Weakly Supervised Photo Cropping

Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe

***Abstract*—Photo cropping is widely used in the printing industry, photography, and cinematography. Conventional photo cropping methods suffer from three drawbacks: 1) the semantics used to describe photo aesthetics are determined by the experience of model designers and specific data sets, 2) image global configurations, an essential cue to capture photos aesthetics, are not well preserved in the cropped photo, and 3) multi-channel visual features from an image region contribute differently to human aesthetics, but state-of-the-art photo cropping methods cannot automatically weight them. Owing to the recent progress in image retrieval community, image-level semantics, *i.e.*, photo labels obtained without much human supervision, can be efficiently and effectively acquired. Thus, we propose weakly supervised photo cropping, where a manifold embedding algorithm is developed to incorporate image-level semantics and image global configurations with graphlets, or, small-sized connected subgraph. After manifold embedding, a Bayesian Network (BN) is proposed. It incorporates the testing photo into the framework derived from the multi-channel post-embedding graphlets of the training data, the importance of which is determined automatically. Based on the BN, photo cropping can be casted as searching the candidate cropped photo that maximally preserves graphlets from the training photos, and the optimal cropping parameter is inferred by Gibbs sampling. Subjective evaluations demonstrate that: 1) our approach outperforms several representative photo cropping methods, including our previous cropping model that is guided by semantics-free graphlets, and 2) the visualized graphlets explicitly capture photo semantics and global spatial configurations.**

***Index Terms*—Photo cropping, Weakly supervised, Bayesian network, Image-level Semantics**

I. Introduction

Photo cropping refers to removing unwanted subjects or irrelevant details from a photo, changing its aspect ratio, or adjusting its overall composition. Conventional photo cropping methods have been applied in many fields. For example, in printing industry, the visual attractiveness of a photo can be increased by cropping it from a panoramic one; in telephoto photography, an image is cropped to magnify the primary subject; and in cinematography, film footage can be cropped to change its aspect ratio, without stretching the image or filling with the blank bars. However, photo cropping is still a challenging problem due to the following three reasons:

- Semantics is an important cue to describe photo aesthetics, but state-of-the-art photo cropping models cannot exhibit semantics effectively. Typically, a cropping system only employs a small number of manually defined semantics based on a specific data set. They are defined by determining whether photos in the data set are covered with sky, vegetation, water, etc. Additionally, the semantics is usually detected using an auxiliary object detector, *e.g.*, a human face detector. There is no guarantee that all the pre-specified semantic objects can be accurately discovered.
- Global spatial configurations, which reflects the spatial arrangements of all components in a photo, play an important role in photo aesthetics, but existing cropping models cannot well preserve them. As shown in Fig. 1, the relative displacement of water, sky, and sailboats determines the photo global layout, but it cannot be explicitly captured by the existing cropping models.
- Multi-channel visual features from an image region influence differently on human aesthetics. For example, the texture channel is perceptually less dominant for a textureless image region. Unfortunately, existing cropping methods cannot automatically adjust the importance of multi-channel visual features from an image region.

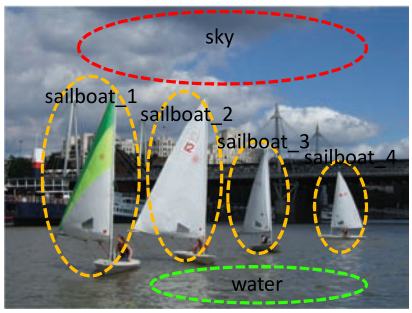


Fig. 1. Preserving the relative displacement of sky, water, and four sailboats implicitly maintains global spatial configuration.

To resolve the above mentioned problems, we propose a weakly supervised photo cropping method, which integrates the strategy of transferring from image-level semantics to region-level semantics, image global spatial configurations, and multi-channel visual features weighting scheme, into a graphlet-guided photo cropping framework. As shown in Fig. 2, to capture the local composition of each photo, we construct graphlets to model the spatial arrangements of local atomic regions. To incorporate semantics and global spatial

 This work was supported in part by National Natural Science Foundation of China(61170142), National Key Technology R&D Program (2011BAG05B04), International Science & Technology Cooperation Program of China (2013DFG12840), National High Technology Research and Development Program of China (2013AA040601),and the Fundamental Research Funds for the Central Universities.

L.Zhang and M. Song are with the College of Computer Science, Zhejiang University, China.

Yi Yang is with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia.

Q. Zhao is with the Department of Electrical and Computer Engineering,National University of Singapore, Singapore.

C. Zhao is with the School of Electrical Engineering and Computer Science, Peking University, Beijing, China.

N. Sebe is with the Department of Information Engineering and Computer Science, University of Trento, Italy.

configurations into graphlets, a manifold embedding algorithm is derived to maximally preserve the image-level semantics of each photo and the Golub-Werman distances between all pairwise graphlets in each photo. After sampling a number of candidate cropped photos, we obtain the multi-channel post-embedding graphlets from each candidate cropped photo. Thereafter, we form a BN to measure the quality of each candidate cropped photo, where the importance of each channel visual feature is adjusted automatically. Based on the new photo quality measure, we cast photo cropping as seeking the parameter of the candidate cropped photo with the maximum posterior probability, and Gibb sampling is applied for parameter inference.

The contributions of this paper are three-fold:

- Weakly supervised photo cropping, a new approach to improve photo cropping performance using image-level semantics;
- Manifold graphlet embedding, a new algorithm to encode image-level semantics and photo global spatial configurations into graphlets;
- A BN which automatically weights multi-channel visual cues in the post-embedding graphlets transferring process.

II. Related Work

A typical photo cropping algorithm contains three steps: sampling a number of candidate cropped photos and scoring the quality of each one based on some photo quality measure; the most qualified one will then be selected. In such an algorithm, candidate cropped photo evaluation is an essential and indispensable procedure in the cropping process. In recent years, several photo cropping and photo assessment approaches have been proposed. Among them, two research topics closely relate to the proposed method [1].

A. Global Features-Based Approaches

Global features-based approaches design different types of global low-level and high-level visual features to represent photo aesthetics. These global features are typically concatenated into a long vector and used to train a classifier or regression function for measuring photo quality. Luo *et al*. [1] proposed a number of high-level semantic features based on the division of the subjects and background. Ke *et al*. [2] designed a group of high-level image features, such as image simplicity based on spatial distribution of edges, to imitate people's perception of photo quality. Datta *et al*. [3] proposed 58 low-level visual features, such as shape convexity, to capture photo aesthetics. Wong *et al*. [4] proposed three types of global features, *i.e.*, low-level features such as exposure extracted from the overall image and the salient regions, as well as the difference between low-level features extracted from subject and background regions. Dhar *et al*. [9] proposed a set of high-level attribute-based predictors to evaluate photo aesthetics. Three types of attribute-based predictors are proposed, *i.e.*, compositional attributes, content attributes, and sky illumination attributes. In [10], Luo *et al*. proposed a Gaussian mixture model (GMM)-based hue distribution and a prominent line extraction-based texture distribution to represent the global composition of each photo. To describe photo's local composition, three regional features respectively describing human faces, region clarity, and complexity are developed. It is worth noting the limitations of the above global feature-based approaches: First, Luo *et al*. [1]'s approach relies heavily on a blur detection technique to identify the foreground object's boundary, precluding its application to photos taken by point-and-shoot cameras. Second, Luo *et al*. [10]'s approach adopts a category-dependent regional feature extraction, which has the prerequisite that photos are 100% accurately classified into one of the seven categories. This prerequisite is infeasible in real applications. Third, the attributes proposed in Dhar *et al*. [9]'s approach are designed manually and are data set dependent, thus have difficulty in generalizing to different data sets. Fourth, all these global low-level and high-level visual features are designed heuristically, there is short of evidence that they effectively capture the photo aesthetics, such as the spatial interaction between the water and the sailboat in Fig. 1.

B. Probabilistic Local Patches Integration-Based Approaches

To describe the spatial interaction of image patches, probabilistic local patch integration based approaches is proposed. These approaches extract local patches within each candidate cropped photo, and then probabilistically integrate them into a quality measure to select the cropped photo. In [7], Nishiyama *et al*. first detected multiple subject regions in an image, where each subject region is a bounding rectangle containing the salient part of each object. A SVM classifier is then trained for each subject region. The quality of each candidate cropped photo is computed by probabilistically combining the scores of the SVM classifier corresponding to the cropped photo's internal subject regions. Although multiple subjects are considered in [7], their spatial interactions, such as whether the sky is below or above the sea, are ignored. In [6], Cheng *et al*. proposed omni-range context, *i.e.*, spatial distributions of arbitrary pairwise image patches, to model the photo compositions. The learned omni-range context priors are combined with the other cues, such as the patch number, to form a posterior probability for measuring the quality of each candidate cropped photo. It is noticeable that, the omni-range context only captures the binary spatial interactions of image patches. Higher-order spatial interactions, such the four linearly arranged sailboats in Fig. 1, cannot be captured. To describe the high-order spatial interactions of image patches, Zhang *et al.* [38] introduced graphlets and further designed a probabilistic model to transfer them from the training photos into the cropped photo. However, graphlets reflect no photo semantics and photo global spatial configurations, which are essential cues to be exploited in a cropping model. Besides, the color and texture channel visual features are identically weighted in the graphlet transferring process, which is not consistent with human aesthetics.

To address the above problems, we extract graphlets to describe the high-order spatial interaction of atomic regions in

[1]We suggest readers refer to Zhang *et al.* [38]'s work for a more comprehensive overview of the representative photo cropping methods.

