Perception-Guided Multimodal Feature Fusion for Photo Aesthetics Assessment

Luming Zhang[†], Yue Gao[†], Chao Zhang[‡], Hanwang Zhang[†], Qi Tian[◇], Roger Zimmermann[†]
[†]School of Computing, National University of Singapore, Singapore
[‡]Department of Computer Science, University of Illinois at Urbana-Champaign, USA
[◇]Department of Computer Science, University of Texas at San Antonio, USA
{zhanglm,gaoyue,hanwang,rogerz}@comp.nus.edu.sg, czhang82@illinois.edu, qitian@cs.utsa.edu

ABSTRACT

Photo aesthetic quality evaluation is a challenging task in multimedia and computer vision fields. Conventional approaches suffer from the following three drawbacks: 1) the deemphasized role of semantic content that is many times more important than low-level visual features in photo aesthetics; 2) the difficulty to optimally fuse low-level and high-level visual cues in photo aesthetics evaluation; and 3) the absence of a sequential viewing path in the existing models, as humans perceive visually salient regions sequentially when viewing a photo.

To address and solve these challenges, we propose a new aesthetic descriptor that mimics the way humans sequentially perceive visually/semantically salient regions in a photo. In particular, a weakly supervised learning paradigm is developed to project the local aesthetic descriptors (graphlets in this work) into a low-dimensional semantic space. Thereafter, each graphlet can be described by multiple types of visual features, both at low-level and in high-level. Since humans usually perceive only a few salient regions in a photo, a sparsity-constrained graphlet ranking algorithm is proposed that seamlessly integrates both the low-level and the high-level visual cues. The top-ranked graphlets are the discerned visually/semantically prominent graphlets in a photo. They are sequentially linked into a path that simulates the process of humans actively viewing. Finally, we learn a probabilistic aesthetic measure based on such actively viewing paths (AVPs) from the training photos that are marked as aesthetically pleasing by multiple users. Experimental results show that: 1) the AVPs are 87.65% consistent with real human gaze shifting paths, as verified by the eye-tracking data; and 2) our photo aesthetic measure outperforms many of its competitors.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

MM '14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654903.

Keywords

Weakly supervised; multimodal; sparsity; actively viewing; gaze shifting; path

1. INTRODUCTION

Photo aesthetic quality evaluation is a useful technique in multimedia applications. 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 prediction 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 three reasons.

- Semantics is an important cue to describe photo aesthetics, but the state-of-the-art models cannot exploit semantics effectively. Typically, a photo aesthetic model [**?**] only employs a few heuristically defined semantics according to a specific data set. They are determined by whether photos in a data set are covered by objects like sky, water, and *etc*. In addition, the semantics are typically detected using a set of external object detectors, *e.g.*, a human face detector. There is no guarantee that all the pre-specified semantic objects can be accurately discovered.

- Eye tracking experiments [**?**] have shown that humans allocate gazes to important regions in a sequential manner. As shown in Fig. 4, most people start with viewing the player, and then shift gazes to the grass, and finally to the audiences and the red track. Existing photo aesthetic models, however, fail to encode such a gaze shifting sequence.

- Psychophysics studies [**?**] have shown that both the bottom-up and the top-down visual features draw the attention of human eye. It is generally accepted that an aesthetic model should integrate both the low-level and the high-level visual cues. However, current models typically fuse multiple types of features in a linear or nonlinear way, where the cross-feature information is not well utilized. Even worse, these integration schemes cannot emphasize the visually/semantically salient regions within a photo.

To solve these problems, a sparsity-constrained ranking algorithm jointly discovers visually/semantically important graphlets along the human gaze shifting path, based on which a photo aesthetic model is learned. An overview of our proposed aesthetic model is

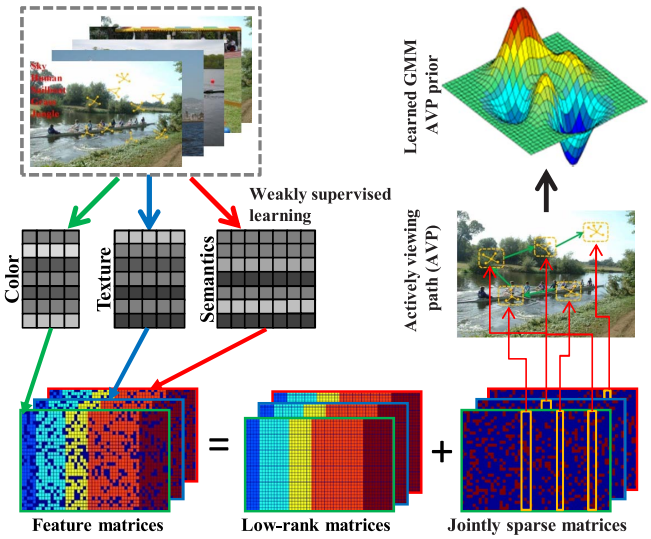

Figure 1: The pipeline of the proposed photo aesthetic model.

presented in Fig. 1. By transferring semantics of image labels into different graphlets of a photo, we represent each graphlet by a couple of low-level and high-level visual features. Then, a sparsity-constrained framework is proposed to integrate multiple types of features for calculating the saliency of each graphlet. Particularly, by constructing the matrices containing the visual/semantic features of graphlets in a photo, the proposed framework seeks the consistently sparse elements from the joint decompositions of the multiple-feature matrices into pairs of low-rank and sparse matrices. Compared with previous methods that linearly/non-linearly combine multiple global aesthetic features, our framework can seamlessly integrate multiple visual/semantic features for salient graphlets discovery. These discovered graphlets are linked into a path, termed actively viewing path (AVP), to simulate a human gaze shifting path. Finally, we employ a Gaussian mixture model (GMM) to learn the distribution of AVPs from the aesthetically pleasing training photos. The learned GMM can be used as the aesthetic measure, since it quantifies the amount of AVPs that are shared between the aesthetically pleasing training photos and the test image.

The main contributions of this paper are two-fold:

- A sparsity-constrained ranking framework that discovers visually/semantically important graphlets that draw the attention of the human eye, by seamlessly combining a few low-level and high-level visual features;
- Actively viewing path (AVP), a new aesthetic descriptor that mimics the way humans actively allocate gazes to visually/semantically important regions in a photo.

2. RELATED WORK

In recent years, many photo aesthetic quality evaluation methods have been proposed. Roughly, these methods can be divided into two categories: global feature-based approaches and local patch integration-based approaches.

Global feature-based approaches design global low-level and high-level visual features that represent photo aesthetics in an implicit manner. Ke *et al.* [**?**] developed a group of high-level visual features, such as an image simplicity based on the spatial distribution of edges, to imitate human perception of photo aesthetics. Datta *et al.* [**?**] proposed 58 low-level visual features, *e.g.*, shape convexity, to capture photo aesthetics. Dhar *et al.* [**?**] proposed a set of high-level attribute-based predictors to evaluate photo aesthetics. In [**?**], Luo *et al.* adopted a GMM-based hue distribution and a prominent lines-based texture distribution to represent the photo global composition. To capture the photo local composition, regional features describing human faces, region clarity, and region complexity were developed. In [**?**], Marchesotti *et al.* proposed using generic descriptors, *i.e.*, the bag of visual words and the Fisher vector, to access photo aesthetics. Experiments shown that the two generic descriptors outperform many specifically designed photo aesthetic descriptors. It is worth noting the limitations of the above approaches: 1) Luo *et al.* [**?**]'s approach relies on category-dependent regional feature extraction, requiring that photos can be 100% accurately classified into one of the seven categories. This prerequisite is infeasible in practice; 2) the attributes proposed by Dhar *et al.* [**?**] are designed manually and are data set dependent. Thus, it is difficult to generalize them to different data sets; and 3) all these global low-level and high-level visual features are designed heuristically. There is no strong indication that they can capture the complicated spatial configurations of different photos.

