Flickr Circles: Aesthetic Tendency Discovery by Multi-View Regularized Topic Modeling

Richang Hong, *Member, IEEE*, Luming Zhang, *Member, IEEE*, Chao Zhang, and Roger Zimmermann, *Senior Member, IEEE*

***Abstract*—Aesthetic tendency discovery is a useful and interesting application in social media. In this paper we propose to categorize large-scale Flickr users into multiple circles, each containing users with similar aesthetic interests (e.g., landscapes). We notice that: 1) an aesthetic model should be flexible as different visual features may be used to describe different image sets; 2) the numbers of photos from different Flickr users vary significantly and some users may have very few photos; and 3) visual features from each Flickr photo should be seamlessly integrated at both low-level and high-level. To meet these challenges, we propose to fuze color, textural, and semantic channel features using a multi-view learning framework, where the feature weights are adjusted automatically. Then, a regularized topic model is developed to quantify each user's aesthetic interest as a distribution in the latent space. Afterward, a graph is constructed to describe the discrepancy of aesthetic interests among users. Apparently, densely connected users are with similar aesthetic interests. Thus, an efficient dense subgraph mining algorithm is adopted to group Flickr users into multiple circles. Experiments have shown that our approach performs competitively on a million-scale image set crawled from Flickr. Besides, our method can enhance the transferal-based photo cropping [40] as reported by the user study.**

***Index Terms*—Aesthetic, Flickr circle, multi-view, tendency.**

I. Introduction

FLICKR is a well-known photo and video hosting website. There are many groups in Flickr where users with the same aesthetic interest can share photos and exchange opinions. There are a variety of popular public groups (e.g., "architecture" and "silhouette") with thousands of members and over one million shared photos. Users can freely join in/leave a public group or launch a new group. However, the grouping mechanism is less intelligent since the groups are constructed and maintained manually. In practice, we want a system that automatically categorizes Flickr users into multiple communities based on users' aesthetic tendency. Generally, an effective photo enhancement framework should be built upon the experiences of professional photographers. For example, to produce an attractive landscape picture, it is necessary to leverage the knowledge of those landscape photographers whom are experienced in optimizing the relative positions between foreground and background sceneries. In practice, however, it is difficult to find photographers with each particular aesthetic tendency (e.g., landscape or portrait). However, establishing such a system is challenging due to the following reasons.

1) In many computational aesthetic models [1], [2], different visual features are employed to represent different image sets. For example, if an image set contains portraits, then the active shape model [3] can be used to localize human faces. This requires the designed system to be highly extensible. Thereby different visual features can be encoded flexibly.
2) The number of photos belonging to different users varies dramatically. Some users have uploaded over 50 000 photos while others may have only less than 10 photos. This will brings severe overfitting problem in the model training stage.
3) It is generally accepted that visual features at low-level and high-level describe photo content collaboratively. However, effectively describing the semantics of each Flickr photo is a difficult task. In addition, many models fuze multiple visual features in a linear/nonlinear way, where the cross-feature information might not be optimally utilized.

To solve these problems, a regularized Gaussian mixture model (GMM) is proposed to predict Flickr users' aesthetic tendency, by integrating multiple visual features intelligently. We first use color and texture to describe each photo at low-level, and the semantics channel is described by engineering image tags in the latent space (Component 1). The three channel visual features are then fuzed through a multiview learning framework (Component 2). Next, a manifold-based GMM is designed to seamlessly integrate these low&high-level features and further represent user's aesthetic interest by a distribution of latent topics (Component 3). To alleviate the overfitting caused by the scarcity of photos of some users, a regularized term is added into the GMM. Using KL-divergence to measure the distribution between users, an affinity graph is constructed to describe the similarity of aesthetic interests among users. Users

Manuscript received October 10, 2015; revised January 20, 2016 and March 14, 2016; accepted April 05, 2016. Date of publication May 11, 2016; date of current version July 15, 2016. This work was supported in part by the National Basic Research Program under Grant 2013CB336500, in part by the National High-Tech Development Program under Grant 2014AA015104, in part by the Fundamental Research Funds for the Central Universities, in part by the Program for New Century Excellent Talents under Grant NCET-13-0764, in part by the Project from National Nature Science Foundation of China under Grant 61472116 and Grant 61572169, in part by the Anhui Fund for Distinguished Young Scholars under Grant 1508085J04, and in part by the National University of Singapore (Suzhou) Research Institute, Suzhou Industrial Park, Jiang Su, China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chengcui Zhang.

R. Hong and L. Zhang are with the Department of CSIE, Hefei University of Technology, Hefei 230009, China (e-mail: hongrc.hfut@gmail.com; zglumg@gmail.com).

C. Zhang is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA (e-mail: chaozhangzju@gmail.com).

R. Zimmermann is with the School of Computing, National University of Singapore, Singapore 117417 (e-mail: rogerz@comp.nus.edu.sg).

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

Digital Object Identifier 10.1109/TMM.2016.2567071

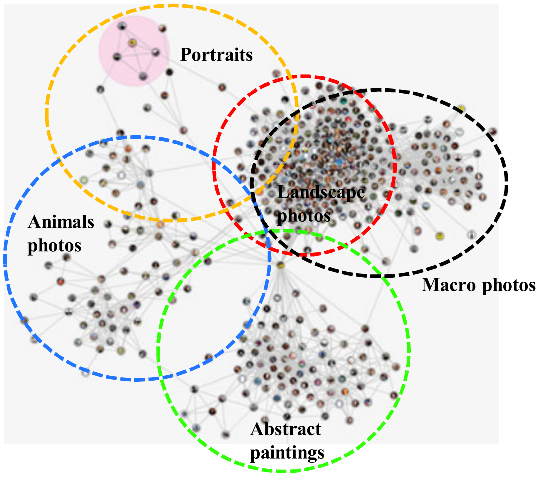


Fig. 1. Generating Flickr circles (differently colored) from a number of users (the red points denote Flickr users and each edge links users with similar aesthetic interests). Macro photography is extreme close-up photography, usually of very small subjects, in which the size of the subject in the photography is greater than the life size.

with similar aesthetic interests are densely distributed on the graph. They are finally categorized into different Flickr circles (as exemplified in Fig. 1) based on a dense subgraphs mining algorithm (Component 4).

Part of this work has been accepted by a preliminary conference ICME 2015 [31]. Our work is a fundamental improvement from its conference version.

1) We propose a multiview learning algorithm to optimally integrate visual features from color, textural and semantic channels, where the semantic channel is captured by engineering image tags in the latent space.
2) A manifold-guided regularizer is incorporated onto GMM to avoid overfitting. It assumes that Flickr photos reside on a sub-manifold of the ambient space. Comparatively, the regularizer used in the conference paper assumes that Flickr photos are distributed in the original feature space. Quantitative results confirm the clear advantage of the manifold data modeling.
3) A more detailed experimental analysis on a much larger Flickr image set is presented to evaluate our method. Enhancement of photo retargeting by leveraging the discovered Flickr circles is validated by the user study.

The reminder of this paper is organized as follows. In Section II we briefly review the related works. Sections III, IV and V introduce the proposed method, including low&high-level visual features extraction, manifold-guided regularized GMM for aesthetic interest modeling and dense subgraphs mining for Flickr circles discovery. Comprehensive experiments in Section VI demonstrate the effectiveness of the proposed method. Section VII concludes.

II. Related Work

A. Computational Aesthetic Models

There are a rich number of computational aesthetic models in multimedia [25], [37], [38] and computer vision [9], [26], [27]. Datta *et al.* [4] proposed 58 low-level visual features to capture photo aesthetics. Wong *et al.* [5] proposed three global aesthetic features: 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. Luo *et al.* [1] proposed a hue distribution and a prominent line-based texture distribution to represent the photo global composition. Dhar *et al.* [6] proposed a set of high-level attributes to evaluate photo aesthetics. To capture the process of human viewing images, Zhang *et al.* [23] proposed to learn human gaze shifting pathes for evaluating photo aesthetics. Cheng *et al.* [7] proposed omni-range context, i.e., spatial distributions of arbitrary pairwise image patches, to model photo aesthetics. Zhang *et al.* [8] introduced graphlets and designed a probabilistic model to transfer them from the training photos into the cropped one. Further, Zhang *et al.* [2] proposed to optimally fuze visual features from multiple channels to access photo aesthetics. It is worth emphasizing that these methods can only access the aesthetics of a single image. They cannot measure the aesthetic discrepancy of image sets belonging to different Flickr users.

