sailboats, and two sailors

Fig. 1. An example of the local aesthetics extracted from a photo.

- Both the spatial layout of locally and globally distributed regions in a scene play important roles in determining photo aesthetics. As seen from Fig. 1, the spatial interaction of the four linearly arranged sailboats captures the regional aesthetics; while the relative displacement of the sailboats, the water, and the sky reflects the global aesthetics. Existing rules can only heuristically define what spatial distribution among the salient image regions is aesthetically pleasing. For example, rule of the thirds [6] favors salient regions locating near the evenly 3 × 3 intersections of a photo. Although these aesthetic rules are convenient to use, they cannot reflect the specific spatial structure in photo aesthetics assessment, *e.g.*, the linearity and the triangularity among salient regions.
- Multi-channel visual cues collaboratively describe photo local aesthetics. That is to say, only visually salient regions with a particular color and texture distribution can arouse viewers' aesthetic perception. Yet it is difficult to determine the importance of each visual cue.

Manuscript received May 10, 2013; revised September 22, 2013, December 15, 2013, and January 9, 2014; accepted January 10, 2014. Date of publication January 29, 2014; date of current version February 18, 2014. This work was supported in part by the Singapore National Research Foundation through the International Research Centre at Singapore Funding Initiative, in part by the IDM Programme Office, in part by the National Natural Science Foundation of China under Grant 61125106, and in part by the Shaanxi Key Innovation Team of Science and Technology under Grant 2012KCT-04. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bulent Sankur. *(Corresponding author: Y. Gao.)*

L. Zhang and R. Zimmermann are with the School of Computing, National University of Singapore, Singapore 119613 (e-mail: zglumg@gmail.com; rogerz@comp.nus.edu.sg).

Y. Gao is with Tsinghua University, Beijing, China 100086 (e-mail: kevin.gaoy@gmail.com)

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitianx@utsa.edu).

X. Li is with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

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.2014.2303650

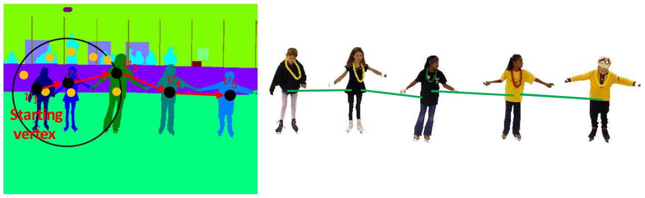


Fig. 2. Graphlet extraction procedure (the yellow marked atomic regions are locally distributed because they are within the circle.)

To address the aforementioned problems, we present a new photo aesthetics evaluation framework. To represent the structure of locally distributed regions of a photo (as shown in Fig. 2), we extract graphlets that can effectively capture the interaction of spatially neighboring atomic regions. Because both atomic regions and their spatial arrangements are essential for describing photo local structures, we represent each graphlet by a matrix that can encode both the properties. Based on the matrix representation, graphlets can be deemed as points on the Grassmann manifold [7], [16], [23]. To preserve the global spatial layout, we propose a manifold embedding algorithm to preserve all the distances between pairwise graphlets of each photo. At the same time, visual cues from multiple channels are dynamically tuned. More specifically, the weights of the three channel visual features (*i.e.*, color, texture, and visual saliency) are adjusted by optimizing the objective function as formulated in (8). After the embedding, graphlets are transformed into equal-lengthed feature vectors. Then, we integrate them into a probabilistic model for evaluating the aesthetic quality of a test photo. The probabilistic model quantifies the amount of aesthetic features (post-embedding graphlets) that are shared between the training photos (aesthetically pleasing) and the test one.

The rest of the paper is organized as follows: Section II briefly reviews the previous photo quality models. From Section III to Section V, we exploit the local and global structures from multiple channels to represent photo aesthetics, and a probabilistic photo aesthetic model is developed correspondingly. Experimental results in Section VI thoroughly demonstrate the effectiveness of our proposed model and Section VII concludes.

II. Related Work

Recently many photo aesthetics evaluation methods have been proposed. Roughly, these methods can be divided into two groups: global feature-based approaches and local patch integration-based approaches.

Global feature-based approaches [3], [11], [17], [39], [40] design global low-level and high-level visual features to represent photo aesthetics in an implicit manner. Ke *et al.* [11] designed a group of high-level visual features, such as the image simplicity based on the spatial distribution of edges, to imitate human perception of photo aesthetic quality. Datta *et al.* [3] proposed 58 low-level visual features, *e.g.*, shape convexity, to capture photo aesthetics. Dhar *et al.* [39] proposed a set of high-level attribute-based predictors to evaluate photo aesthetics. In [40], Luo *et al.* proposed using the GMM (Gaussian mixture model)-based hue distribution and the prominent lines extraction-based texture distribution to represent the global composition of a photo. To describe local composition of a photo, three regional features respectively describing human faces, region clarity, and region complexity were developed. In [17], Marchesotti *et al.* proposed using generic descriptors, *i.e.*, the bag of visual words and the Fisher vector, to access photo aesthetics. Experimental results demonstrated that the two generic descriptors outperform many specifically designed photo aesthetic descriptors. Ji *et al.* [31] proposed a multi-channel coding based approach for mobile location recognition, in which different channel cues, which can largely ensure the search robustness to achieve the state-of-the-art search accuracy. For the aforementioned global aesthetic features, there is no strong indication that they can effectively capture photo aesthetics, such as the linearly arranged sailors in Fig. 1. This implies that they may perform unsatisfactorily on some photos. In particular, it is worth noting the limitations/shortcomings of the above global feature-based approaches: 1) Luo *et al.* [40]'s approach adopts a category-dependent regional feature extraction, which has the prerequisite that photos can be 100% accurately classified into one of the seven categories. This prerequisite, however, is infeasible in real applications. 2) The attributes proposed in Dhar *et al.* [39]'s approach are designed manually and are data set dependent, and thus prove difficult to generalize to different data sets. Third, all these global low-level and high-level visual features are designed heuristically. They model the statistics of visual descriptors within the whole image. There is no strong indication that they can accurately capture the photo local and global compositions. For example, the co-occurrence of the four sailboats and their linear spatial arrangements as shown in Fig. 1.

Local patch integration-based approaches [1], [2], [21], [29], [37] extract local patches within a photo and then integrate them to measure photo aesthetic quality. In [2], Cheng *et al.* proposed the omni-range context, *i.e.*, the spatial distribution of arbitrary pairwise image patches, to model photo composition. The learned omni-range context priors are combined with the other cues, such as the patch number, to form a posterior probability to measure the aesthetics of a photo. One limitation of Cheng *et al.*'s work is that only the binary correlation between image patches is considered. To describe high-order spatial interactions of image patches, Zhang *et al.* [29] introduced graphlets. And a probabilistic model is proposed to quantify photo aesthetics as the amount of graphlets that can be transferred them from the training photos into the cropped one. However, graphlets cannot reflect photo global spatial configurations, which are essential cues for determining photo aesthetics. Besides, the color and texture channel visual features are assigned with the same weight in the graphlet transferring phase, which is not consistent with human perception of photo aesthetics. In [37], Nishiyama *et al.* first detected multiple subject regions in a photo, where each subject region is a bounding rectangle containing the salient part of an object. Then, an SVM classifier is trained for each subject region. Finally, the aesthetics of each candidate cropped photo is computed by combining the scores of the SVM classifier corresponding to a photo's internal subject regions. One limitation of Nishiyama *et al.*'s approach is that it cannot model the spatial interaction of multiple image regions explicitly. Thus, this method cannot discriminate linearly or triangularly distributed sailboat as shown in Fig. 1. In [21], Nishiyama *et al.* proposed a color harmony-based photo aesthetic evaluation method. A color harmony model is first applied to the patches within a photo to describe their color distribution. The patch-level color distribution is then integrated into a bag-of-patches histogram. The histogram is further classified by an SVM to identify whether a photo is highly or low aesthetic. Note that, Nishiyama *et al.* [21] evaluates photo aesthetics by utilizing visual features in color channel only. Features capturing aesthetics in other channels, such as texture, are neglected. Bhattacharya *et al.* [1] proposed the spatial recomposition to allow users interactively select a foreground object. The system then presents recommendations to indicate an optimal location of the foreground object, which is detected by combining multiple aesthetic cues, such as the relative foreground position and the visual weight ratio. The major shortcoming of Bhattacharya *et al.*'s method is the necessity of human interaction, limiting its application to large scale photo aesthetics evaluation.

