Tag Completion for Image Retrieval

Lei Wu *Member, IEEE*, Rong Jin, Anil K. Jain, *Fellow, IEEE*

Abstract—Many social image search engines are based on keyword/tag matching. This is because tag based image retrieval (TBIR) is not only efficient but also effective. The performance of TBIR is highly dependent on the availability and quality of manual tags. Recent studies have shown that manual tags are often unreliable and inconsistent. In addition, since many users tend to choose general and ambiguous tags in order to minimize their efforts in choosing appropriate words, tags that are specific to the visual content of images tend to be missing or noisy, leading to a limited performance of TBIR. To address this challenge, we study the problem of *tag completion* where the goal is to automatically fill in the missing tags as well as correct noisy tags for given images. We represent the image-tag relation by a *tag matrix*, and search for the optimal tag matrix consistent with both the observed tags and the visual similarity. We propose a new algorithm for solving this optimization problem. Extensive empirical studies show that the proposed algorithm is significantly more effective than the state-of-the-art algorithms. Our studies also verify that the proposed algorithm is computationally efficient and scales well to large databases.

Index Terms—tag completion, matrix completion, tag-based image retrieval, image annotation, image retrieval, metric learning.

1 Introduction

With the remarkable growth in the popularity of social media websites, there have been a proliferation of digital images on the Internet, which have posed a great challenge for large-scale image search. Most image retrieval methods can be classified into two categories: content based image retrieval [41], [36] (CBIR) and keyword/tag based image retrieval [32], [58] (TBIR).

CBIR takes an image as a query, and identifies the matched images based on the visual similarity between the query image and gallery images. Various visual features, including both global features [33] (e.g., color, texture, and shape) and local features [16] (e.g., SIFT keypoints), have been studied for CBIR. Despite the significant efforts, the performance of available CBIR systems is usually limited [38], due to the semantic gap between the low-level visual features used to represent images and the high level semantic meaning behind images.

To overcome the limitations of CBIR, TBIR represents the visual content of images by manually assigned keywords/tags. It allows a user to present his/her information need as a textual query, and find the relevant images based on the match between the textual query and the manual annotations of images. Compare to CBIR, TBIR is usually more accurate in identifying relevant images [24] by alleviating the challenge arising from the semantic gap. TBIR is also more efficient in retrieving relevant images than CBIR because it can be formulated as a document retrieval problem and therefore can be efficiently implemented using the inverted index technique [29].

However, the performance of TBIR is highly dependent on the availability and quality of manual tags. In most cases, the tags are provided by the users who upload their images to the social media sites (e.g., Flickr), and are therefore often inconsistent and unreliable in describing the visual content of images, as indicated in a recent study on Flickr data [47]. In particular, according to [37], in order to minimize the effort in selecting appropriate words for given images, many users tend to describe the visual content of images by general, ambiguous, and sometimes inappropriate tags, as explained by the principle of least effort [25]. As a result, the manually annotated tags tend to be noisy and incomplete, leading to a limited performance of TBIR. This was observed in [44], where, on average, less than 10% of query words were used as image tags, implying that many useful tags were missing in the database. In this work, we address this challenge by automatically filling in the missing tags and correcting the noisy ones. We refer to this problem as the *tag completion* problem.

One way to complete the missing tags is to directly apply automatic image annotation techniques [52], [17], [20], [42] to predict additional keywords/tags based on the visual content of images. Most automatic image annotation algorithms cast the problem of keyword/tag prediction into a set of binary classification problems, one for each keyword/tag. The main shortcoming of this approach is that in order to train a reliable model for keyword/tag prediction, it requires a large set of training images with *clean* and *complete* manual annotations. Any missing or noisy tag could potentially lead to a biased estimation of prediction models, and consequentially suboptimal performances. Unfortunately, the annotated tags for

L. Wu, R. Jin, and A.K. Jain are with department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA. A. K. Jain is also with the Dept. of Brain & Cognitive Engineering, Korea University, Anamdong, Seongbukgu, Seoul 136-713, Republic of Korea. E-mail: leiwu@live.com, rongjin@cse.msu.edu, jain@cse.msu.edu.
Manuscript received August 22, 2011.

most Web images are incomplete and noisy, making it difficult to directly apply the method of automatic image annotation.