B. Communities Detection Using a Probabilistic Model

Probabilistic topic models such as latent Dirichlet allocation (LDA) [10] and its variants [11], [12] have been applied to detect social communities in the past decade. Based on a probabilistic topic model, a social link graph can be considered as a generative process. The model categorizes users into different communities though a sampling process, given the distribution of communities in the graph, the distribution of users in communities, and the distribution of social links among users. Communities are detected based on the fact that users belonging to one community have similar link patterns in the graph. Some work [13], [14] applies probabilistic models to detect communities where each community is considered as a combination of semantic topics. Zhou *et al.* [13] proposed a model that extracts e-communities from email corpus. The model employs social interactions and topical similarity extracted from documents to search communities. A recent work by Yin *et al.* [14] constructs text-associated graphs. The model combines the generation of links between users and words of users to extract communities, where both the link structures and the users' semantic topics are exploited. In [41], Hong *et al.* proposed a unified photo enhancement framework by effectively mining aesthetics-related communities on Flickr. Noticeably, our proposed method is significantly distinguishable from Hong *et al.* [41]'s method in three aspects. First, Hong *et al.* designed a tag-regularized LDA to describe the distribution of photos belonging to each Flickr user, where the potentially missing image tags can be dynamically integrated for aesthetic modeling. In contrast, our method optimally fuzes multi-channel visual features from each Flickr image, and then develops manifold-regularized GMM to describe the distribution of Flickr photos. The theories of the tag-regularized LDA and manifold-regularized GMM are fundamentally different. Second, Hong *et al.*'s method linearly combines multi-channel visual features from each Flickr photo, where features from different channels contribute equally. Our approach upgrade this scheme by adopting a multi-view learning algorithm, where the feature weight can be automatically learned and the cross-feature correlations can be well exploited. Third, Hong *et al.*'s method directly uses the augmented frequency vector to represent the semantics of each Flickr photo. As an improvement, our approach mines the latent semantic topics from the high-dimensional augmented frequency vector, resulting in a succinct and representative set of semantic features.

It is noticeable that the above probabilistic models describe the distribution of each entity using uni-modal features only. In practice, multiple types of visual features should be exploited to describe Flickr photos, potentially requiring an optimal feature fusion scheme in the probabilistic model. Besides, the number of samples belonging to different entities (e.g., Flickr users) may vary significantly. To alleviate the sample imbalance problem, we have to integrate a regularizer onto the probabilistic model.

III. Multi-View Visual Descriptors of Each Flickr Photo

A. Low-Level Visual Descriptors

Color Feature: We use color moment [21] to describe the color distribution of each photo. It uses three central moments of a region's color distribution: mean, standard deviation and skewness, to characterize each photo. For the ith RGB channel, the three moments are

$$\mu_i = \frac{1}{Z}\sum_{j=1}^{Z} I_c(i,j) \tag{1}$$

$$\sigma_i = \left[\frac{1}{Z}\sum_{j=1}^{Z}(I_c(i,j)-\mu_i)^2\right]^{1/2} \tag{2}$$

$$s_i = \left[\frac{1}{Z}\sum_{j=1}^{Z}(I_c(i,j)-\mu_i)^3\right]^{1/3} \tag{3}$$

where μ_i, σ_i and s_i denote the mean, standard deviation and skewness of the ith RGB channel respectively; $I_c(i,j)$ is the value of ith RGB channel at the jth pixel; Z denotes the number of pixels of photo I. Based on this, the color moment of each photo is represented by concatenating the three moments from all the RGB channels into a 9-dimensional feature vector.

Textural feature: We use the well-known histogram of gradient (HOG) [17] to model the texture of each photo, which is robust to the variance of local geometry, rotations and photometric transformations.

Let $I_o(x,y,i)$ and $I_m(x,y,i)$ respectively be the orientation and magnitude of the intensity gradient at pixel (x,y) of photo I in the ith RGB channel, we compute gradients using a finite difference filter $[-1,0,+1]$ and its transpose. Denote l as the RGB channel with the largest gradient magnitude, i.e., $l = \arg\max_i I_m(x,y,i)$, we define $I_o(x,y) = I_o(x,y,l)$ and $I_r(x,y) = I_m(x,y,l)$. Then, the gradient orientation at each pixel is discretized into one of the eight values as

$$I_o^d(x,y) = \text{round}\left(\frac{8\cdot I_o(x,y)}{2\pi}\right)\text{mod}8. \tag{4}$$

We define a pixel-level feature map that specifies a sparse HOG magnitudes at each pixel. Let $k \in \{0,1,\ldots,7\}$ range over the orientation bins. The pixel-level feature map for photo I at (x,y) is

$$I_{o,r}(x,y) = \begin{cases} I_r(x,y), & \text{if } k = I_o^d(x,y) \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Next, we divide each photo into 4×4 sub-regions. For the uvth sub-region, we accumulate a local histogram of eight gradient directions over the pixels into a feature vector $h(u,v)$. To achieve the invariance of illumination and shadowing of each sub-region, we normalize feature vector $h(u,v)$ as: $\bar{h}(u,v) = \frac{h(u,v)}{||h(u,v)||^2}$. Afterward, we concatenate the normalized feature $\bar{h}(u,v)$ from all the 4×4 sub-regions into a $8\cdot 4\cdot 4 = 128$-dimensional feature vector in order to characterize the texture of each photo.

B. High-Level Visual Descriptor

Given a Flickr photo, we use an M-dimensional augmented frequency vector $\overrightarrow{\alpha}$ to represent the distribution of its tags. In particular, to avoid the randomly-occurring noisy image tags, we select the M most frequent tags from the training image set. Then, we treat the tag set of each Flickr photo as a document $\mathcal{D}$, based on which the ith element of vector $\overrightarrow{\alpha}$ is calculated as

$$\overrightarrow{\alpha}(i) = 0.5 + \frac{0.5 * \text{freq}(i,\mathcal{D})}{\max(\text{freq}(j,\mathcal{D}) : j \in \mathcal{D})} \tag{6}$$

where $freq(i,\mathcal{D})$ counts the times of the ith tag from the M most frequent ones occurring in document $\mathcal{D}$, and the denominator functions as a normalization factor. In our implementation, we tune M from 10 to 200 with a step of 10, finally we observe that the best performance is achieved when $M = 100$.

Latent semantic topics learning: The above tag-based frequency vector may be redundant since it fails to detect and simplify multiple tags reflecting the same semantics, such as "snow" "ice" and "white". Toward a succinct representation of photo semantics, we post-process the tag-based vector in the latent topic space.

Given N Flickr photos, they can be represented by N tag-based frequency vectors. Afterward, we column-wise stack them into a matrix $\mathbf{X} = [\overrightarrow{\alpha}_1; \overrightarrow{\alpha}_2; \ldots; \overrightarrow{\alpha}_N] \in \mathbb{R}^{N\times M}$, where each column $\vec{X}_j \in \mathbb{R}^N$ denotes the jth feature vector of all the documents. To obtain the semantic feature of each photo, we adopt a subspace algorithm, latent factor analysis (LFA) [32], which projects the original M-dimensional vector corresponding to each photo into a D-dimensional semantic feature vector (We set $B = 17$ based on cross validation. Particularly, we tune B from 1 to 50 with a step of 1. It is observable that the best performance is achieved when $B = 17$).

LFA assumes that the B bases of the columns of matrix $\mathbf{X}$, i.e., $\{\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_B\}$, are uncorrelated. Each $\vec{u}_b \in \mathbb{R}^N$ has a unit length. Denote $\mathbf{U} = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_B] \in \mathbb{R}^{N\times B}$, we have

$\mathbf{U}^T\mathbf{U} = \mathbf{I}_N$ where $\mathbf{I}_N$ is the identity matrix. It is reasonable to assume that each feature vector $\vec{X}_j$ can be linearly reconstructed by the B bases

$$\mathbf{X}_j = \sum_{b=1}^{B} a_{\mathrm{bj}}\mathbf{u}_b + \epsilon_j. \tag{7}$$

In the matrix form, the above equation can be reorganized into $\mathbf{X} = \mathbf{U}\mathbf{A} + \epsilon$, where $\mathbf{A} = [a_{dj}] \in \mathbb{R}^{B\times M}$ denotes the projection matrix from the tag space to the semantic feature space. We can obtain the projection matrix $\mathbf{A}$ by solving the following optimization which minimizes the rank-B approximation error subject to the orthogonality constraint of matrix $\mathbf{U}$

$$\min\nolimits_{\mathbf{U},\mathbf{A}} ||\mathbf{X} - \mathbf{U}\mathbf{A}||_F^2, \text{ s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}_N \tag{8}$$

where $||\cdot||_F$ denotes the matrix Frobenius norm. The objective function can be solved analytically [32].

The time consumption of the three channel visual features extraction is as follows. The extraction of color moment is linearly increasing with the number of pixels within a Flickr photo, which means that this process is efficient. The HOG-based texture extraction consumes about 0.2 seconds to handle one photo. The tag channel vector calculation is linearly increasing with the number of tags of each photo, which is very fast. In total, it takes about 0.4 seconds to extract the three channel visual features from each photo.

C. Optimal Multiview Feature Fusion

Single channel feature optimization: Our adopted multiview feature fusion is actually a manifold embedding algorithm. The first part of embedding incorporates the distribution of samples (i.e., Flickr photos) from each view. It minimizes the discrepancy of distances in the original feature space and those in the embedded space. For any color, texture and semantic channel, the objective function is

$$\arg\min\nolimits_{\mathbf{Z}} \sum_{\mathrm{ij}}[d_E(z_i, z_j) - d_E(y_i, y_j)]^2 \tag{9}$$

where $\mathbf{Z} = [z_1, z_2, \ldots, z_N]$ denotes the Flickr photos in the embedded space; y_i and y_j are photos from a particular channel in the original feature space.