Local patch integration-based approaches extract patches within a photo and then integrate them to measure photo aesthetic quality. In [**?**], 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 model is that only the binary correlation between image patches is considered. To describe high-order spatial interactions of image patches, Nishiyama [**?**] *et al.* first detected multiple subject regions in a photo, where each subject region is a bounding rectangle containing the salient parts of an object. Then, an SVM classifier is trained for each subject region. Finally, the aesthetics of a test 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. In [**?**], 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 high or low aesthetics. It is noticeable that Nishiyama *et al.* [**?**] evaluates photo aesthetics by utilizing visual features in the color channel only. Features describing photo aesthetics in other channels, such as texture, are neglected. Bhattacharya *et al.* [**?**] proposed a spatial recomposition that allows users to interactively select a foreground object. The system presents recommendations to indicate an optimal location of the foreground object, which is detected by combining multiple aesthetic features, *e.g.*, the relative position of the foreground objects. The major shortcoming of Bhattacharya *et al.*'s method is the necessity of human interaction, limiting its application for large-scale photo aesthetics evaluation.

3. LOW-LEVEL AND HIGH-LEVEL LOCAL AESTHETICS DESCRIPTION

3.1 The Concept of Graphlets

There are usually a number of of components (*e.g.*, the human and the red track in Fig. 2) in a photo. Among these components, a few spatially neighboring ones and their spatial interactions cap-

ture the local aesthetics of a photo. Since graph is a powerful tool to describe the relationships among objects, we use it to model the spatial interactions of components in a photo. Our technique is to segment a photo into a set of atomic regions[1], and then construct graphlets to characterize the local aesthetics of this photo. In particular, a graphlet is a small-sized graph defined as:

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

where $\mathcal{V}$ is a set of vertices representing those spatially neighboring atomic regions; and $\mathcal{E}$ is a set of edges, each of which connects pairwise spatially adjacent atomic regions. We call a graphlet with t vertices a t-sized graphlet. It is worth emphasizing that the number of graphlets within a photo is exponentially increasing with the graphlet size. Therefore, only small graphlets (*i.e.*, vertex number less than 10) are employed.

In this work, we characterize each graphlet in both color and

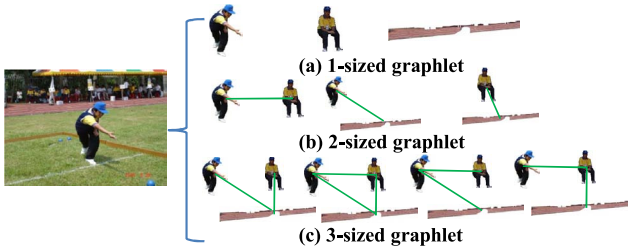

Figure 2: An example of differently sized graphlets extracted from a photo.

texture channels. Given a t-sized graphlet, each row of matrix $\mathbf{M}_r^c$ represents the 9-dimensional color moment [**?**] and each row of matrix $\mathbf{M}_r^t$ denotes the 128-dimensional HOG [**?**] of an atomic region. To describe the spatial interactions of atomic regions, we employ a $t \times t$ adjacency matrix as:

$$\mathbf{M}_s(i,j) = \begin{cases} \theta(R_i, R_j) & \text{if } R_i \text{ and } R_j \text{ are spatially adjacent} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $\theta(R_i, R_j)$ is the horizontal angle of the vector from the center of atomic region R_i to the center of atomic region R_j. Based on the three matrices $\mathbf{M}_r^c$, $\mathbf{M}_r^t$, and $\mathbf{M}_s$, the color and texture channel of a graphlet is described by $\mathbf{M}^c = [\mathbf{M}_r^c, \mathbf{M}_s]$ and $\mathbf{M}^t = [\mathbf{M}_r^t, \mathbf{M}_s]$, respectively. Moreover, similar to many feature fusion algorithms [**?**, **?**] that allow only vector representation of a sample, the color channel matrix $\mathbf{M}^c$ and the texture channel matrix $\mathbf{M}^t$ of a graphlet are further converted into vectors by staking each of them in row-wise.

3.2 Semantically Local Aesthetics Pursuit

In addition to color and texture channels description, high-level semantic cues should also be exploited for photo aesthetics evaluations. In this paper, the semantic cues are integrated based on a weakly supervised paradigm. Particularly, we transfer the semantics of image labels[2] into different graphlets in a photo. This is implemented based on a manifold embedding algorithm described as follows:

$$\begin{aligned} &\arg\min_{\mathbf{Y}} [\sum_{i,j} ||y_i - y_j||^2 l_s(i,j) - \sum_{i,j} ||y_i - y_j||^2 l_d(i,j)] \\ &= \arg\min_{\mathbf{Y}} \text{tr}(\mathbf{Y}\mathbf{R}\mathbf{Y}^T), \end{aligned} \tag{3}$$

where $\mathbf{Y} = [y_1, y_2, \cdots, y_N]$ denotes a collection of d-dimensional post-embedding graphlets; $l_s(\cdot,\cdot)$ and $l_b(\cdot,\cdot)$ are functions measuring the similarity and discrepancy between graphlets; $\mathbf{R} = [\vec{\mathbf{e}}_{N-1}^T, -\mathbf{I}_{N-1}]W_1[\vec{\mathbf{e}}_{N-1}^T, -\mathbf{I}_{N-1}] + \cdots + [-\mathbf{I}_{N-1}, \vec{\mathbf{e}}_{N-1}^T]\mathbf{W}_N[-\mathbf{I}_{N-1}, \vec{\mathbf{e}}_{N-1}^T]$; $\mathbf{W}_i$ is an $N \times N$ diagonal matrix whose h-th diagonal element is $[l_s(h,i) - l_d(h,i)]$.

More specifically, $l_s(\cdot,\cdot)$ and $l_d(\cdot,\cdot)$ are functions measuring

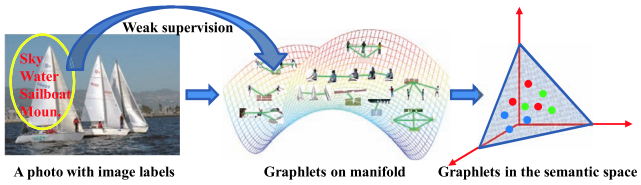

Figure 3: An illustration of embedding graphlets into the semantic space on manifold.

the semantic similarity and difference between graphlets, respectively. Denote $\mathbf{b}_i$ as the C-dimensional row vector containing the multiple labels of the photo from which graphlet $\mathcal{G}_i$ is extracted; and $\vec{\mathbf{N}} = [N^1, \cdots, N^C]^T$ where N^c is the number of photos for category c, then l_s and l_d are defined as:

$$l_s(i,j) = \frac{[\mathbf{b}_i \cap \mathbf{b}_j]\vec{\mathbf{N}}}{\sum_c N^c} \cdot d_{GW}(\mathcal{G}_i, \mathcal{G}_j), \tag{4}$$

$$l_d(i,j) = \frac{[\mathbf{b}_i \oplus \mathbf{b}_j]\vec{\mathbf{N}}}{\sum_c N^c} \cdot d_{GW}(\mathcal{G}_i, \mathcal{G}_j), \tag{5}$$

where $d_{GW}(\mathcal{G}_i, \mathcal{G}_j) = ||\mathbf{M}_i^o - \mathbf{M}_j^o||_2$; $\mathbf{M}_i^o$ and $\mathbf{M}_j^o$ are the orthonormal basis to the matrices of graphlets $\mathcal{G}_i$ and $\mathcal{G}_j$, respectively.

