Image Retagging Using Collaborative Tag Propagation

Dong Liu, Shuicheng Yan, *Senior Member, IEEE*, Xian-Sheng Hua, *Member, IEEE*, and Hong-Jiang Zhang, *Fellow, IEEE*

***Abstract*—Photo sharing websites such as Flickr host a massive amount of social images with user-provided tags. However, these tags are often imprecise and incomplete, which essentially limits tag-based image indexing and related applications. To tackle this issue, we propose an image retagging scheme that aims at refining the quality of the tags. The retagging process is formulated as a multiple graph-based multi-label learning problem, which simultaneously explores the visual content of the images, semantic correlation of the tags as well as the prior information provided by users. Different from classical single graph-based multi-label learning algorithms, the proposed algorithm propagates the information of each tag along an individual tag-specific similarity graph, which reflects the particular relationship among the images with respect to the specific tag and at the same time the propagations of different tags interact with each other in a collaborative way with an extra tag similarity graph. In particular, we present a robust tag-specific visual sub-vocabulary learning algorithm for the construction of those tag-specific graphs. Experimental results on two benchmark Flickr image datasets demonstrate the effectiveness of our proposed image retagging scheme. We also show the remarkable performance improvements brought by retagging in the task of image ranking.**

***Index Terms*—Image retagging, label propagation, multi-graph multi-label learning, semantic correlation.**

I. Introduction

THE prevalence of digital cameras and the growing practice of photo sharing in community-contributed image websites like Flickr [1] and Zooomr [2] have led to a flourish of social images on the Web. Besides plain visual information, such large-scale images are augmented with user-provided tags, which greatly benefit a wide variety of applications such as image search, organization and management.

Despite the high popularity of manually tagging social images, the quality of the tags associated with images is still far from satisfactory. Currently, the image tagging on the image sharing websites solely relies on the manual inputs, which is an extra burden for grassroots Internet users and often prohibits accurate and comprehensive textual descriptions of the visual content. A recent study reported in [3] and [4] reveals that the user-provided tags associated with social images are rather imprecise, with only about 50% precision rate. On the other hand, the average number of tags for each social image is relatively small [5], which is far from the number that can fully describe the content of an image. Without accurate and sufficient tags, social images on the Web cannot be well indexed by search engine and consequently, cannot be well accessed by users. Therefore, effective methods to refine these unreliable tags have become emerging needs.

In this work, we propose an image retagging scheme that aims at improving the quality of the tags. The proposed scheme is motivated by the following three observations from real-world social images. 1) Visually similar images often reflect similar theme and thus are typically annotated with similar tags. The observation, referred to as visual consistency, has been widely explored in visual category learning [6], [7], but the correlations among tags are generally not utilized. 2) The tags associated with social images do not appear in isolation. Instead, they appear correlatively and naturally interact with each other at the semantic level. For example, the presence of tag “dog” often occurs together with the presence of tag “animal” while rarely co-occurs with “vegetable”. We refer it as semantic consistency assumption. It also has been applied in some multi-label or contextual learning algorithms [8], [9] , while typically a unique affinity matrix (i.e., similarity graph) is applied for all the different tags and they often can only handle dataset with clean labels (rather than the social images with noisy tags). 3) With the general knowledge that human-beings share most of common concepts in the semantic space, the user-provided tags of an image, despite imperfect, still reasonably reveal the primary semantic theme of the image content. We can regard these tags as prior information and use them to guide our retagging process.

Actually, image retagging can be regarded as a multi-label learning problem from imperfectly-labeled image set, since each social image is typically associated with multiple (noisy) tags. Graph-based approaches are frequently applied to solve multi-label learning problems, which can be categorized into two main paradigms. In the first paradigm, the multi-label problem is transferred into a set of independent label propagation problems [10]. The drawback of this approach is the lack of consideration of the inherent correlation among the

Manuscript received June 04, 2010; revised September 21, 2010 and January 21, 2011; accepted February 22, 2011. Date of publication March 28, 2011; date of current version July 20, 2011. This research is completed for CSIDM Project No. CSIDM-200803, which was supported in part by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale.

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

S. Yan is with the National University of Singapore, 117576 Singapore (e-mail: eleyans@nus.edu.sg).

X.-S. Hua is with Microsoft Research Asia, Beijing 100080, China (e-mail: 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.2011.2134078

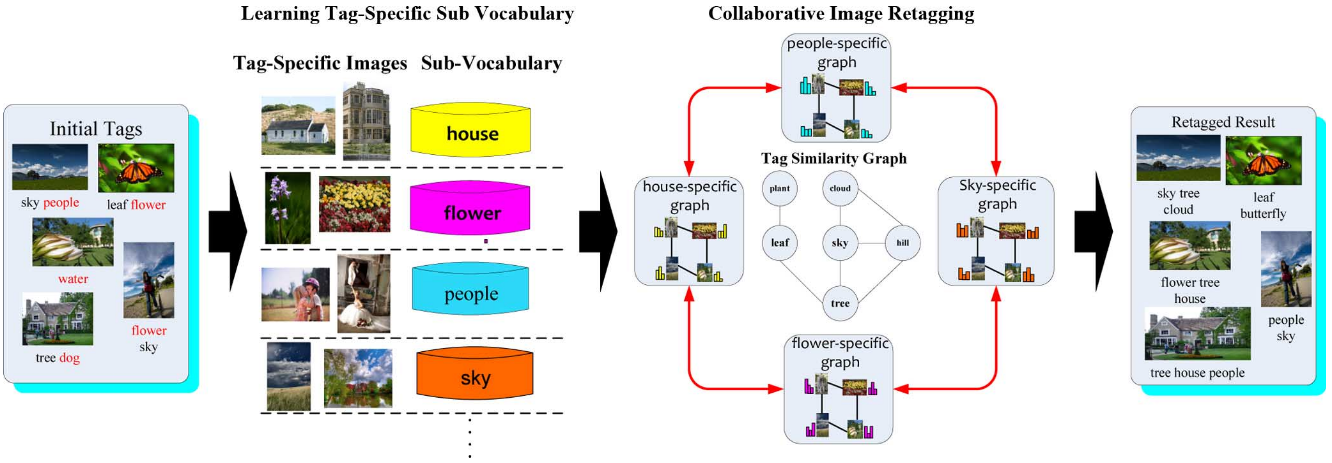


Fig. 1. Schematic illustration of the proposed collaborative image retagging approach. Given a collection of social images with initial user-provided tags, the tag-specific visual sub-vocabulary is first learned for each tag, based on which, the tag-specific image similarity graph is constructed. Then the image retagging is performed across multiple tag-specific similarity graphs in a collaborative way with an extra tag similarity graph. Finally, we obtain the image retagging result.

labels. The second paradigm moves one step ahead and further leverages the correlation among the labels in the propagation procedure [11], [12] . The key issue in the second paradigm, as aforementioned, is the measurement of the vertex similarities, where the existing methods simply infer a unique similarity graph based on low-level features and a similarity function. However, whether two images are similar actually depends on what the semantic tags are that we are caring about. Using a single graph to measure the image similarity is unable to take the tags into account; thus, it may not be able to well capture the desired semantic relationship among the images.

In this paper, we cast the image retagging task into an optimization problem with graph regularization, which simultaneously maximizes the visual and semantic consistency and at the same time minimizes the derivation from initial user-provided tags. Besides explicitly modeling the inter-dependency among the tags, the proposed algorithm addresses the issue in the second paradigm by constructing a set of tag-specific image similarity graphs to precisely reflect the relationships of the images with respect to different tags. The optimization problem is solved with an efficient multiplicative nonnegative iterative procedure, which can be interpreted as a propagation of the tag information among multiple graphs. Specifically, the tag information is propagated in two different dimensions: one is within-graph propagation, where the confidence scores for a specific tag are propagated among the images on the corresponding tag-specific graph; the other is cross-graph propagation, where the inherent correlation among the tags can be sufficiently exploited. By doing so, this work tries to propagate multiple tags on multiple graphs in an integrated manner and makes them benefit from each other in a collaborative way.

