Semi-Automatic Tagging of Photo Albums via Exemplar Selection and Tag Inference

Dong Liu, Meng Wang, *Member, IEEE*, Xian-Sheng Hua, *Member, IEEE*, and Hong-Jiang Zhang, *Fellow, IEEE*

***Abstract*—As one of the emerging Web 2.0 activities, tagging becomes a popular approach to manage personal media data, such as photo albums. A dilemma in tagging behavior is the users' manual efforts and the tagging accuracy: exhaustively tagging all photos in an album is labor-intensive and time-consuming, and simply entering tags for the whole album leads to unsatisfying results. In this paper, we propose a semi-automatic tagging scheme that aims to facilitate users in photo album tagging. The scheme is able to achieve a good trade-off between manual efforts and tagging accuracy as well as to adjust tagging performance according to the user's customization. For a given album, it first selects a set of representative exemplars for manual tagging via a temporally consistent affinity propagation algorithm, and the tags of the rest of the photos are automatically inferred. Then a constrained affinity propagation algorithm is applied to select a new set of exemplars for manual tagging in an incremental manner, based on which the performance of the tag inference in the previous round can be estimated. If the results are not satisfying enough, a further round of exemplar selection and tag inference will be implemented. This process repeats until satisfactory tagging results are achieved, and users can also stop the process at any time. Experimental results on real-world Flickr photo albums have demonstrated the effectiveness and usefulness of the proposed scheme.**

***Index Terms*—Exemplar selection, photo album, semi-automatic tagging, tag propagation.**

I. Introduction

With the popularity of digital cameras, recent years have witnessed a rapid growth of personal photo albums. People capture photos to record their lives and share them on the web. For example, Flickr [1], the earliest and the most popular photo sharing website, hosts over 3.6 billion personal photos [2].

Tagging has proved to be a popular approach to facilitate the management and the sharing of photos. By providing tags to describe the content of photos, many manipulations can be easily accomplished, such as indexing, browsing, and search. Intuitively, the most convenient approach to generate tags is to investigate automatic tagging (or annotation) techniques [3], [4].

Manuscript received January 13, 2010; revised May 18, 2010 and August 12, 2010; accepted September 28, 2010. Date of publication October 18, 2010; date of current version January 19, 2011. This work was performed at Microsoft Research Asia. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

D. Liu is with the Harbin Institute of Technology, Harbin 150001, China (e-mail: dongliu.hit@gmail.com).

M. Wang and X.-S. Hua are with Microsoft Research Asia, Beijing 100080, China (e-mail: mengwang@microsoft.com; xshua@microsoft.com).

H.-J. Zhang is with the Microsoft Advanced Technology Center, Beijing 100080, China (e-mail: hjzhang@microsoft.com).

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.2010.2087744

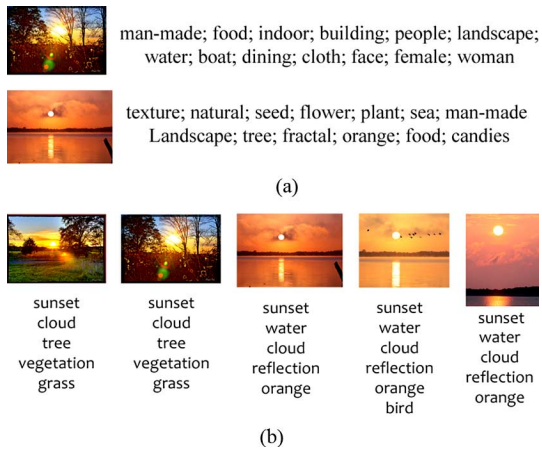

Fig. 1. (a) Tags obtained by the ALIPR system [5] for two photos, and we can see that many of them are inaccurate. (b) Several photos and their associated tags from a personal album on Flickr, where many of them are both visually and semantically close.

However, although great advances have been achieved in automatic tagging, currently these methods can hardly obtain satisfactory performance for real-world photos that contain significantly varying content. As an example, Fig. 1(a) illustrates two photos and the tags predicted by ALIPR [5], a state-of-the-art image annotation system introduced in [3]. From the results, we can see that many of the predicted tags are incorrect. Actually, nowadays, most photo sharing websites adopt the manual tagging approach, i.e., allowing users to manually enter tags to describe their uploaded photos. There are mainly two kinds of manual tagging approaches on the photo sharing websites. The first one is exhaustively tagging, in which users provide tags for each individual photo in the album. This approach tends to result in relative high tagging accuracy, but the drawback is its labor cost: a simple study in [6] shows that typically a user needs 6.8 s to enter tags for an image. Therefore, exhaustively tagging all the photos in a large album will be a labor-intensive task. In fact, many photos in a personal album are usually captured continuously to record one or more events and thus many of them are close to each other [7], [8], such as the examples illustrated in Fig. 1(b). Hence, a large part of manual efforts will be redundant in the exhaustive tagging manner. The second manual tagging approach in most photo sharing websites is batch tagging, in which users can assign tags to a suite of continuously uploaded photos. However, directly applying this approach to a whole album will introduce significant imprecise tags for many photos. Based on the above observations on the existing manual tagging approaches, we argue that there is a dilemma between manual efforts and tagging accuracy.

In this paper, we propose a semi-automatic photo tagging scheme that is able to modulate the manual efforts and the tagging performance in a flexible way. The proposed tagging scheme works in a semi-automatic manner in which the users only need to manually tag several selected exemplars and the tags of the rest of the photos are inferred automatically. In practice, due to the fact that different photo owners have distinct expectations of the tagging accuracy and are willing to make different levels of manual efforts, we design the scheme in an incremental manner. For the users who want to assign precise tags to their photos, the tagging process needs to be implemented intensively until the user established accuracy threshold is achieved. On the other hand, the users can also choose to terminate the tagging process freely at any time they do not want to continue. Specifically, the proposed tagging scheme is performed as follows. Given an album, a set of exemplary photos are first selected for manual tagging and the tags of the rest of the photos are inferred automatically. Then we further select an additional set of exemplars for manual tagging. With the user provided tags as ground truths, the performance of tag inference in the first round can be estimated on these newly selected exemplars. If the performance has not achieved the requirement established by users or they want to continue tagging, the process can proceed, and otherwise we perform the last round of tag inference by employing all exemplars selected in the tagging process as labeled data, and output the result as the final tagging result. A more detailed process will be illustrated in Section III.