Inspired by many manifold algorithms [**?**], we assume a linear approximation of the graphlet embedding process. Therefore, we suppose $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$ and thus (3) can be reorganized into:

$$\begin{aligned} &\arg\min_{\mathbf{U}} \text{tr}(\mathbf{U}^T\mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{U}) \\ &= \arg\min_{\mathbf{U}} \text{tr}(\mathbf{U}^T\mathbf{L}\mathbf{U}) \;\; s.t. \;\; \mathbf{U}^T\mathbf{U} = \mathbf{I}_d, \end{aligned} \tag{6}$$

where $\mathbf{X}$ is obtained by row-wise stacking matrix $[\mathbf{M}_r^c, \mathbf{M}_r^t, \mathbf{M}_s]$ into a vector, $\mathbf{U}$ is the linear projection matrix, and $\mathbf{L} = \mathbf{X}\mathbf{R}\mathbf{X}^T$. The above objective function (6) is a basic optimization problem that can be solved using the Lagrangian multiplier. The optimal solution is the d eigenvectors associated with the d smallest eigenvalues of matrix $\mathbf{L}$.

4. SPARSITY-CONSTRAINED SALIENT GRAPHLETS DISCOVERY

In a human vision system, usually only the distinctive sensory information is selected for further processing. From this perspective, only a few visually/semantically salient graphlets within a photo are usually perceived by humans. These salient graphlets are significantly different from those non-salient ones, either in their appearances or in their semantics. Inspired by this, a sparsity-constrained ranking scheme is developed to discover salient graphlets, by exploring color, texture, and semantic channels collaboratively. More

[1]The atomic regions are superpixels segmented using SLIC [**?**]. The codes are publicly available. Experiments show SLIC is efficient and the generated superpixels are neatly adherent to object boundaries.

[2]With the advances of supervised image retrieval, nowadays image labels are cheaply available. They can be efficiently and accurately acquired by many existing models, *e.g.*, SPM [**?**] and its variants.

specifically, the ranking algorithm can be formulated as follows:

Formulation: Let $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$ be the three feature matrices in color, texture, and semantic channels respectively, where the columns in different matrices with the same index correspond to the same graphlet. The size of each $\mathbf{X}_i$ is $d_i \times N$, where d_i is the feature dimension and N is the number of graphlets. Then, the task is to find a weighting function to each graphlet $S(\mathcal{G}_i) \in [0, 1]$ by integrating the three feature matrices $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$.

Based on the theory of visual perception [**?**], there are usually strong correlation among the non-salient regions in a photo. That is to say, the non-salient graphlets can be self-represented. This analysis suggests that feature matrix $\mathbf{X}$ ($\mathbf{X}$ can be any one of matrices $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$) can be decomposed into a salient part and a non-salient part, *i.e.*,

$$\mathbf{X} = \mathbf{X}\mathbf{Z}_0 + \mathbf{E}_0, \tag{7}$$

where $\mathbf{X}\mathbf{Z}_0$ denotes the non-salient part that can be reconstructed by itself, $\mathbf{Z}_0$ denotes the reconstruction coefficients, and $\mathbf{E}_0$ denotes the remaining part corresponding to the salient targets.

Without a constraint, there are an infinite number of possible decompositions with respect to $\mathbf{Z}_0$ and $\mathbf{E}_0$. Toward a unique solution that indicates those salient graphlets, we need some criteria for characterizing matrices $\mathbf{Z}_0$ and $\mathbf{E}_0$. Aiming at this, two observations are made. On one hand, motivated by many approaches in computer vision, we assume that only a small fraction of graphlets are salient, *i.e.*, matrix $\mathbf{E}_0$ is sparse. The connection between sparsity and saliency is also consistent with the fact that only a small subset of sensory information is selected for further processing in a human vision system. On the other hand, the strong correlation among the background graphlets suggests that matrix $\mathbf{Z}_0$ is with low rankness. Based on the above analysis, we can infer the salient graphlets by adding a sparsity and low-rankness constraint to (7), thereby the graphlet saliency detection can be formulated as a low-rank representation (LRR) [**?**] problem:

$$\min_{\mathbf{Z}_0, \mathbf{E}_0} ||\mathbf{Z}_0||_* + \lambda||\mathbf{E}_0||_{2,1}, \;\; s.t. \;\; \mathbf{X} = \mathbf{X}\mathbf{Z}_0 + \mathbf{E}_0, \tag{8}$$

where $||\cdot||_*$ denotes the matrix nuclear norm that is a convex relaxation of the rank function, parameter $\lambda > 0$ is used to balance the effects of the two parts, and $||\cdot||_{2,1}$ is the $l_{2,1}$ norm defined as the sum of the $l_{2,1}$ norms of the columns of a matrix:

$$||\mathbf{E}_0||_{2,1} = \sum_i \sqrt{\sum\nolimits_j (\mathbf{E}_0(j,i))^2}. \tag{9}$$

It is noticeable that the minimization of the $l_{2,1}$ norm encourages the columns of $\mathbf{E}_0$ to be zero, and hence it fits well with our saliency detection problem. For a column corresponding to the i-th graphlet $\mathcal{G}_i$, a larger magnitude implies that the corresponding graphlet is more salient in drawing the attention of the human eye.

Let $\mathbf{E}_0^*$ be the optimal solution (with respect to $\mathbf{E}_0$) to problem (7). To obtain the saliency value of graphlet $\mathcal{G}_i$, we quantify the response of the sparse matrix as follows:

$$S(\mathcal{G}_i) = ||\mathbf{E}_0^*(:,i)||_2 = \sqrt{\sum\nolimits_i (\mathbf{E}_0^*(j,i))^2}. \tag{10}$$

where $||\mathbf{E}_0^*(:,i)||_2$ denotes the l_2 norm of the i-th column of $\mathbf{E}_0^*(:,i)$. A larger score of $S(\mathcal{G}_i)$ means that graphlet $\mathcal{G}_i$ has a higher probability to be salient.