The tag-specific visual similarity measurement is crucial for the good performance of our retagging algorithm since it directly determines the underlying structure of each tag-specific graph. The traditional distance metric learning approaches, although effective in learning tag-specific distance/similarity measure, are often sensitive to the noisy tags in the social image collection. Instead, we pursue tag-specific low-level representations that are robust to tag noises in data. Specifically, we propose a novel algorithm to learn the tag-specific visual sub-vocabulary for each tag, which consequently results in tag-specific image similarity graph better revealing the relationship of the images with respect to the specific tag.

Fig. 1 illustrates the pipeline of our approach. The scheme consists of two steps: tag-specific visual sub-vocabulary learning and collaborative image retagging. In the first step, tag-specific visual vocabulary is optimized for each tag by leveraging a collection of images with noisy tags. Based on the learned vocabulary for each tag, we construct tag-specific similarity graph to characterize the relations of the images with respect to a certain tag. A retagging process is then implemented across the multiple tag-specific graphs in a collaborative way. The entire framework is fully automatic without any user interaction. Furthermore, we can also implement the two steps iteratively to boost the retagging performance if the time complexity and computational cost are not a concert. It is also worth noting that our approach can not only refine the imperfect tags associated with those social images but also indicate their relevance scores with respect to the associated images. Therefore, we can achieve the purpose of image ranking, which orders images based on their relevance with respect to a specific query tag.

The main contributions of this paper are as follows: 1) we propose a graph-based optimization algorithm to improve the quality of tags by modeling visual consistency of the images over multiple tag-specific similarity graphs, semantic consistency of the tags as well as prior information provided by users. Comparing with existing graph-based algorithms, we perform the tag propagation over the multiple tag-specific similarity graphs in a collaborative way; (2) to construct tag-specific similarity graph, we propose a robust optimization algorithm to learn tag-specific visual sub-vocabulary from a collection of social images with noisy tags.

II. Related Work

Our work is first related to the automatic image annotation (tagging), a process to predict the tags associated with an image. Many algorithms have been proposed for this task, varying from building classifiers for individual semantic labels [13], [14], [15] to learning relevance models between images and keywords [16], [17] . Much of this excitement centers around utilizing the machine learning techniques to learn the mapping between the image contents and semantic tags based on a collection of precisely labeled training images and then use the learnt model to predict the tags of those unlabeled images. On the contrary, the original goal of our work here is not to predict tags for the unlabeled images, but rather to refine the imprecise tags provided by the users. Moreover, we do not have precisely labeled training images in the refinement procedure. Instead, the only available supervision to learn the improved tag assignment in our scenario is the imprecise, incomplete or subjective tags provided by the users. Therefore, we argue that image retagging is a more challenging task.

There are some efforts on social tagging in the literature. Kennedy *et al.* [3] evaluated the performance of the classifiers trained with Flickr images and their associated tags and demonstrated that tags provided by Flickr users contain many noises. Liu *et al.* [18] proposed to rank the tags according to their relevance with respect to the associated images. Weinberger *et al.* proposed a method to analyze the ambiguity of tags [19]. Despite the fact that these works have shown encouraging results, they focus on directly utilizing the tags as a knowledge source or simply analyzing the relation between images and their associated tags, whereas there is still a lack regarding improving the quality of tags.

The previous research on improving unreliable descriptive keywords of images to date has focused on annotation refinement, i.e., identifying and eliminating the imprecise annotation keywords produced by the automatic image annotation algorithms. As a pioneering work, Jin *et al.* [20] used WordNet [21] to estimate the semantic correlation among the annotated keywords and then those weakly-correlated ones are removed. However, this method can only achieve limited success since it totally ignores the visual content of the images. To address this problem, Wang *et al.* [22] proposed a content-based approach to re-rank the automatically annotated keywords of an image and only reserve the top ones as the refined results. Despite these efforts, existing methods typically focus on selecting a coherent subset of keywords from the automatically annotated keywords. This is reasonable since the underlying assumption for annotation refinement is that the generative likelihoods between visual content and annotation keywords have been maximized by the automatic annotation algorithms. On the other hand, the tags associated with social images are often imprecise and incomplete, and thus, their descriptive capability to the visual content cannot be guaranteed. Therefore, the reranking-and-removing strategy in annotation refinement methods is not applicable in the image retagging scenario. To address this difficulty, Liu *et al.* [23] proposed to refine the tags based on the visual and semantic consistency residing in the social images, which assigns similar tags to visually similar images. However, it still uses the holistic image similarity to model the image relationship and thus only results in limited performance improvements (see Section IV-B).

Recently, several approaches have been proposed to construct discriminative visual sub-vocabularies, which explicitly incorporate the category-specific information. For example, Liu *et al.* [24] proposed to unify the discriminative visual codebook generation with the classifier training in the task of object category recognition. Moosmann *et al.* [25] accomplished the task through randomized clustering forests. Zhang *et al.* [26] proposed to learn descriptive visual words to represent images under different concepts, which results in good performance in a series of image applications. However, despite these efforts, there are difficulties in directly applying the existing methods in our scenario, due to the fact that our training images are collected from the Internet based on tag queries, and thus, many of the returned images do not even contain the desired targets. Handling such mistagged data is beyond the capability of existing methods. Therefore, we need to design a noise-tolerant algorithm to accomplish the task.

There are also some “multi-graph” label propagation works [27], [28] , but the focus of these works is to combine multiple graphs to obtain a complementary graph, which, however, is essentially different from this work where each graph is constructed to characterize the particular relationship among the images with respect to a certain tag.

III. Collaborative Image Retagging

This section presents our proposed scheme for image retagging. We first describe the problem formulation and then introduce an efficient multiplicative nonnegative iterative procedure for the optimization. Finally, we present the tag-specific visual vocabulary learning algorithm.

A. Problem Formulation

Denote by $\mathcal{I} = \{x_1, x_2, \ldots, x_n\}$ a social image collection, where n is the size of the image set. All unique initial tags appearing in this collection form a tag set $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, where m denotes the total number of unique tags. The initial tag membership for the whole image collection can be presented in a binary matrix $\hat{\mathbf{Y}} \in \{0,1\}^{n\times m}$, whose element $\hat{Y}_{ij}$ indicates the presence of tag t_j in image x_i (i.e., if t_j is associated with image x_i, then $\hat{Y}_{ij} = 1$ and $\hat{Y}_{ij} = 0$ otherwise). To represent the final retagging results, we define another matrix $\mathbf{Y}$ whose element $Y_{ij} \geq 0$ denotes confidence score of assigning tag t_j to image x_i.

The entire formulation of the image retagging algorithm consists of three components: a loss term $E_l(\mathbf{Y})$ and two regularization terms $E_v(\mathbf{Y})$ and $E_s(\mathbf{Y})$. Specifically, $E_l(\mathbf{Y})$ corresponds to the deviation of the retagging results from the initial user-provided tags; $E_v(\mathbf{Y})$ is a regularizer to enforce the visual consistency over multiple tag-specific similarity graphs while $E_s(\mathbf{Y})$ is another regularizer to enforce the semantic consistency. Based on these three terms, the image retagging problem can be formulated as minimizing the following objective function:

$$\min_{\mathbf{Y}} F(\mathbf{Y}) = E_v(\mathbf{Y}) + \alpha E_s(\mathbf{Y}) + \beta E_l(\mathbf{Y}) \qquad (1)$$

where α and β are two tunable parameters that balance the latter two terms.