The above objective function can be reorganized as

$$\arg\min\nolimits_{\mathbf{Z}} \sum_{\mathrm{ij}}[d_E(z_i, z_j) - d_E(y_i, y_j)]^2 = \arg\max\nolimits_{\mathbf{Z}} \ \mathrm{tr}(\mathbf{Z}\mathbf{K}\mathbf{Z}^T) \tag{10}$$

where $\mathbf{K} = -\mathbf{R}_N\mathbf{S}\mathbf{R}_N/2$, $(\mathbf{S})_{\mathrm{ij}} = (\mathbf{D})_{\mathrm{ij}}^2$, $\mathbf{D}$ is an $N \times N$ matrix whose entries are the Euclidean distance between Flickr photos, $\mathbf{R} = \mathbf{I}_N - \vec{e}_N\mathbf{S}\vec{e}_N^T/N$ is the centralization matrix, $\mathbf{I}_N$ is a $N \times N$ identity matrix and $\vec{e}_N = [1, \ldots, 1]^T \in \mathbb{R}^N$; and $d_E(z_i, z_j)$ denotes the Euclidean distance between vector z_i and z_j.

Multiple channel feature fusion: To jointly describe each Flickr photo from the perspective of color, texture and semantics, nonnegative weights $\alpha = \{\alpha_1, \alpha_2, \alpha_3\}$ are imposed on the part optimization. Accordingly, the objective function of the multiview embedding is given as

$$\arg\max\nolimits_{\mathbf{Z},\alpha} \sum_{i=1}^{3} \alpha_i \mathrm{tr}(\mathbf{Z}\mathbf{K}_i\mathbf{Z}^T). \tag{11}$$

The solution of α is $\alpha_i = 1$ corresponding to the minimum $\mathrm{tr}(\mathbf{Z}_i\mathbf{K}_i\mathbf{Z}_i^T)$ over different views. This solution means that only one view is selected. Therefore, its performance is equivalent to the one from the best view. Obviously, this solution is not satisfactory since it fails to exploit the complementary property of multiple views to achieve an optimal embedding.

To avoid this phenomenon, we adopt the trick in [33]. We set $\alpha_i \leftarrow \alpha_i^r$ with $r > 1$. Therefore, $\sum_{i=1}^{3} \alpha_i^r = 1$ achieves its minimum when $\alpha_i = 1/3$ under the condition that $\sum_{i=1}^{3} \alpha_i = 1$, $\alpha_i > 0$. More specifically, the objective function can be written as

$$\arg\max\nolimits_{\mathbf{Z},\alpha} \sum_{i=1}^{3} \alpha_i^r \mathrm{tr}(\mathbf{Z}\mathbf{K}_i\mathbf{Z}^T),$$
$$\text{s.t. } \mathbf{Z}\mathbf{Z}^T = \mathbf{I}_N, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0. \tag{12}$$

To solve (12), an iterative algorithm is derived by optimizing $\mathbf{Z}$ and α alternatively. First, we fix $\mathbf{Z}$ and update α. By using a Lagrange multiplier λ to take the constraint into account, we obtain the Lagrange function

$$l(\alpha, \lambda) = \sum_{i=1}^{3} \alpha_i^r \mathrm{tr}(\mathbf{Z}\mathbf{K}_i\mathbf{Z}^T) - \lambda\left(\sum_{i=1}^{3} \alpha_i - 1\right). \tag{13}$$

By setting the derivative of $l(\alpha, \lambda)$ with respect to α_i and λ to zero, we obtain

$$\alpha_i = \frac{(1/\mathrm{tr}(\mathbf{Z}\mathbf{K}_i\mathbf{Z}^T))^{1/(r-1)}}{\sum_i (1/\mathrm{tr}(\mathbf{Z}\mathbf{K}_i\mathbf{Z}^T))^{1/(r-1)}}. \tag{14}$$

Next, we fix α and update $\mathbf{Z}$. The optimization in (12) then becomes

$$\arg\max\nolimits_{\mathbf{Z}} \ \mathrm{tr}(\mathbf{Z}\mathbf{K}\mathbf{Z}^T), \ \text{s.t. } \mathbf{Z}\mathbf{Z}^T = \mathbf{I}_N \tag{15}$$

where $\mathbf{K} = \sum_i \alpha_i\mathbf{K}_i$. (15) is a quadratic problem with a quadratic constraint that can be solved using an eigenvalue decomposition, whose time complexity is $\mathcal{O}(N^3)$.

IV. Manifold-Guided Regularized GMM

After representing each Flickr photo by fuzing color, textural and semantic features, we describe the aesthetic interest of each Flickr user. We notice that: 1) the numbers of photos belonging to different Flickr users are highly imbalanced. Some users may have thousands of photos while others may have fewer than 10. This brings severe overfitting problem in model learning; 2) both statistical and empirical studies [34], [35] shown that, modeling the distribution of multimedia data on manifold is usually a better choice than that in the Euclidean space.

A. Standard GMM

We incorporate a manifold-guided regularizer onto GMM to model the aesthetic interest of each Flickr user. First, we

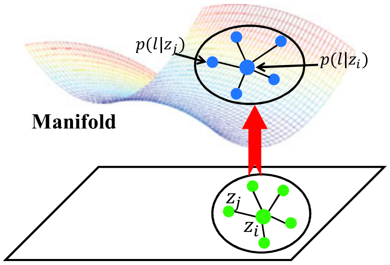


Fig. 2. Illustration of the locality preservation property of the condition probabilities.

introduce the standard GMM, which is defined as

$$p(z|\Theta) = \sum_{i=1}^{K} \beta_i p_i(z|\theta_i) \tag{16}$$

where the parameters are $\Theta = (\beta_1, \ldots, \beta_K, \theta_1, \ldots, \theta_K)$; $\sum_{i=1}^{K} \beta_i = 1$ and each p_i is a Gaussian density function parameterized by θ_i.

To calculate the parameters Θ, typically the maximum likelihood (ML) estimation is adopted, i.e.

$$\begin{aligned}\eta(\Theta) =& \log p(z|\Theta) = \log \prod\nolimits_{i=1}^{N} p(z_i|\Theta) \\ =& \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \beta_i p_j(z_i|\theta_j) \right). \end{aligned} \tag{17}$$

The above objective function is difficult to optimize since it contains the log of sum. To simplify the likelihood expression, let $l_i \in \{1, 2, \ldots, K\}$ denotes which Gaussian component x_i is from and $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$. Suppose we know the value of l_i, the likelihood then becomes

$$\eta(\Theta) = \log \prod_{i=1}^{N} p(\mathcal{Z}, \mathcal{L}|\Theta) = \sum_{i=1}^{N} \log(\beta_{l_i} p_{l_i}(z_i|\theta_{l_i})) \tag{18}$$

where $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$. This objective function can be optimized using a variety of techniques such as the EM [36].

B. Manifold-Guided GMM Regularizer

Based on the locality of points on manifold [34], if pairwise samples z_i and z_j are close in the intrinsic geometry, then their conditional probabilities $p(l|z_i)$ and $p(l|z_j)$ should be similar, as the example shown in Fig. 2. That is to say, the conditional probability $p(l|\cdot)$ varies smoothly along the geodesics in the intrinsic geometry. Given K components in a GMM and denote $f_k(z) = p(l|z)$ as the conditional probability distribution function, we use $||f_k||_{\mathcal{M}}$ to measure the smoothness of f_k, i.e.

$$||f_k||_{\mathcal{M}} = \int_{z \in \mathcal{M}} -div\nabla(f) f_k dp_{\mathcal{Z}}(z) \tag{19}$$

where $-div\nabla(f)$ is the Laplace–Beltrami operator on manifold $\mathcal{M}$. By minimizing $||f_k||_{\mathcal{M}}$, we can obtain a sufficiently smooth probability distribution function.

However, in practice, the data manifold is usually unknown. To calculate $||f_k||_{\mathcal{M}}$, it is necessary to model the geometrical structure of manifold $\mathcal{M}$. Particularly, we construct a nearest neighbor graph. For each data z_i, we find its H nearest neighbors and put an edge between z_i and each neighbor. Based on this, a weight matrix $\mathbf{S}$ is obtained, where each element is defined as

$$\mathbf{S}_{\text{ij}} = \begin{cases} \exp\left(-\frac{||z_i - z_j||^2}{\sigma^2}\right), & \text{if } z_i \text{ and } z_j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases}. \tag{20}$$

Define matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix whose entries are column-wise sum of matrix $\mathbf{S}$. We call matrix $\mathbf{L}$ the graph Laplacian, based on which $||f_k||_{\mathcal{M}}^2$ can be discretely approximated by

$$\begin{aligned} r_k =& \frac{1}{2} \sum_{\text{ij}} (p(l|z_i) - p(l|z_j))^2 \mathbf{S}_{\text{ij}} \\ =& \mathbf{f}_k^T \mathbf{D} \mathbf{f}_k - \mathbf{f}_k^T \mathbf{S} \mathbf{f}_k = \mathbf{f}_k^T \mathbf{L} \mathbf{f}_k \end{aligned} \tag{21}$$

where $\mathbf{f}_k = (f_k(z_1), f_k(z_2), \ldots, f_k(z_N))$.