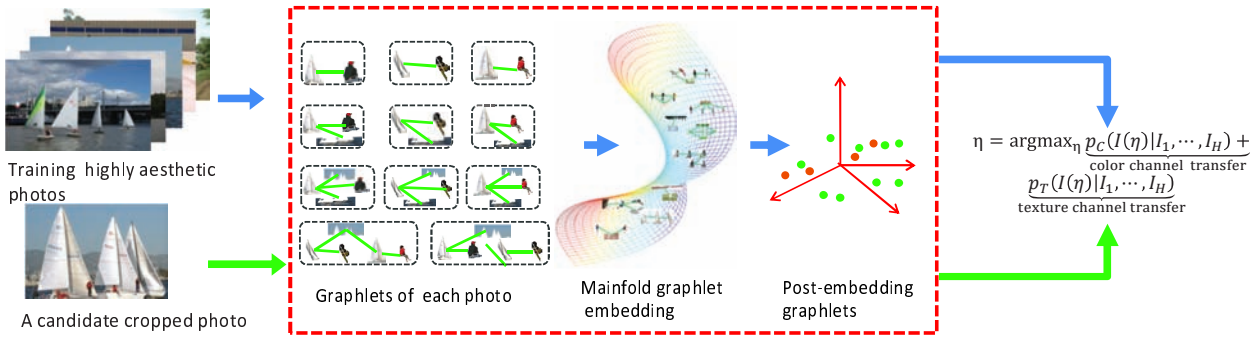


Fig. 2. The pipeline of the proposed approach

a photo. Then, a weakly supervised framework is proposed to integrate image-level semantics and photo global spatial layout into graphlets. Finally, a probabilistic model is derived to transfer training post-embedding graphlets into the cropped photo, where multi-channel graphlets are automatically weighted.

III. Transferring Image-Level Semantics into Graphlets

A. Defining Graphlet under Spatial Pyramid Framework

There are usually tens to hundreds of objects in a photo. Among these components, a few spatially adjacent ones as well as their correlations determine the local composition of a photo. The local composition may reflects the regional aesthetics in a photo, thus it is essential to exploit them in a cropping model. To this end, we ameliorate the graphlet in Zhang *et al.* [38] by re-defining it under the spatial pyramid framework.

In Zhang *et al.* [38]'s work, two atomic regions are considered as adjacent if they are spatially connected. This criteria is too strict in practice. For example, although the three neck-in-neck horses in Fig. 3, which are three atomic regions, are closely located and aesthetically pleasing, they are not spatially connected. Thus, there are deemed as non-adjacent in the previous model. To solve this problem, inspired by spatial pyramid [15], which uses cells from multi-level spatial pyramid to label the location of each local feature, we construct a three-level spatial pyramid to label the location of each atomic region. As shown in Fig. 3, an atomic region's corresponding cell denotes the cell into which it can be maximally divided, in a coarse-to-fine manner. The two left horse and their riders can be maximally divided into cell ϕ_{21}^2 while the right horse and its rider corresponds to cell ϕ_{22}^2, where the upper index represents the level in the pyramid. Unlike local feature location labeling, it is difficult to completely group an atomic region into a cell because each atomic region usually contains hundreds of pixels and some may stick out of the cell. In this work, if 90% of the pixels in an atomic region are overlapped with a cell, we consider that this atomic region can be grouped into this cell. After the labeling process, two regions are spatially adjacent if their corresponding cells are identical or neighboring.

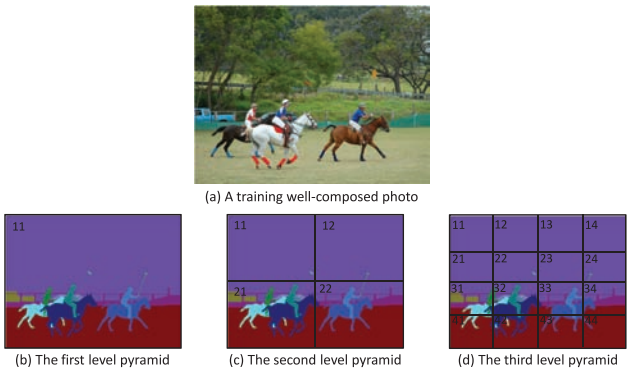


Fig. 3. An illustration of the newly defined spatially adjacent regions.

Following the above spatial pyramid-based adjacent region identification, we define the graphlet $\mathcal{G}$ to formularize the local composition of each photo, that is,

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

where V denotes a small set of vertices, each representing an atomic region obtained via multiple unsupervised fuzzy clusterings (UFCs) [14], and E denotes a set of edges, each connecting a pair of spatially adjacent atomic regions. UFC algorithm is an improved clustering algorithm, it guarantees the less consuming time and good clustering precision. Moreover, there is no need to know the cluster number and it can cluster arbitrary -shaped cluster. When adopting UFC on image segmentation, the advantage is that, prior knowledge of the number of segmented atomic regions is not required, and its tolerance bound is flexible to tune. Second, each photo is segmented five times under different tolerance bounds of UFC, *i.e.*, the tolerance bound is tuned from 0.1 to 0.5 with a step of 0.1.

The graphlet size denotes the number of vertices in a graphlet. Noticeably, the number of graphlets from a photo is exponentially increasing with its size. As shown in Fig. 4, suppose the left three segmented regions are spatially neighboring. There will be three 1-sized graphlets, three 2-sized graphlets and four 3-sized graphlets. Thus, the total number of resulting graphlets is $C_3^1+C_3^2+C_3^3 = 2^3-1 = 7$. And straightforwardly, we four spatially neighboring segmented regions are considered, there will be $2^4 - 1 = 15$ different graphlets. Therefore, toward an effective cropping system, only small-sized graphlets are adopted. Because the color and the texture channels are generally complementary to each

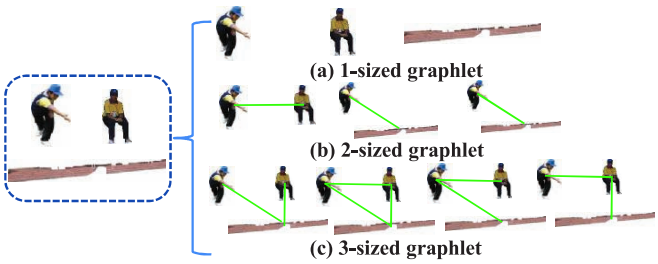


Fig. 4. A graphical illustration of the exponential number of graphlets in a photo.

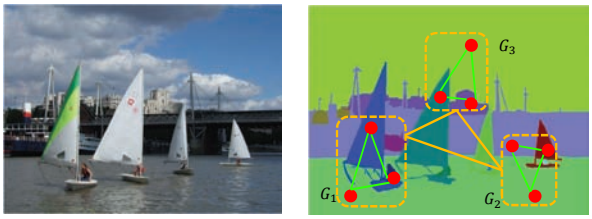


Fig. 5. An example of preserving pairwise graphlets' Golub-Werman distances

other in measuring the appearances of each atomic region, we describe each atomic region in color and textural channels, which are implemented as 9-dimensional color moment [16] and 128-dimensional histogram of gradient (HOG) [17] respectively.

B. Manifold Graphlet Embedding

Both of a graphlet's atomic regions and structure could be represented by appearance feature vectors, but it is natural to represent a graphlet by concatenating appearance feature vectors since they collaboratively contribute to photo aesthetics. First, we define two matrices to symbolize the atomic regions and structure. Given a t-sized graphlet in color channel, we characterize all its atomic regions by a matrix $M_C \in \mathbb{R}^{t\times 9}$, each row of which denotes a 9-dimensional feature vector signifying the color moment of an atomic region. To represent the structure of a graphlet or the spatial correlation between atomic regions in a graphlet, we adopt a $t \times t$-sized adjacent matrix as:

$$M_S(i,j) = \begin{cases} 1 & \text{if } R_i \text{ and } R_j \text{ are spatially adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where R_i and R_j denote atomic regions corresponding to the i-th and the j-th vertex in a graphlet respectively.

With M_C and M_S, we represent a t-sized graphlet by a matrix of $t \times (9+t)$ as:

$$M = [M_C, M_S] \tag{3}$$

Following [20], each matrix can be deemed as a point on the Grassmann manifold, and the Golub-Werman distance [21] between identical-sized matrices is defined as:

$$d_{GW}(M, M') = \|M_O - M_O'\|_2 \tag{4}$$

where M_O and M_O' denote the orthonormal basis of M and M' respectively.

To incorporate image-level semantics and image spatial configuration with graphlets, we propose a manifold embedding [40], [41] algorithm with the objective function as:

$$\arg\min_Y \underbrace{\sum\nolimits_h \sum\nolimits_{ij} [d_{GW}(M_i^h, M_j^h) - d_E(y_i^h, y_j^h)]^2}_{\textit{Preserve pairwise graphlets Golub–Werman distances}} + \underbrace{\sum\nolimits_{ij} \|y_i - y_j\|^2 l_s(i,j) - \sum\nolimits_{ij} \|y_i - y_j\|^2 l_d(i.j)}_{\textit{Represent image–level semantics}} \quad s.t.\ YY^T = \mathbf{I}_d \tag{5}$$

where $Y = [y_1, y_2, \cdots, y_N]$, in which y_i^h and y_j^h are column vectors standing for the d-dimentional representations of the i-th and the j-th graphlets from the h-th photo. Our goal is to seek a Y that characterize the image spatial configuration, and consider the image-level sematics as well. For the former purpose, the first term in (5) preserves all pairwise graphlets' Golub Werman distances, which reflects the global spatial layout of a photo, as shown in Fig. 5 and we explain it as follows. The distance between pairwise graphlets reflects their relative displacement. As shown in the right of Fig. 5, the Golub-Werman distance d_{GW} between graphlet G_1 and G_3 reflects the relative position between two residential areas covered by G_1 and G_3. Straightforwardly, if we preserve all the pairwise distances between graphlets in the embedding process, all their relative positions are kept. And this operation can implicitly kept the global spatial layout. As shown in Fig. 5, preserving three relative distances between (G_1, G_2), (G_1, G_3) and (G_2, G_3) roughly capture the global spatial layout, and intuitively, when more graphlets are considered, more accurate global spatial layout can be kept. For the latter one, we add the photo category information as the second term.