III. Graphlet-Based Local Structure Descriptor

There are usually tens to hundreds of components within a photo, such as the sailboats and sailors in Fig. 1. Among these components, a few spatially neighboring ones and their interactions capture photo local aesthetics. Since graph is a powerful tool to describe the relationships between objects, we use graph to model the spatial interactions between image components. Our technique is to segment a photo into a set of atomic regions using unsupervised fuzzy clustering (UFC) [26], where each atomic region denotes the segmented image patch. Based on this, we extract graphlets to characterize the local aesthetics of a photo. Graphlet is a small-sized connected graph defined as:

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

where V is a set of vertices representing locally distributed atomic regions (as the example in Fig. 2); and E is a set of edges, each of which connects pairwise spatially adjacent atomic regions.[1] We call a graphlet with t vertices a t-sized graphlet. Because the number of graphlets within a photo exponentially increases with graphlet size,[2] only small-sized graphlets are employed.

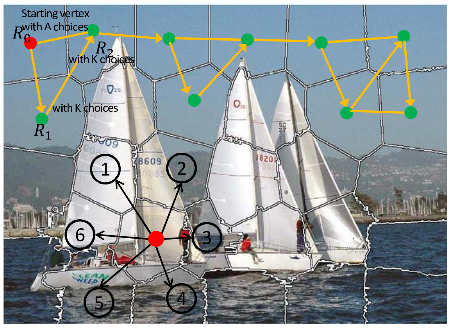


Fig. 3. An illustration of the degree of a superpixel as well as the random walking process.

As a purely data-driven segmentation algorithm, UFC produces numerous imperfectly segmented regions.[3] To maximally preserve optimally segmented regions, we generate a large number of atomic regions and then remove those are imperfect. Five times segmentation under UFC tolerance bounds {0.1, 0.2, 0.3, 0.4, 0.5} is applied firstly. Then tiny-sized regions (less than 50 pixels) are removed. Finally, segmented regions with low correlation to photo categories ($corr(R) < 1/13$) are abandoned since they reflect little semantics. The category correlation of segmented region is calculated by an LDA [14]-like measure. It shows that a higher discriminative segmented region is more correlated with the photo semantics:

$$corr(R) = \frac{\sum_{c(R_i)=c(R)} ||F(R_i) - F(R)||_{l_2}}{\sum_{c(R_i)\neq c(R)} ||F(R_i) - F(R)||_{l_2}}, \tag{2}$$

where $F(R)$ is a vector that combines the HOG [5] (128-dimensional) and the color moment [22] (9-dimensional) from segmented region R. $c(R_i)$ indicates the category of photo from which segmented region R_i is extracted. It is computed from the 13-class SVM trained from Feifei *et al.* [15]'s scene data set.

The graphlet extraction can be illustrated as a probabilistic walking process. As shown in Fig. 2, we first choose a starting vertex with a probability of $\frac{p(A)}{A}$, where $p(A)$ is the probability of A atomic regions existing in photo I. The spatially adjacent atomic regions are then visited one-by-one. The probability of visiting a spatially adjacent vertex is decided by the degree (*e.g.*, the superpixel marked as red in Fig. 3 is with degree 6.) of the current vertex, *i.e.*, $\frac{1}{\sum_d p_d(R_l)d(R_l)}$, where $d(R_l)$ denotes the degree of the current atomic region R_l and $p_d(R_l)$ is the corresponding probability. The visiting process stops when the maximum graphlet size is reached. Based on the above description, the probability of extracting a t-sized graphlet G from photo I is calculated by:

$$p(G|I) \propto \frac{p(A)}{A} \prod\nolimits_l^{t-1} \frac{1}{\sum_d p_d(R_l)d(R_l)}, \tag{3}$$

where $p(A)/A$ denotes the probability of choosing a starting vertex. As shown in Fig. 3, the probability of A vertices existing in a photo is $p(A)$, and the probability of selecting a vertex from these A atomic regions is $1/A$. Thus, the probability of choosing a starting vertex in a photo is $p(A)/A$. $\frac{1}{\sum_d p_d(R_l)d(R_l)}$ reflects the probability of choosing a vertex in the l-th step of random walking. It is noticeable that $\sum_d p_d(R_l)d(R_l)$ denotes the expectation degree of superpixel R_l. Due to the number of graphlets is exponentially increasing with graphlet size, it is computationally intractable to adopt all the t-sized graphlets

[1] Based on the definition of graphlets, as shown in Fig. 2, each vertex denotes an atomic region. Thus, we use "vertex" and "an atomic region“ indiscriminately in this paper.

[2] The number is $A * K^{t-1}/t!$ where K is the average degree of atomic regions; A counts atomic regions in an image; and t is the graphlet size.

[3] Imperfection means some segmented regions partially cover one or multiple semantic components.

into the proposed aesthetic model. Therefore, we sample 500 graphlets from each photo.

Because visual features from multiple visual channels collaboratively contribute to photo aesthetics, a 9-dimensional color moment [22], a 128-dimensional histogram of gradient (HOG) [5], and a 64-dimensional quantized visual saliency histogram, are used to describe each atomic region. Visual features from the three channels are used here because color and texture are generally complementary in describing the appearance of an atomic region, and the saliency channel indicates which atomic region is visually attractive. In this work, the visual saliency descriptor is implemented as the graph-based visual saliency (GBVS) [9]. We choose GBVS because: 1) compared with high-level saliency models that are manually designed and are data dependent, GBVS relies completely on the low-level visual features, making it more adaptable to real-world applications; 2) GBVS is among the top performers of the purely low-level visual feature-based saliency models. It is worth emphasizing that, for each atomic region, GBVS only outputs its pixel-level saliency map. In our approach, a K-means-based quantization is adopted for fixed-length vector representation.

The above three visual features result in three matrices $\mathbf{M}_R^C$,$\mathbf{M}_R^T$, and $\mathbf{M}_R^S$, describing the atomic regions of a graphlet in color, texture, and visual saliency channels respectively. Given a t-sized graphlet, each row of matrix $\mathbf{M}_R^C \in \mathbb{R}^{t\times 9}$ represents a 9-dimensional color moment of an atomic region ($\mathbf{M}_R^T$ and $\mathbf{M}_R^S$ are defined similarly). To capture the spatial interactions of atomic regions, we adopt a $t \times t$-sized adjacency matrix as:

$$\mathbf{M}_S(i,j) = \begin{cases} \theta(i,j) & \text{if } R_i \text{ and } R_j \text{ are spatially adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\theta(i,j)$ is the horizontal angle of the vector from the center of atomic region R_i to that of atomic region R_j. Based on $\{\mathbf{M}_R^C, \mathbf{M}_R^T, \mathbf{M}_R^S\}$ and $\mathbf{M}_S$, three matrices $\mathbf{M}^C = [\mathbf{M}_R^C, \mathbf{M}_S]$, $\mathbf{M}^T = [\mathbf{M}_R^T, \mathbf{M}_S]$, and $\mathbf{M}^S = [\mathbf{M}_R^S, \mathbf{M}_S]$ are constructed, which describe a graphlet in color, texture, and visual saliency channels respectively.