Besides the classification approaches, several advanced machine learning approaches have been applied to image annotation, including annotation by search [54], tag propagation [38], probabilistic relevant component analysis (pRCA) [33], distance metric learning [19], [46], [49], [28], Tag transfer [15], and reranking [61]. Similar to the classification based approaches for image annotation, to achieve good performance, these approaches require a large number of well annotated images, and therefore are not suitable for the tag completion problem.

The limitation of current automatic image annotation approaches motivates us to develop a new computational framework for tag completion. In particular, we cast tag completion into a problem of matrix completion: we represent the relation between tags and images by a *tag matrix*, where each row corresponds to an image and each column corresponds to a tag. Each entry in the tag matrix is a real number that represents the relevance of a tag to an image. Similarly, we represent the partially and noisy tagged images by an observed tag matrix, where an entry (i, j) is marked as 1 if and only if image i is annotated by keyword/tag j. Besides the tag information, we also compute the visual similarity between images based on the extracted visual features. We search for the optimal tag matrix that is consistent with both the observed tag matrix and the pairwise visual similarity between images. We present an efficient learning algorithm for tag completion that scales well to large databases with millions of images. Our extensive empirical studies verify both the efficiency and effectiveness of the proposed algorithm in comparison to the state-of-the-art algorithms for automatic image annotation.

The rest of this paper is organized as follows. In Section 2, we overview the related work on automatic image annotation. Section 3 defines the problem of tag completion and provides a detailed description for the proposed framework and algorithm. Section 4 summarizes the experimental results on automatic image annotation and tag based search. Section 5 concludes this study with suggestions for future work.

2 Related Work

Numerous algorithms have been proposed for automatic image annotation (see [18] and references therein). They can roughly be grouped into two major categories, depending on the type of image representations used. The first group of approaches are based upon global image features [31], such as color moment, texture histogram, etc. The second group of approaches adopts the local visual features. [30], [43], [48] segment image into multiple regions, and represent each region by a vector of visual features. Other approaches [22], [56], [45] extend the bag-of-features or bag-of-words representation, which was originally developed for object recognition, for automatic image annotation. More recent work [34], [27] improves the performance of automatic image annotation by taking into account the spatial dependence among visual features. Other than predicting annotated keywords for the entire image, several algorithms [11] have been developed to predict annotations for individual regions within an image. Despite these developments, the performance of automatic image annotation is far from being satisfactory. A recent report [38] shows that the-state-of-the-art methods for automatic image annotation, including Conditional Random Fields (CRM) [52], inference network approach (infNet) [17], Nonparametric Density Estimation (NPDE) [7], and supervised multi-class labeling (SML) [20], are only able to achieve $16\% \sim 28\%$ for average precision, and $19\% \sim 33\%$ for average recall, for key benchmark datasets Corel5k and ESP Game. Another limitation of most automatic image annotation algorithms is that they require fully annotated images for training, making them unsuitable for the tag completion problem.

Several recent works explore multi-label learning techniques for image annotation that aim to exploit the dependence among keywords/tags. Ramanan et al. [13] proposed a discriminative model for multi-label learning. Zhang et al. [40] proposed a lazy learning algorithm for multi-label prediction. Hariharan et al. [8] proposed max-margin classifier for large scale multi-label learning. Guo et al. [55] applied the conditional dependency networks to structured multi-label learning. An approach for batch-mode image retagging is proposed in [32]. Zha et al. [60] proposed a graph based multi-label learning approach for image annotation. Wang et al. [26] proposed a multi-label learning approach via maximum consistency. Chen et al. [21] proposed an efficient multi-label learning based on hypergraph regularization. Bao et al. [9] proposed a scalable multi-label propagation approach for image annotation. Liu et al. [59] proposed a constrained non-negative matrix factorization method for multi-label learning. Unlike the existing approaches for multi-label learning that assume complete and perfect class assignments, the proposed approach is able to deal with noisy and incorrect tags assigned to the images. Although a matrix completion approach was proposed in [1] for transductive classification, it differs from the proposed work in that it applies Euclidean distance to measure the difference between two training instances while the proposed approach introduces a distance metric to better capture the similarity between two instances.