The objective function (7) calculates the saliency of a graphlet based on one type of visual feature, which is suboptimal since multiple visual features determine graphlet saliency collaboratively. To combine together visual features in color, texture, and semantic channels, we generalize the objective function (7) into a multimodal version:

$$\min_{\substack{\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3 \\ \mathbf{E}_1,\mathbf{E}_2,\mathbf{E}_3}} \sum_{i=1}^{3} ||\mathbf{Z}_i||_* + \lambda||\mathbf{E}||_{2,1}, \;\; s.t. \;\; \mathbf{X}_i = \mathbf{X}_i\mathbf{Z}_i + \mathbf{E}_i, \tag{11}$$

where $\mathbf{E} = [\mathbf{E}_1; \mathbf{E}_2; \mathbf{E}_3]$ is formed by vertically concatenating $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ together along a column. The integration of multiple features is seamlessly performed by minimizing the $l_{2,1}$ norm of $\mathbf{E}$. That is, we enforce the columns of $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ to have jointly consistent magnitude values.

Let $\{\mathbf{E}_1^*, \mathbf{E}_2^*, \mathbf{E}_3^*\}$ be the optimal solution to (11), to obtain a saliency score for graphlet $\mathcal{G}_i$, we quantify the response of the sparse matrices as follows:

$$S(\mathcal{G}_i) = \sum\nolimits_{j=1}^{3} ||\mathbf{E}_j^*(:,i)||_2, \tag{12}$$

where $||\mathbf{E}_j^*(:,i)||_2$ denotes the l_2 norm of the i-th column of $\mathbf{E}_j^*$. A larger score of $S(\mathcal{G}_i)$ means that graphlet $\mathcal{G}_i$ has a higher probability to be salient. Algorithm 1 summarizes the procedure of our proposed multimodal graphlet saliency detection. The details of solving (11) are illustrated in the Appendix.

Algorithm 1 Multimodal Salient Graphlets Discovery

input: Graphlets from a labeled photo, the projection matrix $\mathbf{U}$;
output: A number of graphlets ranked by their saliency values;

1) Compute the feature matrices $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ in color, texture, and semantic channels to describe each graphlet;
2) Obtain the sparsity matrices $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ in color, texture, and semantic channels respectively, by solving the objective function (11);
3) Compute the graphlet saliency based on (12), and a few top-ranked graphlets are deemed as the salient ones.

5. PERCEPTION-GUIDED PROBABILISTIC AESTHETIC MEASURE

Based on the above discussion, the top-ranked graphlets are the salient ones that can draw the attention of the human eye. That is, humans first fixate on the most salient graphlet in a photo, and then shift their gazes to the second salient one, and so on. Inspired by the scan path used in human eye-tracking experiments, we propose an actively viewing path (AVP) to mimic the sequential manner biological species perceive a visual scene. The procedure of generating an AVP from a photo is described in Fig. 4. It is noticeable that all the AVPs from a data set are with the same number of graphlets K. Typically, we set K to 4 and its influence on photo aesthetics prediction is evaluated in our experiments.

Given a set of aesthetically pleasing training photos $\{I^1, \cdots, I^H\}$ and a test image I^*, they are highly correlated through their respective AVPs $\mathcal{P}$ and $\mathcal{P}^*$. Thus, a probabilistic graphical model is utilized to describe this correlation. As shown in Fig. 5, the graphical model contains two types of nodes: observable nodes (blue color) and hidden nodes (gray color). More specifically, it can be divided into four layers. The first layer represents all the training photos, the second layer denotes the AVPs extracted from the training photos, the third layer represents the AVP of the test photo, and the last layer denotes the test photo. The correlation between the first and the second layers is $p(\mathcal{P}|I^1, \cdots, I^H)$, the correlation between the second and the third layers is $p(\mathcal{P}^*|\mathcal{P})$, and the correlation between the third and the fourth layers is $p(I^*|\mathcal{P}^*)$.

According to the formulation above, photo aesthetics can be

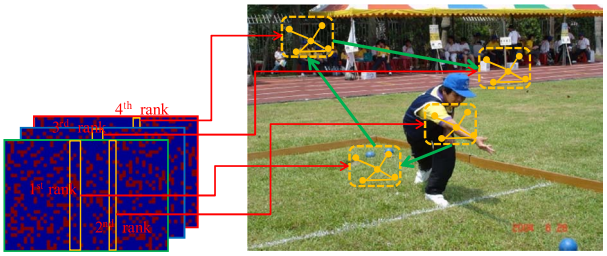

Figure 4: An illustration of AVP generation based on the top-ranked graphlets.

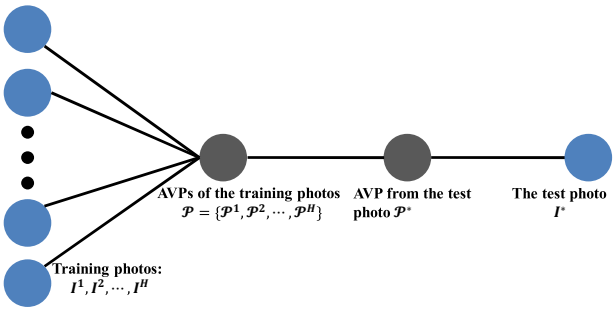

Figure 5: An illustration of the probabilistic graphical model.

quantified as the similarity between the AVPs from the test photo and those from the training aesthetically pleasing photos. The similarity is interpreted as the amount of AVPs that can be transferred from the training photos into the test image. That is, the aesthetic quality of the test photo I^* can be formulated as:

$$\begin{aligned}\gamma(I^*) &= p(I^*|I^1, \cdots, I^H) \\ &= p(I^*|\mathcal{P}^*) \cdot p(\mathcal{P}^*|\mathcal{P}) \cdot p(\mathcal{P}|I^1, \cdots, I^H). \end{aligned} \quad (13)$$

The probabilities $p(I^*|\mathcal{P}^*)$, $p(\mathcal{P}^*|\mathcal{P})$, and $p(\mathcal{P}|I^1, I^2, \cdots, I^H)$ in (13) are computed respectively as:

$$\begin{aligned} p(I^*|\mathcal{P}^*) &= \prod\nolimits_{\mathcal{G}^* \in \mathcal{P}^*} p(I^*|\mathcal{G}^*) \\ &= \prod\nolimits_{\mathcal{G}^* \in \mathcal{P}^*} \frac{p(\mathcal{G}_1^*, \cdots, \mathcal{G}_T^*|I^*)p(I^*)}{p(\mathcal{G}_1^*, \cdots, \mathcal{G}_T^*)} \\ &\propto \prod\nolimits_{\mathcal{G}^* \in \mathcal{P}^*} p(\mathcal{G}_1^*, \cdots, \mathcal{G}_T^*|I^*)p(I^*) \\ &= \prod\nolimits_{\mathcal{G}^* \in \mathcal{P}^*} \prod\nolimits_{i=1}^{T} \prod\nolimits_{j=1}^{Y_t} p(\mathcal{G}_t^*(j)|I^*), \end{aligned} \quad (14)$$

where T is the maximum graphlet size, Y_i is the number of i-sized graphlets in the test photo I^*, and $\mathcal{G}_t^*(j)$ is the j-th t-sized graphlet of AVP from the test photo. $p(\mathcal{G}_t^*(j)|I^*)$ denotes the probability of extracting graphlets $\mathcal{G}_t^*(j)$ from the test photo I^*, which is calculated as described next.