IV. Pursuing Global Spatial Layout on Manifold

The above matrix-form graphlets are descriptive, but they are still not ready for evaluating photo aesthetic quality. First, although different-sized graphlets have comparable aesthetic properties, *e.g.*, four and five linearly arranged skaters are aesthetically similar, their distance cannot be directly calculated as their corresponding matrices are with different sizes. Second, global composition plays an important role in photo aesthetics. However, as the number of graphlets is exponentially increasing with their size, only small-sized graphlets are employed. In this case, the small graphlet size limits the descriptive ability of graphlets to photo global spatial layout.

A. Manifold Embedding to Preserve Global Layout

It can be observed that, spatially neighboring graphlets in a photo are partially overlapping. This indicates that it is beneficial to exploit the local structure [16], [23] among graphlets. Therefore, we project the matrix-form graphlets onto a manifold, thereby the Golub-Werman distance [24] between identical-sized matrices is:

$$d_{GW}(\mathbf{M}, \mathbf{M}') = ||\mathbf{M}_O - \mathbf{M}'_O||_2, \tag{5}$$

where $\mathbf{M}_O$ and $\mathbf{M}'_O$ denote the orthonormal basis of $\mathbf{M}$ and $\mathbf{M}'$ respectively; and $||\cdot||_F$ denotes the Frobenius norm.

Inspired by the patch alignment framework [28], we propose a graphlet embedding algorithm to 1) transform different-sized graphlets from multiple visual channels into equal-lengthed vectors, and 2) encode the global spatial layout of each photo into its constituent graphlets (*i.e.* graphlets extracted from a photo). The graphlet embedding algorithm contains two parts. As shown on the right of (5), the first part incorporates the global spatial layout of each photo. That is, it minimizes the discrepancy between the distances between graphlets on the manifold and those in the Euclidean space. For graphlets in color/texture/saliency channel, the objective function is:

$$\arg\min_{\mathbf{Y}^h} \sum\nolimits_{ij} [d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h) - d_E(y_i^h, y_j^h)]^2, \tag{6}$$

where $\mathbf{M}_i^h$ and $\mathbf{M}_j^h$ are $t \times (F+t)$-sized matrices (F denotes the feature dimension in color/texture/saliency channel) to the i-th and the j-th identical-sized graphlets, from the h-th photo; y_i^h and y_j^h are their d-dimensional vectors; $d_{GW}(\cdot,\cdot)$ and $d_E(\cdot,\cdot)$ are the Golub-Werman distance [24] and Euclidean distance between identical-sized matrices respectively. (6) is an objective function that minimizes the Golub-Werman distance between graphlets and the Euclidean distance between post-embedding graphlets. The Golub-Werman distance between graphlets is $d_{GW}(M_i^h, M_j^h)$. The Euclidean distance between graphlets is $d_E(y_i^h, y_j^h)$. The global spatial layout of a photo can be considered as the relative position of all pairwise graphlets in a photo. If we preserve all these relative distances in the graphlet embedding, the global spatial layout can be preserved.

Based on the derivation in the Appendix, the above objective function can be reorganized as:

$$\begin{aligned} &\arg\min_{\mathbf{Y}^h} \sum\nolimits_{ij} [d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h) - d_E(y_i^h, y_j^h)]^2 \\ &\qquad = \arg\max_{\mathbf{Y}^h} tr(\mathbf{Y}^h \mathbf{Z}^h (\mathbf{Y}^h)^T), \end{aligned} \tag{7}$$

where $\mathbf{Y}^h = [y_1^h, y_2^h, \ldots, y_N^h] \in \mathbb{R}^{d\times N_h}$ denotes the matrix containing all the post-embedding graphlets from the h-th photo, $\mathbf{Z}^h = -\mathbf{R}_{N_h}\mathbf{S}_{GW}^h\mathbf{R}_{N_h}/2$.

By summing the graphlet embedding from all the photos in color, texture, and saliency channel, the second part embedding is given as:

$$\begin{aligned} &\arg\max_{\mathbf{Y},\alpha} \sum\nolimits_{h=1}^{H} \sum\nolimits_{k=1}^{3} \alpha_k^r \mathrm{tr}(\mathbf{Y}\mathbf{S}^h\mathbf{Z}_k^h(\mathbf{S}^h)^T\mathbf{Y}^T) \\ &\quad = \arg\max_{\mathbf{Y},\alpha} \sum_{k=1}^{3} \alpha_k^r \mathrm{tr}(\mathbf{Y}\mathbf{Z}_k\mathbf{Y}^T) \quad s.t.\ \mathbf{Y}\mathbf{Y}^T = \mathbf{I}_d, \sum\nolimits_k \alpha_k = 1, \end{aligned} \tag{8}$$

where $\mathbf{Y} = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{d\times N}$ is a matrix containing all the post-embedding graphlets; $\mathbf{Z}_k = \sum_{h=1}^{H} \mathbf{S}^h\mathbf{Z}_k^h(\mathbf{S}^h)^T$;

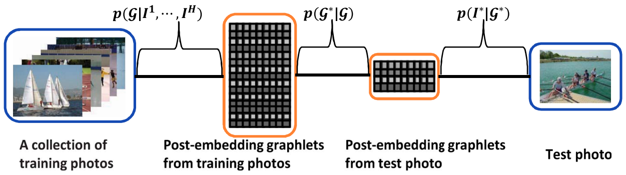

$p(\mathcal{G}|I^1,\cdots,I^H)$
$p(\mathcal{G}^*|\mathcal{G})$
$p(I^*|\mathcal{G}^*)$
A collection of training photos
Post-embedding graphlets from training photos
Post-embedding graphlets from test photo
Test photo

Fig. 4. The probabilistic model for aesthetic evaluation.

the constraint $\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_d$ uniquely determines the embedding $\mathbf{Y}$; and $r > 1$ determines the complementary property of the multiple visual channels, as detailed in the appendix.

V. A Probabilistic Aesthetics Measure

The post-embedding graphlets capture both the local and global spatial layouts from multiple visual channels in a photo. To effectively leverage them for photo aesthetics evaluation, a probabilistic model is proposed.

Given a set of training photos $\{I^1,\ldots,I^H\}$ and a test one I^*, they are highly correlated through their respective graphlets $\mathcal{G}$ and $\mathcal{G}^*$. Thus, a probabilistic graphical model is utilized to describe this correlation. As shown in Fig. 4, the graphical model contains two types of nodes: observable nodes (blue rectangle) and hidden nodes (orange rectangle). The probabilistic graphical model can be divided into four layers. The first layer represents all training photos, the second layer denotes the post-embedding graphlets from the training photos, the third layer represents all the post-embedding graphlets from the test photo, and the last layer denotes the test photo. The correlation between the first and second layers is $p(\mathcal{G}|I^1,\ldots,I^H)$, the correlation between the second and third layers is $p(\mathcal{G}^*|\mathcal{G})$, and the correlation between the third and fourth layers is $p(I^*|\mathcal{G}^*)$.[4]