There are two challenges in our proposed tagging scheme. The first one is the selection of exemplars. Actually it further contains two problems: 1) how to integrate multiple information clues in exemplar selection; 2) how to select exemplars in an incremental way. Personal photos are captured at greatly varied conditions with different photography skills, and these factors make the exemplar selection challenging. On the other hand, time is an important information clue for personal photos since photos that are temporally close will have high probability to record an identical scene or event [7]–[9]. As to be shown later in this paper, we propose a *temporally consistent affinity propagation* algorithm to group a photo album into a set of clusters and one exemplar is selected from a cluster. To realize the exemplar selection in an incremental way, we propose a *constrained affinity propagation* algorithm which is able to perform incremental exemplar selection conditioned on the existing exemplars selected in the previous rounds.

The second challenge is an effective tag inference algorithm, which utilizes the manually tagged photos to predict the tags of the rest of the photos. To accomplish this task, we construct a graph between the photos, where photos are linked to each other with their similarities. Once the graph is created, we will apply a graph-based semi-supervised learning approach [10] to propagate tags of the exemplary photos to the other non-exemplars.

The contributions of this paper can be summarized as follows: 1) we propose a semi-automatic photo tagging scheme that is able to modulate manual efforts and tagging accuracy in a flexible way. It can be applied in either online photo sharing or desktop photo management services; 2) we investigate the exploration of multiple clues in photo exemplar selection; 3) we propose a constraint affinity propagation algorithm that can realize exemplar selection in an incremental manner. In each round, exemplars can be selected with consideration of the existing exemplars selected in the previous rounds.

The organization of the rest of this paper is as follows. We provide a short review on the related work in Section II and describe the overview of our proposed semi-automatic tagging scheme in Section III. In Sections IV and V, we introduce the exemplar selection algorithm and the tag inference algorithm, respectively. Empirical study is presented in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

In this section, we will review related work along three threads, including photo tagging, active learning, and interactive photo album tagging.

A. Photo Tagging

Extensive research efforts have been dedicated to photo tagging. Ames *et al.* [11] have explored the motivation of tagging on the Flickr website and claimed that most users tag photos to make them better accessible to the general public. Kennedy *et al.* [12] have evaluated the performance of the classifiers trained with Flickr photos and their associated tags, and demonstrated that tags provided by Flickr users contain many noises. Liu *et al.* [13] have revealed the fact that the tags associated with Flickr images are in a random order and then proposed a tag relevance learning method to rank the tags according to their relevance levels. Yan *et al.* [6] proposed a model that is able to predict the time cost of manual image tagging. Tag recommendation is an intensively studied approach to help users tag photos more efficiently [14], [15]. By recommending a set of potentially relevant keywords in the tagging process, users can directly select the correct ones instead of entering them and it can effectively reduce the labor cost. Different from these existing efforts, this work adopts a different approach to facilitate users in tagging. For an album, only a set of selected photos are manually tagged, and the tags of the other photos are automatically inferred. In this way, the manual efforts can be significantly reduced and we will demonstrate that fairly high tagging accuracy can still be maintained.

B. Active Learning

In the computer vision and machine learning communities, active learning is a widely applied approach to reduce human efforts in labeling training samples [16]–[18]. Typically, the active learning approaches work on a predefined concept set and try to build a classifier for each concept with an initial set of training data. Then they iteratively annotate a set of elaborately selected samples so that the expected generalization error for the classifiers can be minimized in each step. It is clear that the active learning approaches work on a predefined ontology. On the contrary, we focus on an ontology-free scenario in which any textual keywords may be utilized as tags, and thus, it is impractical to build a fixed set of classifiers and minimize their generalization errors. In addition, our proposed scheme does not need any initial labeled training data.

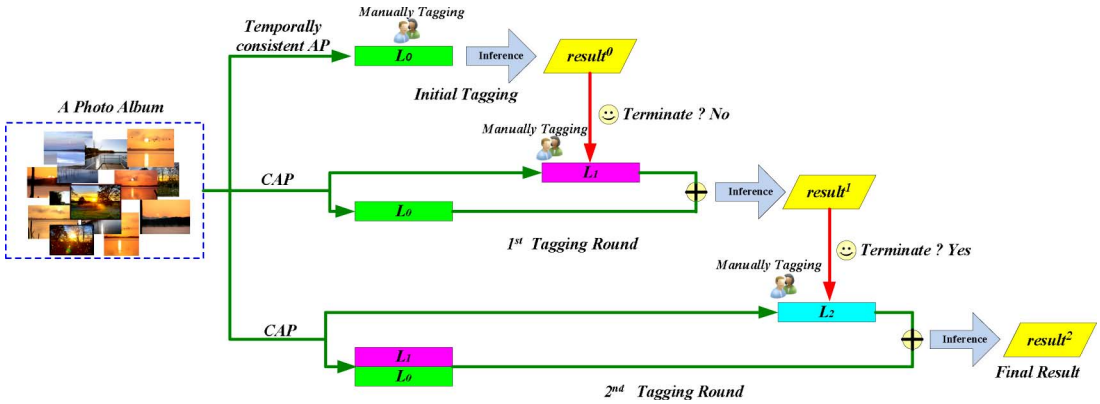

Fig. 2. Process of our proposed semi-automatic photo tagging scheme.

C. Interactive Photo Album Tagging

There exist some interactive photo album tagging systems in the literature [19], [20], which typically use the idea of clustering to partition the whole albums into smaller clusters and then ask the users to simultaneously label the photos in a cluster in one operation. For example, Suh *et al.* [19] have proposed an interactive photo album tagging system to annotate albums with people appearance. They extracted torso color based on face detection results, and then clustered the photos in an album based on torso color of people's clothes. However, it is sensitive to human cloth color and can only cluster photos from the same day. Cui *et al.* [20] further proposed to cluster faces or photos with similar scene together, then applied a contextual reranking procedure to boost the browsing experience of manual labeling. Although these systems can reduce manual efforts at some extent, they heavily rely on the clustering results obtained. If the clustering results are not convincing, especially when the visual distributions within a photo album are complex, these tagging systems will tend to fail. In the experiments, we will also show that the tagging performance of this naive method is poor.[1] Different from these existing systems, our proposed scheme asks the users to label the representative exemplar images and then propagates tags to the unlabeled images, which, on the one hand, reduces the sensitivity to the clustering results, and on the other hand, provides sufficient labeled images (the sufficiency of the labeled images can be controlled by the termination expectation of the users) to the follow-up tag inference algorithm.

III. Semi-Automatic Photo Tagging Scheme

Fig. 2 shows the work flow of the tagging process of a personal photo album. The tagging process works in an incremental manner until the satisfactory tagging accuracy is achieved or users stop the process. The whole tagging procedure can be summarized in Algorithm 1.

Algorithm 1: The Procedure of Our Proposed Semi-Automatic Photo Tagging Scheme: **Input**: A given photo album $\mathcal{U}$.