As shown in Fig. 6, the graphlet extraction is based on random walk. We first index all atomic regions in a photo and choose a starting one with probability $\frac{p(Y)}{Y}$, where Y means there are Y atomic regions in photo I and $p(Y)$ is the corresponding probability. We then visit a spatially adjacent larger-indexed vertex[3] with probability $\frac{1}{2\sum_d p_d(R_l)d(R_l)}$, where $d(R_l)$ is the degree of the current atomic region R_l and $p_d(R_l)$ denotes the probability of atomic region R_l with degree $d(R_l)$. In our implementation, $p_d(R_l) \propto \exp\left(-\frac{1}{||d(R_l) - \bar{d}p(R_l)d(R_l)||}\right)$ and $p(Y) \propto \exp\left(-\frac{||Y-\bar{Y}||^2}{\sigma_Y^2}\right)$. The random walk process stops when the maximum graphlet size is reached. Therefore, we obtain

[3]It is with the same probability of visiting a larger-indexed vertex or visiting a smaller-indexed one.

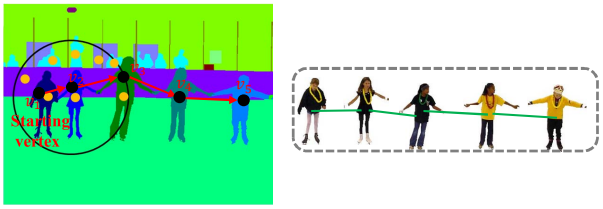

Figure 6: Graphlet extraction based on a random walk model.

$$p(\mathcal{G}|I) \propto \frac{p(Y)}{Y} \prod\nolimits_{l=1}^{T-1} \frac{1}{2\sum_d p_d(R_l)d(R_l)}, \quad (15)$$

The probability $p(\mathcal{P}^*|\mathcal{P})$ measures the similarity between AVPs from the training photos and that from the test photo:

$$p(\mathcal{P}^*|\mathcal{P}) = \prod\nolimits_{i=1}^{H} \prod\nolimits_{\substack{\mathcal{G}^* \in \mathcal{P}^* \\ \mathcal{G} \in \mathcal{P}^i}} p(\mathcal{G}^*|\mathcal{G}), \quad (16)$$

where $p(\mathcal{G}^*|\mathcal{G})$ measures the similarity between the same-indexed graphlets from the test and the training photos, respectively. In our implementation, $p(\mathcal{G}^*|\mathcal{G})$ is implemented as a Gaussian kernel.

The probability $p(\mathcal{P}|I^1, \cdots, I^H)$ measures the probabilities of AVPs generated from the training photos, which is calculated as:

$$\begin{aligned} p(\mathcal{P}|I^1, \cdots, I^H) &= \prod\nolimits_{i=1}^{H} p(\mathcal{P}^i|I^i) \\ &= \prod\nolimits_{i=1}^{H} \prod\nolimits_{\mathcal{G} \in \mathcal{P}^i} p(\mathcal{G}|I^i), \end{aligned} \quad (17)$$

In summary, the procedure of our proposed photo aesthetics evaluation algorithm is presented in Algorithm 2.

Algorithm 2 Perception-Guided Photo Aesthetic Model

input: A set of training aesthetically pleasing $\{I^1, \cdots, I^H\}$, the number of graphlets in an AVP: K, and the test photo I^*;
output: The aesthetic score of the test photo I^*;

1) Extract graphlets from each training photo; represent them in both color and texture channels; learn the embedding model and represent each graphlet in semantic channel;
2) Construct the AVPs of each photo by selecting K salient graphlets, based on the sparsity-constrained ranking algorithm described in Sec. 4;
3) Compute the aesthetic score of the test photo I^* using (13).

6. EXPERIMENTS 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 aesthetic models. The second part step-by-step evaluates each component of the proposed approach. Third, based on an eye-tracking experiment, we make a quantitative comparison between the proposed AVP and a real human gaze shifting path. Lastly, we analyze the influence of different parameter settings on our aesthetic model.

6.1 Data Sets and Implementation Details

To the best of our knowledge, there exist three off-the-shelf data sets for evaluating photo aesthetics: the CUHK [**?**], the Photo.net,

Table 1: 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	**0.9059**	**0.8552**	**0.8413**

and the AVA [**?**]. A rough description of the three data sets is as follows: 1) 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. We use a standard split of training/test sets on this data set. 2) The Photo.net [**?**] consists of 3581 images. Only URLs of the original photos are provided. Approximately half of the images have since been removed from the websites, leaving only nearly 1,700 images available. Thus, we extend this data set by online crawling 4,000 photos and name the extended Photo.net data set PNE. The aesthetics of these additionally crawled photos are manually labeled. They are randomly split into equal partitions, one for training and the rest for testing. 3) The AVA [**?**] data set contains 25,000 highly- and low-aesthetic photos, 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.

For the classifier-based photo aesthetic models, such as those proposed by Marchesotti *et al.* [**?**] and Nishiyamma *et al.* [**?**], both the highly- and low-aesthetic photos are adopted to learn the model. More specifically, the highly aesthetic photos are used as positive samples and the low aesthetic ones are used negative samples. For those models that are based on transferring aesthetic features (*e.g.*, Cheng *et al.* [**?**]'s model), they employ only the "good" aesthetic features to score a test photo. Thus, it is necessary to assign a weight for each training AVP indicating its "goodness", *i.e.*, a larger weight reflects a higher aesthetic level. The weight is determined by the aesthetics of the photo from which the graphlet is extracted. For the three data sets, different settings are used to assign the weight of each photo. For the CUHK, we use the probabilistic output described in Yan *et al.*'s work [**?**] to rank the aesthetics of each photo. For the PNE, we manually select 674 highly aesthetic photos and leave the rest as the low-aesthetic ones. Then, we extract the aesthetic features based on [**?**], and further use a probabilistic SVM output to score the aesthetics of each photo. For the AVA, each training photo is rated by multiple users. We average the rating scores of a photo as its overall aesthetic score.

All the experiments were carried out on a personal computer with an Intel X5482 processor and 8GB RAM. The algorithm is implemented on the Matlab 2011 platform.

6.2 A Comparative Study

The first experiment compares our approach with five photo aesthetics evaluation methods. The compared algorithms include three global feature-based approaches proposed by Dhar *et al.* [**?**], Luo *et al.* [**?**], and Marchesotti *et al.* [**?**], respectively; and two local patch integration-based methods proposed by Cheng *et al.* [**?**] and Nishiyama *et al.* [**?**], respectively.