Based on the above manifold-guided regularizer and the conventional GMM as defined in (18), we define the objective function of the regularized GMM as

$$l(\Theta) = \sum_{i=1}^{N} \log(\beta_{l_i} p_{l_i}(z_i|\theta_{l_i})) - \lambda \sum_{i=1}^{K} r_i. \tag{22}$$

C. EM-Based GMM Parameters Inference

Due to the complicated form, it is difficult to calculate the parameters of (22) analytically. Instead, an EM algorithm is applied. Generally, the EM iteration alternates between 1) an expectation (E) step creating a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and 2) a maximization (M) step computing parameters that maximize the expected log-likelihood calculated in the E step. These parameter estimates are then used to determine the distribution of the latent variables in the next E step. More specifically:

E-Step: For nth iteration, the EM algorithm evaluates the posterior probabilities using the parameters $\beta_i^{(n-1)}$ and $\theta_i^{(n-1)}$ in the previous iteration

$$p(l|z_i, \Theta^{(n-1)}) = \frac{\beta_i^{(n-1)} p_i(z_i|\theta_i^{(n-1)})}{\sum_{j=1}^{K} \beta_j^{(n-1)} p_j\left(z_j|\theta_j^{(n-1)}\right)}. \tag{23}$$

Then, we try to find the expected value of the log-likelihood $p(\mathcal{Z}, \mathcal{L}|\Theta)$ with respect to the unknown data $\mathcal{L}$ given the observed data $\mathcal{Z}$. We define

$$Q(\Theta, \Theta^{(n-1)}) = E\left[\log p(\mathcal{Z}, \mathcal{L}|\Theta) | \mathcal{Z}, \Theta^{(n-1)}\right] \tag{24}$$

where $Q(\Theta, \Theta^{(n-1)})$ is the function defined by ourselves.

M-Step: We solve the maximization of (24), which has no analytic solution. By taking the first and second derivatives of $\mathbf{f}_k^T \mathbf{L} \mathbf{f}_k$ with respect to $p(l|z_i)$, we obtain the following updating scheme of $p(l|z_i)$:

$$p(l|z_i) \leftarrow (1 - \gamma) p(l|z_i) + \gamma \frac{\sum_{j=1}^{N} \mathbf{S}_{\text{ij}} p(l|z_j)}{\sum_{j=1}^{N} \mathbf{S}_{\text{ij}}} \tag{25}$$

where $\gamma \in [0, 1]$ is the step parameter. γ is fixed to 0.8 by cross validation. Noticeably, the weight matrix $\mathbf{S}$ is highly sparse, which makes the updating process efficient.

After the conditional probability function is smoothed, we can maximize the log-likelihood function $Q(\Theta, \Theta^{(n-1)})$ of the standard GMM. Thereby, the following updating scheme for Gaussian parameters is obtained:

$$\beta_i = \frac{1}{m} \sum\nolimits_{j=1}^{N} p(i|z_j) \tag{26}$$

$$\mu_i = \frac{\sum_{j=1}^{N} z_j p(i|z_j)}{\sum_{j=1}^{N} p(i|z_j)} \tag{27}$$

$$\Sigma_i = \frac{\sum_{j=1}^{N} p(i|z_j)(z_j - \mu_i)(z_j - \mu_i)^T}{\sum_{j=1}^{N} p(i|z_j)}. \tag{28}$$

The E-step and M-step are carried out iteratively until convergence. It is worth emphasizing that the initialization of parameter $\Theta^{(0)}$ is important. In our implementation, K-means is adopted to initialize $\Theta^{(0)}$. Empirical results shown that much faster convergence is achieved compared with the random initialization of $\Theta^{(0)}$.

The EM-based GMM parameter inference is carried out iteratively. Generally, it takes about $80 \sim 120$ iterations before convergence and the time consumption is about $70 \sim 110$ seconds.

V. Flickr Circles Discovery Based on Dense Subgraphs Mining

1) Affinity Graph Construction: To construct an affinity graph describing the aesthetic similarity between Flickr users, a similarity measure is required. Based on the regularized GMM, user's aesthetic interest is represented by a mixture of Gaussian distribution. To measure the similarity between distributions, KL-divergence [16] $d_{\mathrm{KL}}(\mathcal{N}||\mathcal{N}')$ is adopted. $\mathcal{N}$ and $\mathcal{N}'$ denote the learned aesthetic distribution of two users respectively.

Due to the non-symmetry of KL-divergence, it is difficult to integrate it into a semi-definite matrix for grouping task (e.g., spectral clustering [18]). Instead, we use the square root of Jensen–Shannon divergence [16], a metric derived form KL-divergence

$$d_{\mathrm{JS}}^{1/2}(\mathcal{N}||\mathcal{N}') = \sqrt{\frac{1}{2}\left(d_{\mathrm{KL}}(\mathcal{N}||\mathcal{N}') + d_{\mathrm{KL}}(\mathcal{N}'||\mathcal{N})\right)}. \tag{29}$$

The above metric reflects the aesthetic similarity between Flickr users. It is integrated into a dense graph mining framework for detecting users with similar aesthetic interests. First, we construct an affinity matrix $\mathbf{W}$ where the ijth element is calculated as

$$\mathbf{W}_{\mathrm{ij}} = \exp\left(-\frac{d_{\mathrm{JS}}(\mathcal{N}_i||\mathcal{N}_j)}{2\psi^2}\right) \tag{30}$$

where $\mathcal{N}_i$ and $\mathcal{N}_j$ denote aesthetic distribution of the ith and the jth users; ψ denotes the standard deviation of the Gaussian distribution. On the basis of the affinity matrix, we construct an affinity graph.

2) Graph-Shift-Based Flickr Circles Discovery: Flickr users with similar aesthetic interests are densely distributed in the affinity graph. To effectively discover those dense subgraphs, two conditions are required. 1) *Compatibility With Graph Representation*: Many similarity metrics are defined based on binary relation. However, only graph-based clustering can utilize the pairwise relation directly. 2) *Robustness to Outliers*: A few Flickr users have very particular aesthetic interests (e.g. skull photos) and they may not belong to any circles. Methods insisting on partitioning all users into coherent circles without outliers may fail to preserve the structure of multiple circles.

Conventional clustering algorithms are not suitable for discovering circles from Flickr users, as they insist on partitioning all the input data. Comparatively, graph shift [19], which is efficient and robust for graph mode seeking, is suitable for mining the densely distributed Flickr users. It directly works on graph, supports arbitrary number of clusters, and leaves the outlier points ungrouped. Formally, we generate an affinity graph $\mathbf{G} = (\mathbf{V}, \mathbf{W})$, where $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$ is a set of vertices denoting the Flickr users, and $\mathbf{W}$ is a symmetric matrix detailed in (30). The modes of a graph $\mathbf{G}$ are defined as local maximizers of graph density function $g(p) = p^T \mathbf{A} p$, $p \in \Delta^N$, where $\Delta^N = \{p \in \mathbb{R}^N : p \geq 0 \text{ and } ||p||_1 = 1\}$. More specifically, the similarity between users is expressed as the edge weights of graph $\mathbf{G}$. Those strongly connected subgraphs correspond to large local maxima of $g(p)$ over simplex, which is an approximate measure of the average affinity score of these subgraphs.

The target patterns are the local maximizers of $g(p)$. They represent the users in each Flickr circle, which are calculated by solving a quadratic optimization as

$$\max\nolimits_p g(p) = p^T \mathbf{A} p, \text{ s.t. } p \in \Delta^N. \tag{31}$$

Note that obtaining an analytic solution of (34) is difficult. Therefore, we employ replicator dynamics to find the local maxima of (34). Given an initialization $p^{(0)}$, the local solution p^* can be iteratively computed by the discrete-time version of the first-order replicator equation, i.e.

$$p_i^{(t+1)} = p_i^{(t)} \frac{(\mathbf{A} p^{(t)})_i}{(p^{(t)})^T \mathbf{A} p^{(t)}} \tag{32}$$

where $p^{(t)}$ denotes the value of p in the tth iteration.

Time complexity of the graph shift-based dense subgraph mining is as follows. Suppose the average number of edges in the subgraph is a and the average number of iterations for the replicator equation is c, then the time complexity of the replicator dynamics procedure is $\mathcal{O}(ac)$. The total time complexity is $\mathcal{O}(acq)$, where q is the number of the shrink/expansion stages in the graph shifting.

By summarizing the techniques described in Sections III, VI and V, we present the procedure of the proposed Flickr circles discovery in Algorithm 1.

VI. Experiments and Analysis

This section evaluates the performance of our proposed Flickr circles discovery based on four experiments. We first compare

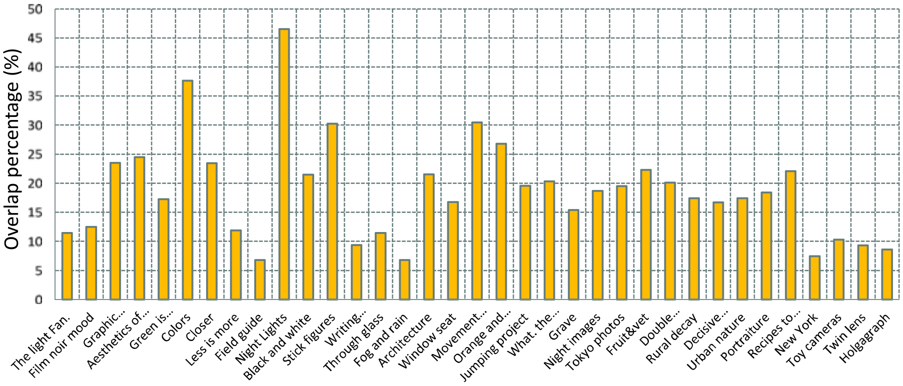


Fig. 3. A graphical illustration of the overlaps among Flickr groups.