Output: Tagging result of the album.

1) Initial tagging.
   a) Select an initial set of exemplars $\mathcal{L}_0$ for manually tagging via temporally consistent affinity propagation algorithm.
   b) Perform 0th tag inference on $\mathcal{U}$ by employing $\mathcal{L}_0$ as the labeled data.
2) Incremental tagging.
   a) Select a set of exemplars $\mathcal{E}_t$ from $\mathcal{U}$ based on constrained affinity propagation algorithm, and manually tag the exemplars in $\mathcal{E}_t$.
   b) Validate the performance of the $(t-1)$th tag inference on $\mathcal{E}_t$.
   c) $\mathcal{L}_t = \mathcal{L}_{t-1} \bigcup \mathcal{E}_t$.
   d) Perform tth tag inference on $\mathcal{U}$ by employing all tagged exemplars in $\mathcal{L}_t$ as labeled data.
   e) If the result in b) meets user's satisfaction or the user wants to terminate the iteration.
      - Output the result as the final tagging result on $\mathcal{U}$;
      - Break;

        Else
      - $t = t + 1$, go to a).

IV. Exemplar Selection

In this section, we will describe the exemplar selection strategy in our proposed photo tagging scheme. We first perform the initial exemplar selection via the temporally consistent affinity propagation algorithm and then propose a constrained affinity propagation algorithm to accomplish the incremental exemplar selection.

A. Initial Exemplar Selection via Temporally Consistent Affinity Propagation

As previously discussed, the exemplar selection at the first round of our proposed semi-automatic tagging scheme is accomplished via a temporally consistent affinity propagation algorithm. We first introduce affinity propagation (AP) algorithm [21], [22], which is a similarity-based clustering algorithm that is able to group a given set of samples into several clusters as well as select an exemplar from each cluster.

Given a set of n data points $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, the algorithm takes as input the pairwise similarity $s(i, j)$ between any

[1] Actually, these clustering-based interactive photo album tagging systems can be abstracted as one uniform framework, where clustering is first implemented, followed by a manual labeling process that assigns tags to the individual clusters. We will name this framework as *naive tag assignment* in the experiments (see Section VI).

two points x_i and x_j in $\mathcal{X}$. The algorithm then works by iterating the following two simple messages until convergence:

$$r(i,k) = s(i,k) - \max_{k' \neq k}[a(i,k') + s(i,k')] \tag{1}$$

$$a(i,k) = \begin{cases} \min\left[0, r(k,k) + \sum_{i' \notin \{i,k\}} \max[0, r(i',k)]\right], & i \neq k \\ \sum_{i' \neq k} \max\left[0, r(i',k)\right], & i = k. \end{cases} \tag{2}$$

The above messages have an intuitive interpretation: the "responsibility" $r(i,k)$ sent from x_i to x_k indicates how well x_k serves as the exemplar of x_i considering other potential exemplars for x_i, and the "availability" $a(i,k)$ sent from x_k to x_i indicates how appropriate x_i chooses x_k as its exemplar considering other potential samples that may choose x_k as their exemplar. The belief that image x_i selects image x_k as its exemplar is derived as the sum of the incoming messages

$$t(i,k) = a(i,k) + s(i,k). \tag{3}$$

After convergence of the message updates, the exemplar of point x_i is decided as x_k according to the criterion

$$k^* = \arg\max_k [t(i,k)]. \tag{4}$$

We choose AP as our exemplar selection algorithm due to its advantages in the following aspects: 1) its effectiveness in clustering has been shown in many tasks; 2) it simultaneously accomplishes the clustering and the selection of exemplars. Several other methods, such as K-means and spectral clustering, only cluster samples, and the centroids of the obtained clusters may not be real samples.

Most existing works model the similarity of two images based on their visual features [23], [24]. However, time is an important information clue for personal photos [9]. Since the photos in an album are captured by the same person, two photos that are temporally close will have high probabilities to record an identical scene or event[2]. Therefore, we integrate the visual and temporal information to compute the similarities of photos.

More specifically, the similarity between photos x_i and x_j is estimated as

$$s(i,j) = \alpha \exp\left(-\frac{\|v_i - v_j\|^2}{\sigma_v^2}\right) + (1-\alpha)\exp\left(-\frac{\|t_i - t_j\|^2}{\sigma_t^2}\right) \tag{5}$$

where v_i and t_i indicate the visual feature vector and timestamp of photo x_i, respectively, α is a weight factor between 0 and 1, and $\|\cdot\|$ denotes ℓ_2-norm. We perform AP with this similarity measure, and we name this method temporally consistent affinity propagation (TCAP) since the selected exemplars will be not only visually representative but also cover widely and diversely in time.

[2]Occasionally, there might be some situations where the photos are from recurring events that are temporally far apart. In this case, we can skip the temporal clue and only utilize the visual clue to estimate the image similarity.

As for the visual feature vector v_i, we observe that the images in a given photo album may have an entirely separate set of visual features that are common among the whole collection, where some images are similar in terms of color feature while other images are similar with respect to texture feature. Representing images with distinct visual features may have a great impact on the image similarity estimation and in turn will affect the results of the clustering and the exemplar selection. To reduce the sensitivity caused by the image representation, we look for different visual features that complementarily describe the visual contents of the images. More specifically, we represent each image with a complementary feature vector composed of the following three kinds of global visual features: 1) 225-dimensional block-wise color moment feature generated from 5-by-5 partition of the image, where the first three moments of three channels of CIE Luv color space are extracted [25]; 2) 128-dimensional Gabor texture feature, where we take eight scales and eight orientations of Gabor transformation and further use their means and standard deviations to represent the image [26]; 3) 75-dimensional edge direction histogram feature, which is taken as the line direction of an image and represents the global shape information of an image [27]. After normalization, we concatenate color, texture, and edge feature vectors into a 428-dimensional feature vector.

B. Incremental Exemplar Selection With Constrained Affinity Propagation

Now, suppose we have obtained a set of exemplars $\mathcal{L}$ for a photo album $\mathcal{U}$ in the early stages of tagging, and we wish to endow our system with the ability to incrementally select a new exemplar set conditioned on the previously selected ones. More specifically, we aim at obtaining a new exemplar set $\mathcal{L}'$ from $\mathcal{U}$.