Here, we explain (5) in detail as follows. M_i^h and M_j^h respectively denote matrices corresponding to the i-th and the j-th identical-sized graphlets from the h-th photo, and $d_E(\cdot,\cdot)$ represents the Euclidean distance. $l_s(\cdot,\cdot)$ is a function measuring the semantical similarity between graphlets, $l_d(\cdot,\cdot)$ is a function measuring the semantical difference between graphlets. Denoting $\vec{N} = [N^1, N^2, \cdots, N^C]^T$ where N^c is the number of photos from the c-th category, and $c(\cdot)$ the photo category of photo from which the graphlet is extracted, then $l_s(i,j) = \frac{[c(G_i) \bigcap c(G_j)]\vec{N}}{\sum_c N^c}$ and $l_d(i,j) = \frac{[c(G_i) \bigoplus c(G_j)]\vec{N}}{\sum_c N^c}$. $YY^T = \mathbf{I}_d$ is a term to uniquely determine Y. Noticeably, different-sized graphlets are embedded independently based on (5).

Denote D_{GW}^h as an $N \times N$ matrix whose ij-th entry is $d_{GW}(M_i^h, M_j^h)$, *i.e.,* the Golub-Werman distance between the i-th and the j-th identical-sized graphlet extracted from the h-th photo. Then, the inner product matrix is obtained by $\tau(D_{GW}^h) = -R_{N_h} S_{GW}^h R_{N_h}/2$, where $(S_{GW}^h)_{ij} = (D_{GW}^h)_{ij}^2$; $R_{N_h} = \mathbf{I}_{N_h} - \vec{e}_{N_h}\vec{e}_{N_h}^T/N$ is the centralization matrix; $\mathbf{I}_{N_h}$ is a $N_h \times N_h$ identity matrix and $\vec{e}_{N_h} = [1, \cdots, 1]^T \in \mathbb{R}^{N_h}$; and N_h is the number of graphlets from the h-th photo. The first term in (5)

can be rewritten as:

$$\begin{aligned}&\arg\min_Y \sum_h \sum_{ij} [d_{GW}(M_i^h, M_j^h) - d_E(y_i^h, y_j^h)]^2 \\ &= \arg\min_Y \sum_h \|\tau(D_{GW}^h) - \tau(D_Y^h)\|^2 \\ &= \arg\max_Y tr(Y\tau(D_{GW}^h)Y^T) \\ &= \arg\max_Y tr(Y\tau(D_{GW})Y^T) \end{aligned} \quad (6)$$

where $\tau(D_{GW})$ is a block diagonal matrix with $H \times H$ blocks, and the h-th diagonal block is $\tau(D_{GW}^h)$.

The second term in (5) can be rewritten as:

$$\begin{aligned}&\arg\min_Y \sum_{ij} \|y_i - y_j\|^2 [l_w(i,j) - l_b(i,j)] \\ &= \arg\max_Y tr(YAY^T) \end{aligned} \quad (7)$$

where $A = [-\vec{e}_{N-1}^T, \mathbf{I}_{N-1}]^T W_1 [-\vec{e}_{N-1}^T, \mathbf{I}_{N-1}] + \cdots + [\mathbf{I}_{N-1}, -\vec{e}_{N-1}^T]^T W_N [\mathbf{I}_{N-1}, -\vec{e}_{N-1}^T]$, and W_i is an $N \times N$ diagonal matrix whose h-th diagonal element is $l_s(h,i) - l_d(h,i)$.

Based on (6) and (7), the objective function (5) can be reorganized into:

$$\begin{aligned}&\arg\max_Y tr(Y(A + \tau(D_{GW})Y^T))) \\ &= \arg\max_Y tr(YZY^T) \quad s.t. \quad YY^T = \mathbf{I}_d \end{aligned} \quad (8)$$

where $Z = A + \tau(D_{GW}) \in \mathbb{R}^{N \times N}$.

C. Incremental Graphlet Embedding Algorithm

The problem in (8) is a quadratic programming with quadratic constraints that can be solved using eigenvalue decomposition, which has a time complexity of $O(N^3)$. However, Z is a large-sized matrix because usually $N > 100,000$, thus it is computational intractable to solve (8) using a global once-for-all eigenvalue decomposition. To solve this problem, we develop an incremental graphlet embedding algorithm. First, we solve an initial embedding using (8) under graphlets extracted from $\{I^1, \cdots, I^{H^{(0)}}\}$ photos, where $H^{(0)} << H$. Then we solve a new embedding under graphlets extracted from $\{I^1, \cdots, I^{H^{(0)}}, \cdots, I^{H^{(1)}}\}$ photos. The objective function is:

$$\arg\max_{Y^{(1)}} tr(Y^{(1)} Z^{(1)} (Y^{(1)})^T) \quad s.t. \quad y_i^{(1)} = y_i^{(0)}, i \in 1, \cdots, P \quad (9)$$

where $Y^{(1)} = [Y_L, Y_U] \in \mathbb{R}^{d \times H^{(1)}}$; $Y_L = \{y_1, \cdots, y_P\}$ is the known embedding in $Y^{(1)}$ from the previous incremental embedding under $\{I^1, \cdots, I^{H^{(0)}}\}$ photos, and $Y_U = \{y_{P+1}, \cdots, y_Q\}$ is the unknown embedding in Y; $Z^{(1)}$ is the matrix constructed from $H^{(1)}$ photos which can be divided into four blocks as:

$$Z^{(1)} = \begin{pmatrix} Z_{LL} & Z_{LU} \\ Z_{UL} & Z_{UU} \end{pmatrix} \quad (10)$$

Denote d as the dimensionality of post-embedding graphlet, the problem in (9) can be decomposed into d sub-problems. Each sub-problem is a quadratic problem with linear constrains that can be iteratively solved. Let $Y_L^i = [f_i(1), \cdots, f_i(d)] \in \mathbb{R}^d, i = 1, \cdots, P$, we can reorganize (9) into:

$$\begin{cases} \arg\max_{X(1)} tr(X(1)Z^{(1)}X(1)^T) \ s.t. \ x_i(1) = f_i(1) \ i = 1, \cdots, P \\ \arg\max_{X(d)} tr(X(d)Z^{(1)}X(d)^T) \ s.t. \ x_i(d) = f_i(d) \ i = 1, \cdots, P \end{cases} \quad (11)$$

where $X(i)$ is a Q-dimensional row feature vector to be solved.

Since each sub-problem in (9) has the same form, we can simplify (11) into:

$$\arg\max_X tr(XZ^{(1)}X^T) \quad s.t. \quad x_i = f_i, \quad i = 1, \cdots, P \quad (12)$$

Converting the hard constrains in (12) into soft constrains and introducing a prediction term, we have the following regularization representation:

$$\arg\max_X XZ^{(1)}X^T + \mu_1 \sum_{i=1}^P (x_i - f_i)^2 + \mu_2 \sum_{i=P+1}^Q (x_i - g_i)^2 \quad (13)$$

where g_i is a predicted value of x_i which is specified by the prediction strategy proposed by Xiang *et al.*[25]. Given M_i and its k neighbors $\{M_1, \cdots, M_{k'}, M_{k'+1}, \cdots, M_k\}$, where the low-dimensional representation of $\{M_1, \ldots, M_{k'}\}$ are known while that of $\{M_{k'+1}, \cdots, M_k\}$ are unknown, we first use kernel PCA (with kernel $k^{PCA}(M, M') = \exp\left(-d_{GW}^2(M, M')\right)$ to transfer M_i and all its neighbors $\{M_1, \cdots, M_{k'}, M_{k'+1}, \cdots, M_k\}$ into a set of d-dimensional feature vectors. Then we learn a function satisfying $f_i = g(m_i), i \in \{1, \cdots, k'\}$, where $g(m)$ is a spline regression function defined as: $g(m) = \sum_{i=1}^{k'} \alpha_i \phi_i(m) + \sum_{i=1}^{\lambda} \beta_i \psi_i(m)$, here m_i is the low-dimensional representation of M_i using kernel PCA; $\phi_i(m) = \|m - m_i\|$ and $\Psi_i(m)_{i=1}^{\lambda}$ constitutes a base of a polynomial space; and the parameter α and β are computed by solving the linear system: $\begin{pmatrix} \Phi & \Psi \\ \Phi^T & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$, where Φ is an $k' \times k'$ matrix with each element $\Phi_{ij} = \phi(\|m_i - m_j\|)$, Ψ is an $k' \times p$ matrix with element $\Psi_{ij} = \psi_i(m_j)$ and $f = [f_1, \cdots, f_{k'}]$.

