Image Tag Refinement With View-Dependent Concept Representations

Jianlong Fu, Jinqiao Wang, Member, IEEE, Yong Rui, Fellow, IEEE, Xin-Jing Wang, Member, IEEE, Tao Mei, Senior Member, IEEE, and Hanqing Lu, Senior Member, IEEE

Abstract—Image tag refinement is the task of refining initial tags of an image such that the refined tags can better reflect the content of the image and, therefore, can help users better access that image. The quality of tag refinement depends on the quality of concept representations that build a mapping from concepts to visual images. While good progress was made in the past decade on tag refinement, the previous approaches only achieved a limited success due to their limited concept representations. In this paper, we show that the visual appearances of a concept consist of both a generic view and a specific view, and therefore we can comprehensively represent a concept by two components. To ensure a clean concept representation, this representation is learned on clean click-through data, where noises are greatly reduced. In the framework, a coarse-to-fine image tag refinement is proposed, which: 1) first generates an efficient star graph to find candidate tags but missing in the initial tag list of an input image and 2) guided by this view-dependent concept representation, formulates a probabilistic objective function to eliminate irrelevant tags. Extensive experiments on two widely used standard data sets (MIRFlickr-25K and NUS-WIDE-270K) demonstrate the effectiveness of our approach.

Index Terms—Concept representations, image tag refinement, image tagging.

I. INTRODUCTION

WHILE the exact number of images in the world is unknown, it is definitely very large. For example, Flickr1 has 1.4 million photos uploaded every day and Instagram2 claims 40 million photos per day in 2013. Such a large number of images demand effective image accessing techniques. As evidenced by commercial search engines’ success, the most effective way for common users to access web images is through text. Image tagging therefore emerged as an active research area in the past 10 years. Before diving into the details of various image tagging techniques, we will first define some key terminologies used throughout this paper.

Fig. 1. Concept Hotel with its two types of visual appearances (i.e., generic view and specific view) in Flickr images. The result is obtained by a user study in Section VI-C3.

1) Image Tagging: Attempts to label an image with one or more human-friendly textual concepts to reflect the visual content of that image [1], [2]. The resultant tags form the tag list for this image. Note that image tagging can be done manually by a human, or automatically by an algorithm [3], or by combining the both [4]. However, the initial tag lists (e.g., the left tag list in Fig. 2) are often imperfect so that a postprocess is usually needed to refine the result.

2) Image Tag Refinement: Tags in a tag list may be imprecise for that image, and some relevant tags may be missing from the tag list. The process of removing imprecise tags and adding incomplete tags is called image tag refinement, and is the focus of this paper. If we refer to Fig. 2, the input to image tag refinement is an image with its initial tag list, and the output is the refined tag list. Note that the three output tag lists on the right are more refined than the initial tag list on the left.

While image tagging and tag refinement are very important to help users access images, how to do it accurately is not easy. As image tagging is the mapping from concepts to images, the accuracy of image tagging and image tag refinement depends on how well we can learn and build the concept representation. Let us take Fig. 1 as an example. The concept Hotel is one of the popular concepts among the billion-scale images in Flickr. It is clear that the Hotel images correspond to two

types of visual appearances (called two views in this paper), i.e., \textit{generic views} and \textit{specific views}. We can observe that:

1) the generic view of the concept \textit{Hotel} has generic and consistent appearances that can be learned by parametric models;

2) the specific view of the concept \textit{Hotel} has diverse appearances and reflects specific aspects of \textit{Hotel}, e.g., customers, waitresses, tables, beverages, and floors.

From the above observation, we can see that there are complex views embedded in a concept, where generic views represent major visual patterns and specific views represent sparse and diverse visual patterns. Therefore, to reduce the imprecise and incomplete tags for an input image in image tag refinement tasks, the visual representation for concepts should comprehensively consist of both the generic view and the specific view. Then, the presence (or not) of a concept for an image can be determined by discriminating if the image belongs to the generic view, the specific view, or other irrelevant visual patterns for the concept.

While the research on image tag refinement has been studied intensively in the past decade, how to effectively learn the concept representation is often overlooked in existing works. Most existing tag refinement approaches \cite{3,5,6,7,8,9} adopt the idea of collective intelligence. They determine the relevance between a concept and an input image by calculating a similarity score between the input image and its visual neighbors in a training set, e.g., images collected from Flickr, where these visual neighbors have been already associated with the concept. These approaches, though they can be used for the specific views that are difficult to be modeled by parametric methods, cannot comprehensively learn the concept representation for the generic views, because the following two strong assumptions in these approaches are often hard to secure in real applications. On the one hand, the training set should be large enough so that we can always find similar enough or even duplicate visual neighbors to ensure a reliable tag propagation to an input image \cite{3,9}. On the other hand, the training set should be well annotated \cite{8}, and thus noisy tags cannot be propagated to the input image. However, the size of a training set is limited. Irrelevant tags can be propagated to an input image through its visual neighbors due to the well-known semantic gap, especially when these visual neighbors are not similar enough to the input image. Besides, most works adopted Flickr as the training set due to its rich vocabulary and easy accessibility. However, Flickr is far less to be a well-annotated training set, as half of the user-contributed tags are noises to image content \cite{11,12}.

To address the above issues, we present in this paper a novel view-dependent concept representation approach with click-through data to refine image tags. First, we specially design a method to discriminate the generic view and specific view of a concept, and utilize model-based components and model-free components to represent the two views, respectively. The model-based components ensure a good generalization ability when we cannot find the similar enough or duplicate visual neighbors in training sets, while the model-free components complement the loss of the discriminative power of the model-based components. Note that this proposed approach is different from training a single model for both the two views merely using model-based components, because the specific view is so diverse and sparse that cannot guarantee a reliable model-based representation. Second, we formulate the model-based and model-free components into a unified probabilistic objective function. Then, the task of image tag refinement is further converted into an energy minimization problem that can be solved efficiently by graph cuts. Third, we leverage the click-through data collected from a commercial search engine as training data, and thus the noises can be greatly reduced.

To the best of our knowledge, our proposed approach is the first: 1) to specially consider the concept representations from generic views and specific views in the task of image tag refinement; 2) guided by this view-dependent representation, to formulate the appearance term of an objective function as a combination of model-based and model-free components to boost the performance of image tag refinement; and 3) to collect click-through data as training data with a total of 0.38 million tags, which is the largest vocabulary set in this field.