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

	DAPC	PPR	SBPC	SCSM	OCBC	GTPC	Our	Score
DAPC	--	5	8	7	8	6	2	36
PPR	25	--	8	12	11	14	9	79
SBPC	22	22	--	12	15	14	10	95
SCSM	23	18	18	--	12	9	7	87
OCBC	22	19	15	18	--	12	9	95
GTPC	24	16	16	21	18	--	8	103
Our	28	21	20	23	21	22	--	135

	DAPC	PPR	SBPC	SCSM	OCBC	GTPC	Our	Score
DAPC	--	16	15	18	19	11	6	85
PPR	14	--	13	12	13	9	7	68
SBPC	15	17	--	17	18	12	9	88
SCSM	12	18	13	--	16	8	4	71
OCBC	11	17	12	14	--	6	3	63
GTPC	19	21	18	22	24	--	9	113
Our	24	23	21	26	27	21	--	142

	DAPC	PPR	SBPC	SCSM	OCBC	GTPC	Our	Score
DAPC	--	21	19	17	16	12	11	96
PPR	9	--	15	17	18	21	10	90
SBPC	11	15	--	14	12	19	12	83
SCSM	13	13	16	--	13	14	8	77
OCBC	14	12	18	17	--	15	9	85
GTPC	18	9	11	16	15	--	6	75
Our	19	20	18	22	21	24	--	124

	DAPC	PPR	SBPC	SCSM	OCBC	GTPC	Our	Score
DAPC	--	3	6	5	7	5	8	34
PPR	27	--	14	17	18	15	11	102
SBPC	24	16	--	21	22	19	8	110
SCSM	25	13	9	--	14	13	8	82
OCBC	23	12	8	16	--	16	7	82
GTPC	25	15	11	17	14	--	6	88
Our	22	19	22	22	23	24	--	132

Figure 7: Cropped photos produced by the compared methods and the preference matrices (OP: the original photo; the red numbers denote the scores of the algorithms).

implement them. Toward a convincing comparative study, we try 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 codes from Li *et al.* [**?**] to extract the attributes from each photo. These attributes are combined with the low-level features proposed by Yeh *et al.* [**?**] to train the aesthetic classifier. For Luo *et al.*'s approach, not only are the low-level and high-level features in their publication implemented, but also the six global features from Getlter *et al.* [**?**]'s work 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 implement it as a simplified version of our approach, *i.e.*, only 2-sized graphlets are employed for aesthetics measure. Notably, for the three probabilistic model-based aesthetics evaluation methods (*i.e.*, Cheng *et al.*'s method, Nishiyama *et al.*'s method, and our model), if the aesthetic score 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, about half of the photos are highly aesthetic.

We present the aesthetics prediction accuracy on the CUHK, the PNE, and the AVA in Table 1. On the three data sets, our approach outperforms Marchesotti *et al.*'s approach by nearly 3%, and exceeds the rest of the compared methods by more than 6%, which demonstrates the effectiveness our approach.

Last but not least, we evaluate the proposed aesthetic model in comparison with several well-known cropping methods. For our approach, the sub-region with the highest aesthetic score in each photo is cropped. The compared cropping methods are sparse coding of saliency maps (SCSM [**?**]), sensation based photo cropping (SBPC [**?**]), omni-range context-based cropping (OCBC [**?**]), personalized photo ranking (PPR [**?**]), describable attribute for photo cropping (DAPC [**?**]), and the graphlet transferring-based photo cropping(GTPC) [**?**], respectively. We conduct a paired comparison-based user study [**?**] and present the preference matrix obtained from the above methods. Each preference matrix is filled by 25∼40

Table 2: Aesthetics prediction accuracy decrement.

	CUHK	PNE	AVA
Graphlet→single atomic region	-4.31%	-3.55%	-4.76%
Remove adj. mat from graphlet	-3.38%	-3.65%	-2.77%
Mani. Grap. emb.→Single-ch. emb.	-3.16%	-2.81%	-2.89%
Mani. Grap. emb.→kernel PCA	-6.21%	-5.12%	-4.33%
Prob. mea. → clasf. mea.	-3.31%	-2.81%	-2.91%

volunteers. As shown in Fig. 7, the result again demonstrates the advantage of the proposed method.

6.3 Step-by-Step Model Justification

The proposed photo aesthetic model includes three main components: the multimodal graphlet representation, the sparsity-constrained graphlet ranking, and the probabilistic aesthetics measure, which are theoretically indispensable and inseparable. To demonstrate the effectiveness of each step, we replace each component by a functionally reduced counterpart and report the corresponding aesthetics prediction accuracy.

- To illustrate the effectiveness of the first component, two experimental settings are applied: 1) reducing the multimodal graphlet representation to a single channel one (Mani.Grap→ Sing-ch), where only the color channel is used. The color channel is preserved here because as shown by several photo aesthetics methods [**?**], 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 calculated as:

$$k(\mathcal{G}, \mathcal{G}') \propto \exp(-d_{GW}(\mathbf{M}, \mathbf{M}')), \tag{18}$$

  where $\mathbf{M} = [\mathbf{M}_r^c, \mathbf{M}_r^t, \mathbf{M}_r^s, \mathbf{M}_s]$ and $d_{GW}(\cdot,\cdot)$ is the Golub-Werman distance between identically sized matrices.

- To justify the usefulness of the sparsity-constrained graphlet ranking, we replace this component by the locality preserving active learning paradigm proposed by Zhang *et al.* [**?**]. In Zhang *et al.*'s model, the weighting matrix (W in Eqn. (12) of Zhang *et al.* [**?**]) is calculated by considering a graphlet and its spatially neighboring graphlets in a photo. As the comparative aesthetic features shown in Fig. 8, our approach prefers to select small graphlets due to the sparsity constraint in our model. More importantly, our approach can explicitly model the gaze shifting sequence of each photo, while the other methods fail.

- To demonstrate the effectiveness of the third component, we replace the probabilistic aesthetics measure by a kernel SVM-based one (Prob.mea.→Clas.mea.), wherein the kernel is computed the same as that in the above kernel PCA.

As shown in Table 2, when replacing one component of the proposed approach by an existing one, the aesthetics prediction accuracy reduces dramatically. This demonstrates that each component of the proposed approach is indispensable and inseparable.

6.4 Visualization of the Active Viewing Paths and the Photos Ranked by Aesthetics

This subsection quantitatively and qualitatively compares the proposed actively viewing path (AVP) with real human gaze shifting path. In particular, we record the eye fixations of five viewers by making use of the eye-tracker EyeLink II, and then link the fixations into a path in a sequential manner. As can be seen from Fig. 9, in most photos, the proposed AVPs are consistent with human gaze shifting paths. In addition, we calculate the proportion of the human gaze shifting path that overlaps with an AVP. The overlapping between our proposed AVP and a real human gaze shifting path is computed as shown in Fig. 10. Given the five real gaze shifting paths, we connect all the segmented regions on the gaze shifting path and then obtain the human gaze shifting path with segmented regions. Based on this, the similarity between an AVP and a gaze shifting path with segmented regions is calculated by:

$$s(P_1, P_2) = \frac{nPixel(P_1 \cap P_2)}{nPixel(P_1) + nPixel(P_2)}, \tag{19}$$

where P_1 and P_2 denote an AVP and the gaze shifting path with segmented regions, respectively, $nPixel$ counts the pixels within an image region, and $P_1 \cap P_2$ denotes the shared region between P_1 and P_2. Based on (19), we observe that the overlapping percentage between an AVP and a real human gaze shifting path is 87.65% on average. This observation shows that the AVP can accurately predict the human gaze shifting process.

Next, we visualize the top-ranked graphlets from images in

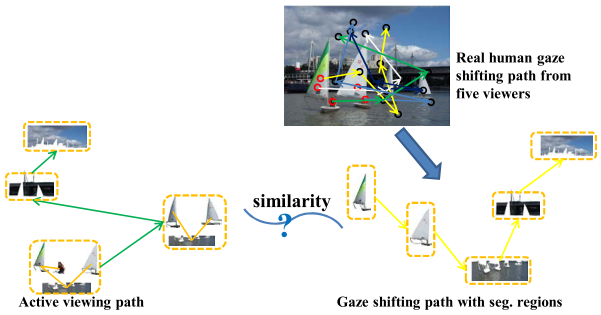

Figure 9: An illustration of the overlapping between an AVP and a human gaze shifting path.

the LHI [**?**] data set. As shown in Fig. 12, for each image, the first four most visually/semantically salient graphlets are presented. The results again demonstrate the importance of mining the spatial interaction of image regions for describing photo aesthetics.