By differentiating the objective function (13) with respect to X_U and setting the derivative to 0, we obtain

$$X_U = \frac{1}{1+\mu_2} X_U (\mathbf{I} - Z_{UU}) - \frac{1}{1+\mu_2} Z_{UL} X_L + \frac{\mu_2}{1+\mu_2} g_U \quad (14)$$

where X_U denotes one dimension of the unknown post-embedding graphlets. Thereby we can update X_U based on the following equation:

$$X_U^{(t+1)} = \frac{1}{1+\mu_2} X_U^{(t)} (\mathbf{I} - Z_{UU}) - \frac{1}{1+\mu_2} Z_{UL} X_L + \frac{\mu_2}{1+\mu_2} g_U \quad (15)$$

The iteration is carried out repeatedly until X_U becomes stable. To obtain a d-dimensional embedding of X_U, the iterative algorithm is carried out d times, and we finally obtain the low-dimensional representation of graphlets $Y^{(1)} = [X(1), \cdots, X(d)]^T$ from this incremental embedding step. In the next incremental embedding step, we solve the new embedding $Y^{(2)}$ from $\{I^1, \cdots, I^{H^{(1)}}, \cdots, I^{H^{(2)}}\}$ photos, where the embedding process is the same as that of the previous incremental embedding step.

IV. Transferring Multi-Channel Graphlets into the Cropped Photo

We believe that photo cropping preserves the post-embedding graphlets through a process of inference, wherein the post-embedding graphlets are used to make the probabilistically best guesses about what should be preserved in the cropped photo. Here we apply a BN to implement photo cropping. Given a test photo I, we define its cropped photo

as $I(\eta)$, where $\eta = (\eta_s, \eta_\theta, \eta_t)$ is the cropping parameter. In particular, η_s is a two-dimensional variable denoting the XY coordinate scale of the cropped photo, $\eta_\theta \in [0, 2\pi]$ is the rotation angle of the cropped photo, and η_t is a two-dimensional variable denoting the translation from the center of the test photo to that of the cropped photo.

For an original photo, its corresponding cropped photo

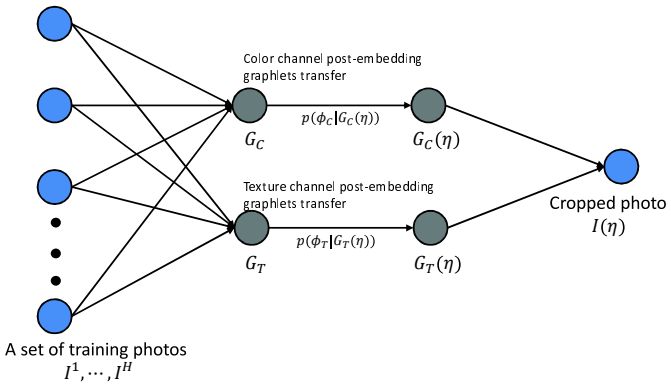


Fig. 6. BN-based photo cropping

should preserve the aesthetic features from the highly aesthetic training photos to the largest possible extent. We notice that the training photos and the cropped one are highly correlated by their aesthetic features. In terms of color channel post-embedding graphlets, there are strong correlations between the following three pairs of variables: 1) $I^1, \cdots, I^H$ and G_C, 2) G_C and $G_C(\eta)$, and 3) $G_C(\eta)$ and $I(\eta)$. In terms of texture channel graphlets, there are also strong correlations between another three pairs of variables: 4) $I^1, \cdots, I^H$ and G_T, 5) G_T and $G_T(\eta)$, and 6) G_T and $I(\eta)$, where $I^1, \cdots, I^H$ denote the states of training photos and $I(\eta)$ the state of the cropped photo; G and $G(\eta)$ respectively denote the states of post-embedding graphlets from the training photos and the cropped photo; $G_C(\eta)$ and $G_T(\eta)$ the states of color and texture channel post-embedding graphlets from the cropped photo. The above six correlations can be represented by a BN [37], [36]. As shown in Fig. 6, the proposed BN contains two types of nodes: observable nodes (blue-colored) and hidden nodes (gray-colored). Directed edges describe the relationships between nodes. These two types of nodes form four layers. The first layer represents all training photos. The second layer denotes the training post-embedding graphlets. The third layer represents post-embedding graphlets from the cropped photo. And the fourth layer represents the cropped photo.

Based on the above BN, we cast photo cropping as a process that maximally transfers the post-embedding graphlets from the highly aesthetic training photos into the cropped photo. This process can be formulated into the following maximum a posterior (MAP) framework:

$$\begin{aligned}\eta &= \arg\max_\eta p(I(\eta)|I^1, I^2, \cdots, I^H) \\ &= \arg\max_\eta \underbrace{p(\phi_C|G_C(\eta))p(I(\eta)|G_C(\eta))p(G_C(\eta)|G_C)p(G_C|I^1, \cdots, I^H)}_{Color\ channel\ graphlets\ transfer} \\ &+ \underbrace{p(\phi_T|G_T(\eta))p(I(\eta)|G_T(\eta))p(G_T(\eta)|G_T)p(G_T|I^1, \cdots, I^H)}_{Texture\ channel\ graphlets\ transfer}\end{aligned} \quad (16)$$

where $p(\phi_C|G(\eta))$ and $p(\phi_T|G(\eta))$ respectively denote the importance of color and texture channel post-embedding graphlets $G(\eta)$ from the cropped photo. Specifically, $p(\phi_C|G_C(\eta)) + p(\phi_T|G_T(\eta)) = 1$ and $p(\phi_T|G_T(\eta))$ is determined by the sparseness of texture channel graphlet $G_T(\eta)$, which is defined it as a logistic function, *i.e.*,

$$p(\phi_T|G_T(\eta)) = \frac{1}{1 - \exp(-aG_T(\eta) + b)} \quad (17)$$

where a and b are parameters obtained from the training data.

As probabilities of graphlets in different channels have the same form, we omit the subscript of G_C,G_T and $G_C(\eta)$,$G_T(\eta)$. To calculate the above three probabilities in (16), we need another three probabilities $p(I(\eta)|G(\eta))$, $p(G(\eta)|G)$, and $p(G|I^1, \cdots, I^H)$. They are detailed as follows.

$$\begin{aligned}p(I(\eta)|G(\eta)) &= \frac{p(G^1(\eta), \cdots, C^T(\eta)|I(\eta))p(I(\eta))}{p(G^1(\eta), \cdots, G^T(\eta))} \\ &\propto p(G^1(\eta), \cdots, C^T(\eta)|I(\eta))p(I(\eta)) \\ &= \prod_{i=1}^{T} p(G^i(\eta)|I(\eta))p(I(\eta)) \\ &= \prod_{i=1}^{T}\prod_{j=1}^{Y_i} p(G_j^i(\eta)|I(\eta))p(I(\eta))\end{aligned} \quad (18)$$

where T is the maximum graphlet size and Y_i the number of i-sized graphlets in I; G^i denotes all the training i-sized graphlets and G_j^i the j-th i-sized training graphlet; $p(G_j^i|I)$ denotes the probability of extracting graphlets G_j^i from photo I. $p(I(\eta))$ is the probability of a photo I cropped using parameter η, which is defined as Gaussian kernel: $p(I(\eta)) = \exp\left(\frac{\|\eta - \bar{\eta}\|^2}{\sigma_\eta^2}\right)$. In this work, the graphlet extraction is based on random walking. We first index all atomic regions 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 (same probability of visiting a larger or smaller-indexed vertex) with probability $\frac{1}{2 * \sum_{de} p_{de}(R_l) de(R_l)}$, where $de(R_l)$ denotes the degree of the current atomic region R_l and $p_{de}(R_l)$ the probability of atomic region R_l with degree $de(R_l)$. In our implementation, $p_{de}(R_l)$ and $p(Y)$ are both defined as Gaussian kernels: $p_{de}(R_l) \propto \exp\left(-\frac{\|de(R_l) - \overline{de}(R)\|^2}{\sigma_{de}^2}\right)$ and $p(Y) \propto \exp\left(-\frac{\|Y - \bar{Y}\|^2}{\sigma_Y^2}\right)$. The random walking process stops when the maximum graphlets size is reached. Therefore, we obtain

$$p(G_j^i|I) \propto \frac{P(Y)}{Y}\prod_{l=1}^{i-1}\frac{1}{2 * \sum_{de} p_{de}(R_l) de(R_l)} \exp\left(-\frac{\|s(R_l) - \bar{s}\|^2}{\sigma_S^2}\right) \quad (19)$$

where the term $\exp\left(-\frac{\|s(R_l) - \bar{s}\|^2}{\sigma_S^2}\right)$ encourages choosing moderate-sized atomic regions. Our model suppresses choosing small-sized, even highly-aesthetic, atomic regions because they are not representative to the original photo. $s(R_l)$ is the number of pixels in atomic region R_l. $\bar{s}$ and σ_s respectively denote the Gaussian center and the covariance of pixel number in atomic regions.

$$\begin{aligned}p(G|I^1, \cdots, I^H) &= p(G^1, \cdots, G^T|I^1, \cdots, I^H) \\ &= \prod_{i=1}^{T} p(G^i|I^1, \cdots, I^H) = \prod_{i=1}^{T}\prod_{j=1}^{Y_i} p(G_j^i|I^1, \cdots, I^H)\end{aligned} \quad (20)$$

where $p(G_j^i|I^1,\cdots,I^H)$ is the probability of graphlet G_j^i extracting from all training photos $I^1,\cdots,I^H$ and is defined as:

$$p(G_j^i|I^1,\cdots,I^H) = 1 - \prod_{h=1}^{H}\left(1 - p(G_j^i|I^h)\right) \quad (21)$$

$p(G(\eta)|G)$ measures the similarity between graphlets from the training photos and those from the cropped photo, *i.e.*,

$$p(G(\eta)|G) \propto \exp\left(-\frac{1}{T * Y_i * Y_i(\eta)}\sum_{t=1}^{T}\sum_{i=1}^{Y_t}\sum_{j=1}^{Y_t(\eta)}\frac{\|G_i^t(\eta) - G_j^t\|^2}{\sigma^2}\right) \quad (22)$$

where $Y_t(\eta)$ denotes the number of t-sized graphlets from the cropped photo; $G_j^t(\eta)$ is the j-th t-sized graphlet from the cropped photo.