The rest of this paper is organized as follows. Section II describes the related works. Section III presents the learning of the view-dependent concept representations. Once these concept representations are learned, a coarse-to-fine framework for image tag refinement is proposed in Section IV. Section V presents the implementation details. Section VI provides extensive evaluations, and we conclude this paper in Section VII.
II. RELATED WORK

A large body of works for concept representation in the field of image tagging and tag refinement proceed along two dimensions, i.e., model-based and model-free approaches. One class of works adopts the idea of collective intelligence, which is a type of model-free approach. For example, Li et al. [13], [14], Wang et al. [15], and Guillaumin et al. [16] directly accumulate neighbor votes received from the visual similar images of the input image for each tag. Later, to estimate the relevance between a tag and an input image, the approaches of content-based tag refinement [17] and tag ranking (TR) [5] propose to use Gaussian function to measure the similarity between the input image and training images who have already been annotated by the tag. This type of measurement also has a very intuitive explanation. Tags of the input image are determined by the soft voting from its visual neighbors. Image retagging [7] and low-rank (LR)-based image tag refinement approaches [6], though effective by imposing some reasonable constraints (e.g., low rank, error sparsity, etc.), still rely on the idea of the sample-based neighbor voting, which inevitably decreases the discriminanceability due to the limited training data. Moreover, the assumption of the two approaches that the tags in Flickr provided by users are reasonably accurate is not justifiable, as previous research has reported that half of the user-contributed tags are noise to image content [11]. Therefore, due to the limited size and heavy noises of the training set in real applications, incomplete and imprecise tags still hinder users’ access to the images after the refinement by these model-free approaches.

One of the representative works of model-based concept representation is visual synset (Vsynset) [18], which shares similar motivation to ours. It applies multiclass one-versus-all linear Support Vector Machine (SVM) models, which are learned from the automatically generated Vsynsets from a large collection of the web images. Compared with the Vsynset, our proposed approach can discriminate different views and represent them with proper view-dependent representations, i.e., model-based components for generic views to enhance generalization ability, while model-free components for specific views as complements since they are too sparse and diverse. Besides, our approach is easily paralleled as the model-based and model-free components are learned per concept, with no dependency on others. Moreover, our approach achieves a larger vocabulary size with an increase of nearly 30% compared with Vsynset. Our approach is more scalable, which can add new concept representations into the existing vocabulary rather than retraining the visual models on the whole vocabulary again compared with using the discriminative models, such as SVM.

III. VIEW-DEPENDENT CONCEPT REPRESENTATIONS

In this section, we first describe the generic view and specific view in a mathematical way, and then describe the learning of concept representation based on the two views. In particular, on the one hand, as the generic view corresponds to generic and consistent appearance, its representation can be easily modeled by model-based representation. On the other hand, the specific view is the sparse appearance of a concept. This view can be better modeled by model-free representation, as any model-based representation will drop in performance because of the sparsity.

Considering both scalability and efficiency, we choose kernel density estimation approach to learn the two types of visual representations for each concept. Specifically, we select Gaussian kernel, as it is a reasonable assumption when we do not know the exact data distribution. Suppose \( \Psi \) is a set of training samples for a given concept. The model-based components are defined by a group of Gaussian kernels \( G_k \)

\[
\Psi_{mb} = \sum_{k=1}^{K_g} G_k(\psi | \omega_k) = \sum_{k=1}^{K_g} \pi_k \left( \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\psi - \mu_k)^T \Sigma_k^{-1} (\psi - \mu_k)} \right). \tag{1}
\]

\( K_g \) is the number of dominant kernels selected from the overall \( K \) kernels, which are jointly learned on the whole training set. \( \omega = \{ \omega_1, \ldots, \omega_K \} = \{ (\pi_1, \mu_1, \Sigma_1), \ldots, (\pi_K, \mu_K, \Sigma_K) \} \) denotes a parameter set. \( D \) is the dimension of an image \( \psi, \pi_k \) is the weight of each component, \( \mu_k = \mu_k^* \geq 0, \) and \( \sum_{k=1}^{K} \pi_k = 1. \mu_k \) is the mean vector, and \( \Sigma_k \) is the covariance matrix of the \( k \)th component. From the definition, we can see that each kernel \( G_k \) in \( \Psi_{mb} \) corresponds to a dense feature subspace with center \( \mu_k \). They are the representations of generic view. Once we have obtained model-based components, the model-free components can be defined by

\[
\Psi_{mf} = \left\{ \psi | \psi \in \Psi \cap \Psi \notin \sum_{k=1}^{K_g} G_k \right\}. \tag{2}
\]

Each \( \psi \) is an image instance in \( \Psi \), but is hard to be modeled by the joint probabilistic distribution \( \sum_{k=1}^{K_g} G_k \). In the following, we will first describe the learning procedure for model-based components, i.e., to determine \( \omega, K, \) and \( K_g \). Then, model-free components can be obtained by (2). We will present the details at the end of this section.

A. Determination of \( \omega \) and \( K \)

We adopt the expectation maximization (EM) algorithm [19] to estimate \( \omega \) on the whole image set \( \Psi \). The EM algorithm is a general method of finding the maximum-likelihood estimation of the parameters of an underlying distribution from a given data set. In each iteration, the algorithm guarantees to increase the loglikelihood \( \log G(\psi | \omega) \) and converge to a local maximum.

As \( K \) affects both the descriptive ability and the computational complexity, we apply an \( n \)-fold cross-validation approach to determine the best \( K \), instead of setting it empirically. For each one of a series of \( K \), we separate training data into \( n \) pieces and pick \( n - 1 \) pieces of data to estimate \( \omega \). Then, we apply the learned model to estimate the loglikelihood on the last one piece of data. The procedure is performed \( n \) times according to the general cross validation and produces an average loglikelihood. Zoran and Weiss [20]
have demonstrated that with the increasing number of $K$, the average loglikelihood increases but seems to be converging to some upper bounds. It also shows that a small number of $K$ is insufficient to achieve good performance. Therefore, we increase $K$ starting from 10 and stop when the difference of loglikelihoods between two successive calculations is smaller than a threshold $T_1$. The iteration process is shown as Algorithm 1.