To accomplish this task, the most straightforward approach is to directly preserve the initial exemplars in $\mathcal{L}$ and then re-run AP to select new exemplars from $\mathcal{U} - \mathcal{L}$. However, as indicated by the experiment results in Section VI-C, this method is sub-optimal since the existing exemplars in $\mathcal{L}$ are overlooked in the selection of new exemplars. Recall that in the AP algorithm [22], the exemplar configuration is determined through a set of hidden binary variables $\{c_{ij}\}_{j=1}^n$ associated with each data point $x_i \in \mathcal{X}$, in which $c_{ij} = 1$ denotes that x_i has selected x_j as its exemplar and $c_{ii} = 1$ indicates that the exemplar of data point x_i is itself. Here we propose a method that constrains the value of these exemplar configuration variable in the message update procedure. To govern the expected set of solutions, we clamp the value of c_{ij} for each data point in $\mathcal{L}$ to be const (i.e., $c_{ij} = 1$ when $i = j$ and $c_{ij} = 0$ when $i \neq j$) and seek for valid configuration of c_{ij} variables for other data points in $\mathcal{U} - \mathcal{L}$. This allows us to explicitly preserve the most representative exemplars obtained in the previous exemplar selection stage as well as to incrementally select new exemplars. In practice, the exemplar configuration constraints can be easily incorporated via an adaptation of the AP model and we refer to this method as constrained affinity propagation (CAP).

1) Optimization Objective of CAP: To facilitate the presentation, let $\mathcal{L} = \{x_{n+1}, x_{n+2}, \ldots, x_{n+m}\}$ and $\mathcal{R} = \mathcal{U} - \mathcal{L} = \{x_1, x_2, \ldots, x_n\}$ denote the set of selected

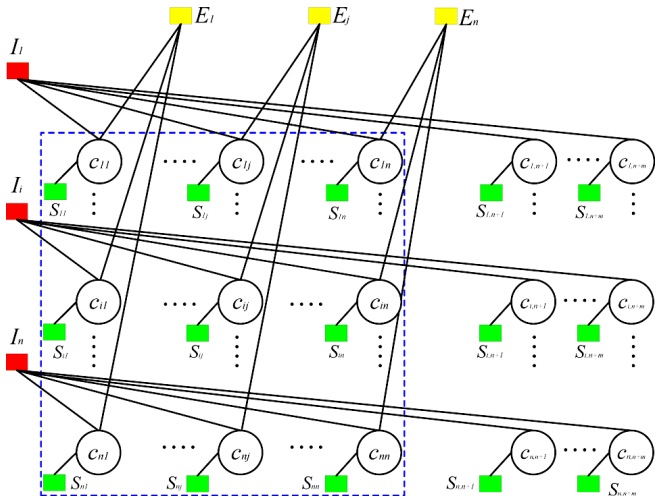

Fig. 3. Factor graph of CAP which is adapted from the factor graph of AP in [22]. The factor graph inside the dotted line is the same as the original AP model.

exemplars and the set of photos that have not been selected as exemplars, respectively. The rule of CAP can be established as that any photo in $\mathcal{R}$ will now be able to select each photo in $\mathcal{U}$ as its exemplar, but the photos in $\mathcal{L}$ can only choose themselves as exemplars.

Fig. 3 shows the graphical model of CAP. Note that the variable nodes c_{ij}'s whose indices of i lie in $\{n+1, n+2, \ldots, n+m\}$ (i.e., those corresponding to any data point $x_i \in \mathcal{L}$) have been removed from the factor graph, since the values of these c_{ij}'s variables are constrained to be constant with the only valid value to be 0 for $j \neq i$ and 1 for $j = i$. The I function nodes, which enforce the property that each point must choose exactly one exemplar, can be formally defined in (6). In addition, the E function nodes, which enforce the constraints that a point can only select its exemplar from those points that identify themselves as exemplars, are not required for data points in $\mathcal{L}$, as by fact, each data point in $\mathcal{L}$ has already selected itself as exemplar. The definition of E function is shown in (7):

$$I_i(c_{i1}, \ldots, c_{i,n+m}) = \begin{cases} -\infty, & \text{if } \sum_{j=1}^{n+m} c_{ij} \neq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$E_j(c_{1j}, \ldots, c_{nj}) = \begin{cases} -\infty, & \text{if } c_{jj} = 0 \text{ and} \\ & \quad \exists i \neq j \ s.t. \ c_{ij} = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Similar to the AP, we can define the overall objective function for CAP as follows:

$$\begin{aligned} S(c_{11}, \ldots, c_{n,n+m}) = & \sum_{i=1}^{n} \sum_{j=1}^{n+m} S_{ij}(c_{ij}) \\ & + \sum_{i=1}^{n} I_i(c_{i1}, \ldots, c_{i,n+m}) \\ & + \sum_{j=1}^{n} E_j(c_{1j}, \ldots, c_{nj}) \end{aligned} \tag{8}$$

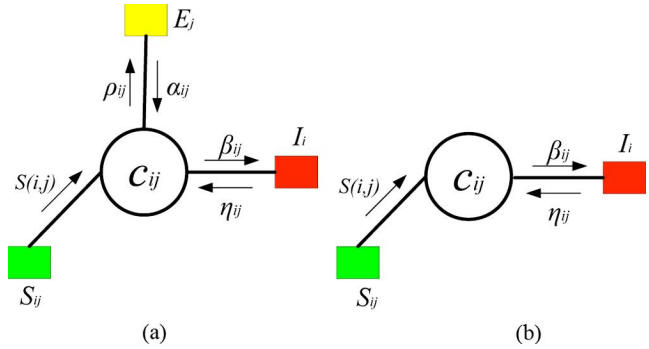

Fig. 4. Message passing between the variable nodes and function nodes, where there are (a) five messages types for the c_{ij} node whose index of j lies in $1, 2, \ldots,$ n and (b) three message types for the c_{ij} node whose index of j lies in $n+1, n+2, \ldots, n+m$.

where

$$S_{ij}(c_{ij}) = \begin{cases} s(i,j), & \text{if } c_{ij} = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

2) Message Propagation of CAP: The message propagation for the new model is similar to the original AP algorithm, whose derivation is based on max-sum algorithm over the factor graph in Fig. 3, in which the messages are passing between variable nodes and function nodes. There are five message types for variable node c_{ij} whose $j \in \{1, 2, \ldots, n\}$ and three message types for the c_{ij} node whose $j \in \{n+1, n+2, \ldots, n+m\}$, which are annotated in Fig. 4(a) and (b), respectively. By applying the message derivation strategy in [22], we can easily obtain the message update rules as follows:

$$\eta_{ij} = -\max_{k \neq j} \beta_{ik} \tag{10}$$

$$\rho_{ij} = s(i,j) + \eta_{ij} \tag{11}$$

$$\beta_{ij} = \begin{cases} s(i,j) + \alpha_{ij}, & j \leq n \\ s(i,j), & j > n \end{cases} \tag{12}$$

$$\alpha_{ij} = \begin{cases} \min\left[0, \rho_{jj} + \sum_{k \notin \{i,j\}} \max(\rho_{kj}, 0)\right], & i \neq j \\ \sum_{k \neq j} \max[\rho_{kj}, 0], & i = j. \end{cases} \tag{13}$$

By merging different messages above, we can obtain the "responsibility" message (i.e., ρ message) and "availability" message (i.e., α message) that are iteratively exchanged between the data points as shown in (14) and (15) at the bottom of the next page.