Besides the classification approaches, several recent works on image annotation are based on distance metric learning. Monay et al. [19] proposed to annotate the image in a latent semantic space. Wu and Hoi et al. [46], [49], [33] proposed to learn a metric

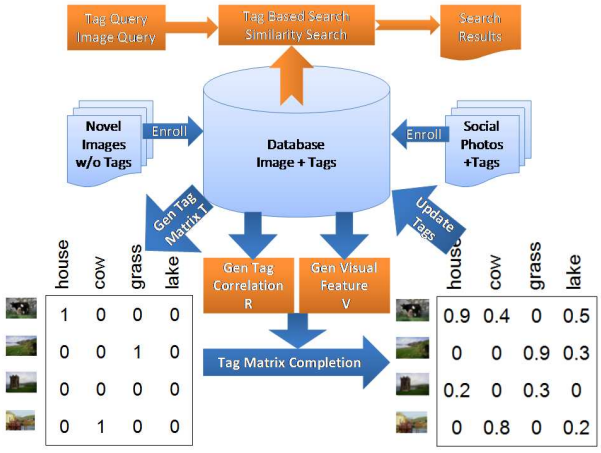


Fig. 1. The framework for tag matrix completion and its application to image search. Given a database of images with some initially assigned tags, the proposed algorithm first generates a tag matrix denoting the relation between the images and initially assigned tags. It then automatically complete the tag matrix by updating the relevance score of tags to all the images. The completed tag matrix will be used for tag based image search or image similarity search.

to better capture the image similarity. Zhuang et al. [28] proposed a two-view learning algorithm for tag re-ranking. Li et al. [53] proposed a neighbor voting method for social tagging. Similar to the classification based approaches, these methods require clean and complete image tags, making them unsuitable for the tag completion problem.

Finally, our work is closely related to tag refinement [24]. Unlike the proposed work that tries to complete the missing tags and correct the noisy tags, tag refinement is only designed to remove noisy tags that do not reflect the visual content of images.

3 Tag Completion

We first present a framework for tag completion, and then describe an efficient algorithm for solving the optimization problem related to the proposed framework.

3.1 A Framework for Tag Completion

Figure 1 illustrates the tag completion task. Given a binary image-tag matrix (*tag matrix* for brief), our goal is to automatically complete the tag matrix with real numbers, that indicate the probability of assigning the tags to the images. Given the completed tag matrix, we can run TBIR to efficiently and accurately identify the relevant images for textual query.

Let n and m be the number of images and unique tags, respectively. Let $\widehat{T} \in \mathbb{R}^{n\times m}$ be the partially observed tag matrix derived from user annotations, where $\widehat{T}_{i,j}$ is set to one if tag j is assigned to image i and zero otherwise. We denote by $T \in \mathbb{R}^{n\times m}$ the completed tag matrix that needs to be computed. In order to complete the partially observed tag matrix $\widehat{T}$, we further represent the visual content of images by matrix $V \in \mathbb{R}^{n\times d}$, where d is the number of visual features and each row of V corresponds to the vector of visual features for an image. Finally, to exploit the dependence among different tags, we introduce the tag correlation matrix $R \in \mathbb{R}^{m\times m}$, where $R_{i,j}$ represents the correlation between tag i and j. Following [10], we compute the correlation score between two tags i and j as follows

$$R_{i,j} = \frac{f_{i,j}}{f_i + f_j - f_{i,j}}$$

where f_i and f_j are the occurrence of tags i and j, and f_{ij} is the co-occurrence of tags i and j. Note that f_i, f_j and $f_{i,j}$ are statistics collected from the partially observed tag matrix $\widehat{T}$. Our goal is to reconstruct the tag matrix T based on the partially observed tag matrix $\widehat{T}$, the visual representation of image data V, and the tag correlation matrix R. To narrow down the solution for the complete tag matrix T, we consider the following three criterions for reconstructing T.