According to the formulation above, photo aesthetics can be quantified as the similarity between post-embedding graphlets from the test photo and those from the training aesthetically pleasing photos. This similarity is interpreted as the amount of graphlets that can be transferred from the training photos into the test one. That is, the aesthetic quality $\gamma(I^*)$ of a test photo I^* is measured as:

$$\begin{aligned}\gamma(I^*) &= p(I^*|I^1,\ldots,I^H)\\ &= p(I^*|\mathcal{G}^*) * p(\mathcal{G}^*|\mathcal{G}) * p(\mathcal{G}|I^1,\ldots I^H), \qquad (9)\end{aligned}$$

The probabilities $p(I^*|\mathcal{G}^*)$, $p(\mathcal{G}^*|\mathcal{G})$, and $p(\mathcal{G}|I^1,I^2,\ldots,I^N)$ in (10) are computed respectively as:

$$\begin{aligned}p(I^*|\mathcal{G}^*) &= p(I^*|\mathcal{G}_1^*,\ldots,\mathcal{G}_T^*)\\ &= \frac{p(\mathcal{G}_1^*,\ldots,\mathcal{G}_T^*|I^*)p(I^*)}{p(\mathcal{G}_1^*,\ldots,\mathcal{G}_T^*)}\\ &= \prod\nolimits_{t=1}^{T}\prod\nolimits_{j=1}^{N_*^t} p(\mathcal{G}_t^*(j)|I^*), \qquad (10)\\ p(\mathcal{G}^*|\mathcal{G}) &= p(\mathcal{G}_1^*,\ldots,\mathcal{G}_T^*|\mathcal{G}_1,\ldots,\mathcal{G}_T)\\ &= \prod\nolimits_{t=1}^{T}\prod\nolimits_{j=1}^{N^t} p(\mathcal{G}_t^*(j)|\mathcal{G}_1,\ldots,\mathcal{G}_T), \qquad (11)\end{aligned}$$

Algorithm 1 Probabilistic Photo Aesthetic Quality Evaluation

```
input: training photos $\{I^1,\cdots,I^H\}$ labeled by their categories;
a test photo $I^*$, and the maximum graphlet size $T$;
output: the aesthetic score $\gamma$ of the test photo $I^*$;
1. Extract $\{1,\cdots,T\}$-sized graphlets from the training and the
test photos in color, texture, and saliency channel respectively;
2. Obtain the matrix of each graphlet, and transform graphlets
into $d$-dimensional feature vector according to (8);
3. Calculate the aesthetic score of test photo $I^*$ based on (10).
```

$$\begin{aligned}p(\mathcal{G}|I^1,\ldots,I^H) &= p(\mathcal{G}_1,\ldots,\mathcal{G}_T|I^1,\ldots,I^H)\\ &= \prod\nolimits_{t=1}^{T}\prod\nolimits_{j=1}^{N^t} p(\mathcal{G}_t(j)|I^1,\ldots,I^H), \qquad (12)\end{aligned}$$

where $\mathcal{G}_t$ denotes all the training t-sized graphlets; $\mathcal{G}_t(j)$ is the j-th training t-sized graphlets; $\mathcal{G}_t^*$ denotes the t-sized test graphlets; $\mathcal{G}_t^*(j)$ is the j-th test t-sized graphlet; N^t is the number of training t-sized graphlets and N_*^t the number of test t-sized graphlets.

To calculate (11), (12) and (13), three probabilities $p(\mathcal{G}_t^*(j)|I^*)$, $p(\mathcal{G}_t^*(j)|\mathcal{G}_1,\ldots,\mathcal{G}_T)$ and $p(G_t(j)|I^1,\ldots,I^H)$ are required. First, $p(\mathcal{G}_t^*(j)|I^*)$ is the probability of extracting graphlet $\mathcal{G}_t^*(j)$ from test photo I^*, which is computed based on (3). Second, $p(\mathcal{G}_t^*(j)|\mathcal{G}_1,\ldots,\mathcal{G}_T)$ is the probability of graphlet $\mathcal{G}_t^*(j)$ existing in $\mathcal{G}_1,\ldots,\mathcal{G}_T$. Inspired by many previous works such as [25], this probability can be defined as a Gaussian kernel:

$$p(\mathcal{G}_t^*(j)|\mathcal{G}_1,\ldots,\mathcal{G}_T) = \exp\left(-\frac{\sum_{G\in\mathcal{G}_1,\ldots,\mathcal{G}_T}||\mathcal{G}_t^*(j)-G||}{|\mathcal{G}_1,\ldots,\mathcal{G}_T|}\right), \qquad (13)$$

Third, $p(\mathcal{G}_t(j)|I^1,I^2,\ldots,I^H)$ is the probability of graphlet $\mathcal{G}_t(j)$ coming from all the training photos $\{I^1,I^2,\ldots,I^H\}$, which is computed as follows:

$$p(\mathcal{G}_t(j)|I^1,\ldots,I^H) = 1-\prod\nolimits_{h=1}^{H}\left(1-p(\mathcal{G}_t(j)|I^h)\right), \qquad (14)$$

This equation is explained as follows: $1-p(G_i(j)|I^h)$ is the probability of graphlet $G_i(j)$ not coming from photo I^h. Straightforwardly, $\Pi_{h=1}^{H}(1-p(G_i(j)|I^h))$ is the probability of $G_i(j)$ not coming from any of $\{I^1,\ldots,I^H\}$. Thus, $1-\Pi_{h=1}^{H}(1-p(G_i(j)|I^h))$ is the probability of $G_i(j)$ coming from $\{I^1,\ldots,I^H\}$.

By summarizing the discussion from Section III to Section V, the pipeline of our probabilistic graphlet-guided photo aesthetics evaluation is summarized in Algorithm 1.

VI. Experimental Results and Analysis

This section evaluates the effectiveness of the proposed method, which can be divided into four parts. The first part compares our approach with well-known photo aesthetics evaluation methods. The second part step-by-step evaluates each component of the proposed approach. In the third part, we discuss the influence of the two free parameters. Lastly, we illustrate the relationships between the proposed method and the four prominent aesthetic rules.

As far as we know, there are three off-the-shelf data sets for evaluating photo aesthetics: the CUHK [11], the Photo.net [3],

[4] To reduce time consumption, our probabilistic model allows for employing a small proportion of training and test graphlets, where $p(\mathcal{G}|I^1,\ldots,I^H)$ and $p(I^*|\mathcal{G}^*)$ are defined in (24) and (22) respectively. If all the graphlets are used, then $p(\mathcal{G}|I^1,\ldots,I^H)=1$ and $p(I^*|\mathcal{G}^*)=1$.

and the AVA [19]. A rough description of the three data sets are given as follows:

- The CUHK contains 12,000 photos collected from DPChallenge.com. These photos have been labeled by ten independent viewers. Each photo is classified as highly aesthetic if more than eight viewers agree on the assessment. For this data set, we use the standard split of training/test image sets.
- The Photo.net consists of 3581 images. Only URLs of the original photos are provided. Approximately half images were removed from the websites, leaving only nearly 1,700 images available. Thus, we extend this data set by online crawling 4,000 photos and naming the extended Photo.net data set PNE. The aesthetics of these additionally crawled photos are manually labeled and are randomly split into equal partitions, one for training and the rest for testing.
- The AVA [19] contains 25,000 highly- and low- aesthetic photos in total, each of which is associated with two semantic tags. The selection criteria is based on the aesthetic quality of each photo, which is scored by 78 to 549 amateur/professional photographers. The training and test photos of the AVA data set are pre-specified.