### Algorithm 1: Determination of $K_g$ on the Overall Training Images

1: Initial $k = 10$
2: for any $n$-1 pieces of training data do
3: \hspace{0.5cm} $\omega$ is determined by EM algorithm
4: \hspace{0.5cm} loglikelihood $l(k)$ is calculated on the last one piece of data
5: end for
6: Average the loglikelihoods of $n$ times as $L(k)$
7: if $L(k) - L(k - 1) < T_1$ then
8: \hspace{0.5cm} stop
9: else
10: \hspace{0.5cm} $k = k + 1$, go back to step 2
11: end if

B. Determination of $K_g$

To obtain $K_g$, we consider that the connection among components in the mixture model is an important cue and apply random walk over these components to select the most dominant ones as the model-based representation. We formulate this connection as a fully connected graph and define a probability matrix $M$ among the $K$ components. Each element $M_{ij}$ in $M$ measures the relationship between two Gaussian components with regard to $i$ and $j$. $M_{ij}$ is defined by Kullback–Leibler divergence [21], which is shown as

$$M_{ij} = \frac{1}{2} \left[ \log \frac{\sum_j \pi_j}{\sum_i \pi_i} + \text{Tr} \left[ \sum_{j} \frac{1}{\pi_i} \sum_i - D \right] + (\mu_i - \mu_j)^T \sum_j (\mu_i - \mu_j) \right].$$  

(3)

Then, a random walk process can be implemented using power iteration [22]

$$\pi M = \hat{\pi}$$  

(4)

where we use weight vector $\pi = \{\pi_1, \ldots, \pi_K\}$ of the original mixture model as an initial vector for iteration. The iterative process will stop if the $L_2$ distance of $\pi$ between two successive iterations is smaller than a threshold $T_2$. At this point, we can obtain a updated weight vector $\pi'$ in stationary state. We rank the above $K$ mixture models according to $\pi'$ in a descending order and select the top $K_g$ models as reliable model-based components by hard truncation at the largest gap.

We have presented the principle of the selection of model-free components in (2). Once we obtain $\omega$, $K$, and $K_g$, we calculate image responses on the learned model-based components for all instances in $\Psi$. Image instances with low responses, e.g., below a threshold $T_3$, are considered as sparse visual appearances. We collect these low-response images together, mark them by flags, and keep each of them as a model-free component in their original feature space. We will describe how to leverage this view-dependent representation to refine image tags in the following sections and demonstrate its effectiveness in experiments.

IV. FRAMEWORK OF IMAGE TAG REFINEMENT

In this section, we propose a coarse-to-fine framework to refine image tags based on the learned view-dependent concept representation. The whole framework is shown in Fig. 3. First, to expand the relevant but missing tags and eliminate the obviously imprecise tags of an input image, we propose an efficient semantic-visual embedded star graph based on the initial user-contributed tags and visual neighbors (in training set) of the input image, as shown in Fig. 3(a) and (b). Then, tags of images in the star graph can determine a candidate vocabulary, which is the coarse refined result presented in Fig. 3(c). Second, given the candidate vocabulary, we propose a probability objective function to measure the alignment between the image and these candidate tags based on the view-dependent concept representations. Noted that we do not introduce additional tags in this step, we just select the most relevant tags from the candidate vocabulary. As shown in Fig. 3(d), the probability formulation is further decomposed into three terms of an energy function to represent the responses on the view-dependent representations (i.e., model-based and model-free terms) and the relationship between tags (i.e., the contextual priority term). The solution of the energy minimization problem is the final image tag refinement result, as shown in Fig. 3(e).

A. Semantic-Visual Embedded Star Graph

In the coarse step, to obtain a candidate vocabulary, we adopt a two-stage tag expansion and pruning scheme in both the semantic and visual spaces as in [15]. We denote an input image $x$ with its associated tag list $T = \{t_1, t_2, \ldots, t_n\}$. The initial tag list $T$ is usually imprecise and incomplete. Our aim is to eliminate and enrich the tags from $T$ to present a more accurate semantic description for image $x$.

For $x$, we first find its corresponding semantic-related image set $\Psi_s$ in the training set by searching the indexes in the inverted index with each tag $t_i$ in $T$ [Fig. 3(a)], shown as

$$\Psi_s = \{\psi \mid t(\psi) = t_i \ \forall t_i \in T\}$$  

(5)

where $\psi$ represents an image indexed in the inverted index whose tag, i.e., the entry in the index, is $t(\psi)$. We further construct a star graph for image $x$, where the leaves are visual $K$-nearest neighbor selected from the semantic-related image set $\Psi_s$ [Fig. 3(b)]. More details about the visual features can be found in Section VI-B. We denote the semantic-related and visual-similar image set as $\Psi_{sv}$, which is represented as

$$\Psi_{sv} = \{\psi \mid \psi \in \Psi_s \cap \psi \in \text{visual KNN} x\}$$  

(6)

where visual KNN represents images in the star graph, which are the visual $K$-nearest neighbors to image $x$. 
Fig. 3. Coarse-to-fine image tag refinement framework. (a) For the input image, we first find its semantic-related images in training sets by searching with its initial user-contributed tags. (b) Second, we build a star graph from the semantic-related images based on visual similarity. Both the semantic-related and visual-similar nearest neighbor images of the input image are marked by yellow rectangles in the image list in (a). (c) Third, a candidate vocabulary is obtained from the tags of images in the semantic-visual embedded star graph. (d) Finally, we apply the novel probability formulation guided by the view-dependent concept representation and (e) acquire the refined output. The coarse step refers to (a)–(c), while the fine step refers to (d) and (e).

To distinguish the original tag set $T = \{t_1, t_2, \ldots, t_n\}$, we use $Y = \{y_1, y_2, \ldots, y_m\}$ to denote the candidate vocabulary that is determined by the semantic-visual embedded star graph [Fig. 3(c)]. $Y$ can be obtained by searching images of $\Psi_{yo}$ in a forward index, shown as

$$Y = \{ \psi | \psi \in \Psi_{yo} \}.$$  

(7)

The merits of our semantic-visual embedded star graph are twofolds. First, we can find more relevant but missing tags via these semantic-related and visual-similar images. For instance, Sea and Sky do not appear in the initial tagging results of the image $x$ in Fig. 3(a), whereas the two relevant tags can be found in the candidate vocabulary shown in Fig. 3(c). The reason is that Sea and Sky are expanded into the candidate vocabulary of the image $x$ via the tags of its neighbor images, such as the third images of Hawaii and the second images of Kauai in the inverted index. Second, our approach guarantees that the images in the graph should be consistent with the visual content of image $x$. Thus, some noisy tags like 2008 are eliminated in the candidate vocabulary. Obviously, the coarse step can introduce some noisy tags due to the semantic gap, and we will resort to the view-dependent representations to solve this problem in the subsequent processing.