There are three important constraints in the matrix completion algorithm to avoid trivial solutions.

First, the complete tag matrix T should be similar to the partially observed matrix $\widehat{T}$. We add this constraint by penalizing the difference between T and $\widehat{T}$ with a Frobensius norm, and we prefer the solution T with small $\|T - \widehat{T}\|_F^2$.

Second, the complete tag matrix T should reflect the visual content of images represented by the matrix V, where each image is represented as a row vector (visual feature vector) in V. However, since the relationship between tag matrix T and the visual feature matrix V is unknown, it is difficult to implement this criterion directly. To address this challenge, we propose to exploit this criterion by comparing image similarities based on visual content with image similarities based on the overlap in annotated tags. More specifically, we compute the visual similarity between image i and j as $\mathbf{v}_i^\top \mathbf{v}_j$, where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the ith and jth rows of matrix V. Given the complete tag matrix T, we can also compute the similarity between image i and j based on the overlap between their tags, i.e., $\mathbf{t}_i^\top \mathbf{t}_j$, where $\mathbf{t}_i$ and $\mathbf{t}_j$ are the ith and jth rows of matrix T. If the complete tag matrix T reflects the visual content of images, we expect $|\mathbf{v}_i^\top \mathbf{v}_j - \mathbf{t}_i^\top \mathbf{t}_j|^2$ to be small for any two images i and j. As a result, we expect a small value for $\sum_{i,j=1}^{n} |\mathbf{v}_i^\top \mathbf{v}_j - \mathbf{t}_i^\top \mathbf{t}_j|^2 = \|TT^\top - VV^\top\|_F^2$.

Finally, we expect the complete matrix T to be consistent with the correlation matrix R, and therefore a small value for $\|T^\top T - R\|_F^2$. Combining the three criterions, we have the following optimization problem for finding the complete tag matrix T.

$$\min_{T\in\mathbb{R}^{n\times m}} \|TT^\top - VV^\top\|_F^2 + \lambda\|T^\top T - R\|_F^2 + \eta\|T - \widehat{T}\|_F^2 \quad (1)$$

where $\lambda > 0$ and $\eta > 0$ are parameters whose values will be decided by cross validation.

There are, however, two problems with the formulation in (1). First, the visual similarity between images i and j is computed by $\mathbf{v}_i^\top \mathbf{v}_j$, which assumes that all visual features are equally important in determining the visual similarity. Since some visual features may be more important than the others in deciding the tags for images, we introduce a vector $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{R}_+^d$, here w_i is used to represent the importance of the ith visual feature. Using the weight vector $\mathbf{w}$, we modify the visual similarity measure as $\mathbf{v}_i^\top A \mathbf{v}_j$, where $A = \text{diag}(\mathbf{w})$ is a diagonal matrix with $A_{i,i} = w_i$. Second, the complete tag matrix T computed by (1) may be dense in which most of the entries in T are non-zero. But, on the other hand, we generally expect that only a small number of tags will be assigned to each image, and as a result, a sparse matrix for T. To address this issue, we introduce into the objective function an L_1 regularizer for T, i.e., $\|T\|_1 = \sum_{i=1}^n \sum_{j=1}^m |T_{i,j}|$. Incorporating these two modifications into (1), we have the final optimization problem for tag completion

$$\min_{T \in \mathbb{R}^{n \times m}, \mathbf{w} \in \mathbb{R}_+^d} \quad \mathcal{L}(T, \mathbf{w}) \tag{2}$$

where

$$\begin{aligned}\mathcal{L}(T, \mathbf{w}) &= \|TT^\top - V\text{diag}(\mathbf{w})V^\top\|_F^2 \\ &+\lambda\|T^\top T - R\|_F^2 + \eta\|T - \widehat{T}\|_F^2 + \mu\|T\|_1 + \gamma\|\mathbf{w}\|_1\end{aligned}$$

Note that in (2) we further introduce an L_1 regularizer for $\mathbf{w}$ to generate a sparse solution for $\mathbf{w}$.