Based on (16), we use Gibbs sampling to compute the optimal cropping parameter, where the procedure is the same as in our previous work [20]. We present the procedure of the proposed weakly supervised photo cropping in Algorithm 1.

Algorithm 1 Weakly Supervised Photo Cropping

input:a set of category-labeled training photos $I^1,\cdots,I^H$; a test photo I; and maximum graphlet size T;
output: a cropped photo $I(\eta)$;
1. Apply UFC-based segmentation to decompose each photo into atomic regions; extract $\{1,\cdots,T\}$-sized multi-channel graphlets fromtraining photos based on random walking.
2. Adopt the manifold embedding to transform color and texture channel graphlet into d-dimensional feature vectors based on (5).
3. Gibbs sampling for optimal cropping parameter selection based on (16), and output the corresponding cropped photo.

V. Experimental Results and Analysis

In this section, we evaluate the effectiveness of the proposed weakly supervised photo copping. The first part shows the effectiveness of the post-embedding graphlets in capturing photo aesthetics. The second part evaluates the performance of the proposed approach by comparing with representative cropping methods. The third part step-by-step evaluates the effectiveness of each component in the proposed approach. In the fourth part, we discuss the influences of the free parameters on the cropping results. The last part presents the cropping results under different segmentation results[2].

A. Data Collection and Preprocessing

To the best of our knowledge, there are no standard data sets released for evaluating cropping performance. Therefore, we compile our own photo cropping data set. The total training data contains approximately 6000 highly-ranked as well as 6000 low-ranked photos, which are crawled from two online photo sharing websites Photosig and Flicker. Because both 4:3 aspect ratio and panoramic photos are used in the previous cropping experiments, towards a comprehensive comparative study, we construct two groups of test photos. The first group contains 337 badly-composed photos with an aspect ratio of 4:3. We intend to obtain a well-composed photo by cropping a sub-region from the original photo. The second group contains 313 well-composed panoramic photos. We intend to maximally preserve the composition from the original panoramic photos into the cropped 4:3 aspect ratio photos. Due to space limitation, we only present the cropping results using test photos with a 4:3 aspect ratio[3].

To obtain the image-level semantics of each photo in our data set, we represent each photo by a 512-dimensional Gist descriptor [25] and then classify it by a 13-class SVM classifier. We use the probabilistic SVM [35] and set the semantics of each photo as the three highest probable predicted labels. The 13-class SVM classifier is trained from the scene data set published by Feifei *et al*. [26].

B. Aesthetics Captured by the Post-Embedding Graphlets

In this experiment, we evaluate the effectiveness of our post-embedding graphlets in capturing the photo aesthetics. We compare the post-embedding graphlets with six aesthetic features proposed by Luo *et al.* [1], Luo *et al.* [10], Ke *et al*. [2], Yeh *et al.* [8], and Zhang *et al.* [38]. The saliency model proposed by Itti *et al*. is also used employed for comparison. In particular, we experiment on the data set collected by Yeh *et al.*, which contains 6000 highly aesthetic as well as 6000 low aesthetic photos collected from DPChallenge[4]. Most images from this data set contain one single object and the background is typically clear. To compare the effectiveness of the five features, we use each feature to predict whether a test photo is highly aesthetic or low aesthetic. We use the same split of training and test sets as in the program provided by Yeh *et al.*, and then train a binary SVM classifier based on the five features. Noticeably, as local descriptors, multi-channel post-embedding graphlets cannot be directly used to train an image-level photo quality classifier. Inspired by graph kernel [33] which measures the similarity between pairwise graphs by comparing all their respective subgraphs, we construct an image-level kernel to measure the similarity between photos, *i.e.*,

$$k^L(I,I') = \frac{1}{N_I * N_{I'}}\sum\nolimits_{G\in I, G'\in I'} k(G,G') \quad (23)$$

where $k(\cdot,\cdot)$ is the basis kernel which is set linear in our experiment; $\frac{1}{N_I * N_{I'}}$ is a normalization factor wherein N_I and $N_{I'}$ respectively denote the numbers of graphlets in I and I'. Note that, two graph kernels k_{CM}^L and k_{HOG}^L are constructed in both color and texture channels. To integrate the two kernels together as an input to kernel SVM, we use multiple kernel learning (MKL)[5] [34] which automatically learns the weights of each kernel and linearly combine them.

To quantitatively compare the proposed approach with the six competitors, we calculate the precision recall curve is calculated via: $Precision = \frac{tp}{tp+fp}$ and $Recall = \frac{tp}{tp+fn}$, where tp and fp are true positive and false positive respectively, and fn and tn are false negative and true negative respectively. As shown in Fig. 7, the proposed method significantly outperforms its competitors. The reasons are given

[2]This part is presented in the supplemental material.

[3]Cropping results from panoramic test photos are given in the supplemental material.

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

[5]http://www.di.ens.fr/~fbach/path/.

Fig. 8. Asthetics captured by four top-ranked graphlets based on Zhang *et al.* [38]'s approach (green rectangle) and the proposed approach (blue rectangle). We present the four most discriminative graphlets from each photo.

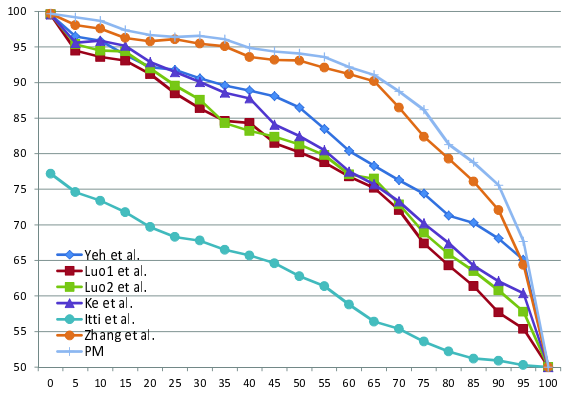


Fig. 7. Precision-Recall curve of the seven compared features (PM denotes the proposed method)

as follows. First, Luo *et al.*'s and Yeh *et al.*'s approaches are partially based on the assumption that the foreground and background of the photos taken by SLR cameras can be easily discriminated. Unfortunately, this assumption does not hold in the experimental data set. Second, there is lack of evidence that the concatenated global low-level and high-level features effectively capture the photo aesthetics, since they are defined intuitively. Third, the worst performance is achieved by the saliency model from Itti *et al.*. This is because the saliency map only tells the conspicuity of each pixel and it fails to capture important aesthetic features of a photo, such as color and texture information. Fourth, the graphlet used in our previous work only capture local spatial compositions. Global spatial configurations and photo semantics, which are also essential cues for predicting photo aesthetics, are not considered.

To further demonstrate the advantage of our approach over the previous work, we compare the discriminative graphlets produced by the proposed method and those obtained from the previous work, on the LHI data set [39]. This data set contains 10 sports event categories collected from the Internet. The experimental settings of the previous work are the same as in [38]. That is, in each category, we use half the photos for training and leave the rest for testing. We set the maximum size of graphlet T to 5. In each category, we calculate the discrimination of a graphlet based on:

$$g(M) = \frac{1}{N}\sum_{i=1}^{N} \alpha k(M_i, M) \tag{24}$$

where M denotes the matrix obtained from the graphlet and N the number of training graphlets, and α is the leading eigenvector in solving the linear discriminate analysis [31].

For the proposed method, we manually assign each photo with multiple image-level labels. The discrimination of each graphlet is computed based on the second term in (5). We present the comparative discriminative graphlets in Fig. 8. The proposed approach has the following advantages.

1) The proposed method preserves more semantic objects in the graphlets. For instance, in the "Rowing" category, the graphlets produced by the proposed method capture the spatial interactions between the waterman, the water,

and the trees, which are important semantics that should be preserved in the cropped photo. However, in the previous work, only the spatial correlation among the watermen are captured. In the "Ice-skate" category, the graphlets obtained by the proposed method well describe the spatial correlation among the hand-in-hand skaters, the ice, and the architecture. Nevertheless, in the previous model, only spatial interactions of the skaters are considered.

2) In addition to photo semantics, the global spatial configurations are more appropriately captured by the graphlets from the proposed approach. As shown in the "Sailing" category, the graphlets produced by the new approach well describe the spatial interaction among the sky, the linear arranged sailboats, and the water, which reflects the global configuration of this photo. In contrast, the previous work only captures the local composition, such as the linearly arranged sailboats. Moreover, as shown in the "Badminton" category, the spatial interaction among the wall, the floor, and the four players are well captured by the proposed method. However, in the previous work, the wall is neglected.