B. Unified Probability Formulation

In this section, we will specify the tag selection procedure from the candidate vocabulary with a combination of both model-based and model-free components. Let $z_i$ be a binary variable, which is related to $y_i \cdot z_i = 1$ if $y_i$ is a correct tag for image $x$ and otherwise $z_i = 0$. The problem with tag refinement is to infer the correctness of $y_i$ in $Y$. We call the refined result a hypothesis as $Z^* = \{z_1, z_2, \ldots, z_m\}$. Given an image $x$ with weight parameters $\theta$ for different components, the tag refinement task equals finding a hypothesis $Z^*$ that maximizes the probability $P(Z^*|x, \theta)$. The concrete form of $\theta$ will be presented in Section V. By the Bayesian rule [23], we can obtain

$$\arg \max_{Z^*} P(Z^*|x, \theta) \propto \arg \max_{Z^*} P(x|Z, \theta)P(Z|\theta)$$  

(8)

where $P(x|Z, \theta)$ measures the visual alignment of image $x$ given hypothesis $Z^*$ with the parameter $\theta$, which we refer to as an appearance term. $P(Z|\theta)$ is a prior term, which models the relationship between tags. For example, Sky and Clouds and Bridge and Water are likely to appear in the same image. We name the prior term as contextual priority.

The appearance likelihoods of $x$ corresponding to different tags are assumed to be independently and identically
distributed, hence $P(x|Z, \theta)$ can be decomposed as

$$P(x|Z, \theta) \approx \prod_{i=1}^{m} p(x|z_i, \theta). \tag{9}$$

The above decomposition is reasonable. For instance, some visual clues can be extracted from image $x$, which indicate the attribute of Sky because of the blue background. Meanwhile, we can also find a kind of visual appearance from image $x$, which reveals the attribute of Clouds due to the white color. The Sky model may prefer a blue component, while the Clouds model may prefer white in the color space. Therefore, though Sky and Clouds are semantically related, there is little evidence to show that their appearance models are related too.

In the contextual priority term, we take pairwise interactions as the basic element and higher order interactions are modeled indirectly through these pairwise interactions. For a hypothesis with $m$ tags, the contextual priority term can be decomposed as

$$P(Z|\theta) = \prod_{i=1}^{m} \prod_{j=1}^{m} p(z_i, z_j|\theta). \tag{10}$$

The pairwise interaction $P(z_i, z_j|\theta)$ indicates the joint probability that $y_i$ and $y_j$ appear in the image simultaneously.

By substituting (9) and (10) into (8), and applying the logarithm operation, the MAP problem equals the following energy maximization problem:

$$\arg \max_{Z, z_i \in [0, 1]} \sum_{i=1}^{m} f(x, z_i) + \sum_{i=1}^{m} \sum_{j=1}^{m} f(z_i, z_j) \tag{11}$$

where $f(x, z_i) = \log p(x|z_i, \theta)$ and $f(z_i, z_j) = \log p(z_i, z_j|\theta)$.

To present the appearance likelihoods of $x$ for a given tag, we measure the responses to both model-based and model-free components of the tag. Thus, the appearance term is further divided into two terms, and (11) is decomposed as

$$\arg \max_{Z, z_i \in [0, 1]} \sum_{i=1}^{m} f_{mb}(x, z_i) + \sum_{i=1}^{m} f_{mf}(x, z_i) + \sum_{i=1}^{m} \sum_{j=1}^{m} f(z_i, z_j) \tag{12}$$

where $f_{mb}(x, z_i)$ indicates the responses to model-based components of tag $y_i$ (denoted as MB_term), while $f_{mf}(x, z_i)$ reflects the responses to model-free components (denoted as MF_term). The two terms construct a comprehensive mapping from tag $y_i$ to image $x$ no matter $x$ is an image of generic views or specific views. We will give the concrete forms of the two terms in Section V. So far the results of image tag refinement can be obtained by solving the above optimization problem.

Besides, our framework can not only predict the presence (or not) of a tag but also discriminate the visual type of the input image with respect to the tag. The evaluation standard is defined as

$$z_i = 1 \text{ AND } MB_{\text{term}} > MF_{\text{term}}; \quad y_i \text{ is relevant and } x \text{ is generic view}$$

$$z_i = 1 \text{ AND } MB_{\text{term}} < MF_{\text{term}}; \quad y_i \text{ is relevant and } x \text{ is specific view}$$

$$z_i = 0; \quad y_i \text{ is irrelevant.} \tag{13}$$

V. IMPLEMENTATION DETAILS FOR IMAGE TAG REFINEMENT

In this section, we will present the concrete forms of each term in (12) via model-based and model-free components.

A. Representation of the Appearance Term

To obtain the concrete form of MB_term and MF_term, we apply Bayesian rule to (11) and obtain

$$f(x, z_i) = \log p(x|z_i, \theta) = \log \frac{p(z_i|x, \theta)p(x|\theta)}{p(z_i|\theta)} = \log \frac{p(x|\theta)}{p(z_i|\theta)} + \log p(z_i|x, \theta) \tag{14}$$

where the first and second terms are the concrete forms of $f_{mb}(x, z_i)$ (MB_term) and $f_{mf}(x, z_i)$ (MF_term), respectively. Then, the MB_term can be expanded as

$$f_{mb}(x, z_i) = \log \sum_{k} p(x|G_k, \theta)p(G_k) = \log \sum_{k} p(x|G_k, \theta)p(G_k)\tag{15}$$

where $G_k$ is the $k$th model-based component of tag $y_i$. $p(G_j)$ indicates the importance of each model-based component, which can be measured by the updated weight vector $\pi'$. $p(x|G_k, \theta)$ presents the response of image $x$ to model-based component $G_k$. $p(z_i|G_k, \theta)$ is estimated by the responses of the large set of training images of tag $y_i$ to the model-based component $G_k$. It can be measured as a bias of the model-based representation. In (15), we assume the parameter $\theta$ is a linear coefficient that reflects a weight of the responses to model-based components. The weight is specifically defined as $\theta_{mb}$. Therefore, the MB_term is further represented as

$$f_{mb}(x, z_i) = \theta_{mb} \left( \log \sum_{k} p(x|G_k)p(G_k) - \log \sum_{k} p(z_i|G_k)p(G_k) \right) \tag{16}$$