3.2 Optimization

To solve the optimization problem in (2), we develop a subgradient descent based approach (Algorithm 1). Compared to the other optimization approaches such as Newton's method and interior point methods [39], the subgradient descent approach is advantageous in that its computational complexity per iteration is significantly lower, making it suitable for large image datasets.

The subgradient descent approach is an iterative method. At each iteration t, given the current solution T_t and $\mathbf{w}_t$, we first compute the subgradients of the objective function $\mathcal{L}(T, \mathbf{w})$. Define

$$G = T_t T_t^\top - V\text{diag}(\mathbf{w}_t)V^\top,\ H = T_t^\top T_t - R$$

We compute the subgradients as

$$\nabla_T \mathcal{L}(T_t, \mathbf{w}_t) = 2GT_t + 2\lambda T_t H + 2\eta(T_t - \widehat{T}) + \mu\Delta \tag{3}$$

$$\nabla_{\mathbf{w}} \mathcal{L}(T_t, \mathbf{w}_t) = -2\text{diag}(V^\top G V) + \gamma\delta \tag{4}$$

where $\Delta \in \mathbb{R}^{n \times m}$ and $\delta \in \mathbb{R}^d$ are defined as

$$\Delta_{i,j} = \text{sgn}(T_{i,j}),\ \delta_i = \text{sgn}(w_i)$$

Here, $\text{sgn}(z)$ outputs 1 when $z > 0$, -1 when $z < 0$, and a random number uniformly distributed between -1 and $+1$ when $z = 0$. Given the subgradients, we update the solution for T and w as follows

$$\begin{aligned}T_{t+1} &= T_t - \eta_t \nabla_T \mathcal{L}(T_t, \mathbf{w}_t) \\ \mathbf{w}_{t+1} &= \pi_\Omega\left(\mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} \mathcal{L}(T_t, \mathbf{w}_t)\right)\end{aligned}$$

where η_t is the step size of iteration t, and $\Omega = \{\mathbf{w} \in \mathbb{R}_+^d\}$ and $\pi_\Omega(\mathbf{w})$ projects a vector $\mathbf{w}$ into the domain Ω to ensure that the learned weights are non-negative.

One problem with the above implementation of the subgradient descent approach is that the immediate solutions $T_1, T_2, \ldots,$ may be dense, leading to a high computational cost in matrix multiplication. We address this difficulty by exploring the method developed for composite function optimization [12]. In particular, we rewrite $\mathcal{L}(T, \mathbf{w})$ as $\mathcal{L}(T, \mathbf{w}) = A(T, \mathbf{w}) + \gamma\|\mathbf{w}\|_1 + \mu\|T\|_1$, where

$$\begin{aligned}A(T, \mathbf{w}) = \quad &\|TT^\top - V\text{diag}(\mathbf{w})V^\top\|_F^2 + \\ &\lambda\|T^\top T - R\|_F^2 + \eta\|T - \widehat{T}\|_F^2\end{aligned}$$

At each iteration t, we compute the subgradients $\nabla_T A(T_t, \mathbf{w}_t)$ and $\nabla_{\mathbf{w}} A(T_t, \mathbf{w}_t)$, and update the solutions for T and $\mathbf{w}$ according to the theory of composite function optimization [3]

$$T_{t+1} = \arg\min_T \frac{1}{2}\|T - \widehat{T}_{t+1}\|_F^2 + \mu\eta_t\|T\|_1 \tag{5}$$