In our experiment, for the classifier-based photo aesthetic models, such as those proposed by Marchesotti *et al.* [17] and Nishiyama *et al.* [37], both the highly- and low-aesthetic training photos are adopted to learn the model. Particularly, the highly-aesthetic photos function as the positive samples while the low-aesthetic ones as the negative samples. For those models that are based on transferring aesthetic features, such as Cheng *et al.* [2]'s model, they employ those "good" aesthetic features to evaluate a test photo. Thus, it is necessary to assign a weight for each graphlet that denotes its aesthetics, that is, a larger weight reflects a higher aesthetic level. And the weight is determined by the aesthetics of the photo from which the graphlet is extracted. For the three data sets, different experimental settings are used to assign the weight of each photo. For the CUHK, we use the probabilistic output from Yan *et al.*'s work [11] to rank the aesthetics of each photo. For the PNE, we manually selected 674 highly-aesthetic photos and leave the rest as the low-aesthetic ones. Then, we extracted the aesthetic features based on [11], and further used a probabilistic SVM output to score the aesthetics of each photo. For the AVA, each training photo is rated according to their aesthetics on a scale of $\{0.1, 0.2, \ldots, 1\}$. We average the rating scores as the aesthetics of each photo. The aesthetics of these additionally crawled photos are manually labeled by 23 students from the department of computer science at Zhejiang University. Most of them are experienced with photography.

All the experiments were carried out on a personal computer with an Intel E8500 processor and 4GB RAM. The algorithm was implemented on the Matlab 2011 platform.

A. Comparison With the Existing Aesthetic Evaluation Models

In this subsection, we compare our approach with five photo aesthetics evaluation methods: 1) three global features-based approaches respectively proposed by Dhar *et al.* [39], Luo *et al.* [40], and Marchesotti *et al.* [17]; and 2) two local patch integration-based methods proposed by Cheng *et al.* [2] and Nishiyama *et al.* [37] respectively.

TABLE I

COMPARISON OF AESTHETICS PREDICTION ACCURACIES

	CUHK	PNE	AVA
Dhar *et al.*	0.7386	0.6754	0.6435
Luo *et al.*	0.8004	0.7213	0.6879
Marchesotti *et al.*(FV-Color-SP)	0.8767	0.8114	0.7891
Cheng *et al.*	0.8432	0.7754	0.8121
Nishiyama *et al.*	0.7745	0.7341	0.7659
The proposed method	0.9031	0.8302	0.8324

In the comparative study, we notice that the source codes of the above five compared methods are not provided and some experimental details are not mentioned, therefore it is difficult to strictly implement them. Toward a convincing comparative study, in our implementation, we tend to strengthen some components of the compared methods. Based on this, we adopt the following implementation settings: For Dhar's approach, we use the public code from Li *et al.* [12] to extract the attributes from each photo. These attributes are combined with the low-level features proposed by Yeh *et al.* [38] to train the aesthetics classifier. For Luo *et al.*'s approach, not only the low-level and high-level features in their publication are implemented, but also the six global features from Getlter *et al.* [8] are used to strengthen the aesthetic prediction ability. For Marchesotti *et al.*'s approach, similar to the implementation of Luo *et al.*'s method, the six additional features are also adopted. For Cheng *et al.*'s approach, we implemented it as a simplified version of our approach, *i.e.*, only 2-sized graphlets are employed for aesthetics measure. Noticeably, for the three probabilistic model-based aesthetic evaluation methods respectively proposed by Cheng *et al.*, Nishiyama *et al.*, and us, given a test photo, if the aesthetics probability calculated by (9) is larger than 0.5, then this photo is deemed as highly aesthetic, and vice versa. We choose 0.5 as the threshold because for each of the three data sets, half of the photos are highly aesthetic.

We present the aesthetics prediction accuracy on the CUHK, the PNE, and the AVA in Table I. On the three data sets, our approach outperforms Marchesotti *et al.*'s approach by nearly 2%, and exceeds the rest of the compared methods by more than 5%, which demonstrates the effectiveness our approach.

B. Discriminative Ability Evaluation

Each image can be represented by a set of graphlets. The extracted graphlets are planar visual features in $\mathbb{R}^2$. Unfortunately, conventional classifier, such as SVM, can only handle 1-D vector form features. Moreover, both the number and the size of the extracted graphlets are different from one image to another. Thus, it would be impractical for a conventional classifier such as SVM to carry out classification directly based on the extracted graphlets. To tackle this problem, a quantization scheme is developed to transform the extracted graphlets into 1-D vectors. Particularly, the quantization is

TABLE II

COMPARISON OF CATEGORIZATION PERFORMANCE ON THE PASCAL VOC 2009

	FV-Color-SP	FV-SIFT-SP	SC-SIFT-SP	Our (sing-seg)	Our (multi-seg)
Aero plane	27.22%	66.67%	63.34%	35.67%	44.56%
Bicycle	18.81%	49.93%	45.57%	31.21%	38.82%
Bird	17.79%	41.21%	37.79%	26.54%	31.54%
Boat	23.45%	48.98%	44.46%	29.89%	36.65%
Bottle	10.03%	28.87%	32.27%	17.57%	22.16%
Bus	30.12%	56.68%	50.03%	36.65%	43.38%
Car	21.96%	51.18%	46.67%	28.81%	32.19%
Cat	20.14%	53.31%	48.89%	26.63%	32.28%
Chair	19.96%	43.65%	42.21%	25.59%	33.47%
Cow	12.28%	37.76%	35.57%	17.72%	23.15%
Dining table	18.89%	41.24%	37.77%	23.39%	28.81%
Dog	17.65%	42.29%	38.79%	24.41%	29.11%
Horse	23.46%	54.48%	50.03%	28.88%	32.67%
Motorbike	25.46%	55.79%	52.23%	33.35%	38.86%
Person	36.68%	63.35%	59.91%	41.19%	47.73%
Potted plant	9.12%	24.43%	20.16%	14.46%	19.90%
Sheep	8.87%	34.48%	30.12%	13.32%	18.87%
Sofa	11.13%	43.38%	37.79%	15.58%	21.36%
Train	34.46%	64.59%	58.96%	39.94%	44.54%
Tv/monitor	13.32%	49.97%	43.37%	17.76%	26.79%
Average	20.40%	45.53%	43.79%	26.43%	32.43%

TABLE III

COMPARISON OF CATEGORIZATION PERFORMANCE ON THE MIT INDOOR 67

	FV-Color-SP	FV-SIFT-SP	SC-SIFT-SP	IFV-PART	Our (sing-seg)	Our (multi-seg)
Ave. Acc.	16.68%	31.24%	33.45%	61.05%	26.78%	32.24%

inspired by graph kernel [10], where each element of the vector $\mathcal{A} = [a_1, a_2, \ldots, a_N]$ is calculated as:

$$a_i \propto \exp\left(-\frac{1}{N \cdot N_i} \sum\nolimits_{y \in I, y' \in I_i} d(y, y')\right), \tag{15}$$

where N and N_i respectively denote the number of graphlets in photo I and I_i; y and y_i are the post-embedding graphlets from photo I and I_i respectively.

On the basis of the feature vector obtained above, a multiclass SVM is trained. That is, for the training images from the p-th and the q-th classes, we construct the following binary SVM classifier:

$$\max_{\alpha \in \mathbb{R}^{N_{pq}}} W(\alpha) = \sum\nolimits_{i=1}^{N_{pq}} \alpha_i - \frac{1}{2} \sum\nolimits_{i=1}^{N_{pq}} \alpha_i \alpha_j l_i l_j k(\mathcal{A}_i, \mathcal{A}_j)$$

$$s.t. \quad 0 \le \alpha_i \le C, \sum\nolimits_{i=1}^{N_{pq}} \alpha_i l_i = 0, \tag{16}$$

where $\mathcal{A}_i \in \mathbb{R}^N$ is the quantized feature vector from the i-th training image; N is the number of training images; l_i is the class label (+1 for the p-th class and -1 for the q-th class) to the i-th training image; α determines the hyper-plane to separate images in the p-th class from those in the q-th class; $C > 0$ trades the complexity of the machine off the number of nonseperable images; and N_{pq} is the number of training images from both the p-th and the q-th classes.