First, we consider the visual regularization term. For a tag $t_k \in \mathcal{T}$, we have a corresponding weighted and undirected graph $\mathbf{G}_k$ where the vertices are the n images in $\mathcal{I}$ and the symmetric edge weight matrix $\mathbf{W}_k = [w_{ij}^k]$ is estimated based on the tag-specific visual similarity for tag t_k, which is very sparse based on k-Neighbor Neighbor graph. Details of generating tag-specific similarity graph will be presented later in Section IV. The node degree matrix of graph $\mathbf{G}_k$ is defined as

$\mathbf{D}_k = diag\{d_1^k, d_2^k, \ldots, d_n^k\}$, where $d_i^k = \sum_{j=1}^n w_{ij}^k$. Furthermore, the normalized graph Laplacian of graph $\mathbf{G}_k$ is defined as $\mathbf{L}^{t_k} = \mathbf{D}_k^{-1/2}(\mathbf{D}_k - \mathbf{W}_k)\mathbf{D}_k^{-1/2} = \mathbf{I} - \mathbf{D}_k^{-1/2}\mathbf{W}_k\mathbf{D}_k^{-1/2}$. Based on the above terminologies, the visual regularization term is formulated as

$$E_v(\mathbf{Y}) = \sum_{t=1}^{m}\sum_{i,j=1}^{n} w_{ij}^t \left(\frac{Y_{it}}{\sqrt{d_i^t}} - \frac{Y_{jt}}{\sqrt{d_j^t}}\right)^2 \tag{2}$$

which is actually a smoothness constraint with the normalized graph Laplacian matrix (normalized using node degrees). Comparing with the formulation of the unnormalized version that directly minimizes $\sum_{t=1}^m \sum_{i,j=1}^n w_{ij}^t (Y_{it} - Y_{jt})^2$, the inclusion of the normalization factors $\sqrt{d_i^t}$ and $\sqrt{d_j^t}$ will penalize a normalized version of within-class edge weights when such an edge connects two nodes with different scaling factors. Such a normalization scheme often leads to better performance in many graph-based learning algorithms [10], [12], and a recent theoretical analysis in [29] further confirms its superiority from the learning theory aspect. It is worth noting that the minimization of (2) actually reflects the visual consistency over m tag-specific similarity graphs. For example, if image i and image j have high visual similarity conditioned on tag t, i.e., the value of w_{ij}^t is large, the values of $Y_{it}/\sqrt{d_i^t}$ and $Y_{jt}/\sqrt{d_j^t}$ tend to be close, indicating that these two images should be assigned with similar confidence scores with respect to tag t. By enforcing such a regularization term, visually similar images are inclined to be annotated with similar tags, leading to the consistent tagging results.

We then introduce the regularization term for the semantic consistency assumption. To this end, we introduce the tag similarity matrix $\mathbf{S}$ whose element $0 \leq s_{kl} \leq 1$ indicates the semantic similarity between tags t_k and t_l. In this work, we adopt a concurrence-based method to estimate this similarity, which is analogous to the Google distance [30]. We first estimate the semantic distance between tags t_k and t_l as

$$d(t_k, t_l) = \frac{\max\{\log g(t_k), \log g(t_l)\} - \log g(t_k, t_l)}{\log R - \min\{\log g(t_k), \log g(t_l)\}} \tag{3}$$

where $g(t_k)$ and $g(t_l)$ are the numbers of images containing tag t_k and tag t_l, respectively, and $g(t_k, t_l)$ is the number of images containing both tag t_k and tag t_l. These numbers can be obtained by performing search by "tags only" on Flickr website using the tags as queries [1]. Moreover, R is the total number of images in Flickr. Then the semantic similarity between t_k and t_l is defined as $s_{kl} = \exp(-d(t_k, t_l))$. After obtaining the semantic similarity, the semantic regularization term is defined as

$$E_s(\mathbf{Y}) = \sum_{i=1}^{n}\sum_{k,l=1}^{m} s_{kl}\left(\frac{Y_{ik}}{\sqrt{\psi_k}} - \frac{Y_{il}}{\sqrt{\psi_l}}\right)^2 \tag{4}$$

where ψ_k is defined as $\psi_k = \sum_{l=1}^m s_{kl}$. Let $\mathbf{\Psi} = diag\{\psi_1, \psi_2, \ldots, \psi_m\}$, we define a normalized graph Laplacian for modeling the semantic correlations among the tags as $\mathbf{L}^s = \mathbf{\Psi}^{-1/2}(\mathbf{\Psi} - \mathbf{S})\mathbf{\Psi}^{-1/2} = \mathbf{I} - \mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$, which is the normalized version of the graph Laplacian for semantic similarity graph among the tags. Obviously, the above formulation imposes a smoothness constraint on the semantic correlation among the tags within the tag similarity graph. For example, if tag t_k and t_l have high semantic similarity, i.e., the value of s_{kl} is high, the confidence score Y_{ik} and Y_{il} are enforced to be close, which indicates the highly semantically correlated tags are inclined to be simultaneously assigned to a same image. Therefore, the formulation in (4) actually models the semantic consistency assumption, which can be used to ensure high performance of image retagging.

Finally, we consider the deviation between the image retagging results and the initial user-provided tags, which is formulated as

$$E_l(\mathbf{Y}) = \sum_{i=1}^{n}\sum_{k=1}^{m}(Y_{ik} - \hat{Y}_{ik})^2 V_{ik} \tag{5}$$

where V_{ik} is a weighting factor characterizing the importance of tag t_k to image x_i, which is estimated from the visual distribution of image x_i on the specific visual vocabulary for tag t_k (detailed in Section III-D). Moreover, all entries V_{ik} $(i = 1, 2, \ldots, n,\ k = 1, 2, \ldots, m)$ form a matrix $\mathbf{V} \in \mathbb{R}^{n \times m}$.

Based on the above three terms, the unified formulation can be written as

$$\begin{aligned}\min_{\mathbf{Y}} \quad & \sum_{i,j=1}^{n}\sum_{t=1}^{m} w_{ij}^t \left(\frac{Y_{it}}{\sqrt{d_i^t}} - \frac{Y_{jt}}{\sqrt{d_j^t}}\right)^2 \\ & + \alpha \sum_{i=1}^{n}\sum_{k,l=1}^{m} s_{kl}\left(\frac{Y_{ik}}{\sqrt{\psi_k}} - \frac{Y_{il}}{\sqrt{\psi_l}}\right)^2 \\ & + \beta \sum_{i=1}^{n}\sum_{k=1}^{m}(Y_{ik} - \hat{Y}_{ik})^2 V_{ik}, \\ s.t. \quad & Y_{it} \geq 0. \end{aligned} \tag{6}$$

To further ease the representation, we vectorize each matrix by stacking their columns into a vector. Denote by $\mathbf{y} = vec(\mathbf{Y})$, $\mathbf{v} = vec(\mathbf{V})$, and $\tilde{\mathbf{y}} = vec(\tilde{\mathbf{Y}})$, the objective function can be rewritten in a more compact form

$$\begin{aligned}\min_{\mathbf{y}} \quad & \mathbf{y}^T\mathbf{L}\mathbf{y} + \alpha(\mathbf{\Gamma}\mathbf{y})^T\mathbf{\Upsilon}(\mathbf{\Gamma}\mathbf{y}) \\ & + \beta(\mathbf{y} - \tilde{\mathbf{y}})^T\mathbf{\Delta}(\mathbf{y} - \tilde{\mathbf{y}}), \\ s.t. \quad & \mathbf{y} \geq 0 \end{aligned} \tag{7}$$

where $\mathbf{\Delta}$ is defined as $diag\{\mathbf{v}\}$ and $\mathbf{\Gamma}$ is an $(m \times n) \times (m \times n)$ binary matrix defined as

$$\mathbf{\Gamma}_{ij} = \begin{cases} 1, & \text{if } j = \lceil \frac{i}{n} \rceil + \big((i-1) mod n\big) \times m, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The operator $\lceil \cdot \rceil$ denotes the ceiling function which gives the smallest integer not smaller than the given value. In addition, $\mathbf{L}$ and $\mathbf{\Upsilon}$ are two block-wise matrices defined as $\mathbf{L} = diag\{\mathbf{L}^{t_1}, \mathbf{L}^{t_2}, \ldots, \mathbf{L}^{t_m}\}$. The block $\mathbf{L}^{t_i}$ denotes the normalized graph Laplacian of images conditioned on tag t_i. The $\mathbf{\Upsilon}$ matrix is diagonal and defined as $\mathbf{\Upsilon} = diag\{\mathbf{L}^s, \mathbf{L}^s, \ldots, \mathbf{L}^s\}$, which is a diagonal matrix with n repetitions of the block $\mathbf{L}^s$, each of which denotes the normalized graph Laplacian for tags.