C. Comparison with the State of the Art Approaches

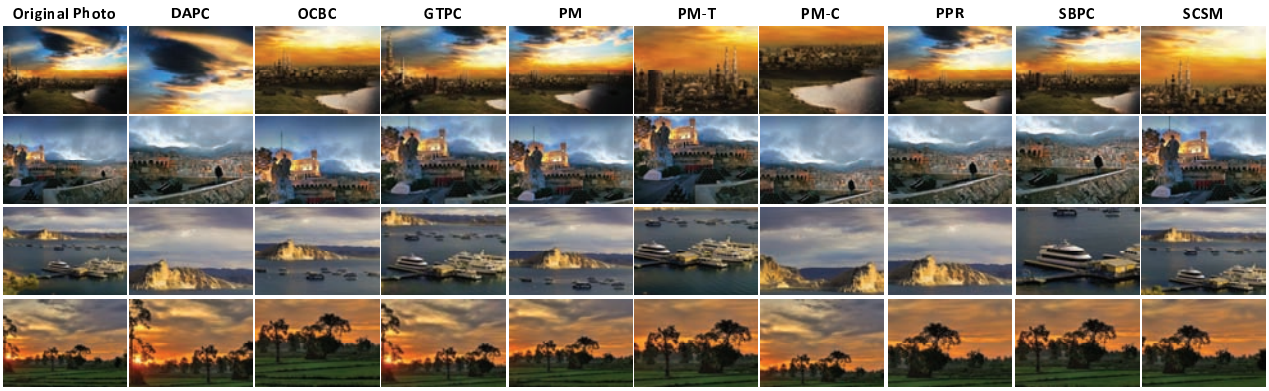
In this subsection, we evaluate the proposed method (PM) in comparison with several well-known cropping methods: sparse coding of saliency maps (SCSM [11]), sensation based photo cropping (SBPC [7]), omni-range context-based cropping (OCBC [6]), personalized photo ranking (PPR [8]), describable attribute for photo cropping (DAPC [9]), and our graphlet transfer-based photo cropping (GTPC [38]). Besides, cropping results obtained only through color channel post-embedding graphlets transferring (PM-C), and only through texture channel post-embedding graphlets transferring (PM-T) are presented also. We follow the experimental settings in the previous work [38].

In Fig. 9, we present the cropping results obtained from the above methods and make the following observations:

1) As shown in the second and the third column, the global features-based photo evaluation methods, PPR and DAPC, are less effective to capture the local details from the original photos. Important local details, such as the sculpture from the second original photo and the scatted yachts from the third original photo, are not preserved by PPR and DAPC.
2) As shown in the fourth column, the saliency model based cropping, SCSM, only crops the most salient region. But salient region is not always consistent with the region that best preserves important visual cues. Particularly, the colorful sunset, the sculpture, and the scattered yachts from the first three original photos are totally ignored by SCSM.
3) As shown in the fifth and the sixth column, the two probabilistic local patch integration based cropping methods, SBPC and OCBC, are competitive but still less effective than our approach. The third cropping result produced by SBPC is less visually balanced than our approach. The first cropping result produced by OCBC not well captures the colorful sunsets.
4) As shown in the seventh column, compared with the previous work [38], the new model crops a photo with more balance between the global spatial layout and the local composition, especially the first and the third photo. Besides, photos cropped using the new approach are more correlated with semantics, as shown in the last two photos. This observation reveals semantics is more effectively preserved in the new model.
5) As shown in the eighth column, our approach preserves both the global spatial configurations and the local details from the original photo. Specifically, the global spatial configurations are: the colorful sunset and lake surface from the first original photo, the architectures and the sky from the second original photo, the yachts, island, and sky from the third original photo, as well as the trees, cultivated field, and sunset from the fourth original photo. The local details are: the architecture areas from the first original photo, the sculpture and arches from the second original photo, the scattered yachts from the third original photo, and the trees from the fourth original photo. Besides, each semantic objects are well preserved in the cropped photo. These observations demonstrate that image-level semantics and image global spatial characteristics are effectively transferred into graphlets.
6) As shown in the last two columns, with only color channel or texture channel graphlet transferring either preserves too much color or texture regions. They are both suboptimal cropping results.

We present the preference matrix corresponding to the above compared methods. Each preference matrix is filled by 30 volunteers at Zhejiang University. The result clearly confirm the advantage of the proposed method.

The time consumption analysis of the proposed method is as follows. All experiments were carried out on a personal computer equipped with Intel E8500 and 4GB RAM. All the six compared methods as well as our approach are implemented on Matlab platform. We present the average time consumption of photos with different aspect ratios in Table I. As shown, the time consumption of our approach is competitive to the compared methods. We give the explanation as follows. First, different from PPR, DAPC, SBPC, and SCSM, which sequentially sample a large number of candidate photos and then one-by-one evaluate their quality, the convergence of Gibbs sampling in our approach is fast. Typically, on a 1024×768-sized photo, the Gibbs sampling takes 70 to 100 iterations to converge, while the above compared methods evaluate more than 1000 candidate cropped photos. Second, we only present the cropping stage time consumption of OCBC in Table I. Practically, it takes hours for the expectation-maximize-based GMM parameter estimation in OCBC. That is, each arbitrary pairwise k-means centers correspond to a five-component GMM, while our approach needs no training time consumption.



	OP	DAPC	OCBC	GTPC	PM	PM-C	PM-T	PPR	SBPC	SCSM	Score
OP	--	24	14	12	8	20	23	17	14	21	153
DAPC	6	--	23	14	14	19	24	11	20	19	150
OCBC	16	7	--	11	10	16	14	11	9	18	112
GTPC	18	16	19	--	7	10	14	17	17	15	133
PM	22	16	20	23	--	25	27	19	21	23	196
PM-C	10	11	14	20	5	--	11	13	12	14	110
PM-T	7	6	16	16	3	19	--	7	9	19	102
PPR	13	19	19	13	11	17	23	--	22	24	161
SBPC	16	10	21	13	9	18	21	8	--	21	137
SCSM	9	11	12	15	7	16	11	6	9	--	96

Preference matrix from the cropping results in the first row

	OP	DAPC	OCBC	GTPC	PM	PM-C	PM-T	PPR	SBPC	SCSM	Score
OP	--	14	14	13	8	14	12	13	15	13	116
DAPC	16	--	12	14	3	14	19	16	13	17	124
OCBC	16	18	--	12	11	17	22	21	14	14	145
GTPC	17	16	18	--	5	10	13	18	16	20	133
PM	22	27	19	25	--	21	26	19	22	18	199
PM-C	16	16	13	20	9	--	19	18	17	17	145
PM-T	18	11	8	17	4	11	--	8	9	7	93
PPR	17	14	9	12	11	12	22	--	15	9	121
SBPC	15	17	16	14	8	13	21	15	--	10	129
SCSM	17	13	16	10	12	13	23	21	20	--	145

Preference matrix from the cropping results in the second row

	OP	DAPC	OCBC	GTPC	PM	PM-C	PM-T	PPR	SBPC	SCSM	Score
OP	--	14	7	13	5	17	18	6	19	14	113
DAPC	16	--	11	13	7	18	17	13	21	16	132
OCBC	23	19	--	16	12	18	21	23	24	17	173
GTPC	17	17	14	--	9	16	14	19	21	20	147
PM	25	23	18	21	--	28	29	25	26	19	214
PM-C	13	12	12	14	2	--	20	12	19	11	115
PM-T	12	13	9	16	1	10	--	8	4	2	75
PPR	24	17	7	11	5	18	22	--	23	20	147
SBPC	11	9	6	9	4	11	26	7	--	9	92
SCSM	16	14	13	10	11	19	28	10	21	--	142

Preference matrix from the cropping results in the third row

	OP	DAPC	OCBC	GTPC	PM	PM-C	PM-T	PPR	SBPC	SCSM	Score
OP	--	17	12	12	9	14	13	11	14	14	116
DAPC	13	--	12	11	8	11	15	12	11	10	103
OCBC	18	18	--	14	11	15	17	19	15	21	148
GTPC	18	19	16	--	11	16	14	19	18	17	148
PM	21	22	19	19	--	21	23	23	19	18	185
PM-C	16	19	15	14	9	--	21	23	16	21	154
PM-T	17	15	13	16	7	9	--	12	11	13	113
PPR	19	18	11	11	7	7	18	--	17	16	124
SBPC	16	19	15	12	11	14	19	13	--	18	137
SCSM	16	20	9	13	12	9	17	14	12	--	122

Preference matrix from the cropping results in the fourth row

Fig. 9. Cropped photos produced by the compared cropping methods and the corresponding preference matrices (OP means the original photo)

TABLE I
COMPARATIVE TIME CONSUMPTION OF THE COMPARED CROPPING METHODS

Size	PPR	DAPC	SBPC	SCSM	OCBC	GTPC	PM	PM-C	PM-T
800×600	14.321s	30.113s	45.541s	23.321s	6.624s	10.45s	11.232s	9.876s	2.124s
1024×768	30.231s	67.785s	93.445s	44.456s	9.343s	14.54s	16.548s	13.342s	3.454s
1600×1200	54.678s	125.435s	197.64s	76.562s	14.541s	20.11s	22.453s	18.883s	5.512s
1000×200	6.564s	12.243s	16.784s	8.563s	2.341s	3.97s	4.451s	3.214s	1.677s
2000×400	25.998s	46.874s	66.453s	31.557s	7.774s	13.21s	14.466s	11.229	3.212s
3000×600	101.334s	186.676s	254.113s	145.336s	13.378s	19.76s	21.334s	17.689s	5.112s

D. Step-By-Step Model Justification

This experiment justifies the effectiveness of the two main components in the proposed approach, manifold graphlet embedding and BN-based photo cropping.