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}_{t+1}\|_F^2 + \gamma\eta_t\|\mathbf{w}\|_1 \tag{6}$$

where η_t is the step size for the t-th iteration and $\widehat{T}_{t+1}$ and $\widehat{\mathbf{w}}_{t+1}$ are given by

$$\widehat{T}_{t+1} = T_t - \eta_t \nabla_T A(T_t, \mathbf{w}_t), \tag{7}$$

$$\widehat{\mathbf{w}}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} A(T_t, \mathbf{w}_t) \tag{8}$$

Using the result in [3], the solutions to (5) and (6) are given by

$$T_{t+1} = \max\left(\mathbf{0}, \widehat{T}_{t+1} - \mu\eta_t \mathbf{1}_n \mathbf{1}_m^\top\right) \tag{9}$$

$$\mathbf{w}_{t+1} = \max\left(\mathbf{0}, \widehat{\mathbf{w}}_{t+1} - \gamma\eta_t \mathbf{1}_d\right) \tag{10}$$

where $\mathbf{1}_d$ is vector of n dimensions with all its elements being 1. As indicated in the equations in Eq. (9) and (10), any entry in T_t ($\mathbf{w}_{t+1}$) which is less than $\mu\eta_t$ ($\gamma\eta_t$ respectively) will become zero, leading to sparse solutions for T and $\mathbf{w}$ by using the theory of composite function optimization.

Our final question is how to decide the step size η_t. There are common choices: $\eta_t = 1/\sqrt{t}$ or $\eta_t = 1/t$. We set $\eta_t = 1/t$, which appears to yield a faster convergence than $\eta_t = 1/\sqrt{t}$. Algorithm 1 summarizes the key steps of the subgradient descent approach.

3.3 Discussion

Although the proposed formulation is non-convex and therefore cannot guarantee to find the global optimal, this however is not a serious issue from the viewpoint of learning theory [5]. This is because as

Algorithm 1 Tag Completion Algorithm (TMC)

1: INPUT:
- Observed tag matrix: $\widehat{T} \in \mathbb{R}^{n \times m}$
- Parameters: γ, η, λ, and μ
- Convergence threshold: ε

2: OUTPUT: the complete tag matrix T
3: Compute the tag correlation matrix $R = \widehat{T}^\top \widehat{T}$
4: Initialize $\mathbf{w}_1 = \mathbf{1}_d$, $T_1 = \widehat{T}$, and $t = 0$
5: **repeat**
6: Set $t = t + 1$ and stepsize $\eta_t = 1/t$
7: Compute $\widehat{T}_{t+1}$ and $\widehat{\mathbf{w}}_{t+1}$ according to (8)
8: Update the solutions T_{t+1} and $\mathbf{w}_{t+1}$ according to (9) and (10)
9: **until** convergence: $\|\mathcal{L}(T_t, \mathbf{w}_t) - \mathcal{L}(T_{t+1}, \mathbf{w}_{t+1})\| \leq \varepsilon \|\mathcal{L}(T_t, \mathbf{w}_t)\|$

the empirical error goes down during the process of optimization, the generalization error will become the leading term in the prediction error. As a result, finding the global optima will not have a significant impact on the final prediction result. In fact, [51] shows that only an approximately good solution would be enough to achieve similar performance as the exact optimal one. To alleviate the problem of local optima, we run the algorithm 20 times and choose the run with the lowest objective function.

The convergence rate for the adopted subgradient descent method is $O(1/\sqrt{t})$, where t is the number of iterations. The space requirement for the algorithm is $O(n \times m)$, where n is the number of images and m is the number of unique tags.

We finally note that since the objective of this work is to complete the tag matrix for all the images, it belongs to the category of transductive learning. In order to turn a transductive learning method into an inductive one, one common approach is to retrain a prediction model based on outputs from the transduction method [2]. A similar approach can be used for the proposed approach to make predictions for out-of-samples.

3.4 Tag Based Image Retrieval

Given the complete tag matrix T obtained by solving the optimization problem in (2), we briefly describe how to utilize the matrix T for tag based image retrieval.