On the other hand, the MF_term is deduced as

$$f_{mf}(x, z_i) = \log \sum_{\psi \in \Psi_{svr}} p(z_i|\psi)p(\psi|x, \theta) \tag{17}$$

where

$$\Psi_{svr} = \{ \psi | \psi \in \Psi_x \cap \psi \in \Psi_{mf} \}. \tag{18}$$
\(\Psi_{x_0}\) is the image set we defined before and it covers semantic-related and visual-similar neighbors of image \(x\) in the star graph. To reflect the similarity with model-free components, \(\psi_t\) is picked up from \(\Psi_{x_0}\) and it should belong to a set of model-free components of tag \(y_i\). These model-free components have been marked by flags, as described in Section III. Here, the parameter \(\theta\) is also assumed as a linear coefficient that reflects a weight of the responses to model-free components. The weight is specifically defined as \(\theta_{mf}\). Therefore, the MF_term is further presented as

\[
f_{mf}(x, z_i) = \theta_{mf} \log \sum_{\psi_t \in \Psi_{mf}} p(z_i | \psi_t) p(\psi_t | x) \tag{19}\]

where \(p(\psi_t | x)\) investigates the similarity between \(\psi_t\) and \(x\). Since each \(\psi_t\) is one of the \(K\)-nearest neighbors of \(x\), we assume each \(\psi_t\) is similar enough to \(x\) and give an identical value to each \(p(\psi_t | x)\), such as one. \(\log \sum_{\psi_t \in \Psi_{mf}} p(z_i | \psi_t)\) reveals the occurrence probability of tag \(y_i\) among model-free components in the star graph, which is approximated using the frequency of \(y_i\) divided by the size of the candidate vocabulary (denoted as \(|Y|\)), then (19) is presented as

\[
f_{mf}(x, z_i) = \theta_{mf} \text{Num}(z_i)/|Y|. \tag{20}\]

In all, \(f_{mf}(x, z_i)\) is obtained by voting from model-free components. It complements the missing links from a concept to its sparse and less significant visual appearances. The larger the \(\text{Num}(z_i)\), the higher probability the tag \(y_i\) is correct. Finally, the appearance term is represented as

\[
f(x, z_i) = f_{mb}(x, z_i) + f_{mf}(x, z_i) \tag{21}\]

where the two aspects measure the alignment on model-based components and model-free components for image \(x\).

### B. Representation of the Contextual Priority Term

Following the principle of Google distance [24], we adopt a cooccurrence-based method to model the tag correlations. In the training set, each image can be associated with multiple tags. We assume if two tags appear simultaneously for a given image, they have a certain correlation. We model tag correlation \(v(i, j)\) between tags \(y_i\) and \(y_j\) and define it based on Google distance as

\[
v(i, j) = \frac{\log \text{Num}(y_i, y_j) - \log \text{Num}(y_i, y_j)}{\max\{\log \text{Num}(y_i), \log \text{Num}(y_j)\} - \log \text{Num}(y_i, y_j)} \tag{22}\]

where \(\text{Num}(y_i), \text{Num}(y_j),\) and \(\text{Num}(y_i, y_j)\) are the numbers of images containing tag \(y_i\), tag \(y_j\), and both of them, respectively. \(M\) is the total number of images in the training set.

For the contextual priority term \(f(z_i, z_j) = \log p(z_i, z_j | \theta)\) in (11), the parameter \(\theta\) is assumed as a linear coefficient that reflects a weight of tag correlation. The weight is specifically defined as \(\theta_c\). The learned tag correlation \(v(i, j)\) is leveraged to present the representation of \(\log p(z_i, z_j)\), and then we can obtain

\[
f(z_i, z_j) = \theta_c \log p(z_i, z_j) = \theta_c v(i, j). \tag{23}\]

### C. Optimization via Graph Cuts

Recalling (12), our aim is to optimize \(Z\) to obtain the maximum appearance likelihood of image \(x\) for candidate tags, and meanwhile consider the tag relationships where tags with a higher contextual priority should be selected or deserted simultaneously. Given an input image \(x\), the number of the possible hypotheses is of exponential order. In our problem, there are \(m\) candidate tags for \(x\) and the number of possible hypothesis is \(2^m\). To solve the problem efficiently, we rewrite (12) into an energy minimization problem and solve it by graph cut [25] algorithm. Specifically, we substitute (16) and (20) into (21), and then combine (21) with (23) to acquire the concrete representation of (12). Finally, the objective function can be rewritten as

\[
\min Z \sum_{i} - \bigg(\theta_{mb} \left(\log \sum_{k} p(G_k | z_i) p(G_k) - \log \sum_{k} p(z_i | G_k) p(G_k)\right)_{\text{MB_term}} + \sum_{i} \theta_{mf} \text{Num}(z_i)/|Y|\bigg)Z_{\text{MF_term}} + \sum_{i,j} \theta_c (z_i - z_j)^2 v(i, j)_{\text{C_term}} \tag{24}\]

where \(Z \in [0, 1]^m\) and \(\text{MB_term}, \text{MF_term},\) and \(\text{C_term}\) are calculated given an image \(x\) and its candidate vocabulary, and then the variable \(Z\) can be optimized. Naturally, the objective function is submodular, and thus the problem can be solved by the graph cuts.

### VI. EXPERIMENT

In this section, we first evaluated the proposed approach on two widely used standard data sets. Second, except for reporting the performance on image tag refinement, we provided examples of generic views and specific views detected by our approach for some concepts. Finally, we presented the complexity analysis of our approach and showed its good scalability for large-scale data.

#### A. Data Sets

The experiments were performed on two widely used large-volume image data sets that were both collected from Flickr. MIRFlickr-25K data set [26] contains 25,000 images with 1386 unique tags, while the NUS-WIDE-270K data set [11] contains 269,648 images with 5018 unique tags. To make fair comparisons, we followed [5] and [6], to perform a preprocessing procedure by matching tags with the Wikipedia thesaurus and removing tags whose occurrence numbers were below 50. As a result, 205 and 521 unique tags were obtained in MIRFlickr-25K and NUS-WIDE-270K, respectively.

We used two click-through data as training set. The first one is a public data set\(^3\) (denoted as Bing_Challenge in this paper) from MSR-Bing Image Retrieval Challenge. The second one

\[^3\text{http://web-ngram.research.microsoft.com/GrandChallenge/Datasets.aspx}\]
is a larger click-through data set (denoted as Bing_Large in this paper). We collected this data set from the data of one week in November 2013 from Bing Image Search. The process was conducted using the above 205 and 521 tags as queries. For the two data sets, an image was eliminated if its click count was less than 10 times. One image can be clicked many times by different queries. We preserved all the associated queries of an image as tags. These tags were stemmed, and stop words were removed. For efficient indexing, we organized training sets by two structures: 1) a forward index and 2) an inverted index. The forward index stored \(<\text{Image}, \text{TagList}>\) pairs, while the inverted index was designed as \(<\text{Tag}, \text{ImageList}>>\). Table I shows the details of the two training sets. In Bing_Large, there are totally 0.61 million images with 0.38 million tags, which to the best of our knowledge is the largest vocabulary in image tag refinement.