To evaluate the effectiveness of the first component, cropping under two experimental settings are adopted. First, we replace the post-embedding graphlet with color SIFT [29]. We use color SIFT because it captures both local color and local texture, which functions similarly to our multi-channel graphlet. In the second column of Fig. 10, color SIFT well captures the color distribution of the original photos, such as the colorful sunset from the first and the last original photo and the blue sky from the second and the third original photo. However, color SIFT fails to encode local structural objects, such as the architectures from the first original photo and the sculpture from the second original photo. Second, we replace the proposed graphlet embedding algorithm with kernel PCA [30] and LDA [31] respectively. For a fair comparison, both the proposed embedding algorithm and kernel PCA transfer each graphlet into the same dimensional feature vector in both color/texture channels. It is worth emphasizing that, due to the large number of training graphlets (usually $N > 100,000$), it is computationally intractable to directly use kernel PCA/LDA on the $N \times N$ kernel matrix. To solve this problem, we randomly sample 5% graphlets from each training photo and use the sampled graphlets to construct the kernel matrix for kernel PCA/LDA. In the third column of Fig. 10, graphlets embedded using kernel PCA yields unsatisfactory cropping results because too many non-semantic regions are retained in the cropped photo. Specifically, only sunsets are

retained in the first and the last cropped photo, while other important regions, such as the architectures from the first original photo and the trees from the last original photo, are totally discarded. The second the third cropped photos respectively ignore the sculpture and the island, which are attractive regions expected to be preserved from their corresponding original photos. In contrast with kernel PCA, better cropping results are achieved when graphlets are embedded using kernel LDA, as shown in the fourth column of Fig. 10. Note that, cropping under kernel LDA still performs worse than that under our graphlet embedding algorithm. For the first original photo, kernel LDA is less capable for color preserving than our graphlet embedding, *i.e.*, less blue sunset region is retained in the cropped photo under kernel LDA. For the third original photo, kernel LDA keeps little details in the cropped photo, *i.e.*, the yachts regions are completely neglected.

To evaluate the effectiveness of the second component, we

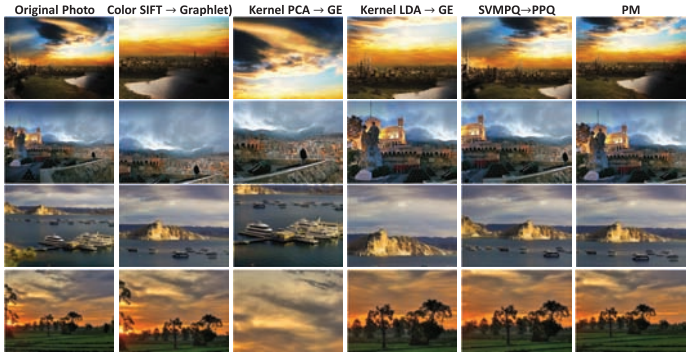


Fig. 10. Performance of replacing one component of the proposed approach with an off-the-shelf component (A→B: replace component B with component A; GE: graphlet embedding;SVMPQ: SVM-based photo quality measure; BNPQ: BN-based quality measure)

replace the BN-based photo quality measure with a kernel SVM-based one, which is based on (23). After that, the quality of each candidate cropped photo is computed based on the posterior probability SVM [35], *i.e.*,

$$p(I \rightarrow \text{highly aesthetic}|I) = \frac{1}{1+\exp(-f(x))} \qquad (25)$$

where $f(x)$ is the linear function of SVM. As seen from the fifth column of Fig. 10, for the first original photo, the visual balance among the sunset, the lake surface, and the architectures in the photo cropped using SVM-based photo quality measure is not as harmonic as that cropped using our BN-based photo quality measure. For the second original photo, the sculpture, as a distinctive architecture, is discarded by the SVM-based photo quality measure. For the third cropped photo, the leftmost island is only half retained, which is obviously suboptimal. For the last original photo, little cultivated filed is preserved in the cropped photo, which influences negatively to the global aesthetics.

E. Parameter Analysis

This experiment studies how free parameters affect the performance of the proposed approach and how to set suitable parameters to achieve a reasonable cropping result. In our approach, there are two sets of free parameters to be tuned: 1) the maximum graphlet size T, and 2) the dimensionalities of post-embedding graphlets in color and texture channels.

To analyze the effects of the maximum graphlet size T for photo cropping, we setup an experiment by varying T. In Fig. 11, we present the cropped photos when the maximum graphlet size T is tuned from 1 to 10. We do not experiment with T larger than 10 because it is computationally intractable. As shown, when T is tuned from one to two, the cropped photo is globally aesthetic but contains few structural regions. This is because 1-sized and 2-sized graphlets are not descriptive enough to capture local composition, such as the architectures in Fig 11. When T is tuned from three to four, the cropped photo includes some structural regions and remains globally aesthetic, but only the low building is not representative to the structure in the overall photo. When T reaches five, the global aesthetics are preserved in the cropped photo. That is, both the tall and low buildings are included in the cropped photo, which means local compositions are well-captured by the cropped photo. When T is larger than five, the cropped photo becomes stable, because few local compositions are captured by graphlet with size larger than five. Thus, we set T to five for this photo.

To evaluate the performance of the proposed approach

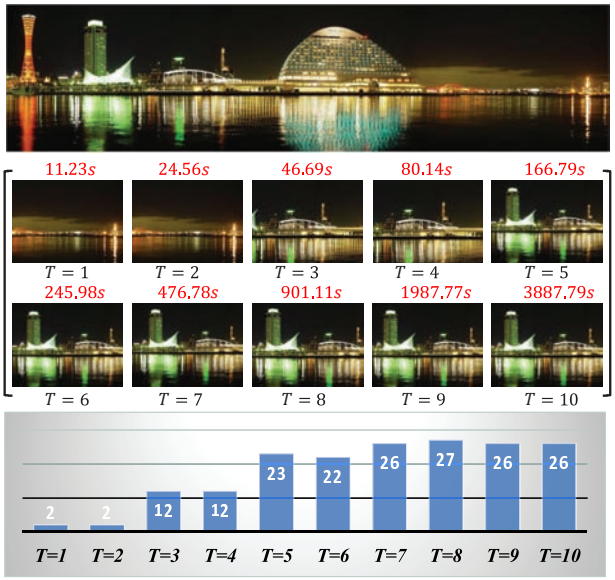


Fig. 11. Performance of the proposed approach under different value of T, the red text denote the time consumption, and the last row shows the votes of most aesthetic photos under 178 viewers.

under different dimensionalities of post-embedding graphlets in color and texture channels, we set the dimensionality in one channel as defaults while tuning that in the other channel. Denote t as the graphlet size, the default post-embedding graphlet dimensionalities in color and texture channels are set to $5t$ and $60t$ respectively. For color channel, the dimensionality of post-embedding graphlets d_{CM} is tuned from t to $9t$ with a step of t. For texture channel, the dimensionality of post-embedding graphlets d_{HOG} is tuned from $10t$ to $120t$ with a step of $10t$. As shown in Fig. 12, first, higher dimensionality of post-embedding graphlet in the color

Fig. 12. Performance under different dimensionalities of post-embedding graphlets in color channel (second row) and texture channel (third row)

channel means more intense-colored regions are preserved in the cropped photo, such as the sunset in the first row of Fig. 12. Second, higher dimensionality of post-embedding graphlet in the texture channel implies more textural regions are persevered in the cropped photo, such as the trees in the second row of Fig. 12. To balance the two visual cues, we set the dimensionality of post-embedding graphlets in color and texture channels respectively to be $8t$ and $20t$.

VI. Conclusion

Conventional photo cropping methods achieved much but are still frustrated by the following three drawbacks: 1) State-of-the-art cropping models cannot incorporate semantics effectively, 2) global spatial configurations are not explicitly captured by the existing cropping models, and 3) the importance of multi-channel visual features cannot be adjusted automatically in the cropping process. Owing to the recent progress in image retrieval community [42], [43], [44], image-level semantics can be efficiently and effectively acquired. Thus, we present weakly supervised photo cropping in this paper. First, a manifold embedding algorithm is derived to integrate image-level semantics and image global spatial configurations into graphlets. Then, a BN is developed to transfer post-embedding graphlets from the training photos into the cropped photo, where the multi-channel visual cues are automatically tuned. Based on the BN, photo cropping can be casted as maximally preserving the post-embedding graphlets from the training photos, and Gibbs sampling is used for parameter inference. Thorough empirical studies demonstrate the effectiveness of our approach in comparison with a group of state-of-the-art photo cropping methods.