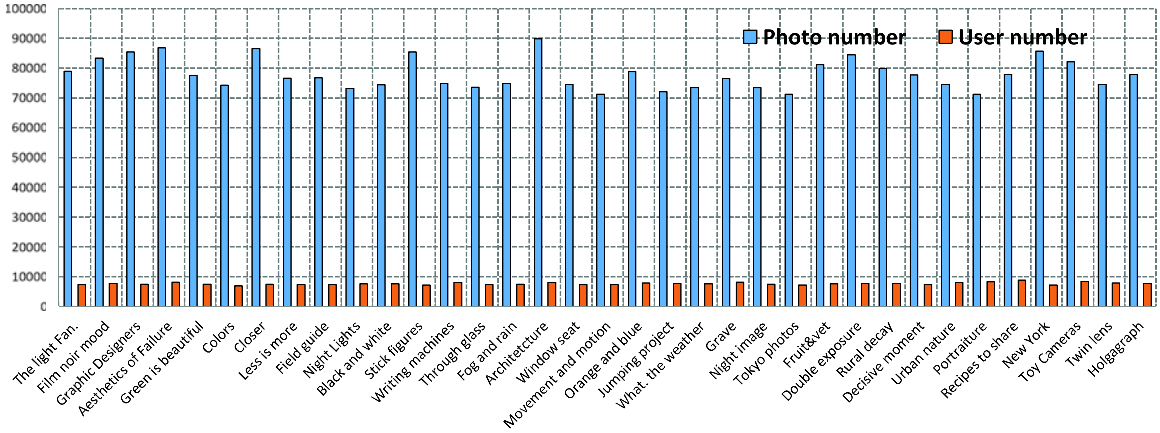


Fig. 4. Groups (the horizontal axis) and the number of Flicker users (the vertical axis) in each of the 35 groups.

Algorithm 1: Flickr Circles Discovery Algorithm.

```
input: a large number of Flickr users with different numbers
of photos; output: multiple detected Flickr circles;
1. For each photo $I$  //feature extraction and fusion
     $z_c(I) = ExtractCM(I)$ based on (1, 2, 3);
     $z_t(I) = ExtractHOG(I)$ using (4, 5);
     $z_s(I) = ExtractLFA(I)$ based on (8);
     $z(I) = MultiFeaFusion(z_c, z_t, z_s)$ using (12);
2. For each Flickr user
     Collect all its photos $Z$;
     $p = \text{regGMM}(Z)$ using (22);  // model the photo distribution
3. Calculate the affinity matrix $\mathbf{W}$ using (30) and obtain the
   corresponding affinity graph $\mathbf{G}$; afterward apply dense
   subgraph mining to detect Flickr circles;
```

our approach with the state-of-the-art communities detection algorithms. Afterward, we analyze the important parameters. Last but not least, we show how to leverage the detect Flickr circles to enhance the transferal-based photo retargeting.

All the experiments were carried out on a workstation equipped with an Intel i7-4940MX CPU and 64GB RAM. Our approach was implemented on the Matlab 2012 platform.

A. *Flickr Image Set Compilation*

Although the proposed method detects Flickr circles in an unsupervised manner, we need the labeled ground-truth data to evaluate its performance. We expended significant time, effort, and resources to crawl photos from 35 well-known public groups from Flickr. As shown in Fig. 3, nearly 20% of the ground-truth groups are relatively independent to the other groups. About 60% of the groups moderately intersect with the other groups. The rest 20% of the groups are highly correlated with the others. For each group, we collected 70 000∼90 000 photos from nearly 7400 Flickr users. The statistics of our data set is given in Fig. 4. For different Flicker groups, the numbers of photos belonging to users varies significantly, as shown in Table I. This is the motivation of our regularized topic model.

B. *A Comparative Study*

We can evaluate the detected Flickr circles $\mathcal{C} = \{C_1, C_2, \ldots, C_A\}$ by comparing them with the ground-truth Flickr circles (i.e., groups) $\bar{\mathcal{C}} = \{\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_{\bar{A}}\}$. Our observation is that for an optimal communities discovery algorithm, the predicted circles should closely align with the ground-truth circles.

To measure the alignment between a predicted circle C and a ground-truth circle $\bar{C}$, we compute the balanced error rate

TABLE I
MAX/MIN/AVE PHOTO NUMBERS OF FLICKER USER FROM THE 35 GROUPS AND THE STANDARD VARIANCE (SV)

Flickr group	Max No.	Min No.	Ave No.	SV
The light Fan.	2132	12	212	43.213
Film noir Mood	1765	22	196	36.764
Graphic designers	2543	7	267	46.541
Aesthetics failure	3567	43	231	24.356
Green is beautiful	2865	21	186	34.251
Colors	5643	41	324	65.674
Closer	3245	6	243	46.784
Less is more	2132	12	134	32.228
Field guide	2657	24	178	46.783
Night lights	4465	3	249	56.887
Black and white	3214	76	147	31.183
Stick figure	2654	16	227	40.654
Writing mach.	1342	11	103	23.355
Through glass	3421	42	245	46.678
Fog and rain	2885	51	215	38.897
Architecture	3146	36	195	37.769
Window seat	2989	11	245	56.782
Movement	4564	32	277	36.689
Orange and blue	3245	9	214	44.325
Jump Project	3105	14	227	43.236
What. the weather	2543	43	121	36.543
Grave	3001	26	95	43.262
Night images	3243	46	94	34.367
Tokyo photos	4454	32	117	47.786
Fruit&vet	5765	34	154	34.445
Double exposure	6231	37	124	37.773
Rural decay	4354	12	69	45.674
Decisive moment	4157	11	56	39.912
Urban nature	3876	21	48	43.458
Portraiture	6544	14	57	54.673
Recipes to share	3543	32	47	43.358
New York	4113	16	54	41.214
Toy cameras	3654	32	63	46.547
Twin lens	4454	34	41	38.897
Holgagraph	3454	32	43	43.213

TABLE II
BER SCORES OF THE SEVEN CLUSTERING ALGORITHMS

Flickr group	KC	HC	LC	CP	LRE	MAC	Ours
The light Fan.	0.6212	0.4213	0.5543	0.5121	0.4214	0.5132	**0.5765**
Film noir Mood	0.6212	0.4879	0.6112	0.5435	0.5231	0.5879	**0.7014**
Graphic designers	0.4456	0.4558	0.4669	0.4873	0.4552	0.4567	**0.5564**
Aesthetics failure	0.6456	0.6776	0.6326	0.6324	0.7116	0.6557	**0.7346**
Green is beautiful	0.4009	0.4557	0.4321	0.4236	0.4668	0.4889	**0.5115**
Colors	0.7668	0.7342	0.7223	0.7243	0.7078	0.7226	**0.8001**
Closer	0.6343	0.6357	0.6448	0.6559	0.6221	0.6557	**0.7245**
Less is more	0.5884	0.6011	0.5664	0.5779	0.5559	0.5794	**0.6347**
Field guide	0.6876	0.7005	0.6558	0.6553	0.6667	0.7211	**0.7331**
Night lights	0.6454	0.6876	0.6643	0.6546	0.6765	0.6443	**0.7443**
Black and white	0.3011	**0.4112**	0.3412	0.3511	0.3431	0.3552	0.3211
Stick figure	0.6543	0.6543	0.6643	0.6511	0.6454	0.6511	**0.7167**
Writing mach.	0.7321	0.7454	0.8011	0.7432	0.7889	0.7711	**0.8343**
Through glass	0.6987	0.7123	0.6565	0.6779	0.6989	0.7232	**0.7445**
Fog and rain	0.3879	0.3821	0.3946	0.4119	0.4228	0.4315	**0.4652**
Architecture	0.6878	0.6987	0.6889	0.6676	0.6643	0.6678	**0.7103**
Window seat	0.7454	0.7668	0.7884	0.7657	0.7675	0.7558	**0.8213**
Movement	0.4668	0.4435	0.5012	0.4898	0.4454	0.4663	**0.5221**
Orange and blue	0.4889	0.5121	0.4886	0.4995	0.4654	0.4121	**0.5664**
Jump Project	0.6221	0.6021	0.5743	0.5987	0.5776	**0.6004**	0.5889
What. weather	0.6112	0.6123	0.5776	0.5997	0.5886	0.5486	**0.6032**
Grave	0.5445	0.5765	0.5676	0.6121	0.6004	0.5995	**0.6321**
Night images	0.6043	0.5765	0.5898	0.5994	0.6114	0.5876	**0.6104**
Tokyo photos	0.5779	0.5886	0.6087	0.5931	0.6115	0.6138	**0.6225**
Fruit&vet	0.6226	0.5997	0.5772	0.6117	0.5823	0.5781	**0.6223**
Double exposure	0.5443	0.5534	0.5465	0.5512	0.5844	0.5411	**0.5984**
Rural decay	0.5668	0.5785	0.5884	0.5994	0.5743	0.5669	**0.5992**
Decisive moment	0.6007	0.6123	0.5997	0.6088	0.6109	0.6132	**0.6321**
Urban nature	0.5889	0.5913	0.6004	0.5879	0.6122	0.6054	**0.6154**
Portraiture	0.5998	0.6127	0.5979	0.6087	0.6123	0.6097	**0.6235**
Recipes to share	0.5435	0.5776	0.5643	0.5886	0.5567	0.5712	**0.6121**
New York	0.5732	0.5568	0.5537	0.5548	0.5889	0.5857	**0.6114**
Toy cameras	0.5335	0.5447	0.5433	**0.5676**	0.5465	0.5443	0.5462
Twin lens	0.5556	0.5434	0.5674	0.5649	0.5574	0.5684	**0.6134**
Holgagraph	0.5998	0.6127	0.5979	0.6087	0.6123	0.6097	**0.6235**
Average	0.5889	0.5879	0.5898	0.5911	0.5662	0.5895	**0.6567**

(BER) [20] between the two circles

$$\mathrm{BER}(C,\bar{C}) = \frac{1}{2}\left(\frac{|C-\bar{C}|}{|C|} + \frac{|C^c-\bar{C}^c|}{|C^c|}\right) \tag{33}$$

where $C - \bar{C}$ denotes the positive instances predicted wrong, C^c is a set of negative instances, and $C^c - \bar{C}^c$ contains the negative instances predicted wrong. The above BER assigns equal importance to false positives and false negatives. On average, trivial or random predictions result in an error of 0.5. Such a measure is preferable to the $0/1$ loss, which assigns extremely low error to trivial predictions.