A practical issue in the above formulation is the number of graphs, which equals to the number of tags in the social image collection. When the number of tags becomes extremely large, we need to involve a large number of tag-specific image similarity graphs into the optimization procedure, which will increase the computational cost tremendously. To tackle this deficiency, we can segment the tag similarity graph into a number of subgraphs via certain graph partition algorithm such as Normalized Cut, where the tags residing within the same subgraph are highly semantically correlated while the tags from different subgraphs are considered to be totally uncorrelated. Starting from the tag subgraphs, the whole retagging problem can be divided into a set of independent subproblems, each of which contains only a small number of image similarity graphs corresponding to those tags that reside within one tag subgraph.

B. *Nonnegative Optimization Procedure*

The formulation in (7) is a quadratic optimization problem with nonnegative constrains. Typically, such an optimization problem can be solved with the standard quadratic programming (QP) software package. However, the number of involved variables in (7) is extremely large, which makes the optimization computationally intractable by the standard QP optimization package. Instead, we propose a scalable iterative optimization procedure with multiplicative nonnegative update rule to derive the solution. Our proposed optimization procedure is scalable in the sense that it is able to handle a large number of Y_{ij} variables in the optimization objective, which is beyond the capability of most of the existing QP methods. As the objective function is quadratic, its derivative with respect to $\mathbf{y}$ is then of first order and

$$\begin{aligned}\frac{\partial F}{\partial \mathbf{y}} &= 2\mathbf{L}\mathbf{y} + 2\alpha(\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma})\mathbf{y} + 2\beta\mathbf{\Delta}(\mathbf{y}-\tilde{\mathbf{y}})\\ &= 2\big(\mathbf{L} + \alpha(\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma}) + \beta\mathbf{\Delta}\big)\mathbf{y} - 2\beta\mathbf{\Delta}\tilde{\mathbf{y}}. \end{aligned} \quad (9)$$

Let $\phi_i \geq 0$ denote the Lagrange multiplier for constraint $y_i \geq 0$, we apply the Karush-Kuhn-Tucker (KKT) condition of $\phi_i y_i = 0$ to the derivative of the Lagrange function and then obtain

$$\big(2(\mathbf{L} + \alpha(\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma}) + \beta\mathbf{\Delta})\mathbf{y}\big)_i y_i - (2\beta\mathbf{\Delta}\tilde{\mathbf{y}})_i y_i = 0. \quad (10)$$

Then we can obtain the following update rule:

$$y_i \leftarrow y_i \times \frac{(2\beta\mathbf{\Delta}\tilde{\mathbf{y}})_i}{\big(2(\mathbf{L} + \alpha(\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma}) + \beta\mathbf{\Delta})\mathbf{y}\big)_i}. \quad (11)$$

Theorem 1: The update rule in (11) will converge and lead to the global minimum of the objective function.

Proof: First, each item of the objective function in (6) is quadratic and the coefficients for squared items are positive; thus, the convexity can be naturally guaranteed. Then the objective function in (7) can be written as

$$F(\mathbf{y}) = \mathbf{y}^T(\mathbf{L} + \alpha\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma} + \beta\mathbf{\Delta})\mathbf{y} - (2\tilde{\mathbf{y}}^T\mathbf{\Delta})\mathbf{y} + \tilde{\mathbf{y}}^T\mathbf{\Delta}\tilde{\mathbf{y}} \quad (12)$$

which is actually in a special form of $\mathbf{y}^T\mathbf{A}\mathbf{y} + b\mathbf{y}$ with $\mathbf{A}$ being positive definite. As proven in [31], the nonnegative update will lead to a global minimum of the objective function. ■

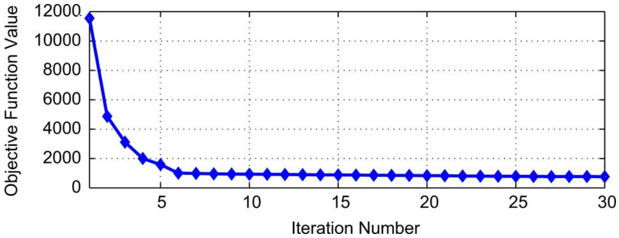


Fig. 2. Convergence process of the objective function over MIRFlickr.

Algorithm 1: Multiplicative updating procedure for collaborative image retagging

```
1: Initialize $\alpha > 0$, $\beta > 0$ and $\mathbf{y}^0 > \mathbf{0}$, set $l = 0$.
2: while not converged do
3:     $l = l + 1$;
4:     Assign $y_i^l \leftarrow y_i^{l-1} \times (2\beta\mathbf{\Delta}\tilde{\mathbf{y}})_i / \big(2(\mathbf{L} + \alpha(\mathbf{\Gamma}^T\mathbf{\Upsilon}\mathbf{\Gamma}) + \beta\mathbf{\Delta})\mathbf{y}^{l-1}\big)_i, \ \forall i$;
5: end while
```

Algorithm 1 summarizes the entire procedure for the optimization. Denote by l the iteration index, the updating rules are performed iteratively until $\|\mathbf{y}^l - \mathbf{y}^{l-1}\|_2 < 10^{-2}$, where $\|\cdot\|_2$ denotes the ℓ_2-norm of a vector. Based on the converged $\mathbf{Y}$, the tags with the top largest values in $[Y_{i1}, Y_{i2}, \ldots, Y_{im}]$ are considered as the retagging result of a social image x_i. In this work, we implement Algorithm 1 on the MATLAB platform of an Intel Xeon X5450 workstation with 3.0-GHz CPU and 16-GB memory and observe that the multiplicative updating iteration converges fast. For example, in our image retagging experiment on MIRFlickr dataset (see Section IV-B), each iterative procedure updating all entries of $\mathbf{y}$ will be processed with 6.3 min. Fig. 2 shows the convergence process of the iterative optimization, which is captured during the image retagging experiment on the MIRFlickr dataset. Here one iteration refers to one round of multiplicative updating for $\mathbf{y}$, which corresponds to steps 2–3 of Algorithm 1 .

C. *Discussions*

The proposed collaborative image retagging algorithm has an intuitive interpretation in the viewpoint of graph-based tag propagation. Based on each tag-specific similarity graph along with the initial user-provided inputs as imperfect “priors”, we propagate the tag information through tag-specific visual similarities among the images. This is actually a within-graph propagation, and the images with higher confidence scores are more likely to be assigned with the given tag. Moreover, the incorporation of semantic correlation among different tags achieves the cross-graph propagation, in which the propagations of different tags are interacted with each other, making them benefit each other in a collaborative way. Note that such a graph-based tag propagation method is essentially different from the existing graph-based propagation methods where only one graph is utilized to model the relationship of the images. In Table I, we summarize

TABLE I
COMPARISON BETWEEN DIFFERENT GRAPH-BASED MULTI-LABEL LEARNING METHODS

Methods	Number of graphs	Considering label correlation ?	Constructing tag-specific graph ?
ML-LGC	1 image similarity graph	No	No
SMSE	1 image similarity graph + 1 label similarity graph	Yes	No
Our method	m image similarity graph + 1 label similarity graph	Yes	Yes

the differences between our proposed collaborative multi-label propagation method with the existing graph-based multi-label propagation methods including 1) multi-label local and global consistency (ML-LGC) method proposed by Zhou *et al.* [10] and 2) Semi-supervised Multi-label learning method by solving a Sylvester Equation (SMSE) proposed by Chen *et al.* [12].