By comparing them with the massage update rules of AP, it can be found that the difference lies on the responsibility message: the previously selected exemplars in $\mathcal{L}$ will now have a direct impact on the values of the messages in CAP. More specifically, as indicated in the formulation of $r(i,k)$ message in (14), the similarity value between data point x_i and the previously selected exemplars in $\mathcal{L}$ will directly determine the value of $r(i,k)$. If x_i has high similarity values with respect to the previous exemplars, its $r(i,k)$ message tends to be small, which

will in turn reduce its possibility of being selected as a new exemplar. Actually, the new message formulations in (14) and (15) will tend to select those data points that have small similarity with the existing exemplars as new exemplars.

3) Exemplar Selection in CAP: The belief that image x_i selects image x_j as its exemplar is derived as the sum of the incoming messages

$$t(i,j) = a(i,j) + s(i,j). \tag{16}$$

Then the exemplar of image x_i is taken as

$$j^* = \underset{j\in\{1,2,\ldots,n+m\}}{\arg\max}\,[t(i,j)]. \tag{17}$$

Hence, the exemplar identification can be determined.

C. Discussion

Now we need to clarify why the proposed CAP algorithm can obtain better results in the exemplar selection task than the re-run of AP algorithm in which the new exemplars are directly selected from the remaining non-exemplar samples without considering the exemplars selected in the previous rounds. It is worth noting that the exemplar selection in the initial exemplar selection stage aims to discover the most representative samples in a collection, which actually reflect the individual recurrent visual patterns in an image collection. Each of the obtained visual patterns summarizes a set of images within a cluster. In the next round, if we simply skip these initially selected exemplars and perform exemplar selection in the remaining samples, the selected exemplars will tend to repeatedly take on the visual patterns obtained previously. The problem of involving such an exemplar selection procedure into our semi-automatic photo tagging scheme is that it may always provide visually similar images for manual labeling, which, on the one hand, introduces a lot of redundant labors, and, on the other hand, significantly affects the visual diversity of the obtained training sets that is critical to the learning speed of the tag inference algorithm. On the contrary, as aforementioned, the exemplars selected by the CAP algorithm will have small visual similarity with respect to the previously selected exemplars. Therefore, the obtained exemplars at each round will be distinct from the existing exemplars, which results in a diversified image training set for the tag inference.

V. Tag Inference

Now we introduce how to infer the tags of the rest of the photos based on all labeled exemplars. Denote by $\Omega = \{t_1, t_2, \ldots, t_m\}$ the set of appeared unique tags in the tagged exemplar set. Denote by $\mathbf{y}(x_i)$ the tag membership vector for photo x_i, in which the kth entry indicates the membership of tag t_k to the photo, i.e., $y_k(x_i) = 1$ if t_k is relevant to x_i and otherwise $y_k(x_i) = -1$. Thus, our task is to estimate $\mathbf{y}(x_i)$. The most intuitive approach is to directly assign the tags of each exemplar to all data points who have identified it as their exemplar (we call it naive tag assignment). However, this method heavily relies on the performance of exemplar selection algorithm and neglects the different distances of photos to the exemplars. Therefore, here we adopt a tag propagation method, which is closely related to a graph-based semi-supervised learning approach [10]. The method works by iteratively propagating the tags of each photo to others and holding the tags of exemplars. The process is illustrated in Algorithm 2.

Algorithm 2: Iterative Tag Propagation Algorithm: **Input**: Similarity matrix $\mathbf{W}$, diagonal matrix $\mathbf{D}$ with $D_{ii} = \sum_j W_{ij}$.
Output: $\mathbf{Y}$.

1) Initialize the tag membership matrix $\mathbf{Y}$.
2) Update matrix $\mathbf{Y} = \mathbf{D^{-1}WY}$.
3) Clamp the tags of exemplars, i.e., let $\mathbf{y}_i = \mathbf{y}(x_i)$ if x_i is an exemplar, where $\mathbf{y}_i$ is the ith row of $\mathbf{Y}$.
4) Repeat from step 2 until $\mathbf{Y}$ converges.

The process will converge to the solution of the following optimization problem [10]:

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j=1}^{n} W_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2 \\ \text{s.t.} \quad & \mathbf{y}_i = \mathbf{y}(x_i) \quad \text{if } x_i \text{ is an exemplar.} \end{aligned} \tag{18}$$

Here we also adopt the similarity measure that explores both visual and temporal clues—see (5)—i.e., $W_{ij} = s(i,j)$. Based on the estimated $\mathbf{Y}$, we can easily obtain the binary tag membership by setting a threshold, i.e., $y_k(x_i) = 1$ if the kth entry in $\mathbf{y}(x_i)$ is above 0, and otherwise $y_k(x_i) = -1$.

VI. Experiments

A. Experimental Settings

To evaluate the proposed tagging scheme, we conduct experiments with 16 different personal albums that are collected from Flickr, each of which is actually a personal photo set that includes a group of images under the same theme and thus can be utilized as realistic personal photo albums. Specifically,

$$r(i,k) = s(i,k) - \max\left[\max_{l\in\{1,\ldots,n\}\setminus\{k\}}[s(i,l)+a(i,l)], \max_{l\in\{n+1,\ldots,n+m\}\setminus\{k\}} s(i,l)\right] \tag{14}$$

$$a(i,k) = \begin{cases} \sum_{l\neq k} \max[r(l,k),0], & i = k \\ \min\left[0, r(k,k) + \sum_{l\notin\{i,k\}} \max[0, r(l,k)]\right], & i \neq k \end{cases} \tag{15}$$

TABLE I
NUMBER OF PHOTOS IN EACH ALBUM

Album Name	Image Numbe
Bombay	719
China	500
Germany	163
Hong Kong	136
Korea	493
London	204
Long Exposure	185
Manasquan	333
Nature	183
New Book	157
New York	117
Paris	265
Shanghai	365
Sunrise	286
Thailand	500
Wildlife	1,221

the photos involved in these 16 photo albums are captured at different locations around the world and contain diverse content, including the records of cityscape, landscape, wildlife, etc. Table I illustrates the number of photos in these albums.

We use the normalized 428-dimensional feature vector described in Section IV-A to represent each image in these albums. In the feature extraction stage, since many of the photos are with very high resolution, we scale each photo to fix its width to 240 pixels to speed up the feature extraction.