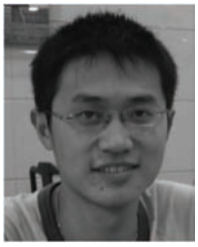
References

[1] Yiwen Luo, Xiaoou Tang, Photo and Video Quality Evaluation: Focusing on the Subject, *in Proc. of ECCV*, pages:386–399, 2008.
[2] Yan Ke, Xiaoou Tang, Feng Jing, The Design of High-Level Features for Photo Quality Assessment, *in Proc. of CVPR*, pages:419–426, 2006.
[3] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang, Studying Aesthetics in Photographic Images Using a Computational Approach, *in Proc. of ECCV*, pages:288–301, 2006.
[4] Lai-Kuan Wong, Kok-Lim Low, Saliency-Enhanced Image Aesthetics Class Prediction, *in Proc. of ICIP*, pages:997–1000, 2009.
[5] Subhabrata Bhattacharya, Rahul Sukthankar, Mubarak Shah, A Framework for Photo-Quality Assessment and Enhancement based on Visual Aesthetics, *ACM Multimedia*, pages: 271–280, 2010.
[6] Bin Cheng, Bingbing Ni, Shuicheng Yan, Qi Tian, Learning to Photograph, *ACM Multimedia*, pages: 291–300, 2010.
[7] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, Imari Sato, Sensation-based Photo Cropping, *ACM Multimedia*, pages: 669–672, 2009.
[8] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, Ming Ouhyoung, Personalized Photograph Ranking and Selection System, *ACM Multimedia*, pages: 211–220, 2010.
[9] Sagnik Dhar, Vicente Ordonez, Tamara L. Berg, High level Describable Attributes for Predicting Aesthetics and Interestingness, *in Proc. of CVPR*, pages: 1657–1664, 2011.
[10] Wei Luo, Xiaogang Wang, Xiaoou Tang, Content-Based Photo Quality Assessment, *in Proc. of ICCV*, pages: 2206–2213, 2011.
[11] Jieying She, Duo Wang, Mingli Song, Automatic Image Cropping Using Sparse Coding, *in Proc. of ACPR*, pages: 490–494, 2007.
[12] Honglak Lee, Alexis Battle, Rajat Raina, Andrew Y. Ng, Efficient Sparse Coding Algorithms, *in Proc. of NIPS*, pages:801–808, 2006.
[13] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, Michael Cohen, Gaze Based Interaction for Semi-Automatic Photo Cropping, *in Proc. of CHI*, pages:771–780, 2006.
[14] Xuejian Xiong, Kap Luk Chan, Towards An Unsupervised Optimal Fuzzy Clustering Algorithm for Image Database Organization, *in Proc. of ICPR*, pages: 897–900, 2000.
[15] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *in Proc. of CVPR*, pages:2169–2178, 2006.
[16] Markus Stricker, Markus Orengo, Similarity of Color Images, *Storage and Retrieval of Image and Video Databases*, pages: 381–392, 1995.
[17] N. Dalal, B.Triggs, Histograms of Oriented Gradients for Human Detection, *in Proc. of CVPR*, pages:886–893, 2005.
[18] Xi Zhou, Kai Yu, Tong Zhang, Thomas S. Huang, Image Classication using Super-Vector Coding of Local Image Descriptors, *in Proc. of ECCV*, pages:141–154, 2010.
[19] Jonathan Harel, Christof Koch, Pietro Perona, Graph-based visual saliency, *in Proc. of NIPS*, pages:545–552, 2007.
[20] Xinchao Wang, Zhu Li, Dacheng Tao, Subspaces Indexing Model on Grassmann Manifold for Image Search, *IEEE T-IP*,20(9), pages:2627–2635, 2011.
[21] Michael Werman , Daphna Weinshall, Similarity and Affine Invariant Distances between 2D Point Sets, *IEEE T-PAMI*,17, pages: 810–814, 1995.
[22] Shiming Xiang, Feiping Nie, Yangqiu Song, Changshui Zhang, Chunxia Zhang, Embedding new data points for manifold learning via coordinate propagation, *Knowlege and Information Systems*, 19(2), pages:159–184, 2008.
[23] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, *Prentice Hall*, 2007.
[24] W.R. Gilks , S. Richardson, David Spiegelhalter, Markov Chain Monte Carlo in Practice, *Chapman & Hall/CRC Interdisciplinary Statistics*, 1996.
[25] Aude Oliva, Antonio Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *IJCV*, 42(3), pages:145–175, 2001.
[26] Fei-Fei Li, Pietro Perona, A Bayesian Hierarchical Model for Learning Natural Scene Categories, *in Proc of CVPR*, pages: 524–531, 2005.
[27] Chris Ding, Xiaofeng He, K-means clustering via principal component analysis, *in Proc. of ICML*, pages: 225–232, 2004.
[28] Laurent Itti, Christof Koch, Ernst Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE T-PAMI*, 20(11), pages: 1254–1259, 1998.
[29] Gertjan J. Burghouts,Jan-Mark Geusebroek, Performance evaluation of local colour invariants, *CVIU*, 113, pages: 48–62, 2009.

[30] Tianhao Zhang, Dacheng Tao, Xuelong Li, Jie Yang, Patch Alignment for Dimensionality Reduction, *IEEE T-KDE*, 21(9), pages: 1299–1313, 2009.
[31] Dacheng Tao, Xuelong Li, Xindong Wu, Stephen J. Maybank, Geometric Mean for Subspace Selection, *IEEE T-PAMI*, 31(2), pages: 260–274, 2009.
[32] Frederic Jurie, Cordelia Schmid, Scale-Invariant Shape Features for Recognition of Object Categories, *in Proc. of ICCV*, pages: 90–96, 2004.
[33] Zaid Harchaoui, Francis Bach, Image Classification with Segmentation Graph Kernels, *in Proc. of CVPR*, pages:1–8, 2007.
[34] Francis R. Bach, Romain Thibaux, Michael I. Jordan, Computing Regularization Paths for Learning Multiple Kernels, *in Proc. of NIPS*, pages:1–8, 2005.
[35] John C. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages: 61–74, 1999, MIT Press.
[36] Mingli Song, Dacheng Tao, Chun Chen, Xuelong Li, Chang Wen Chen, Color to Gray: Visual Cue Preservation, *IEEE T-PAMI*, 32(9), pages: 1537–1552, 2010.
[37] Song Mingli, Dacheng Tao, Chun Chen, Jiajun Bu, Jiebo Luo, Chengqi Zhang, Probabilistic Exposure Fusion, *IEEE T-IP*, 21(1), pages: 341–357, 2011.
[38] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, Chun Chen, Probabilistic Graphlet Transfer for Photo Cropping, *IEEE T-IP*, DOI: 10.1109/TIP.2012.2223226, 2012.
[39] Benjamin Yao, Xiong Yang, SongChun Zhu, Introduction to a Large Scale General Purpose Ground Truth Dataset: Methodology, Annotation Tool, and Benchmarks, *in Proc. of EMMCVPR*, pages: 169-183, 2007.
[40] Yi Yang, Yueting Zhuang, Fei Wu, Yunhe Pan, Harmonizing Hierarchical Manifolds for Multimedia Document Semantics Understanding and Cross-Media Retrieval, *IEEE T-MM*, 10(3), pages: 437–446, 2008.
[41] Zhigang Ma, Feiping Nie, Yi Yang, Jasper Uijlings, Nicu Sebe, Alexander G. Hauptmann, Discriminating Joint Feature Analysis for Multimedia Data Understanding, *IEEE T-MM*, 14(6), pages: 1662–1672, 2012.
[42] Yangxi Li, Bo Geng, Dacheng Tao,Zheng-Jun Zha, Linjun Yang, Chao Xu , Difficulty Guided Image Retrieval Using Linear Multiple Feature Embedding, *IEEE T-MM*, 14(6), pages: 1618–1630, 2012.
[43] Yin-Hsi Kuo, Wen-Huang Cheng, Hsuan-Tien Lin, Winston H. Hsu, Unsupervised Semantic Feature Discovery for Image Object Retrieval and Tag Refinement, *IEEE T-MM*, 14(4), pages: 1079–1090 , 2012.
[44] Lin Chen, Dong Xu, Ivor W. Tsang, Jiebo Luo, Tag-Based Image Retrieval Improved by Augmented Features and Group-Based Refinement, *IEEE T-MM*, 14(4), pages: 1057–1067, 2012.

Luming Zhang received his Ph.D. degree in computer science from Zhejiang University, China. His research interests include visual perception analysis, image enhancement, and pattern recognition.

Mingli Song received the PhD degree in computer science from Zhejiang University, China, in 2006. He is currently an Associate Professor in the College of Computer Science, Zhejiang University. His research interests include visual surveillance, visual perception analysis, image enhancement, and face modeling. He has authored and co-authored more than 60 scientific articles at top venues including IEEE T-PAMI, T-IP, T-MM, PR, CVPR, ECCV, ACM MM, He is a senior member of the IEEE.



Yi Yang received the Ph.D degree in Computer Science from Zhejiang University, Hangzhou, China, in 2010. He is now a DECRA fellow at the University of Queensland. Prior to that, he was a Postdoctoral Research Fellow at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g. multimedia indexing and retrieval, image annotation, video semantics understanding, etc.

Qi Zhao received the PhD degree in computer science from University of California, Santa Cruz. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore. She mainly applies human perceptive mechanism in computer vision and image processing.

Chen Zhao received the B.S. degree in software engineering from Sichuan University, Cheng Du, China, in 2010. She is currently pursuing the Ph.D. degree at School of Electrical Engineering and Computer Science, Peking University, Beijing, China. Also she is a visiting student at University of Washington, Seattle, USA. Her research interests include image/video processing, video coding and video transmission.

Nicu Sebe received the Ph.D. in computer science from Leiden University, Leiden, The Netherlands, in 2001.

Currently, he is with the Department of Information Engineering and Computer Science, University of Trento, Italy,where he is leading the research in the areas of multimedia information retrieval and humancomputer interaction in computer vision applications. He was involved in the organization of themajor conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008, ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, and WIAMIS 2009 and as one of the initiators and a Program Co-Chair of theHuman-Centered Multimedia track of the ACM Multimedia 2007 conference. He is the general chair of ACM Multimedia 2013 and was a program chair of ACM Multimedia 2011. He is a senior member of IEEE and of ACM.