In particular, a close analysis on ML-LGC and SMSE reveals the fact that the objective functions of these two methods are actually the special form of our proposed formulation. This observation can be easily obtained by a simple modification of our proposed objective function in (6). For example, setting α to be 0 while only constructing a single graph to model the holistic image relationship, the objective function will degenerate into that of ML-LGC. Moreover, the objective function will degenerate into the objective function of SMSE if we do not construct the individual tag-specific graphs.

D. Learning Tag-Specific Visual Sub-Vocabulary From Social Images With Noisy Tags

As discussed in Section I, the tag-specific image similarity graph plays an important role in our proposed image retagging scheme. However, learning such similarity measure is a challenging task since the initial tags of the social image collection contains a lot of noises. Here we propose a noise-tolerant algorithm to accomplish the task. Specifically, we adopt the widely-used bag-of-words (BoW) model to represent the visual content of the images and then learn a tag-specific visual vocabulary for each tag.

The purpose of the visual sub-vocabulary related to a target tag is to select the most informative visual words to represent the corresponding tag. Therefore, two criteria are desired for selecting the visual words: 1) the visual words in the tag-specific visual sub-vocabulary should appear more frequently in images labeled with the given tag than in images without the tag; and 2) they should locate on the object/scene, even if the object/scene is surrounded by cluttered background. In the following, we propose an optimization scheme to effectively incorporate these two criteria in the process of learning tag-specific visual sub-vocabulary from a collection of social images annotated with the given tag.

Suppose we have a collection of social images $\mathcal{I}$, in which the images annotated with a given tag t form a subset $\mathcal{X}^t = \{I_1, I_2, \ldots, I_{M_t}\}$ and M_t is the number of images in $\mathcal{X}^t$. Assume we have a universal visual codebook $\Omega = \{vw_1, vw_2, \ldots, vw_N\}$ learned from a set of scale invariant feature transform (SIFT) features [32] extracted from the entire social image collection $\mathcal{I}$, where N denotes the size of the codebook. For any image $I_i \in \mathcal{X}^t$, we first extract its SIFT-features $a_i(j)$'s and then represent I_i by a bag of visual words $\{vw_{a_i(1)}, \ldots, vw_{a_i(j)}, \ldots\}$, where $vw_{a_i(j)}$ is the corresponding visual word for SIFT feature $a_i(j)$.

First, we define $f_i \in [0, 1]$ as the probabilistic score to measure the possibility of a visual word vw_i being descriptive to the target object/scene corresponding to tag t and the scores for all visual words in Ω can be represented as a vector $\mathbf{f} = [f_1, f_2, \ldots, f_N]^T$. In order to combine these probabilities to obtain the probability for image I_i having at least one visual word being descriptive to tag t, we use the softmax function:

$$P_{I_i} = \frac{\sum\limits_{vw \in I_i} f_{k(vw)} e^{f_{k(vw)}}}{\sum\limits_{vw \in I_i} e^{f_{k(vw)}}} \tag{13}$$

where the summation range on the numerator and denominator of (13) covers all the visual words contained in image I_i and $k(vw)$ denotes the index of visual word vw in Ω. The above formulation can also be rewritten as

$$P_{I_i} = \frac{\sum\limits_{j=1}^{N} n_j^{(I_i)} f_j e^{f_j}}{\sum\limits_{j=1}^{N} n_j^{(I_i)} e^{f_j}} \tag{14}$$

where $n_j^{(I_i)}$ denotes the number of occurrences for the jth visual word vw_j in image I_i.

According to the first criterion for tag-specific visual sub-vocabulary, the frequency-of-occurrence information of each visual word is important for identifying its descriptiveness. Besides, spatial co-occurrence information between a pair of visual words is another important clue, since spatially consistent visual words are more likely to locate on the target object/scene corresponding to the tag. Based on these two clues, the task of learning tag-specific visual sub-vocabulary from a noisy social image collection is formulated as

$$\begin{aligned} \min_{\mathbf{f}} \quad & Q(\mathbf{f}) = \sum_{i=1}^{N} (f_i - q_i)^2 \\ & \quad + c_1 \sum_{i,j=1, i \neq j}^{N} o_{ij} (f_i - f_j)^2 - c_2 \sum_{k=1}^{M} P_{I_k}, \\ s.t. \quad & 0 \leq f_i \leq 1, \; i = 1, 2, \ldots, N \end{aligned} \tag{15}$$

where the value of $q_i = freq_i^t / Freq_i$ measures the inherent significance of visual word vw_i and can be adopted as a prior probability to estimate the descriptive capability of visual word vw_i. Here $freq_i^t$ denotes the occurrence frequency of the visual word vw_i in the image subset $\mathcal{X}^t$ and $Freq_i$ denotes the occurrence frequency of visual word vw_i in the whole image collection $\mathcal{I}$. o_{ij} denotes the co-occurrence frequency between visual words vw_i and vw_j, which is defined as $o_{ij} = m(vw_i, vw_j)/M_t$, where $m(vw_i, vw_j)$ denotes the number of images that simultaneously contain visual words vw_i and vw_j.

Note that: 1) the first term in the objective function measures the data fitting capability, namely, the deviation between the estimated probabilistic confidence score and the prior probability; 2) the second term measures the smoothness of confidence score for different visual words, i.e., two visual words with high co-occurrence should also have similar confidence scores of being selected to be descriptive; 3) the third term denotes the negative probability for the kth image to have at least one visual word descriptive to the tag, which penalizes the cases of noisy images without the target object/scene; and 4) each term is normalized by the number of elements in the summation, while c_1 and c_2 are two parameters that control the tradeoff among these three normalized terms. In practice, we can also normalize the second and third terms of (15) by N and M, respectively, making the determination of the values c_1 and c_2 independent of the dataset size.

The optimization problem in (15) is a standard bound-constrained optimization problem that has been well investigated in both theory and practice. Among the existing bound-optimization techniques, the projected gradient method is proven to be simple and effective. Here we also employ the projected gradient method for the optimization. First, the partial derivation of the objective function with respect to f_i is calculated as

$$\frac{\partial Q}{\partial f_i} = 2(f_i - q_i) + 2c_1 \sum_{i,j=1, i\neq j}^{N} o_{ij}(f_i - f_j) - c_2 \sum_{k=1}^{M} \frac{n_i^{(I_k)} e^{f_i} R}{\left(\sum_{j=1}^{N} n_j^{(I_k)} e^{f_j}\right)^2} \tag{16}$$

where $R = \sum_{j=1}^{N} n_j^{(I_k)} e^{f_j} + f_i \sum_{j=1}^{N} n_j^{(I_k)} e^{f_j} - \sum_{j=1}^{N} n_j^{(I_k)} f_j e^{f_j}$.

Denote by $\nabla Q(\mathbf{f}) = [\partial Q/\partial f_1, \ldots, \partial Q/\partial f_N]^\top$ and also assume l denotes the index of iterations. The projected gradient method updates the current solution $\mathbf{f}^l$ to $\mathbf{f}^{l+1}$ by the following rule:

$$\mathbf{f}^{l+1} = P[\mathbf{f}^l - \eta_l \nabla Q(\mathbf{f}^l)] \tag{17}$$

where η_l is the step size and

$$P[f_i] = \begin{cases} f_i, & \text{if } 0 < f_i < 1, \\ 1, & \text{if } f_i \geq 1, \\ 0, & \text{if } f_i \leq 0 \end{cases} \tag{18}$$

maps a point back to the bounded feasible region. The step size η_l plays a key role in the iterative optimization procedure. To find the appropriate parameter setting that ensures the sufficient decrease of the objective function per iteration, we consider a simple and effective strategy proposed by Bersekas *et al.* [33]. Based on the obtained optimization result, the tag-specific visual sub-vocabulary for the given tag t is constructed by selecting the top K codewords with the highest confidence scores. Fig. 3 illustrates some examples in which each image is represented with the tag-specific visual sub-vocabulary learned by our proposed algorithm. From this figure, we can conclude that tag-specific visual sub-vocabulary is descriptive for the visual representation of the specific tag.