<table>
<thead>
<tr>
<th></th>
<th>Image Number</th>
<th>Tag Number</th>
<th># Image / Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing_Challenge</td>
<td>0.28 million</td>
<td>0.20 million</td>
<td>2512</td>
</tr>
<tr>
<td>Bing_Large</td>
<td>0.61 million</td>
<td>0.38 million</td>
<td>3637</td>
</tr>
</tbody>
</table>

B. Experiment Settings

Compared Approaches: The following approaches were compared for performance evaluation.

1) User Tags (UTs): The initial user-contributed tags collected from Flickr.
2) Visual and Semantic Consistency (VSC) [7]: VSC performs tag refinement based on the consistency between visual and semantic similarity.
3) LR [6]: An effective approach to optimize UTs by LR, content consistency, tag correlation, and error sparsity.
4) Random Walk With Restart (RWR) [17]: RWR formulates the refinement process as a Markov process and reranks the candidate annotations by a content-based method.
5) TR [5]: TR first estimates relevance scores based on probability density estimation, and then performs a random walk over a tag similarity graph to refine.
6) Vsynset [18]: Vsynset builds visual-similar and semantic-related synsets with weighted annotations, and uses linear SVM to predict the synset membership for unseen images.
7) Tag Completion (TC) [27]: TC performs the refinement problem as a search for the optimal tag matrix consistent with the observed tags and the visual similarity.
8) Augmented Feature SVM (AFSVM) [28]: AFSVM performs tag refinement by modeling images with augmented features, i.e., scores of an adapted classifier built by prelearned SVM classifiers.

Note that we selected these baselines, because they can represent two main streams in tag refinement approaches and some of them have achieved the strongest performance. One class of them (e.g., VSC, LR, RWR, TR, and TC) adopted the idea of collective intelligence. This type of approach cannot comprehensively learn a concept representation for the generic views. The other class (e.g., Vsynset and AFSVM) did not separate the two views and represent them by only one model-based component. This type of approach ignored the fact that the specific view is so diverse and sparse that cannot ensure a reliable model-based representation.

We followed [6] to extract the same feature representations, including color, texture, and edge distribution of 428 dimensions, for all compared approaches. For fair comparisons, we conducted RWR, TR, Vsynset, and AFSVM based on the results produced by our coarse step instead of the initial user-contributed tags, as these approaches cannot solve the problem of incomplete tags.

Evaluation Metrics: We adopted \(F\)-score as [6] for a detailed evaluation on the 18 (in MIRFlickr-25K) and 81 (in NUS-WIDE-270K) tags, which are provided as ground truth. \(F\)-score is defined as

\[
F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Besides, as relevant tags are more useful when appearing higher in a ranked list, we adopted Precision\(\@k\) at the position of one to five as measurement in NUS-WIDE-270K. Note that the two metrics were evaluated on the tags provided with ground truth. We also presented more accurate tags by case studies to show the superiority of the proposed approach.

C. Evaluation on MIRFlickr-25K

1) Parameter Setting: We first examined the coarse step performance, which relied on the semantic-visual embedded star graph. In coarse step, the recall is a more important criterion than precision. Because if we cannot find the missing tags at this stage, there is no chance to find them in the following steps. The noisy tags introduced in this step, however, can be eliminated by the proposed probability formulation in fine step. We selected a subset of MIRFlickr-25K including 5000 images to test the influences of the number of the semantic-visual neighbors. For the 18 tags in MIRFlickr-25K, the figure of average precision, recall, and \(F\)-score with the changes of this number is shown in Fig. 4(a). As our expectations, with the number increasing, we have more chances to find the missing but relevant tags, and consequently the recall becomes higher. However, the more neighbors, the more imprecise tags are introduced in the coarse result. Therefore, the precision shows a completely different trend. We can observe from the figure, 50 yields a modest result. Therefore, we empirically set the neighbor number as 50 in the following experiments for both efficiency and effectiveness.

In the fine step, there are three thresholds. We empirically set \(T_1\) as \(10^{-2}\), \(T_2\) as \(10^{-5}\), and \(T_3\) as \(10^{-3}\). Experiments show that the best model number in the mixture model varies from 25 to 40 for most tags. There are also three weights \(\theta_{mb}, \theta_{mf},\text{ and } \theta_e\) in the proposed probability formulation. We explored the influence of different parameter settings on the above selected subset. We empirically fixed \(\theta_e = 0.1\) and investigated the average \(F\)-score of tag refinement by tuning \(\theta_{mb}\) and \(\theta_{mf}\) from 0 to 1 with the step of 0.01. \(\theta_{mb}\) and \(\theta_{mf}\) control the percentage of the MB_term and MF_term as well.
as how much the appearance term incorporates the contextual priority term. Fig. 4(b) shows the impact of $\theta_{mb}$ and $\theta_{mf}$. An $F$-score of 0.507 is the best performance when $\theta_{mb} = 0.80$ and $\theta_{mf} = 0.61$, which are used in the following experiment.

2) Evaluation of Image Tag Refinement: Tables II and III show the performances of different approaches on different training sets. We can first observe that the results produced by refinement approaches consistently outperform the initial user-contributed results whichever training sets are used, which proves the effectiveness of image tag refinement.

We further verified the importance of clean training data in image tag refinement tasks by showing the results on different training sets. We can first observe that the results produced by refinement approaches consistently outperform the initial user-contributed results whichever training sets are used, which proves the effectiveness of image tag refinement.

We further verified the importance of clean training data in image tag refinement tasks by showing the results on different training sets. To achieve this goal, we selected four approaches all adopted the idea of collective intelligence, and implemented them using Flickr, Bing_Challenge, and Bing_Large as training sets, respectively. As shown in Table II, the four approaches (i.e., VSC, LR, RWR, and TR) showed the worst results when trained on noisy Flickr images. It shows the difficulty to discover a correct concept representation from the noisy training set, because these approaches heavily rely on tag correctness of most training samples. When leveraging the two clean click-through data sets, Bing_Challenge and Bing_Large, as the training data, the performances relatively increased 29.8%, 34.9% for VSC, 14.6%, 25.0% for LR, 17.3%, 28.4% for RWR, and 16.2%, 22.0% for TR. The significant improvement verified the effectiveness of click-through-based training sets in image tag refinement tasks. Furthermore, the results on Bing_Large are better than that on Bing_Challenge, as the larger training set for a concept and more comprehensive semantic relationship can construct a more reliable concept representation.