Given a quantized feature vector $\mathcal{A} \in \mathbb{R}^N$ obtained from a test image, its label (p or q) is classified by:

$$\mathrm{sgn}\left(\sum\nolimits_{i=1}^{N_{pq}} l_i \alpha_i k(\mathcal{A}_i, \mathcal{A}) + b\right), \tag{17}$$

where the bias $b = 1 - \sum_{i=1}^{N_{pq}} l_i \alpha_i k(\mathcal{A}_i, \mathcal{A}_s)$ and $\mathcal{A}_s$ is a support vector with class label +1. During testing, classification is conducted $C(C-1)/2$ times and the voting rule is utilized to get the final decision. Each binary classification can be deemed to be a voting process wherein votes can be cast for $\mathcal{A}$, and $\mathcal{A}$ is assigned to a class with the maximum number of votes.

Based on the above kernel SVM, a multi-class SVM is trained for image categorization. We experiment on the PASCAL VOC 2009 [18], and the training/validation/test splits are set as defaults. We compare the proposed kernel with FV-Color-SP in Marchesotti *et al*'s work, FV-SIFT-SP as illustrated in Chatfield *et al.*'s work, and SC-SIFT-SP proposed by Yang *et al.* [45]. For our approach, both a single segmentation and multiple segmentations are adopted to decompose each image into numerous atomic regions. As shown in Table II, our approach significantly outperforms Marchesotti *et al*'s method, which is in line with the aesthetics prediction performance in Table I. Besides, our approach is less effective than the SIFT pyramid. This is because there are a huge number of graphlets within an image, some of which contribute slightly or even negatively to the categorization performance. Toward a better categorization performance, a graphlet selection can be adopted in the future.

Lastly, we compare the categorization performance of our approach on the MIT 67 [36] indoor scenes data set. In addition to the above compared methods, we incorporate a

TABLE IV
AESTHETICS PREDICTION ACCURACY DECREMENT

	CUHK	PNE	AVA
Graphlet→single atomic region	-3.34%	-3.04%	-4.05%
Remove adj. mat from graphlet	-3.21%	-3.14%	-2.23%
Mani.Grap. emb.→Single-ch. emb.	-2.11%	-1.87%	-2.01%
Mani.Grap. emb.→kernel PCA	-6.45%	-5.69%	-4.79%
Prob. mea. → clasf. Mea.	-2.36%	-1.89%	-1.98%
Abandon color channel	-9.52%	-7.88%	-7.46%
Abandon texture channel	-4.33%	-3.79%	-3.46%
Abandon visual saliency channel	-4.77%	-4.57%	-4.12%

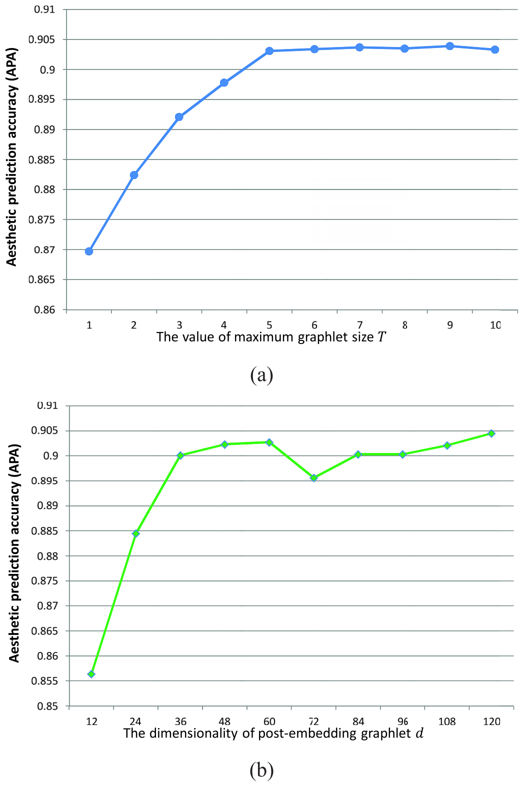
well-known part-based model proposed by Juneja *et al.* [35]. As shown in Table III, the categorization performance is consistent with that on the PASCAL VOC 2009.

C. Step-by-Step Model Justification

This experiment justifies the effectiveness of the three main components in our graphlet-based photo aesthetics model: graphlet-based local compositional descriptor extraction, multi-channel graphlet embedding, and probabilistic aesthetics measure.

- To evaluate the effectiveness of the first component, two experimental settings are adopted: 1) reducing graphlet to a single atomic region that captures no contextual cues (Graphlet→single atomic region); and 2) removing the adjacent matrix term from graphlets (Remove adj. mat from graphlet), which abandons the spatial cues of graphlets.
- To testify the effectiveness of the second component, two experimental settings are applied: 1) reducing the multi-channel graphlet embedding to single channel one (Mani.Grap. emb.→Single-ch. emb.), where only the color channel is used. Color channel is preserved because as shown by many photo aesthetics methods [21], it is the most important channel for representing photo aesthetics; and 2) replacing the manifold graphlet embedding by kernel PCA (Mani.Grap. emb.→kernel PCA), where the kernel is computed as: $k(G, G') \propto \exp(-d_{GW}(\mathbf{M}, \mathbf{M}'))$, $\mathbf{M} = [\mathbf{M}_R^C, \mathbf{M}_R^T, \mathbf{M}_R^S, \mathbf{M}_S]$ and $d_{GW}(\cdot, \cdot)$ is the Golub-Werman [24] distance between identical-sized matrices.
- To demonstrate the effectiveness of the third component, we replace the probabilistic aesthetics measure by a kernel SVM-based one (Prob. mea. → clasf. Mea.), wherein the kernel is computed based on (2).
- Finally, to demonstrate the importance of the three visual cues: color, texture, and visual saliency, we report the aesthetic prediction accuracy by abandoning each of the three cues. As shown in Table IV, when the color channel visual cue is removed, we observe the highest performance decrease of the aesthetic evaluation. This clearly confirms the importance of color in aesthetics prediction, which is consistent with the results reported in Marchesotti *et al.* [17]'s work.

As shown in Table IV, when replacing one component of the proposed approach with an existing one, aesthetics prediction accuracy reduces dramatically. This implies that each component of the proposed approach is indispensable and inseparable. In addition, the performance decrement reflects the importance of each component. As can be seen, manifold graphlet embedding, the key contribution of the proposed approach, plays the most important role in the proposed aesthetics model.



(a)

(b)

Fig. 5. Performance of photo aesthetics evaluation under different parameters. (a) Performance under different maximum graphlet sizes. (b) Performance under different dimensionalities of post-embedding graphlets d.

D. Parameter Analysis

This experiment evaluates the influence of the graphlet size T and the dimensionality of post-embedding graphlets d, on the performance of the proposed approach.

To analyze the effects of the maximum graphlet size T on evaluating photo aesthetics, we set up an experiment by varying T continuously. In Fig. 5(a), we present the aesthetics prediction accuracy (APA) when the maximum size of graphlet is tuned from 1 to 10. As can be seen, prediction accuracy increases moderately when $T \in [1, 5]$ but remains stable when $T \in [6, 10]$. This observation implies that 6-sized graphlets are sufficient for capturing the local composition of images from the CUHK. Also in Fig. 5(b), we present the performance of our model when the dimensionality of post-embedding graphlets is tuned from 12 to 120 with a step size of 12. As can be seen, the prediction accuracy increases steadily when $d \in [12, 36]$ but remains stable when $d \in [36, 120]$.