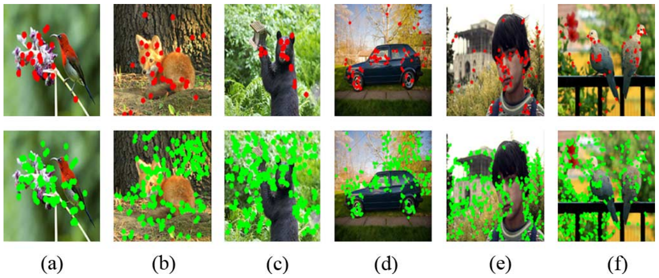


Fig. 3. Some exemplary images from Flickr. The interest points marked in red circle are obtained with the tag-specific visual sub-vocabularies (top row). The interest points marked in green circle are found by the Harris-Laplace detectors which are used for determining the salient points in an image (bottom row). As can be seen, most of the red points in each image locate at the target objects under consideration, which indicates that the learnt tag-specific visual sub-vocabularies are descriptive to the visual representations of the individual tags. On the contrary, the green points merely correspond to visual salient regions in the images and thus are not descriptive to the underlying semantic concepts. (a) Flower. (b) Fox. (c) Bear. (d) Car. (e) People. (f) Bird.

With the learned tag-specific visual sub-vocabulary for tag t, we then present image $x_i \in \mathcal{I}$ in a vector $\mathbf{h}_i^t = \{h_i^t(1), h_i^t(2), \ldots, h_i^t(K)\}$. Each element in $\mathbf{h}_i^t$ corresponds to a descriptive visual word for the given tag t and is simply a count of the number of features indexed with this visual word in image x_i. Such a feature vector can be seen as a tag-specific visual representation of the image, in which the visual words are descriptive to the specific tag. After obtaining the tag-specific visual representation for each image, we define the tag-specific visual similarity between x_i and x_j as

$$w_{ij}^t = \frac{\sum_{l=1}^{K} \min\{h_i^t(l), h_j^t(l)\}}{\sum_{l=1}^{K} \frac{\{h_i^t(l) + h_j^t(l)\}}{2}} \tag{19}$$

which is the number of visual words shared between these two images divided by their average number of visual words. Based on it, the tag-specific visual similarity graph can be constructed. Furthermore, the value of V_{it} in (6), which modulates the importance of tag t to image x_i, is defined as

$$V_{it} = \frac{n(\mathbf{h}_i^t)}{K} \tag{20}$$

where $n(\mathbf{h}_i^t)$ denotes the number of nonzero elements in the tag-specific visual representation $\mathbf{h}_i^t$ and K is the size of the tag-specific visual sub-vocabulary.

IV. Experiments

A. Dataset

As the initial user-provided noisy/missing tags are not available in the benchmark datasets such as COREL [34] and MSRC [35], we employ the real-world social images with human annotated tags to evaluate the performance of the proposed algorithm. Specifically, two publicly available Flickr image datasets, NUS-WIDE [4] and MIRFlickr [36], are used for the experiments. The NUS-WIDE dataset contains a total of 269 648 images and has averagely two tags per image. We select a subset of this dataset, focusing on images containing at least two tags and obtain a collection of 24 300 images with 73 608 tags. The second dataset, MIRFlickr, comprises 25 000 images with 223 537 tags and the average number of tags per image is 8.94.

We notice that the tags in the above two collections are rather noisy and many of them are misspelling or meaningless words. Hence, a pre-filtering process is performed for these tags. We match each tag with entries in a Wikipedia thesaurus and only the tags with coordinates in Wikipedia are retained. In this way, 11 370 and 10 026 unique tags are obtained in total for NUS-WIDE and MIRFlickr, respectively. To avoid sample insufficiency issue in learning tag-specific visual vocabulary, we further remove those tags whose occurrence numbers are below 50 and employ the remaining 749 and 205 tags as the tag vocabulary for these two datasets. It is worth noting that although the numbers of images in the above two social image datasets are moderate, the involved tag numbers (even after removing) are still quite large, leading to a challenge task for the visual content analysis and modeling.

In this work, the images are represented with the BoW image representation. We use Harris-Laplace detectors [37] to determine local regions and extract the SIFT features as the region descriptors. We then perform k-means clustering to obtain a visual codebook containing 500 visual words. Then each detected feature is mapped to an integer (visual word index) between 1 and 500. Thus, an image is represented as a histogram of visual words of size 500.

B. Evaluating Image Retagging

To evaluate the performance of our image retagging algorithm, we need to manually label the ground-truth tags for all images in the collection. However, this is a labor-intensive task considering the large number of images and tags. Instead, we evaluate the performance of our algorithm on the 81 tags in NUS-WIDE and the 18 tags in MIRFlickr where the ground-truth annotations of these tags have been provided. We adopt F-score to measure the image retagging results for each tag and then average them as the final evaluation measurement.

Firstly, we construct the tag-specific visual sub-vocabulary for each tag by employing the optimization algorithm in Section III-D. As aforementioned, our proposed noise-tolerant tag-specific visual sub-vocabulary (NTTSVSV) learning algorithm is able to well handle the images with noisy tags that may degrade the performance of the existing tag-specific visual sub-vocabulary learning algorithms. To confirm this superiority, we involve the state-of-the-art descriptive visual words (DVW) learning algorithm proposed by Zhang *et al.* in [26] as comparison. This algorithm learns the descriptive visual sub-vocabulary for each tag through a random walk procedure, in which the *term frequency-inverse document frequency* (tf-idf) measurements of the individual visual words are employed to estimate their descriptive power. Comparing with our proposed algorithm, the DVW algorithm does not provide any mechanism for handling the images with noisy tags. For the parameter settings for c_1 and c_2 in (15), we set $c_1, c_2 \in \{2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}$ and try various pairs of (c_1, c_2) in the optimization. For each parameter pair, we examine the value of the objective function that equals to $\sum_{i=1}^{N}(f_i-q_i)^2+\sum_{i\neq j} o_{ij}(f_i-f_j)^2-\sum_{k=1}^{M} P_{I_k}$ [see (15)] and the one with minimum residual will be picked as the optimal parameter setting for c_1 and c_2.[1] In the experiments, the optimal parameter settings of (c_1, c_2) on the NUS-WIDE and MIRFlickr are $(2^{-2}, 2^{1})$ and $(2^{1}, 2^{3})$, respectively. Furthermore, we set the size of each tag-specific visual sub-vocabulary K to be 200. For the parameters of the DVW learning algorithm, we choose them according to the suggested setting strategies in the original work.

Based on the learnt tag-specific visual sub-vocabulary for each tag obtained with each of the two tag-specific visual sub-vocabulary learning algorithms discussed above, we further conduct experiments to evaluate the performance of our proposed collaborative image retagging algorithm. The grid search strategy is employed to set the value of α and β in (7). Specifically, we set $\alpha, \beta \in \{2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}$ and seek the (α, β) pair with the best image retagging performance. In the experiments, the values of (α, β) are determined as $(2^{3}, 2^{5})$ and $(2^{-1}, 2^{5})$ on the NUS-WIDE and MIRFlickr dataset, respectively. To confirm the effectiveness of our proposed method, we compare the following six algorithms.[2]

- Baseline, i.e., the original tags provided by the users.
- Separated ReTagging (SRT). In this method, the tag information is propagated separately within the tag-specific graph without considering the collaboration among the tags. This can be achieved by setting the parameter α in (7) as zero and then implementing the multiplicative nonnegative iterative optimization. This is essentially the ML-LGC algorithm that converts the multi-label learning problem into a number of independent single label propagation problems. For fair comparison, the appropriate parameter setting for β is also determined via grid search from the interval of $\{2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}$, which is the same as the parameter setting strategy in our proposed collaborative retagging algorithm.
- Naive Collaborative ReTagging (NCRT). Rather than learning tag-specific visual sub-vocabulary for each tag, the general BoW model is adopted to represent the images. Consequently, the collaborative retagging algorithm is implemented within a unique graph that connects all the images in the collection. This can be realized by replacing the first term in (7) with a single graph while setting the parameter β to be 0. This is essentially the SMSE algorithm which performs multi-label propagation by taking advantages of label correlation. The parameter setting for β is also determined from the $\{2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}$ via grid search.
- Tag Refinement based on Visual and Semantic Consistency (TRVSC) [23]. In this method, the tag refinement task is

[1] Actually, the three terms in (15) (without multiplying c_1 and c_2) can be seen as three individual criteria for evaluating the obtained confidence vector $\mathbf{f}$. For any given $\mathbf{f}$, we substitute it (15) and calculate the summation of the three terms. The smaller the summation is, the better the obtained $\mathbf{f}$ is, which further indicates that the given parameter pair c_1 and c_2 are appropriate for generating reliable $\mathbf{f}$.