To demonstrate the superiority of the proposed view-dependent approach, we conducted another experiment by comparing three representative approaches with ours. Noted that all methods are implemented on click-through-based training sets. We selected TC as a representative of collective intelligence methods, as it is recently published and can almost achieve the best performance for this type of method. For the other class, we selected Vsynset and AFSVM as representatives, because of their strongest performances in this type.

Several observations can be obtained from Table III. First, compared with Vsynset and AFSVM, the collective intelligence-based method (TC) can achieve relative increases of 8.9% and 6.1% trained on Bing_Challenge, respectively. The improvements are 11.6% and 8.1% trained on Bing_Large. It reveals that when using large and clean training sets, this nonparametric method can find similar-enough neighbors and propagate correct tags to test samples. Vsynset and AFSVM, however, cannot ensure a reliable concept representation especially for diverse and sparse specific views, which are intrinsically hard to be modeled by parametric methods. Second, considering comprehensive concept representations, our view-dependent-based approach achieves better performance than all the baselines. The $F$-score relatively increases 15.4%, 12.4%, and 6.0% against Vsynset, AFSVM, and TC, respectively, which are trained on Bing_Challenge, and the improvements are 22.3%, 18.4%, and 9.6%, respectively, which are trained on Bing_Large. The superior performance verifies that the view-dependent concept representation can discover the comprehensive mapping between concepts and visual contents with proper ways. Therefore, our approach can effectively reduce the false positives and false negatives and achieve higher accuracy.
We also conducted experiments on some of the compared approaches to show the detailed performance for each tag. The result is shown in Fig. 5. To show the impact of different terms (i.e., MB_term, MF_term, and C_term) in our probabilistic objective function, we conducted experiments on Bing_Large with different settings. The result is shown in Table IV. Compared with the full model, MB + C drops dramatically, which verifies the key role of the model-free term in concept representations. The setting of MB + MF is comparable with the full model, which shows the more dominant role of the appearance term than the contextual priority term.

Fig. 6 shows our tag refinement result by case studies. Compared with previous works using Flickr as training sets, our approach can effectively correct and enrich tags from click-through data via view-dependent concept representations. Therefore, we can introduce some meaningful and accurate tags that do not belong to the original concept vocabulary in Flickr. For instance, the tags Tourism and Aquamarine in Fig. 6(a) and Preserve in Fig. 6(b) can well reflect the image content.

3) Analysis of the View-Dependent Representation: The analysis in this section is to provide more evidence to support the good performance of the proposed approach.

First, to show that any concept generally corresponds to two types of correct visual appearances, i.e., generic views and specific views, we conducted a user study on 50 randomly selected concepts from the tag list in MIRFlickr-25K [26]. There are about totally 10K images collected from the MIRFlickr-25K data set, where images are uniformly distributed over these concepts. Ten volunteers, including seven males and three females, were employed to evaluate the visual content under each concept. They were from different education backgrounds, including computer science, mathematics, physics, business, management science, art, and design, aging from 20 to 30 years old. For example, each volunteer was asked to understand the definition of the two views and evaluate all the images of the concept Hotel with three choices including relevant and generic, relevant and specific, and irrelevant. The resultant images corresponding to the three choices were considered as generic views, specific views, and irrelevant images. For each image, the choice with the highest votes was considered as the final evaluation result.

We observed that the images with generic view can form several visual clusters; however, there are no strong relations between sparse specific view samples. We show the generic views by clusters and the specific views by image instances in Fig. 1. Besides, we have also obtained the percentages of the generic views and specific views, which are 71% and 18% calculated on the 50 concepts in average.
views show concentrated visual patterns, while the specific views are sparse and diverse. Furthermore, Fig. 7 shows the ratio between the numbers of detected generic views and specific views by our approach. The ratio for the 18 tags in MIRFlickr-25K data set is 3.45 in average, which is very close to the ratio (71%/18% ≈ 3.94) of the manual statistics for the 50 randomly selected tags in the same data set. The statistic verifies that the proposed approach can discriminate different views of a concept and further demonstrates the effectiveness of the modeling method of the view-dependent concept representation. This accurate concept representation can help achieve a good performance for image tag refinement.

D. Evaluation on NUS-WIDE-270K

We evaluated the performance of image tag refinement in NUS-WIDE-270K data set, as it was one of the most typical large-scale data sets. Tables V and VI show the effect of training data and the comparison of three representative approaches with ours, respectively. We skipped VSC and RWR, as they cannot be applied to large-scale data sets. From the two tables, we can observe a consistently superior performance as that evaluated on MIRFlickr-25K. The F-score relatively increases 21.1%, 11.5%, and 8.6% against Vsynset, AFSVM, and TC, respectively, which are all trained on Bing Challenge. In addition, the improvements are 15.7%, 12.9%, and 9.2%, respectively, which are all trained on Bing Large. The comparison demonstrates the superiority of the view-dependent concept representation in this large-scale data set. We also conducted a detailed comparison for each tag using some of the compared approaches. The result is shown in Fig. 9.

Besides, as relevant tags are more important when they locating higher in a ranked list, we used Precision@k as measurement evaluated on this data set. We ranked tags having ground truth according to the score of MB_term + MF_term. The result is shown in Table VII. We have achieved a consistently better performance from Precision@1 to Precision@5 than the three representative compared approaches. In particular, the relative increases of Precision@1 are 12.0%, 6.6%, and 5.6% against Vsynset, AFSVM, and TC, respectively, which are trained on Bing Challenge. The improvements are 10.4%, 10.0%, and 5.4%, respectively, which are trained on Bing Large.

E. Complexity Analysis

The experiment was implemented on a server machine with a dual core 2.80-GHz Intel Xeon processors and 48-GB RAM. The time consumption of different modules of our approach is summarized in Table VIII. We can observe that the average runtime for one image, including feature extraction, generating candidate vocabulary, and solving for the probability formulation, is less than 1 s. Clearly, the process is fast and can provide almost instant response. Moreover, our approach is highly scalable as the view-dependent representation can be learned per concept, with no dependency on others. Therefore, new concept representations can be easily organized.
into the existing vocabulary rather than retraining the whole vocabulary again.