In the experiment, we also report the F_1 score [39]

$$F_1(C,\bar{C}) = 2 \cdot \frac{\mathrm{precision}(C,\bar{C}) \cdot \mathrm{recall}(C,\bar{C})}{\mathrm{precision}(C,\bar{C}) + \mathrm{recall}(C,\bar{C})}. \tag{34}$$

By treating $\bar{C}$ as a set of "relevant" documents and C as a set of "retrieved" documents, the precision and recall in (34) can be defined as

$$\mathrm{precision}(C,\bar{C}) = \frac{|C \cap \bar{C}|}{|C|} \tag{35}$$

$$\mathrm{recall}(C,\bar{C}) = \frac{|C \cap \bar{C}|}{|\bar{C}|}. \tag{36}$$

We compare our approach with seven well-known clustering algorithms, including those considering only the graph/network structure, those exploring only the profile information, and those combining the both. 1) K-means clustering (KC); 2) hierarchical clustering (HC) [15]; 3) link clustering (LC) [22]; 4) clique percolation (CP) [21]; 5) low-rank embedding (LRE) [24]; and 6) multi-assignment clustering (MAC) [28]. KC and HC are built-in functions in Matlab 2012b. Matlab codes of the LC algorithm are provided also.[1] The Matlab codes of the CP are given.[2] The LRE is implemented in the package.[3] Finally, the MAC is implemented.[4]

We compare our approach with the seven baseline clustering algorithms described above. For all the algorithms, we fix the cluster number to 20. The low&high-level visual features of the seven algorithms are the same as ours. As shown in Table II, the following observations can be made. 1) For 32 out of the 35 Flickr groups, our approach outperforms all its competitors,

[1] [Online]. Available: http://barabasilab.neu.edu/projects/linkcommunities/

[2] [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/34202-k-clique-algorithm

[3] [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/38422-improved-nystrom-kernel-low-rank-approximation

[4] [Online]. Available: http://www.mariofrank.net/MACcode/

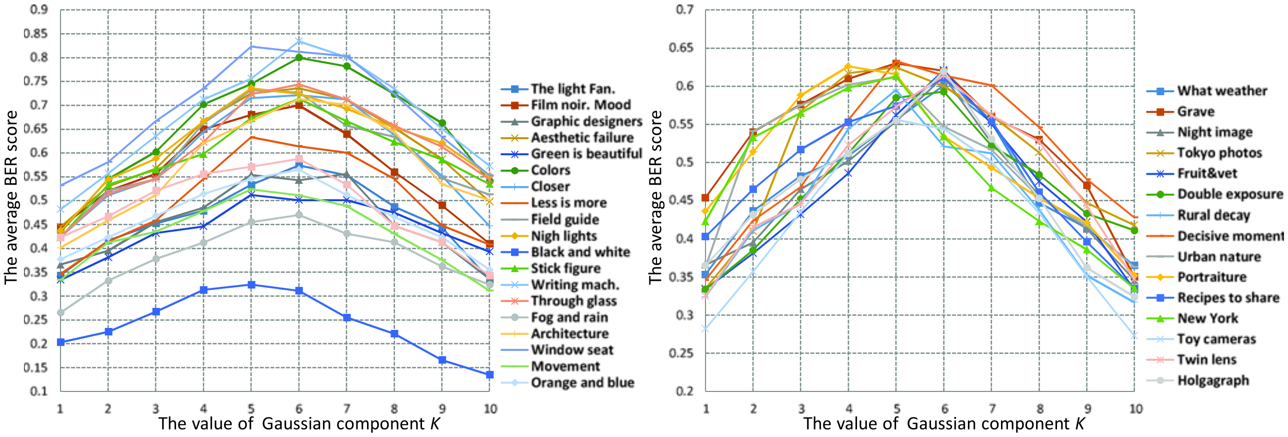


Fig. 5. BER values under different values of k (the top subfigure shows the first 20 Flickr groups, and bottom subfigure displays the rest 15 Flickr groups).

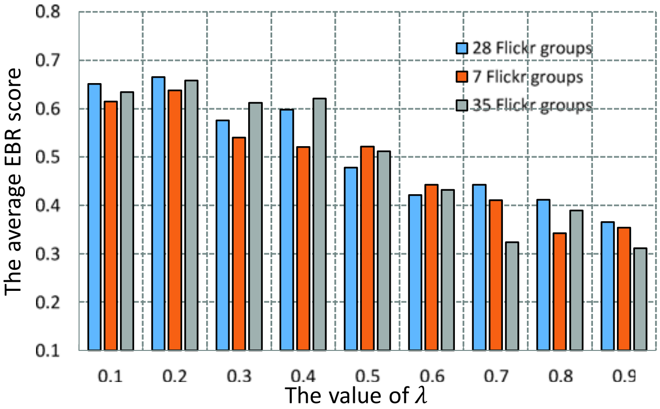


Fig. 6. Average BER values of the three circles containing users with 7 very few photos, the average BER values of the rest 28 circles, and the average BER value over the entire 35 circles.

as the corresponding BER and F_1 scores are the highest. This observation shows the clear advantage of our regularized topic model, which can optimally capture users' aesthetic interest. 2) For Flickr circles describing specific concepts, e.g., the architecture and the window seat, they can be more accurately detected. Comparatively, for circles describing abstract concepts, e.g., the jump project, all the algorithms are difficult to detect the circles. This is because photos with abstract concepts do not have a regular visual appearance. Therefore, our adopted low&high-level features cannot well describe them. 3) Our proposed method performs remarkably better on Flickr groups containing users with very few photos (e.g., the graphic designers and the night lights). This again demonstrates the necessity to use a regularized term to model the aesthetic interests of users with few photos.

C. Parameter Analysis

In retrospect to our proposed Flickr circles detection, there are two important parameters to be tuned: 1) the number of the Gaussian components K and 2) the weight of the regularizer λ. First, to evaluate the performance of our approach under different values of K, we present the corresponding BER values on the 35 Flickr groups. As shown in Fig. 5, on all the 35 groups, the BER values increase and then peak at $V = 6$. Afterward, the BER values decrease to a low level. Based on this observation, we set $V = 6$ in the experiment. Second, to evaluate the performance under different regularizer weights, we tune λ from 0.1 to 0.9 with a step of 0.1. For each triplet in Fig. 6, we present the average BER score of the eight circles containing users with few photos, the average BER score of the remaining 27 circles, and the average BER score over all the 35 circles. As can be seen, the best performance is achieved when $\lambda = 0.2$.

D. Enhancing Transferal-Based Photo Cropping

Each discovered Flickr circle includes users with a particular aesthetic interest, which can be adopted to enhance the transferal-based photo cropping [40]. As shown in Fig. 7, the Flickr circles can be naturally treated as an intermediate layer in the probabilistic model. Generally, the model contains four layers. The first layer represents a collection of training well-aesthetic Flickr photos $I_1, I_2, \ldots, I_L$, the second layer denotes the Flickr circles assigned to each training photo, the third layer is the GMM learned from the visual features of the training photos, and the last layer denotes the cropped photo from a poorly framed one.