[2] For each algorithm that relies on tag-specific visual sub-vocabulary, we employ both our proposed RTSVSV learning algorithm and the DVW learning algorithm to generate the visual sub-vocabulary and then perform image retagging with the specific algorithm.

addressed by maximizing the consistency between "visual similarity" of the images and the "semantic similarity" of the tags through an iterative bound optimization procedure. There is only one tradeoff parameter in the algorithm. For fair comparison, we also use grid search method to search the appropriate parameter setting, where the parameter interval is also set as $\{2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}$.

- Content-based Annotation Refinement (CBAR) [22]. We choose this algorithm as comparison since it is the first work for the content-based image annotation refinement. Note that this algorithm is actually a Markov process which mainly relies on the a set of transition matrices between the tags. Once these matrices are calculated, the algorithm does not have any parameter. It is also worth noting that the CBAR algorithm along with the TRVST algorithm are actually the state-of-the-art algorithms for image tag refinement.
- Our proposed Tag-aware Collaborative ReTagging (TCRT). We construct tag-specific graph for each tag, respectively, and then perform the collaborative image retagging algorithm over the graphs. As aforementioned, the grid search method is employed to determine the appropriate parameter setting for α and β.

Note that each algorithm above can produce confidence scores for all tags. Therefore, we rank the tags of each image based on their confidence scores and then keep the top 5 tags as the image retagging result for each image. Table II shows the performance comparison of the above six algorithms on the NUS-WIDE and MIRFlickr datasets. From the results, we have the following observations. 1) The proposed algorithm achieves much better performance compared to the user-provided baseline. This clearly demonstrates the effectiveness of our proposed image retagging algorithm. 2) The algorithms based on our proposed robust tag-specific tag sub-vocabulary learning method outperform the DVW-based counterparts. This clearly demonstrates that our proposed visual sub-vocabulary learning method can alleviate problems caused by noisy labeled images. 3) The collaborative image retagging algorithms including NCRT and TCRT outperform the SRT algorithm. It indicates the effectiveness of enforcing the collaboration of different tags in the image retagging process. 4) TCRT algorithm results in better performance than the NCRT, CBAR, and TRVSC algorithms which only rely on the general BoW image representation, benefiting from the incorporation of tag-aware image representation. 5) The proposed algorithm outperforms the CBAR algorithm, which shows that our proposed algorithm is more appropriate for refining the imprecise user-provided tags associated with the social images. 6) Our proposed retagging algorithm shows up better performance than TRVSC algorithm, which owes to its explicit exploration on tag correlation and tag-specific image similarity. Fig. 4 illustrates the image retagging results for some exemplary images produced by our proposed algorithm. From the results, we can see that the refined tags are quite "computer-vision friendly" tags that have high correspondence with the visual content. This is reasonable since all content analysis techniques including our proposed image retagging scheme can only handle content-related tags. However, one may argue that those content-unrelated tags such as "canada", "madrid" are also useful. To handle this issue, we can choose to keep both the original tags and the ones obtained from retagging to support both specific and general image search.

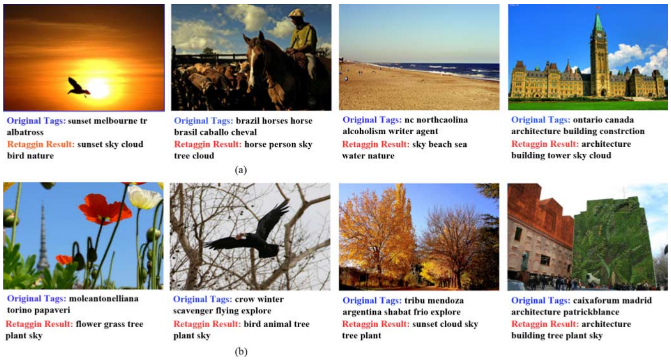

Original Tags: sunset melbourne tr albatross
Retaggin Result: sunset sky cloud bird nature
Original Tags: brazil horses horse brasil caballo cheval
Retaggin Result: horse person sky tree cloud
Original Tags: nc northcaolina alcoholism writer agent
Retaggin Result: sky beach sea water nature
Original Tags: ontario canada architecture building constrction
Retaggin Result: architecture building tower sky cloud
(a)
Original Tags: moleantonelliana torino papaveri
Retaggin Result: flower grass tree plant sky
Original Tags: crow winter scavenger flying explore
Retaggin Result: bird animal tree plant sky
Original Tags: tribu mendoza argentina shabat frio explore
Retaggin Result: sunset cloud sky tree plant
Original Tags: caixaforum madrid architecture patrickblance
Retaggin Result: architecture building tree plant sky
(b)

Fig. 4. Several exemplary images and their top-ranked tags produced by our proposed image retagging algorithm. (a) NUS-WIDE. (b) MIRFlickr.

TABLE II
PERFORMANCE COMPARISON (F-SCORE) OF DIFFERENT IMAGE RETAGGING ALGORITHMS ON NUS-WIDE DATASET AND MIRFLICKR DATASET, WHERE SRT AND TCRT ARE FURTHER COMPARED USING DIFFERENT TAG-SPECIFIC VISUAL SUB-VOCABULARY LEARNING ALGORITHMS

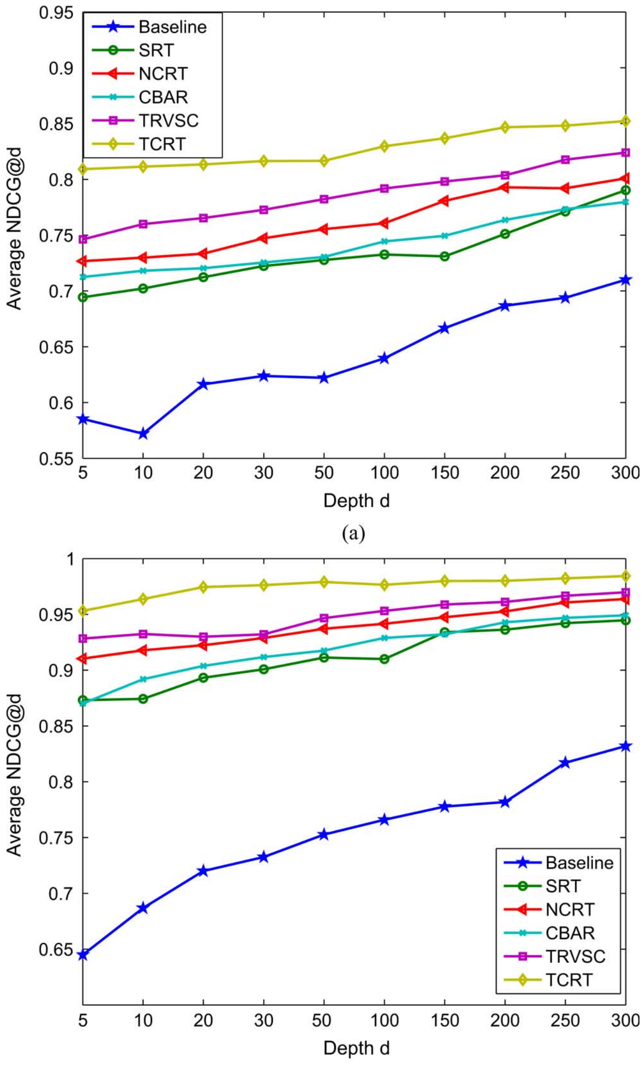
Method	NUS-WIDE	MIRFlickr
Baseline	0.45	0.20
SRT + DVW	0.47	0.27
SRT + NTTSVSV	0.50	0.31
NCRT	0.54	0.37
CBAR	0.49	0.31
TRVSC	0.55	0.37
TCRT + DVW	0.57	0.44
TCRT + NTTSVSV	**0.61**	**0.47**

C. Improving Image Ranking

Our proposed image retagging algorithm is able to improve the quality of the tags; hence, tag-based image search can be improved. In addition, our algorithm also introduces a ranking scheme for tag-based image search since it not only refines the tags but also assigns them confidence scores. We can regard the scores as the indication of relevance levels between the tags and images and provide a relevance-based image ranking.