E. Photo Ranking Results

This subsection presents the photos of the three data sets that are ranked by our probabilistic photo aesthetics measure.

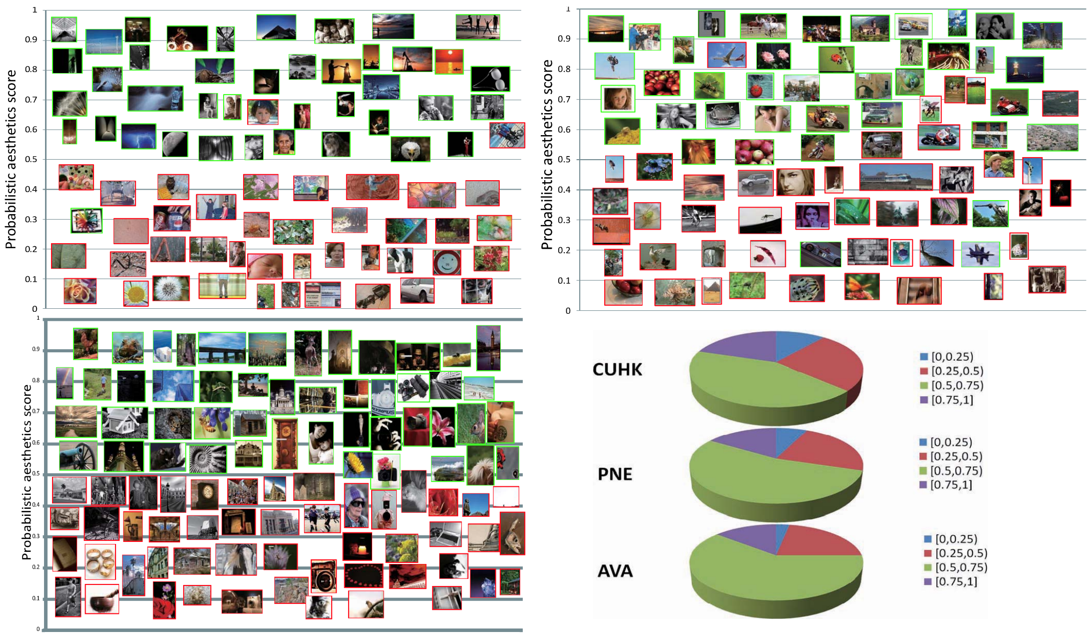


Fig. 6. Ranking results on the CUHK (top left), the PNE (top right), and the AVA (bottom left). Photos with the green rectangles indicate highly aesthetic test photos, and those with red rectangles are deficiently aesthetic test photos. The three pie charts denote the statistics of images from three data sets, according to the proposed model.

As can be seen from Fig. 6, we made the following three observations.

- As shown in the photos ranked between 0.8 and 1, highly aesthetic photos with multiple interacting objects are ranked with very high scores, demonstrating that the post-embedding graphlets effectively capture both local and global aesthetics of a photo.
- As seen from the photos ranked between 0.5 and 0.8, highly aesthetic photos with a single object are also appreciated by the proposed aesthetic model. This is because graphlets are naturally local composition descriptors, and they influence the proposed photo aesthetics based on the proposed probabilistic model.
- Objects from the photos ranked between 0 and 0.5 are either spatially disharmonically distributed or blurred. Thus, these photos are considered as aesthetically low by our model.

VII. Conclusion

By discovering both the local and the global spatial structure among image regions, this paper presents a probabilistic model for photo aesthetics evaluation. In particular, we first extract graphlets which represent photo local composition. Then, these graphlets are projected onto the Grassmann manifold, based on which a manifold embedding algorithm encodes global layout and multi-channel visual features into graphlets. Finally, these post-embedding graphlets are integrated to form a probabilistic measure for evaluating photo aesthetics. Experimental results demonstrate the proposed approach outperforms its competitors. The visualized cropping results confirm photo aesthetics are appropriately captured.

In the future, we plan to develop a more general and comprehensive photo aesthetics evaluation model that includes not only the spatial interaction of image regions, but also other important photography elements such as exposure, contrast, *etc*. In addition, we want to propose a weakly supervised learning paradigm to transfer the image-level semantics into graphlets.

Appendix

A. Derivation from (6) to (7)

Denote $\mathbf{D}^h_{GW} = [d_{GW}(\mathbf{M}^h_i, \mathbf{M}^h_j)]$ as a matrix whose ij-th element is the Golub-Werman distance between the i-th and the j-th graphlets from the h-th photo. Then, the inner product matrix is obtained by $\mathbf{Z}^h = -\mathbf{R}_{N_h}\mathbf{S}^h_{GW}\mathbf{R}(N_h)^T$, wherein $(\mathbf{S}^h_{GW})_{ij} = (\mathbf{D}^h_{GW})^2_{ij}$, $\mathbf{R}_{N_h} = \mathbf{I}_{N_h} - \vec{\mathbf{e}}_{N_h}\mathbf{S}^h_{GW}\vec{\mathbf{e}}^T_{N_h}/N$ is the centralization matrix.

Based on the above formulation, (6) can be reorganized into:

$$\arg\min_{\mathbf{Y}^h} ||\mathbf{Z}^h - (\mathbf{Y}^h)^T\mathbf{Y}^h||^2 = \arg\min_{\mathbf{Y}^h} \mathrm{tr}(\mathbf{Z}^h(\mathbf{Z}^h)^T - 2\mathbf{Y}^h\mathbf{Z}^h(\mathbf{Y}^h)^T + (\mathbf{Y}^h)^T\mathbf{Y}^h(\mathbf{Y}^h)^T\mathbf{Y}^h). \qquad (18)$$

By assuming that $(\mathbf{Y}^h)^T\mathbf{Y}^h$ is a constant matrix, (29) can be rewritten as:

$$\arg\max_{\mathbf{Y}^h} \mathrm{tr}(\mathbf{Y}^h\mathbf{Z}^h(\mathbf{Y}^h)^T). \qquad (19)$$

B. Illustration of the Parameter r in (9)

If we ignore the parameter r (or set $r = 1$), then the solution to α in (9) is $\alpha_k = 1$ when maximizing $\mathrm{tr}(\mathbf{Y}^h_k\mathbf{Z}^h_k(\mathbf{Y}^h_k)^T)$, or $\alpha_k = 0$ otherwise. That means only one channel visual features is finally selected when $r = 1$. Obviously, this solution does not meet our objective on exploring the complementary properties of visual features from multiple channels.

We adopted the trick used in [41] to avoid this phenomenon, *i.e.*, we set $\alpha_i \leftarrow \alpha_i^r$ with $r > 1$. In this way, $\sum_k \alpha_k^r = 1$ achieves its maximum value when $\alpha_i = 1/3$. Straightforwardly, to maximize $\sum_k \alpha_k^r \mathrm{tr}(\mathbf{Y}\mathbf{Z}_k\mathbf{Y}^T) = 1$, α_i of different views will be obtained by setting $r > 1$, which means that each view has a particular contribution to the final low-dimensional embedding $\mathbf{Y}$. Also, we found that r determines the complementary property of different channels: rich complementation implies a larger r. In our experiment, we fix the value of r to 2.

References

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

[2] B. Cheng, B. Ni, S. Yan, and Q. Tian, “Learning to photograph,” in *Proc. Int. Conf. Multimedia*, 2010, pp. 291–300.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Studying aesthetics in photographic images using a computational approach,” in *Proc. ECCV*, 2006, pp. 288–301.