The flowchart of our cropping model can be described as follow. By categorizing the test photo to the pre-specified Flickr circles $\mathcal{C}$, the training photos $I_1, I_2, \ldots, I_L$ belonging to the same circles are selected to learn a GMM that describes the aesthetic distribution of these Flickr circles $\mathcal{C}$. The learned GMM are adopted as a posterior probability to search the optimal subregion I^* in the test photo. Mathematically, this procedure can be formulated as

$$\begin{aligned}
&\max_{I^*} p(I^*|I_1, I_2, \ldots, I_L) \\
&\quad = \max_{I^*} p(I^*|x^*) p(x^*|\mathcal{C}) \cdot p(\mathcal{C}|I_1, I_2, \ldots, I_L) \\
&\quad = \max_{I^*} p(x^*|\mathcal{C}) \qquad (37)
\end{aligned}$$

where $p(x^*|\mathcal{C})$ denotes the posterior probability, which is calculated based on the learned GMM, and x^i denotes the visual

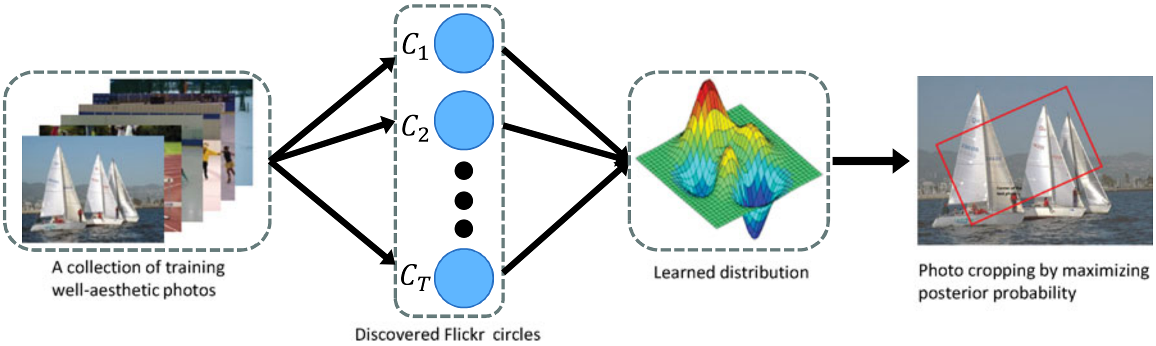

A collection of training well-aesthetic photos
C_1
C_2
C_T
Discovered Flickr circles
Learned distribution
Photo cropping by maximizing posterior probability

Fig. 7. Pipeline of the transferal-based photo cropping by leveraging the discovered Flickr circles.

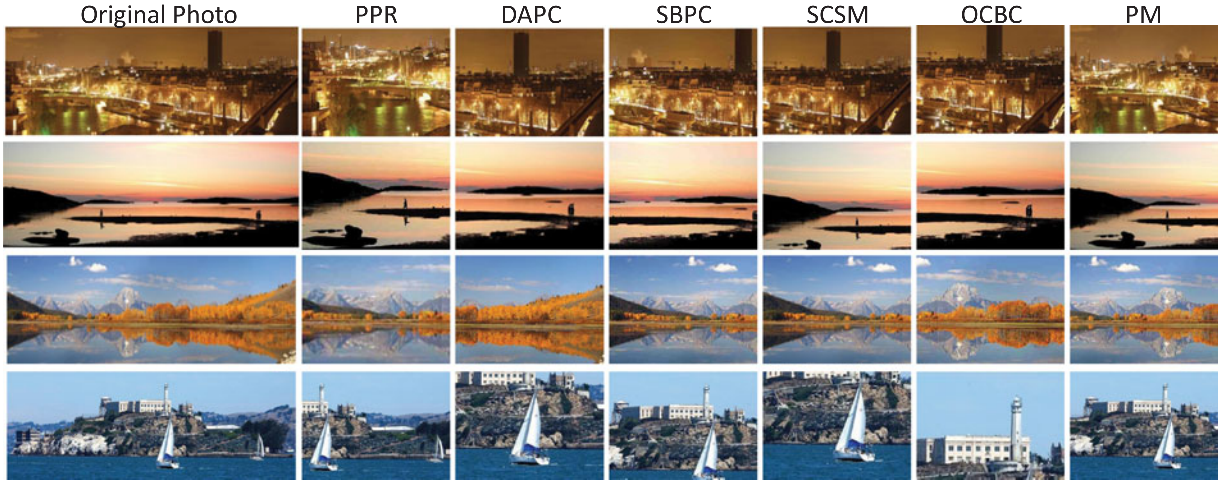
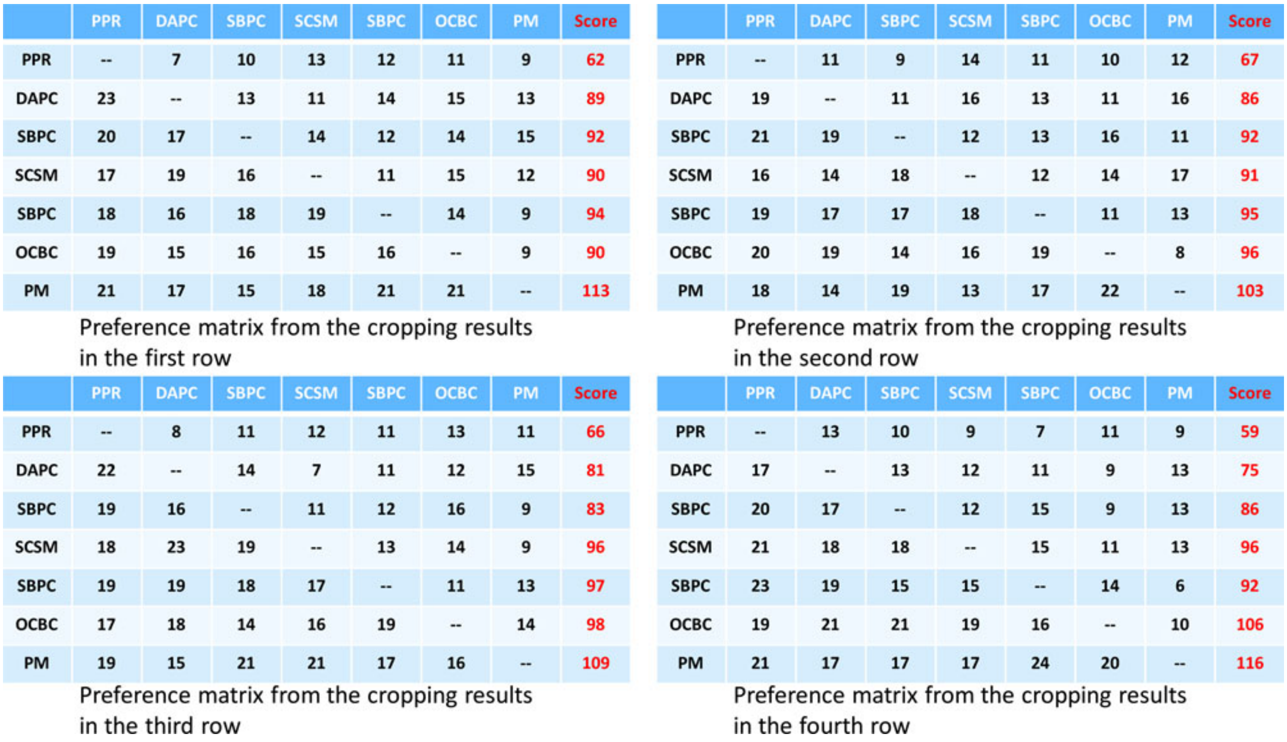


Fig. 8. Comparative cropped photos different different cropping algorithms.



	PPR	DAPC	SBPC	SCSM	SBPC	OCBC	PM	Score
PPR	--	7	10	13	12	11	9	62
DAPC	23	--	13	11	14	15	13	89
SBPC	20	17	--	14	12	14	15	92
SCSM	17	19	16	--	11	15	12	90
SBPC	18	16	18	19	--	14	9	94
OCBC	19	15	16	15	16	--	9	90
PM	21	17	15	18	21	21	--	113

Preference matrix from the cropping results in the first row

	PPR	DAPC	SBPC	SCSM	SBPC	OCBC	PM	Score
PPR	--	11	9	14	11	10	12	67
DAPC	19	--	11	16	13	11	16	86
SBPC	21	19	--	12	13	16	11	92
SCSM	16	14	18	--	12	14	17	91
SBPC	19	17	17	18	--	11	13	95
OCBC	20	19	14	16	19	--	8	96
PM	18	14	19	13	17	22	--	103

Preference matrix from the cropping results in the second row

	PPR	DAPC	SBPC	SCSM	SBPC	OCBC	PM	Score
PPR	--	8	11	12	11	13	11	66
DAPC	22	--	14	7	11	12	15	81
SBPC	19	16	--	11	12	16	9	83
SCSM	18	23	19	--	13	14	9	96
SBPC	19	19	18	17	--	11	13	97
OCBC	17	18	14	16	19	--	14	98
PM	19	15	21	21	17	16	--	109

Preference matrix from the cropping results in the third row

	PPR	DAPC	SBPC	SCSM	SBPC	OCBC	PM	Score
PPR	--	13	10	9	7	11	9	59
DAPC	17	--	13	12	11	9	13	75
SBPC	20	17	--	12	15	9	13	86
SCSM	21	18	18	--	15	11	13	96
SBPC	23	19	15	15	--	14	6	92
OCBC	19	21	21	19	16	--	10	106
PM	21	17	17	17	24	20	--	116

Preference matrix from the cropping results in the fourth row

Fig. 9. Statistics of the paired-comparison-based user study on the four sets of cropped photos in Fig. 8.

features extracted from test image I^*. Practically, the GMM is learned from visual features of the training images belonging to the selected Flickr circles.

To evaluate the performance of the above photo cropping by leveraging the detected Flickr circles, we compare it with several well-known cropping algorithms as introduced in [40], including sparse coding of saliency maps (SCSM), sensation based photo cropping (SBPC), omni-range context based cropping (OCBC), personalized photo ranking (PPR) and describable attribute for photo cropping (DAPC). SCSM selects the cropped region that can be decoded by the dictionary learned from training saliency maps with the minimum error. SBPC selects the cropped region with the maximum quality score, which is computed by probabilistically integrating the SVM scores corresponding to the detected subjects in a photo. OCBC integrates the prior of spatial distribution of two arbitrary image patches into a probabilistic model to score each candidate cropped photo, and the candidate cropped photo with the maximum score is selected as the cropped photo. We experiment on the panoramic photos due to the difficulty to select an optimal sub-region, which is a good platform to testify the robustness of different cropping methods. As shown in Fig. 8, our method performs competitively by obtaining the most aesthetically pleasing sub-region from each original photo.