Further, we visualize photos of the AVA data set that are ranked by our probabilistic photo aesthetics measure. As can be seen from Fig. 11, we made the following three observations.

- As shown in the photos whose aesthetics are ranked between 0.8 and 1, highly aesthetic photos with multiple interacting objects are assigned with very high scores, which shows that our proposed AVPs well predict how humans perceive local/global aesthetics in these beautiful photos.

- As seen from the photos whose aesthetics are ranked between 0.5 and 0.8, highly aesthetic photos with a single object are also appreciated by the proposed aesthetics model. This is because graphlets are naturally local composition descriptors, and they influence photo aesthetics by making use of the proposed probabilistic model.

- Objects from the photos whose aesthetics are ranked between 0 and 0.5 are either spatially disharmoniously distributed or blurred. Thus, these photos are considered as aesthetically low by our model.

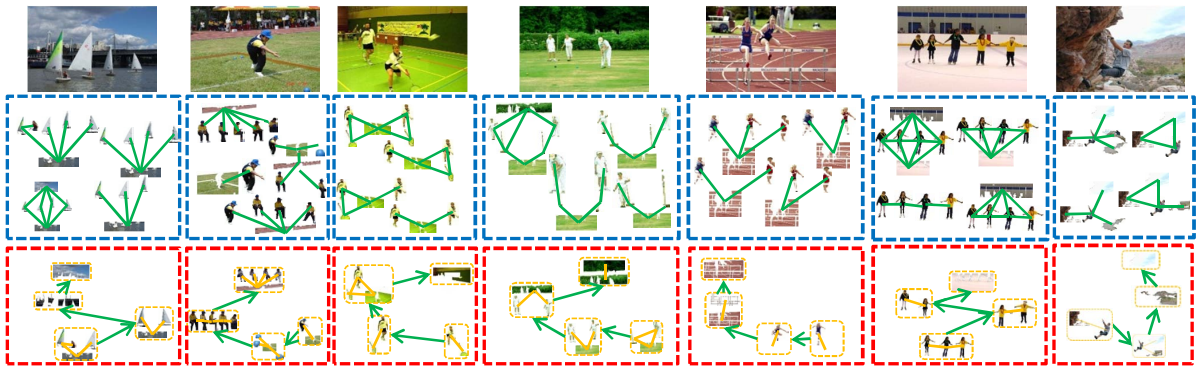

Figure 8: A comparison of representative graphlets (blue rectangles) and the AVPs (red rectangles).

Figure 10: Comparison of gaze shifting paths from five observers (differently colored) and the proposed AVPs.

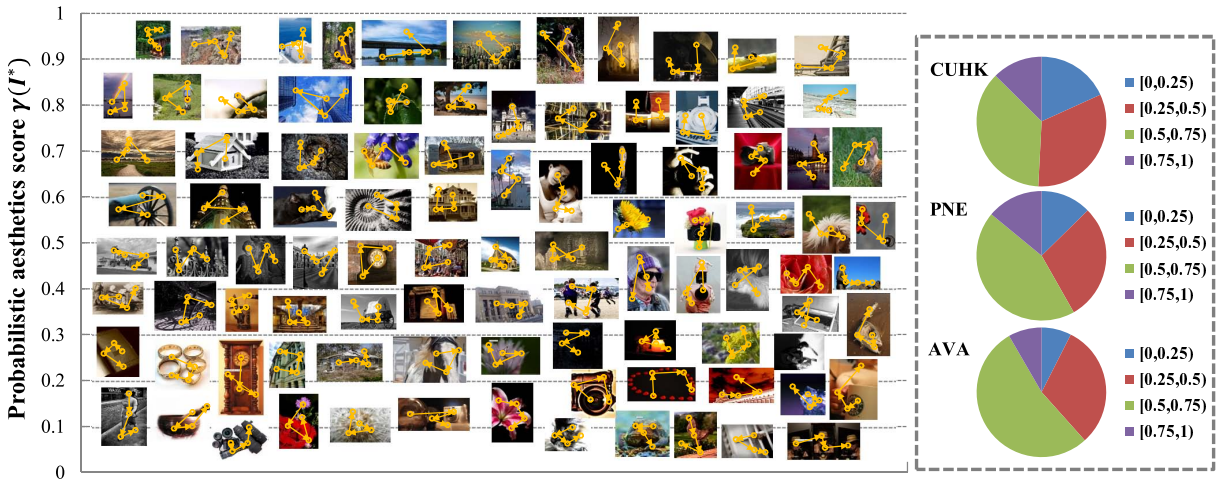

Figure 11: Ranking results on the AVA data set. The yellow paths denote the AVPs predicted by our model, where each circle indicates the location of a graphlet. The three pie charts show statistics of photos from the three data sets based on our model.

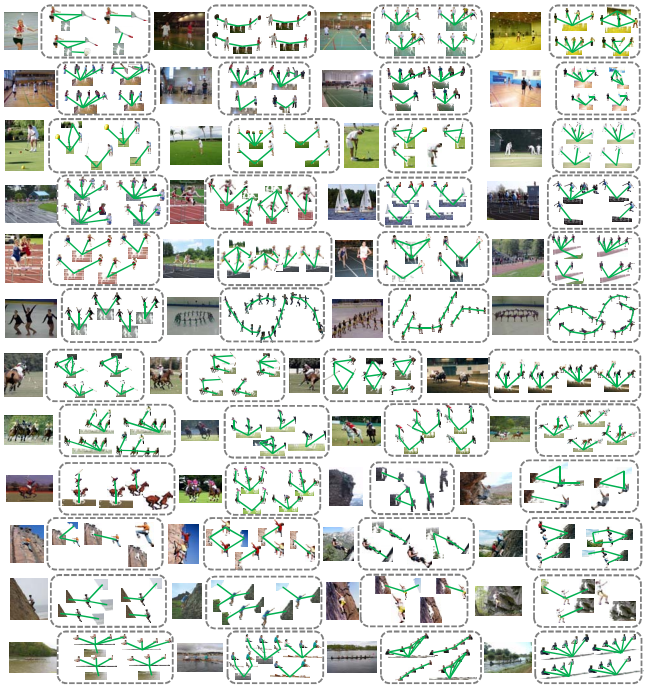

Figure 12: Visualized highly ranked graphlets of the LHI data set.