To evaluate our algorithm on an image ranking task, we perform tag-based image search with the 81 tags in NUS-WIDE and the 18 tags in MIRFlickr as query keywords and then produce the ranking list for each query based on the confidence scores. Specifically, we involve the ranking results produced by different image retagging algorithms as comparison. In addition, we adopt the ranking results obtained based on initial user-provided tags solely as our baseline, in which the relevance score of a social image x with respect to the query tag t is defined as the average semantic similarity between tag t and all other tags associated with image x. Here the semantic similarity is also estimated with (3).

We use the popular normalized discount cumulative gain (NDCG) [38] as the performance evaluation measurement for image relevance ranking. Based on the obtained relevance scores for a query tag, we rank the images in a descending order. Then each image is manually labeled as one of three



(a)

(b)

Fig. 5. Performance comparison of the image ranking results using different algorithms on (a) NUS-WIDE and (b) MIRFlickr datasets.

levels with respect to the query tag: Most relevant (score 2), Partially relevant (score 1), and Irrelevant (score 0). Given a query tag, the NDCG score at the depth d in the ranked image list is defined as $NDCG@d = Z_d \sum_{i=1}^{d}(2^{r(i)} - 1)/log(1+i)$, where $r(i)$ is the relevance level of the ith ranked image and Z_d is a normalization constant such that the NDCG score for the optimal ranking is 1. After obtaining the NDCG measure for each query, we average them as the final evaluation metric. Fig. 5 illustrates the NDCG measurements at different return depths for different algorithms. From the results, we can have the following observations. 1) All algorithms outperform the baseline result significantly, which shows the effectiveness of the retagging-based image relevance ranking. 2) Our proposed TCRT algorithm produces consistent performance improvements over the other algorithms at variant return depths, and this confirms the superiority of the proposed algorithm over the other algorithms. 3) The performances of NCRT, CBAR, and TRVST outperform SRT, which owes to the fact that these algorithms have taken the correlations among the tags into consideration.

V. Conclusions

In this paper, we have introduced a novel image retagging scheme that aims at refining the quality of the tags associated with social images. We formulate the problem as an optimization scheme that simultaneously takes into account visual consistency of images over multiple tag-specific similarity graphs, semantic consistency of tags, and user-provided prior knowledge. It is solved with a collaborative tag propagation algorithm. A tag-specific visual sub-vocabulary learning algorithm is also proposed to construct those tag-specific similarity graphs. Experiments on two benchmark social image datasets have demonstrated its advantages over classical methods. Although we have put more emphasis on Flickr in this work, the proposed framework is flexible and can be easily extended to deal with a variety of online media repositories, such as Zooomr as well as any other image databases with noisy and incomplete tags.

References

[1] Flickr. [Online]. Available: http://www.flickr.com.

[2] Zooomr. [Online]. Available: http://www.zooomr.com.

[3] 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 Int. Workshop Multimedia information retrieval*, 2006, pp. 249–258.

[4] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2009, pp. 1–9.

[5] 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.

[6] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large dataset for non-parametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[7] H. Zhang, A. Berg, M. Maire, and M. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2126–2136.

[8] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. 21st National Conf. Artificial Intelligence*, 2006, pp. 421–426.

[9] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2009, pp. 951–958.

[10] D. Zhou, O. Bousqeut, T. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Advances in Neural Information Processing Systems*, 2004, pp. 321–328.

[11] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1719–1726.

[12] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a Sylvester equation," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 410–419.

[13] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Sep. 2003.

[14] E. Chang, K. Goh, G. Sychay, and G. Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.

[15] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[16] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2003, pp. 119–126.

[17] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2004, pp. 1063–6919.

[18] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proc. ACM Int. Conf. World Wide Web*, 2009, pp. 351–360.

[19] K. Weinberger, M. Slaney, and R. Zwol, "Resolving tag ambiguity," in *Proceeding ACM Int. Conf. Multimedia*, 2008, pp. 111–120.

[20] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotation by combining multiple evidence & WordNet," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 706–715.

[21] C. Fellbaum*, Wordnet: An Electronic Lexical Database*. Cambridge, MA: Bradford, 1998.

[22] C. Wang, F. Jing, L. Zhang, and H.-J Zhang, "Content-based image annotation refinement," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[23] D. Liu, X.-S. Hua, M. Wang, and H.-J Zhang, "Image retagging," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 491–500.

[24] Y. Liu, R. Jin, R. Sukthankar, and F. Jurie, "Unify the discriminative visual codebook generation with classifier training for object category recognition," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[25] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Proc. Advances in Neural Information Processing Systems*, 2006, pp. 985–992.

[26] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 75–84.

[27] D. Zhou, S. Zhou, K. Yu, X. Song, B. Tseng, H. Zha, and C. Giles, "Learning multiple graphs for document recommendations," in *Proc. ACM Int. Conf. World Wide Web*, 2008, pp. 141–150.

[28] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 1159–1166.

[29] R. Johnson and T. Zhang, "On the effectiveness of Laplacian normalization for graph semi-supervised learning," *J. Mach. Learn. Res.*, vol. 8, no. 7, pp. 1489–1517, 2007.

[30] R. Cilibrasi and P. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[31] F. Sha, Y. Lin, L. K. Saul, and D. Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Comput.*, vol. 19, no. 8, pp. 2004–2031, 2007.

[32] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[33] D. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *IEEE Trans. Autom. Control*, vol. AC-21, no. 2, pp. 174–184, Apr. 1976.

[34] H. Müller, S. M. Maillet, and T. Pun, "The truth about Corel–evaluation in image retrieval," in *Proc. Int. Conf. Image and Video Retrieval*, 2002, pp. 38–49.

[35] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation," in *Proc. Eur. Conf. Computer Vision*, 2006, pp. 1–15.

[36] M. Huiskes and M. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Information Retrieval*, 2008, pp. 39–43.

[37] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[38] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

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.

Shuicheng Yan (M'06–SM'09) is currently an Assistant Professor in the Department of Electrical and Computer Engineering at National University of Singapore and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). His research areas include computer vision, multimedia, and machine learning, and he has authored or co-authored about 200 technical papers over a wide range of research topics.

Dr. Yan is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and has been serving as a Guest Editor of special issues for IEEE TRANSACTIONS ON MULTIMEDIA and *Computer Vision and Image Understanding*. He received the Best Paper Awards from ACM MM'10, ICME'10, ICIMCS'09, and PREMIA 2008 & 2010 Best Student Paper Award, the winner prize of the classification task in PASCAL VOC2010, and the honorable mention prize of the detection task in PASCAL VOC2010.

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.

Dr. Hua serves as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and is 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. degree 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, China. 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 is currently on the editorial board of the PROCEEDINGS OF THE IEEE. He is a fellow of the ACM.