[4] S. Dhar, V. Ordonez, and T. L. Berg, “High level describable attributes for predicting aesthetics and interestingness,” in *Proc. CVPR*, 2011, pp. 1657–1664.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. CVPR*, 2005, pp. 886–893.

[6] T. Grill and M. Scanlon, *Photographic Cmposition*. New York, NY, USA: Amphoto Books, 1990.

[7] N. Guan, D. Tao, Z. Luo, and B. Yuan, “Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent,” *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, Jul. 2011.

[8] P. Gehler and S. Nowozin, “On feature combination for multiclass object classification,” in *Proc. 12th ICCV*, 2009, pp. 221–228.

[9] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. NIPS*, 2007, pp. 545–552.

[10] Z. Harchaoui and F. Bach, “Image classification with segmentation graph kernels,” in *Proc. CVPR*, 2007, pp. 1–8.

[11] Y. Ke, X. Tang, and F. Jing, “The design of high-level features for photo quality assessment,” in *Proc. CVPR*, 2006, pp. 419–426.

[12] 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. NIPS*, 2010, pp. 1378–1386.

[13] W. Luo, X. Wang, and X. Tang, “Content-based photo quality assessment,” in *Proc. ICCV*, 2011, pp. 2206–2213.

[14] Y. Li, S. Gong, and H. Liddell, “Kernel discriminant analysis,” *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 5, pp. 745–770, 1998.

[15] F.-F. Li and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *Proc. CVPR*, 2005, pp. 524–531.

[16] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, “Manifold regularized multitask learning for semi-supervised multilabel image classification,” *IEEE Trans. Image Process.*, vol. 32, no. 9, pp. 523–536, Feb. 2013.

[17] M. Veringham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *IJCV*, vol. 2, no. 88, pp. 303–338, 2010.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge(VOC2009)*, Oct. 2009.

[19] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *Proc. CVPR*, 2012, pp. 2408–2415.

[20] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, “Sensation-based photo cropping,” *ACM Multimedia*, 2009, pp. 669–672s.

[21] M. Nishiyama, T. Okabe1, I. Sato, and Y. Sato, “Aesthetic quality classification of photographs based on color harmony,” in *Proc. CVPR*, 2011, pp. 33–40.

[22] M. Stricker and M. Orengo, “Similarity of color images,” in *Proc. Storage Retr. Image Video Databases*, 1995, pp. 381–392.

[23] D. Song and D. Tao, “Biologically inspired feature manifold for scene classification,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.

[24] M. Werman and D. Weinshall, “Similarity and affine invariant distances between 2D point sets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 810–814, Aug. 1995.

[25] X. Wang and E. Grimson, “Spatial latent Dirichlet allocation,” in *Proc. NIPS*, 2007, pp. 1577–1584.

[26] X. Xiong and K. L. Chan, “Towards an unsupervised optimal fuzzy clustering algorithm for image database organization,” in *Proc. 15th ICPR*, 2000, pp. 897–900.

[27] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, “Personalized photograph ranking and selection system,” in *Proc. ACM Multimedia*, 2010, pp. 211–220.

[28] T. Zhang, D. Tao, X. Li, and J. Yang, “Patch alignment for dimensionality reduction,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[29] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, “Probabilistic graphlet transfer for photo cropping,” *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2887–2897, Feb. 2013.

[30] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, “Probabilistic graphlet cut, exploiting spatial structure cue for weakly supervied image segmentation,” *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 1908–1915, Jun. 2013.

[31] R. Ji, L.-Y. Duan, J. Chen, H. Yao, Y. Rui, S.-F. Chang, *et al.*, “Towards low bit rate mobile visual search with multiple channel coding,” in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 573–582.

[32] Y. Gao, M. Wang, Z. J. Zha, Q. Tian, Q. Dai, and N. Zhang, “Less is more: Efficient 3-D object retrieval with query view selection,” *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1071–1018, Oct. 2011.

[33] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu, “Visual-textual joint relevance learning for tag-based social image search,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[34] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, “Discovering discriminative graphlets for aerial image categories recognition,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5071–5084, Dec. 2013.

[35] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, “Blocks that shout: Distinctive parts for scene classification,” in *Proc. CVPR*, Jun. 2013, pp. 923–930.

[36] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *Proc. CVPR*, 2009, pp. 1–8.

[37] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, “Sensation-based photo cropping,” in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 669–672.

[38] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, “Personalized photograph ranking and selection system,” in *Proc. Int. Conf. Multimedia*, 2010, pp. 211–220.

[39] S. Dhar, V. Ordonez, and T. L. Berg, “High level describable attributes for predicting aesthetics and interestingness,” in *Proc. CVPR*, 2011, pp. 1657–1664.

[40] W. Luo, X. Wang, and X. Tang, “Content-based photo quality assessment,” in *Proc. ICCV*, 2011, pp. 2206–2213.

[41] T. Xia, D. Tao, T. Mei, and Y. Zhang, “Multiview spectral embedding,” *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.

[42] R. Ji, H. Yao, W. Liu, X. Sun, and Q. Tian, “Task dependent visual codebook compression,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2282–2293, Apr. 2012.

[43] R. Ji, L.-Y. Duan, H. Yao, L. Xie, Y. Rui, and W. Gao, “Learning to distribute vocabulary indexing for scalable visual search,” *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 153–166, Jan. 2013.

[44] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, “The devil is in the details: An evaluation of recent feature encoding methods,” in *Proc. BMVC*, 2011, pp. 1–12.

[45] J. Yang, K. Yu, Y. Gong, and T. S. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *Proc. CVPR*, 2009, pp. 1794–1801.

Luming Zhang is a Post-Doctoral Research Fellow with the School of Computing, National University of Singapore. His research interests include multimedia analysis, image enhancement, and pattern recognition.

Yue Gao (M'13) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China. His research interests include large scale multimedia retrieval and live social media analysis.

Roger Zimmermann (S'93–M'99–SM'07) received the M.S. and Ph.D. degrees from the University of Southern California in 1994 and 1998. He is currently an Associate Professor with the Department of Computer Science, National University of Singapore (NUS). He is a Deputy Director with the Interactive and Digital Media Institute at NUS and Co-Director of the Centre of Social Media Innovations for Communities. His research interests are in the areas of streaming media architectures, distributed and peer-to-peer systems, mobile and geo-referenced video management, collaborative environments, spatio-temporal information management, and mobile location-based services. He has co-authored a book, six patents, and more than 150 conference publications, journal articles, and book chapters. He is a member of ACM.

Qi Tian (M'96–SM'04) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the M.S. degree from Drexel University, Philadelphia, PA, USA, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, in 1996 and 2002, respectively, both in electrical and computer engineering. He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio, San Antonio. He was on a one-year faculty leave with Microsoft Research Asia from 2008 to 2009. He has authored or co-authored over 160 refereed journal and conference papers. His current research interests include multimedia information retrieval and computer vision. He was a recipient of the Faculty Research Award from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Laboratories, the Best Student Paper Award at ICASSP 2006, the Best Paper Candidate Award at PCM 2007, the 2010 ACM Service Award, the Top 10% Paper Award at MMSP 2011, and the Best Paper Award at ICIMCS 2012. He is the Guest Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, the *Journal of Computer Vision and Image Understanding*, *Pattern Recognition Letter*, the *EURASIP Journal on Advances in Signal Processing*, and the *Journal of Visual Communication and Image Representation*. He is on the editorial board of the IEEE TRANSACTIONS ON CIRCUIT AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal of Multimedia*, and the *Journal of Machine Visions and Applications*. His research projects were funded by NSF, ARO, DHS, SALSI, CIAS, and UTSA.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.