To quantitatively demonstrate the advantage of our cropping method, a paired-comparison-based user study is conducted. Paired comparison is to present each subject with a pair of cropped photos from two different approaches. Participants are then required to indicate a preference, for one of the two cropped photos. Evaluation results are stored in the preference matrix. For example, considering the first preference matrix from Fig. 9, the entry in column SBPC and row DAPC has a value of 13, indicating that 13 subjects prefer the cropped photo produced from DAPC than that produced by SBPC. The participants in the user study are 30 Ph.D./master students from the computer sciences department from National University of Singapore. The rightmost column of each matrix describes the scores of all the compared cropping methods. For example, the entity 94 from the first matrix (corresponding to the row SBPC) reflects that SBPC is the second best performer for the first set of cropping photos in Fig. 8. The four matrices show the best cropping method is always achieved by our method (entity 113 in the first matrix, 103 in the second matrix, 109 in the third matrix and 116 in the last matrix).

VII. Conclusion and Future Work

This paper proposes to discover circles from a large number of Flickr users, where each circle contains users with similar aesthetic interests. We first fuze visual features from color, textural and semantic channels to describe each Flickr photo. Afterward, a manifold-guided regularized term is incorporated into the standard GMM to describe the aesthetic interest of each user. Next, an affinity graph is constructed to describe the aesthetic relationships among users. Finally, users densely distributed on the affinity graph are categorized into multiple circles using the dense subgraphs mining technique.

In the experiment, we evaluated the performance of our method comprehensively. We first compared our approach with the state-of-the-art community detection algorithms. Then, a step-by-step evaluation was conducted to testify each component, which demonstrates their usefulness and in inseparability. We also shown how to leverage the detect Flickr circles to enhance the transferal-based photo retargeting.

One limitation of our approach is that it performs unsatisfactorily on abstract photos, since these photos do not have a regular appearance. Thus, we plan to develop a more general and comprehensive model which considers not only the above low&high-level visual features but also other important photography elements such as exposure, contrast, *etc.*

Appendix

The references [34], [35] are famous publications about manifold theories. Their citations are over 5000 in total. We cite them in our paper to elaborate the necessity and advantage of exploring the distribution of samples on manifold. In brief, the two publications assume the "locality" property of samples on manifold, i.e., each sample can be linearly reconstructed by a set of its neighboring ones. Based on this, the authors developed dimensionality reduction algorithms which can effectively handle human face recognition, visualization, and document retrieval.

Comparatively, the manifold theory developed in our work is fundamentally different from [34], [35] in the following aspects: 1) [34], [35] describe manifold-based dimensionality reduction algorithms, while our work leverage the concept of manifold to design a regularizer which solves the overfitting problem in the GMM learning stage; 2) [34], [35] are discriminative models while our designed GMM regularizer is a probabilistic model. We attached paper [34], [35] as the supplementary material in this round submission, and it is obvious that their formulations are completely different from ours; 3) the parameter inference stage between [34], [35] and our manifold-based regualizer is apparently different. The dimensionality reduction algorithms in [34], [35] have analytic solutions, which can be solved conveniently based on eigenvalue decomposition. In contrast, the regularized GMM in our work has no analytic solution, therefore we developed an EM algorithm which calculates parameters β_i, μ_i, and Σ_i alternatively; 4) the input and output between algorithms [34], [35] and ours are completely different. The dimensionality reduction algorithms in [34], [35] convert the feature vector extracted from each sample (image, document, etc.) into a low dimensional one. Comparatively, our work converts the feature vectors extracted from a collection of Flickr photos into a distribution that models the aesthetic tendency of each Flickr user.

References

[1] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," presented at the Int. Conf. Comput. Vis., Barcelona, Spain, 2011.

[2] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1419–1429, Mar. 2014.

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[4] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," presented at the 9th European Conf. Comput. Vis., Graz, Austria, 2006.

[5] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," presented at the 16th IEEE Int. Conf. Image Process., Cairo, Egypt, 2009.

[6] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," presented at the IEEE Conf. Comput. Vis. Pattern Recog., Providence, RI, USA, 2011.

[7] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 291–300.

[8] L. Zhang *et al.*, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.

[9] L. Zhang *et al.*, "Perception-guided multimodal feature fusion for photo aesthetics assessment," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 237–246.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[11] G. Costa and R. Ortale, "A Bayesian hierarchical approach for exploratory analysis of communities and roles in social networks," *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 194–201.

[12] E. Yao *et al.*, "Probabilistic text modeling with orthogonalized topics," presented at the Special Interest Group Inform. Retrieval, Gold Coast, Australia, 2014.

[13] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 173–182.

[14] Z. Yin, L. Cao, Q. Gu, and J. Han, "Latent community topic analysis: Integration of community discovery with topic modeling," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, 2012, Art. no. 63.

[15] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[16] W. Press, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[18] X. Li, W. Hu, C. Shen, A. Dick, and Z. Zhang, "Context-aware hypergraph construction for robust spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2588–2597, Oct. 2014.

[19] H. Liu and S. Yan, "Robust graph mode seeking by graph shift," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 671–678.

[20] Y.-w. Chen, "Combining SVMs with various feature selection strategies," *Feature Extraction*. Berlin, Germany: Springer-Verlag, 2005.

[21] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[22] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[23] L. Zhang *et al.*, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, Jan. 2014.

[24] T. Yoshida, "Toward finding hidden communities based on user profile," presented at the IEEE Int. Conf. Data Mining Workshops, Sydney, NSW, Australia, 2010.

[25] Z. Ma *et al.*, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[26] B. Ni *et al.*, "Learning to photograph: A compositional perspective," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1138–1151, Aug. 2013.

[27] Y. Li *et al.*, "Difficulty guided image retrieval using linear multiple feature embedding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1618–1630, Dec. 2012.

[28] M. Frank, A. P. Streich, D. Basin, and J. M. Buhmann, "Multi-assignment clustering for Boolean data," *J. Mach. Learn. Res.*, vol. 13, pp. 459–489, 2012.

[29] H. Zhang, R. Edwards, and L. Parker, "Regularized probabilistic latent semantic analysis with continuous observations," presented at the IEEE 11th Int. Conf. Mach. Learn. Appl., Boca Raton, FL, USA, 2012.

[30] L. Zhang *et al.*, "Retargeting semantically-rich photos," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1538–1549, Sep. 2015.

[31] L. Zhang and R. Zimmermann, "Flickr circles: Mining socially-aware aesthetic tendency," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun.-Jul. 2015, pp. 1–6.

[32] X. Chen, Y. Qi, B. Bai, Q. Lin, and J. G. Carbonell, "Sparse latent semantic analysis," presented at the SIAM Int. Conf. Data Mining, Mesa, AZ, USA, 2011.

[33] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 862–871.

[34] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inform. Process. Syst. Conf.*, 2003, pp. 153–160.

[35] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacian faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[36] A. P. Dempster, N. M. Laird, and D. B Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B*, vol. 39, no. 1, pp. 1–38, 1977.

[37] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[38] Z.-J. Zha *et al.*, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, Feb. 2012.

[39] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 37–63, 2011.

[40] L. Zhang *et al.*, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.

[41] R. Hong, L. Zhang, and D. Tao, "Unified photo enhancement by discovering aesthetic communities from Flickr," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1124–1135, Mar. 2016.

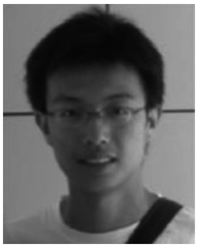
Richang Hong (M'14) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008.

He was a Research Fellow with the School of Computing, National University of Singapore, Singapore, from September 2008 to December 2010. He is currently a Professor with Hefei University, Hefei, China. He has coauthored more than 70 publications in the areas of his research interests, which include multimedia content analysis and social media.

Prof. Hong is a Member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He served as an Associate Editor of the *Information Sciences and Signal Processing*, and the Technical Program Chair of the MMM 2016. He was the recipient of the Best Paper Award in the ACM Multimedia 2010, the Best Paper Award in the ACM ICMR 2015, and the Honorable Mention of the IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award.

Luming Zhang (M'14) received the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2012.

He is currently a Faculty Member with the Hefei University of Technology, Hefei, China. His research interests include visual perception analysis, image enhancement, and pattern recognition.



Chao Zhang received the B.S. and M.S. degrees from Zhejiang University, Zhejiang, China, in 2010 and 2013, respectively, and the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2016.

His research interests include spatiotemporal data mining, social media analysis, and applied machine learning.

Roger Zimmermann (S'94–M'97–SM'07) received the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, CA, USA, in 1994 and 1998, respectively.

He is currently an Associate Professor with the Department of Computer Science, National University of Singapore (NUS), Singapore. He is also the Deputy Director with the Interactive and Digital Media Institute, NUS, and a Co-Director of the Centre of Social Media Innovations for Communities, NUS. He has coauthored a book, six patents, and more than 150 conference publications, journal articles, and book chapters. His research interests include streaming media architectures, distributed and peer-to-peer systems, mobile and geo-referenced video management, collaborative environments, spatio-temporal information management, and mobile location-based services.

Prof. Zimmermann is a Member of ACM.