VII. CONCLUSION

In this paper, we have studied the problem of image tag refinement to help users better access images. To achieve this goal, we have proposed a view-dependent concept representation. Such representations could accurately and comprehensively describe a concept by its visual appearances with proper forms, i.e., a model-based component and a model-free component. Based on this representation, the coarse-to-fine tag refinement framework demonstrated the superiority to the state-of-the-art methods. Note that the influence of domain gap is limited in this task, as we have observed statistically significant improvement using Bing data. Since the larger the size of training data, the higher probabilities we can find sufficiently visually similar images. As our two training data are million scale, the content discrepancy between training and test data can be greatly reduced. We will continue to study the effective concept representations in image tagging and tag refinement and release these models in the near future. Furthermore, we will investigate to leverage the click-through data more intensively and introduce ranking-based solutions to this problem. Besides, it is interesting to explore how to learn the generic and specific views from noisy data. We will also study this topic in our future work.

TABLE VIII

<table>
<thead>
<tr>
<th>modules</th>
<th>time complexity</th>
<th>average runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>The collecting of click-through data (offline)</td>
<td>—</td>
<td>~5.5 h</td>
</tr>
<tr>
<td>The learning of view-dependent representation (offline)</td>
<td>—</td>
<td>~16.75 s/tag</td>
</tr>
<tr>
<td>The coarse step (online)</td>
<td>linear</td>
<td>~0.48 s/tag</td>
</tr>
<tr>
<td>The fine step (online)</td>
<td>polynomial</td>
<td>~0.35 s/image</td>
</tr>
</tbody>
</table>

REFERENCES

**FU et al.: IMAGE TAG REFINEMENT**


Jianlong Fu received the B.E. degree from University of Science and Technology Beijing, Beijing, China, in 2010. He is currently working toward the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He was a Research Intern with Knowledge Mining Group, Microsoft Research, Beijing, from 2012 to 2014. His research interests include computer vision, knowledge mining, and multimedia, in particular, content-based image/video retrieval, analysis, management, and sharing.

Jinqiao Wang (M’09) received the B.E. degree from Hebei University of Technology, Tianjin, China, in 2001; the M.S. degree from Tianjin University, Tianjin, in 2004; and the Ph.D. degree in pattern recognition and intelligence systems from National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2008. He is an Associate Professor with the Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, multimedia, and intelligent video surveillance.

Yong Rui (F’10) received the B.S. degree from Southeast University, Nanjing, China, the M.S. degree from Tsinghua University, Beijing, China; and the Ph.D. degree from University of Illinois at Urbana–Champaign, Champaign, IL, USA.

He is the Senior Director and a Principal Researcher with Microsoft Research, Beijing, where he is leading research effort in the areas of multimedia search and mining, knowledge mining, and social computing. He holds a Microsoft Leadership Training Certificate from the Wharton Business School, University of Pennsylvania, Philadelphia, PA, USA. He is recognized as a leading Expert in his research areas. He has authored 16 books and book chapters, and over 100 referred journal and conference papers. He holds 56 issued U.S. and international patents.

Dr. Rui is a fellow of the International Association for Pattern Recognition and the International Society for Optics and Photonics, and a Distinguished Scientist of the Association for Computing Machinery (ACM). He is the Editor-in-Chief of *IEEE Multimedia Magazine*, an Associate Editor of *ACM Transactions on Multimedia Computing, Communication and Applications*, a Founding Editor of *International Journal of Multimedia Information Retrieval*, and a Founding Associate Editor of *IEEE ACCESS*. He was an Associate Editor of *IEEE TRANSACTIONS ON MULTIMEDIA* from 2004 to 2008, *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGIES* from 2006 to 2010, *ACM/Springer Multimedia Systems* from 2004 to 2006, and *International Journal of Multimedia Tools and Applications* from 2004 to 2006. He is on the Organizing Committees and Program Committees of numerous conferences, including ACM Multimedia, ACM International Conference on Multimedia Retrieval (ICMR), the IEEE International Conference on Multimedia and Expo (ICME), SPIE ITCom, and the International Conference on Pattern Recognition. He was the General Co-Chair of ACM Multimedia in 2009 and 2014, ACM ICMR in 2006 and 2012, and the International Conference on Internet Multimedia Computing and Service in 2010, and the Program Co-Chair of ACM Multimedia in 2006, Pacific Rim Multimedia (PCM) in 2006, and the IEEE ICME in 2009. He is on the Steering Committees of ACM Multimedia, ACM ICMR, the IEEE ICME, and PCM. He is an Executive Member of the ACM Special Interest Group on Multimedia, and the Founding Chair of its China Chapter.

Xin-Jing Wang (M’09) received the Ph.D. degree in electronics engineering from Tsinghua University, Beijing, China, in 2005.

She was a Researcher with Microsoft Research, Beijing. She is currently a Senior Applied Scientist with Bing Core Ranking Team and a Visiting Professor with Xi’an University of Posts and Telecommunications, Xi’an, China. Her research interests include large-scale image understanding and knowledge mining from massive datasets.

Dr. Wang serves as a Program Committee Member for many related conferences.
Tao Mei (M’07–SM’11) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a lead Researcher with Microsoft Research, Beijing, China. He has authored or co-authored over 100 papers in journals and conferences, 10 book chapters, and edited three books. He holds 11 U.S. granted patents and over 20 that are pending. His research interests include multimedia information retrieval and computer vision. Dr. Mei has received several paper awards from prestigious multimedia journals and conferences, including the IEEE Circuits and Systems Society Circuits and Systems for Video Technology Best Paper Award in 2014, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2013, the Best Student Paper Award at the IEEE Visual Communications and Image Processing in 2012, and the best paper awards at ACM Multimedia in 2009 and 2007. He was also a recipient of the Microsoft Gold Star Award in 2010, and the Microsoft Technology Transfer Awards in 2010 and 2012. He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, Multimedia Systems, and Neurocomputing, and a Guest Editor of six international journals. He was the General Co-Chair of the International Conference on Internet Multimedia Computing and Service in 2013, the Program Co-Chair of the IEEE Conference on Magnetism and Magnetic Materials in 2013, and the Workshop Co-Chair of the IEEE International Conference on Multimedia and Expo in 2012 and 2014. He is a Senior Member of the Association for Computing Machinery. He will be the Program Co-Chair of the IEEE International Workshop on Multimedia Signal Processing in 2015.

Hanqing Lu (SM’06) received the B.E. and M.E. degrees from Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China, in 1992. He is the Deputy Director of the National Laboratory of Pattern Recognition with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored over 200 papers. His research interests include image and video analysis, medical image processing, and object recognition.