The performance of our proposed tagging scheme is evaluated by comparing the tagging results with the ground-truth tags. The ground-truth tags of the photos are established by ten volunteers as follows: for each album, the photos are exhaustively tagged by a volunteer. In this way, there are 7.62 ground-truth tags associated with each photo in average. For each photo, we estimate precision, recall, and F1-measure measurements of the tags obtained with the proposed tagging scheme. Then we average the F1-measure of all photos in an album to evaluate the tagging performance on the album. Finally, we average the F1-measure of all albums and it is adopted as the performance evaluation measurement in this work.

B. Evaluation of Exemplar Selection and Tag Inference

In this experiment, we aim to justify the effectiveness of 1) TCAP in the exemplar selection task; and 2) tag propagation in the tag inference task. We select a set of photos for manual tagging via different exemplar selection methods and then infer the tags of the rest of the photos automatically. For comparison, we also ask the volunteers to implement the naive batch tagging for each album, i.e., entering a set of tags for the whole album, and the tagging performance obtained with this method will be adopted as the baseline result.

We compare the following five exemplar selection methods.

- TCAP. We select the n samples with maximal $t(i,j)$ values [see (3)] as the exemplars obtained by TCAP. The parameter σ_v is empirically set to the median value of the pairwise Euclidean distances of all samples, and the parameter σ_t is empirically set to 1 h [see (5)]. The parameter α is simply set to 0.5.
- AP. We employ the similarity estimated with only visual features, and the n samples with maximal $t(i,j)$ values are employed as exemplars.
- Random exemplar selection. We randomly select n samples from the photo album as exemplars.
- K-means clustering. For each cluster, the sample that is closest to the mean vector is selected as exemplar and the number of clusters is set to n.
- Spectral clustering [28]. For each cluster, the sample that is closest to the mean vector is selected as exemplar and the number of clusters is set to n.

We apply each method to obtain n exemplary photos and then perform manual tagging for each photo. The parameter n is set to be 20; thus, a modest manual cost can be maintained.

For tag inference, we compare the following three methods.

- Naive tag assignment, i.e., directly assign the tags of each exemplar to all images that has identified it as their exemplars. As aforementioned, this method essentially summarizes the methodology of state-of-the-art interactive photo album tagging systems.
- Tag propagation, where the graph-based label propagation algorithm is adopted to infer the tags of unlabeled images.
- SVM. We employ the exemplar images labeled with/without a certain tag as positive/negative training samples and then train an SVM classifier with RBF kernel as the prediction model of the given tag. The libSVM toolkit [29] is utilized to implement the classification, where the best parameter settings for γ and C are determined from the interval of $[2^{-5}, \ldots, 2^{5}]$.

According to different combinations of the exemplar selection and tag inference methods, we will compare 15 methods (3×5) in all in the experiments. Table II illustrates the performance comparison. From the results, we can see that the AP method outperforms the K-means and spectral clustering methods with either naive tag assignment, SVM, and tag propagation. In addition, the best result is obtained by the TCAP together with tag propagation, and this indicates the effectiveness of tag propagation and the integration of temporal information. It is also worth noting that the tagging performance for the baseline method is too low to be accepted although this approach needs the least labor cost. Note that the selected exemplars (20 photos) only occupy a very small portion of each photo album (5.5% in average). Thus, the proposed tagging approach significantly reduces the human efforts in comparison with the exhaustive tagging, and fairly high tagging performance can still be maintained (average F1-measure: 0.7066).

C. Evaluation of CAP

An important property of our proposed tagging scheme is its flexibility. Specifically, the tagging accuracy can be gradually improved if users perform the tagging procedure iteratively. In this subsection, we evaluate the performance of the semi-automatic tagging scheme when the iterative tagging processes are employed. The purpose of this experiment is two-fold: 1) we demonstrate that the incrementally selected exemplars obtained by CAP algorithm can be well utilized as representative photos to benefit the follow-up tag inference, and 2) we also justify the performance improvement by introducing the iterative tagging into the semi-automatic tagging scheme. We compare CAP-based incremental exemplar selection with K-means clustering-based, spectral clustering-based, re-running of

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT EXEMPLAR SELECTION AND TAG INFERENCE METHODS

Method	Ave. F1-measure
TCAP + Tag Propagation	***0.7066***
AP + Tag Propagation	0.6831
K-means + Tag Propagation	0.6573
Spectral Clustering + Tag Propagation	0.6523
Random + Tag Propagation	0.6324
TCAP + SVM	0.6939
AP + SVM	0.6701
K-means + SVM	0.6619
Spectral Clustering + SVM	0.6432
Random + SVM	0.6111
TCAP + Naive Tag Assignment	0.6216
AP + Naive Tag Assignment	0.6185
K-means + Naive Tag Assignment	0.5607
Spectral Clustering + Naive Tag Assignment	0.6178
Random + Naive Tag Assignment	0.5230
Baseline	0.2340

TCAP (RTCAP)-based, and randomly incremental exemplar selection methods, respectively. For the last four methods, the incremental exemplars are selected from the photo album in which the previously selected exemplars have been excluded.

Note that the incremental exemplar selection procedure in our proposed semi-automatic tagging scheme is intertwined with the tag inference model. To fairly evaluate the performance of CAP in the incremental exemplar selection task, we use both SVM and the graph-based tag propagation as the tag inference model, respectively. Only if the CAP algorithm shows good performance on both SVM model and tag propagation model can we draw a conclusion that the CAP algorithm is effective in the incremental exemplar selection task. In this experiment, we utilize the tagging results obtained in Section VI-B as the initial tagging results for different methods, i.e., the results of random selection, K-means, spectral clustering, and TCAP along with tag propagation/SVM are applied as the initial results before iterative tagging procedure.

Then at each round of iteration, we select and manually tag ten new exemplars, and then predict the tagging performance of the previous iteration on this incrementally selected exemplar set. If the performance is not satisfying, the newly selected exemplars will be incorporated into the previously tagged exemplars and a tag inference process will be implemented. Such a round will be repeated step-by-step.

Fig. 5 illustrates the tagging performance obtained with different incremental exemplar selection methods in terms of average F1-measure on all photo albums, where Fig. 5(a) is based on graph-based tag propagation and Fig. 5(b) is based on SVM. Specifically, we limit the round of iteration to be ten and report the tagging performance at each step. It is interesting to note that, whatever the tag inference model (SVM or tag propagation) is applied, CAP has the best performance compared with other exemplar selection methods among the ten iterations, especially at the early stage of iteration. This is due to the fact that the CAP algorithm has taken the interaction between previously selected exemplars and potentially new exemplars into account, and this confirms the effectiveness of CAP-based incremental exemplar selection method. Fig. 6 illustrates some sample images and their tags after the iterative tagging procedure.