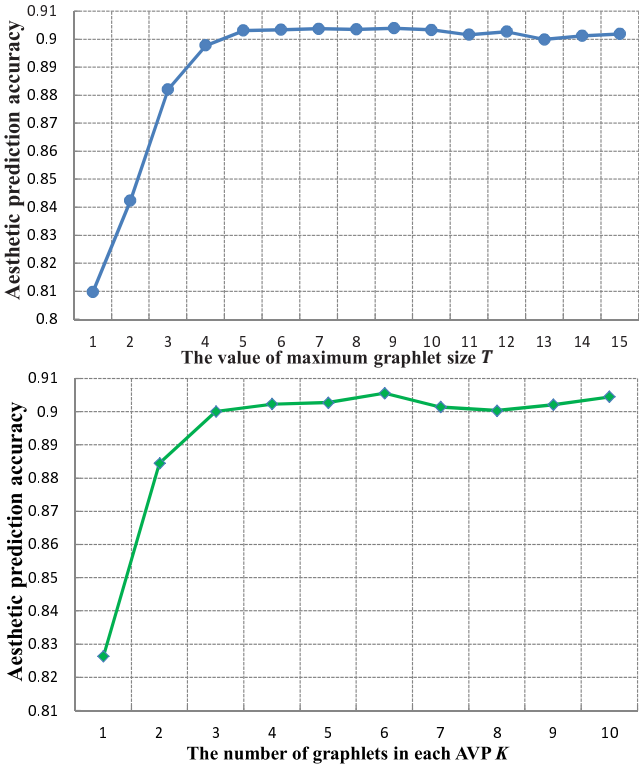

Figure 13: Photo aesthetics evaluation performance under different parameters.

We compare the AVPs predicted by our approach, under both low aesthetic and high aesthetic photos. As shown in Fig. 11, neither low nor high aesthetic photo have particular path geometry, such as the angle between pairwise shifting vectors (yellow arrows). However, for high aesthetic photos, the fixation points (yellow circles) are aesthetically pleasing and the objects along the path are harmoniously distributed.

6.5 Parameter Analysis

The experiment evaluates the influence of the graphlet size T and the number of graphlets of each AVP K on the performance of the proposed photo aesthetics model.

To analyze the effects of the maximum graphlet size T on predicting photo aesthetics, we set up an experiment by varying T continuously. In the top graph of Fig. 13, we present the aesthetics prediction accuracy (on the CUHK data set) when the maximum size of the graphlet is tuned from 1 to 10. As can be seen, prediction accuracy increases moderately when $T \in [1, 5]$ but remains almost unchanged when $T \in [6, 10]$. This observation implies that 5-sized graphlets are sufficient for capturing the local composition of photos in the CUHK data set. In the bottom graph of Fig. 13, we present the performance when the number of graphlet in an AVP (K) is tuned from 1 to 10. As can be seen, the prediction accuracy increases quickly when $K \in [1, 4]$ but remains stable when $K \in [5, 10]$.

7. CONCLUSIONS

Image/video aesthetics quality assessment is a useful technique in multimedia field [**?**, **?**, **?**]. In this paper, a new model is proposed to evaluate the aesthetics of a photo by simulating the process of humans sequentially perceiving semantics of a photo. By discovering visually/semantically salient graphlets using a sparsity-constrained ranking paradigm, an active viewing path is constructed to mimic the process where humans actively look at important components in a photo. Thereafter, we develop a probabilistic model that quantifies photo aesthetics as the amount of AVPs that can be transferred from a set of aesthetically pleasing training photos into the test image. Extensive experiments demonstrate the effectiveness of our model.

In the future, we will apply our model to video aesthetics evaluation. The aesthetics of a video clip is determined by accumulating the aesthetic scores of all its constituent frames, using a probabilistic model.

8. APPENDIX

Problem (11) is convex and can be optimized efficiently. We first convert it into the following equivalent problem:

$$\begin{aligned} &\min_{\mathbf{J}_i, \mathbf{Z}_i, \mathbf{E}_i} \sum_{i=1}^{3} ||\mathbf{J}_i||_* + \lambda ||\mathbf{E}||_{2,1}, \\ &s.t.\ \ \mathbf{X}_i = \mathbf{X}_i \mathbf{Z}_i + \mathbf{E}_i, \mathbf{Z}_i = \mathbf{J}_i, \end{aligned} \tag{20}$$

This problem can be solved with the ALM method which minimizes the following augmented Lagrange function:

$$\begin{aligned} \mathcal{L} = \lambda ||\mathbf{E}||_{2,1} + \sum_{i=1}^{3} (||\mathbf{J}_i||_* + \langle \mathbf{Y}_i, \mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i - \mathbf{E}_i \rangle \\ + \langle \mathbf{W}_i, \mathbf{Z}_i - \mathbf{J}_i \rangle + \frac{\mu}{2} ||\mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i - \mathbf{E}_i||_F^2 + \frac{\mu}{2} ||\mathbf{Z}_i - \mathbf{J}_i||_F^2) \end{aligned} \tag{21}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ are Lagrange multipliers and $\mu > 0$ is a penalty parameter. The inexact ALM method, which is also called the alternating direction method, is illustrated in Algorithm 3. Note that the subproblems of the algorithm are convex

Algorithm 3 Inexact ALM-based solution of (11)

input: Data matrices $\{\mathbf{X}_i\}$, parameter λ;
output: The optimal solution $\mathbf{E}^*$;

while not converged **do**
1) Fix the others and update $\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3$ by:
$\mathbf{J}_i = \arg\min_{\mathbf{J}} \frac{1}{\mu}||\mathbf{J}||_* + \frac{1}{2}||\mathbf{J}_i - (\mathbf{Z}_i + \frac{\mathbf{W}_i}{\mu})||_F^2$.
2) Fix the others and update $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ by:
$\mathbf{Z}_i = \mathbf{M}(\mathbf{X}_i^T(\mathbf{X}_i - \mathbf{E}_i) + \mathbf{J}_i + \frac{\mathbf{X}_i^T\mathbf{Y}_i - \mathbf{W}_i}{\mu})$
where $\mathbf{M} = (\mathbf{I} + \sum_{i=1}^3 \mathbf{X}_i^T\mathbf{X}_i)^{-1}$.
3) Fix the others and update $\mathbf{E} = [\mathbf{E}_1; \mathbf{E}_2; \mathbf{E}_3]$ by
$\mathbf{E} = \arg\min_{\mathbf{E}} \frac{\lambda}{\mu}||\mathbf{E}||_{2,1} + \frac{1}{2}||\mathbf{E} - \mathbf{G}||_F^2$,
where $\mathbf{G}$ is formed by vertically concatenating the matrices
$\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i + (\mathbf{Y}_i/\mu)$, $i = 1, 2, 3$ together along column.
4) Update the multipliers
$\mathbf{Y}_i = \mathbf{Y}_i + \mu(\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i)$; $\mathbf{W}_i = \mathbf{W}_i + \mu(\mathbf{Z}_i - \mathbf{J}_i)$;
5) Update the parameter μ by
$\mu = \min(\rho\mu, 10^{10})$
where the parameter ρ controls the convergence speed. It is set as $\rho = 1.1$ in all experiments.
6) Check the convergence condition: $\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i \to 0$ and $\mathbf{Z} - \mathbf{J}_i \to 0$, $i = 1, 2, 3$;
end while

and they have closed-form solution. Step 1 is solved via the singular value thresholding operator [**?**], whereas Step 3 is solved via [**?**].

Acknowledgements

This research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office through the Centre of Social Media Innovations for Communities (COSMIC). Dr. Tian was also supported for this work in part by ARO grant W911NF-12-1-0057, Faculty Research Award by NEC Laboratories of America, and in part by the National Science Foundation of China (NSFC) under contract No. 61128007, respectively.