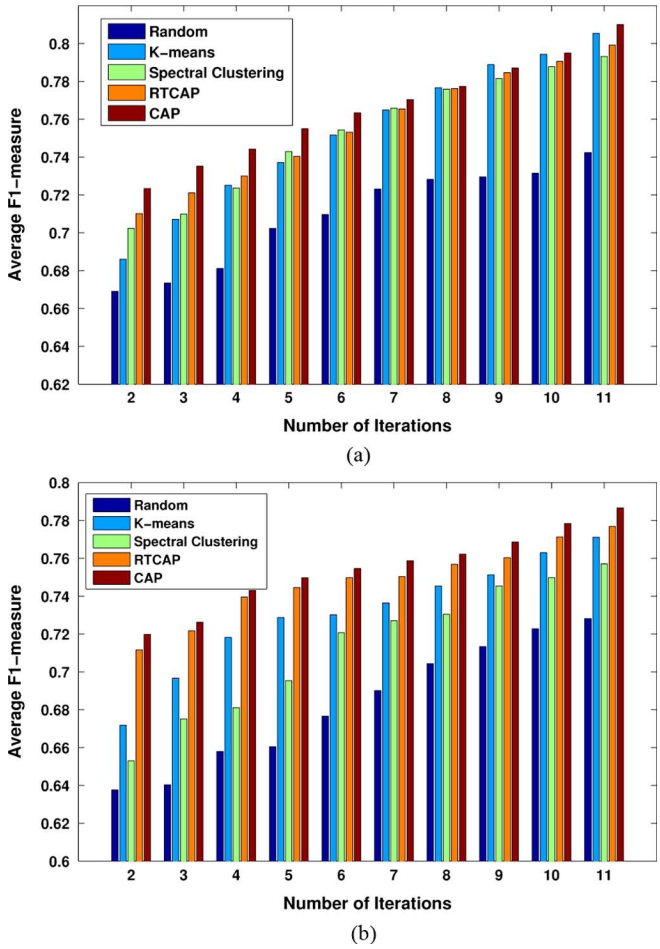

Fig. 5. Tagging performance at each round in the iterative tagging procedure, where (a) graph-based tag propagation and (b) SVM are utilized as tag inference model, respectively. (a) Graph-based tag propagation. (b) SVM.

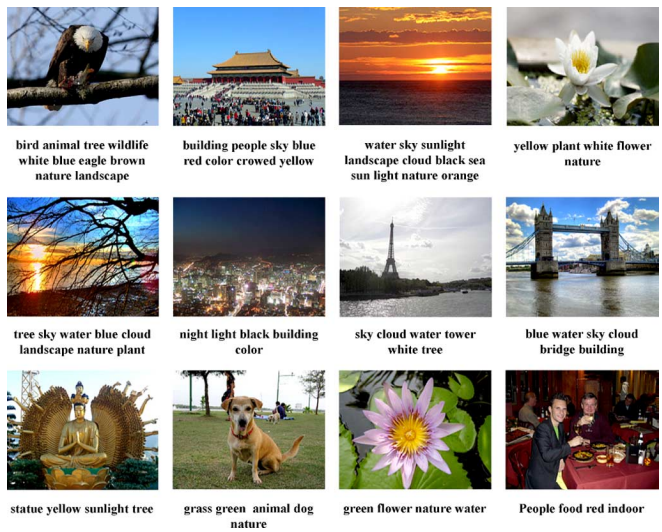

Fig. 6. Several exemplary tagging results after the iterative tagging procedure.

We also report the prediction accuracy of automatic tag inference on the incrementally selected exemplars at each round, and the results can be illustrated in Fig. 7. We can find that the

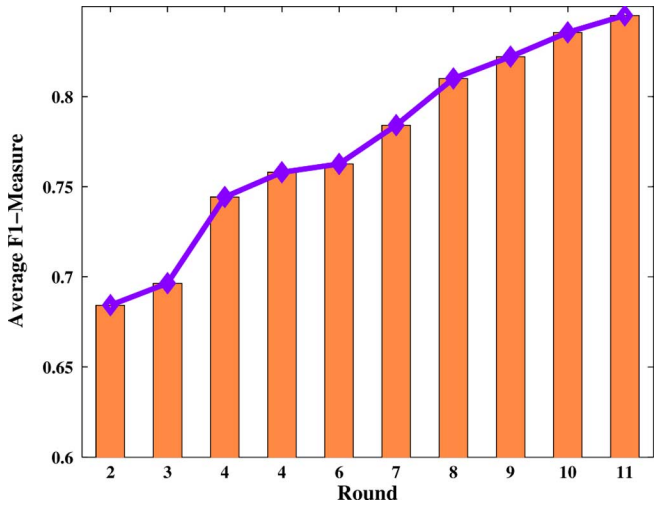

Fig. 7. Tagging performance on the incremental exemplar set at each iteration.

TABLE III
DIFFERENCE BETWEEN TAG ACCURACY ON EXEMPLAR SET AND ON WHOLE PHOTO ALBUM IN TERMS OF AVERAGE F1-MEASURE

Round	|Difference|
2	0.0391
3	0.0388
4	0.0001
5	0.0029
6	0.0008
7	0.0137
8	0.0326
9	0.0350
10	0.0406
11	0.0349

prediction accuracy on each incremental exemplar set keeps increasing as the tagging round proceeds, and the trend is very similar to the changing of tagging accuracy on the whole photo album in Fig. 5. We also calculate the difference between tagging performance on the incremental exemplar set and on the whole photo album in Table III. As can be observed, the difference between two kinds of tagging performance is small. This confirms that the tagging accuracy on the exemplar set can be utilized as the prediction of tagging performance on the whole photo album, and validates the effectiveness of the termination strategy for the iterative tagging process.

D. Computational Cost

Computational cost is also one of the main considerations for the web scenarios. The cost in each iteration of our proposed semi-automatic tagging scheme consists of the following four parts: 1) feature extraction; 2) similarity estimation; 3) temporally consistent AP-based exemplar selection; and 4) tag propagation.

In our practice, it costs about 0.062 s to extract features from each photo. For an album, the similarity estimation, exemplar selection, and tag propagation averagely cost 1.7, 8.5, and 2.1 s, respectively. All these time costs are recorded on a PC with Pentium 4 3.0-G CPU and 1-G memory. We can see that the costs are fairly low, and our user study results demonstrate that they are tolerable (in fact, these time costs will be much less than the procedure of uploading photos if we apply the tool on photo sharing websites, and the feature extraction and similarity estimation can be implemented during the uploading process).

VII. CONCLUSION

We have proposed a semi-automatic tagging scheme for personal photo albums, which achieves a good trade-off between manual efforts and tag performance. The proposed scheme is flexible in the sense that the tagging performance can be dynamically adjusted according to users' satisfaction. In this scheme, a set of exemplary photos are first selected from a personal photo album for manual tagging via a temporally consistent affinity propagation algorithm, and the tags of the rest of the photos can be automatically inferred. Then an additional set of exemplars are incrementally selected for manual tagging with the proposed constrained affinity propagation algorithm, based on which the tagging performance at the previous step can be estimated. A further round of tag inference will be implemented when the tagging performance cannot meet users' requirements. This process repeats until a satisfactory tagging performance is achieved. Our empirical results on multiple photo albums have demonstrated the effectiveness of the proposed scheme.

REFERENCES

[1] Flickr. [Online]. Available: http://www.flickr.com.
[2] [Online]. Available: http://en.wikipedia.org/wiki/Flickr.
[3] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.
[4] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2003, pp. 119–126.
[5] ALIPR. [Online]. Available: http://alipr.com/.
[6] R. Yan, A. Natsev, and M. Campbell, "A learning-based hybrid tagging and browsing approach for efficient manual image annotation," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
[7] L. Cao, J. Luo, and T. Huang, "Annotating photo collections by label propagation according to multiple similarity cues," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 121–130.
[8] L. Cao, J. Luo, H. Kautz, and T. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 208–219, Feb. 2009.
[9] J. Jia, N. Yu, and X.-S. Hua, "Annotating personal albums via web mining," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 459–468.
[10] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Machine Learning*, 2003, pp. 912–919.
[11] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems*, 2007, pp. 971–980.
[12] L. Kennedy, S.-F. Chang, and I. Kozintsev, "To search or to label? Predicting the performance of search-based automatic image classifiers," in *Proc. ACM Workshop Multimedia Information Retrieval*, 2006, pp. 249–258.
[13] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proc. 18th ACM Int. Conf. World Wide Web*, 2009, pp. 351–360.
[14] B. Sigurbjörnsson and R. Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th ACM Int. Conf. World Wide Web*, 2008, pp. 327–336.
[15] H. Chen, M. Chang, P. Chang, M. Tien, W. Hsu, and J. Wu, "Sheepdog-group and tag recommendation for Flickr photos by automatic search-based learning," in *Proc. 15th ACM Int. Conf. Multimedia*, 2008, pp. 737–740.
[16] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. 9th ACM Int. Conf. Multimedia*, 2001, pp. 107–118.

[17] D. Liu, X.-S. Hua, L. Yang, and H.-J. Zhang, "Multiple-instance active learning for image categorization," in *Proc. 15th Int. Multimedia Modeling Conf.*, 2009, pp. 239–249.
[18] W. Jiang, S.-F. Chang, and A. Loui, "Active context-based concept fusion with partial user label," in *Proc. IEEE Int. Conf. Image Processing*, 2008.
[19] B. Suh and B. Bederson, "Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition," *Interact. Comput.*, vol. 19, no. 4, pp. 524–544, 2007.
[20] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang, "Easyalbum: An interactive photo annotation system based on face clustering and re-ranking," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2007, pp. 367–376.
[21] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 319, no. 5963, pp. 972–976, 2007.
[22] I. Givoni and B. Frey, "A binary variable model for affinity propagation," *Neural Computat.*, vol. 21, no. 6, pp. 1589–1600, 2009.
[23] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua, "Finding image exemplars using fast sparse affinity propagation," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 639–642.
[24] W. Chu and C. Lin, "Automatic selection of representative photo and smart thumbnailing using near-duplicate detection," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 829–832.
[25] M. Stricker and M. Orengo, "Similarity of color images," *Proc. SPIE, Storage and Retrieval for Image and Video Databases*, pp. 381–392, 1995.
[26] B. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
[27] A. Vailaya, A. Jain, and H.-J. Zhang, "On image classification: City images vs. Landscapes," *Pattern Recognit.*, vol. 31, no. 12, pp. 3–8, 1998.
[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[29] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/ libsvm.

Dong Liu is currently pursuing the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.

From January 2009 to November 2010, he worked as a research engineer in the Department of Electrical and Computer Engineering, National University of Singapore. Prior to this, he worked as a research intern in the Internet Media Group at Microsoft Research Asia for two years. His research interests include multimedia information retrieval and analysis, machine learning, and computer vision. He has published over ten technical papers in the above areas.

Mr. Liu won the Microsoft Research Asia Fellowship in 2009–2010.

Meng Wang (M'09) received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 1999 and 2008, respectively.

From 2008 to 2010, he was with Microsoft Research Asia as a research staff member. From 2010, he joined AKiiRA Media Systems Inc. as a research scientist. His current research interests include multimedia content analysis, management, search, mining, and large-scale computing. He has authored over 70 papers and book chapters in these areas.

Dr. Wang served as an editorial board member, guest editor, or program committee member of numerous international journals and conferences. He received the best paper award in the ACM International Conference on Multimedia 2009 and the best paper award in the International Multimedia Modeling Conference 2010. He is a member of ACM.

Xian-Sheng Hua (M'05) received the B.S. and Ph.D. degrees in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively.

Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a lead researcher with the Internet Media Group. His current research interests include video content analysis, multimedia search, management, authoring, sharing, and advertising. He has authored more than 130 publications in these areas and has more than 30 filed patents or pending applications. He is an adjunct professor at the University of Science and Technology of China, Hefei, China.

Dr. Hua serves as an associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA and an editorial board member of *Multimedia Tools and Applications*. He won the Best Paper Award and the Best Demonstration Award at ACM Multimedia 2007 and also won the TR35 2008 Young Innovator Award from the MIT Technology Review. He is a member of the ACM.

Hong-Jiang Zhang (M'91–SM'97–F'03) received the B.S. in electrical engineering from Zhengzhou University, Henan, China, in 1982 and the Ph.D. degree in electrical engineering from the Technical University of Denmark, Lyngby, in 1991.

From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a research manager at Hewlett-Packard Labs, Palo Alto, CA, where he was responsible for research and development in the areas of multimedia management and intelligent image processing. In 1999, he joined Microsoft Research, where he is currently the managing director of the Advanced Technology Center in Beijing. He has coauthored/coedited four books, more than 350 papers and book chapters, numerous special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as more than 60 granted patents.

Dr. Zhang currently is on the editorial board of the PROCEEDINGS OF IEEE. He is a fellow of the ACM.