We first consider the simplest scenario when the query consists of a single-tag. Given a query tag j, we simply rank all the gallery images in the descending order of their relevance scores to tag j, corresponding to the jth column in matrix T. Now consider the general case when a textual query is comprised of multiple tags. Let $\mathbf{q} = (q_1, \ldots, q_m)^\top \in \{0, 1\}^m$ be a query vector, where $q_i = 1$ if the ith tag appears in the query and $q_i = 0$ otherwise. A straightforward approach is to compute the tag based similarity between the query and the images by $T\mathbf{q}$. A shortcoming of this similarity measure is that it does not take into account the correlation between tags. To address this limitation, we refine the similarity between the query and the images by $TW\mathbf{q}$, where $W = \pi_{[0,1]}(T^\top T)$ is the tag correlation matrix estimated based on the complete tag matrix T. Here, $\pi_{0,1}(A)$ projects every entry of A into the range between 0 and 1.

4 Experiments

We evaluate the quality of the completed tag matrix on two tasks: automatic image annotation and tag based image retrieval.

Four benchmark datasets are used in this study:

- *Corel* dataset [43]. It consists of $4,993$ images, with each image being annotated by at most five tags. There are a total of 260 unique keywords used in this dataset.
- *Labelme* photo collection. It consists of 2,900 online photos, annotated by 495 non-abstract noun tags. The maximum number of annotated tags per image is 48.
- *Flickr* photo collection. It consists of one million images that are annotated by more than 10,000 tags. The maximum number of annotated tags per image is 76. Since most of the tags are only used by a small number of images, we reduce the vocabulary to the first $1,000$ most popular tags used in this dataset, which reduces the database to 897,500 images.
- *TinyImg* image collection. It consists of 79,302,017 images collected from the web, annotated by 75,062 non-abstract noun tags. The maximum number of annotated tags per image is 82. Similar to the Flickr photo collection, we reduce the vocabulary to the first $1,000$ most popular tags in the dataset, which reduces the database size to 997,420 images.

Table 1 summarizes the statistics of the four datasets used in our study.

For the Corel data, we use the same set of features as [38], including SIFT local features and a robust hue descriptor that is extracted densely on multi-scale grids of interest points. Each local feature descriptor is quantized to one of 100,000 visual words that are identified by a k-means clustering algorithm. Given the quantized local features, we represent each image by a bag-of-words histogram. For Flickr and Labelme photo collections, we adopt the compact SIFT feature representation [57]. It first extracts SIFT features from an image, and then projects the SIFT features to a space of 8 dimensions using the Principle Component Analysis (PCA). We then cluster the projected low dimensional SIFT features into 100,000 visual words, and represent the visual content of images by the histogram of the visual words. For TinyImg dataset, since the images are of low resolution, we adopt a

TABLE 1
Statistics for the datasets used in the experiments.

	Corel	Labelme	Flickr	TinyImg
No. of Images	4,993	2,900	897,500	997,420
Vocabulary Size	260	495	1,000	1,000
No. of Tags per Image (mean/max)	3.4/5	10.5/48	12.7/76	14.4/82
No. of Image per Tag (mean/max)	58.6/1,004	67.1/379	416.5/76,890	575.5/87,120

global SIFT descriptor to represent the visual content of each image.

To make a fair comparison with other state-of-the-art methods, we adopt average precision, and average recall [6] as the evaluation metrics. It computes the precision and recall for every test image by comparing the auto-annotations to the ground truth, and then takes the average of precisions and recalls over all the test images as the final evaluation result.

For large scale tag-based image retrieval, since it is very difficult to get the ground truth for evaluating the recall, we adopt the Mean Average Precision (MAP) as the evaluation metric, which can be calculated by manually check the correctness of retrieved images. MAP takes into account the rank of returned images when computing average precision, and consequentially heavily penalizes the retrieval results when the relevant images are returned at low rank.

4.1 Experiment (I): Automatic Image Annotation

We first evaluate the proposed algorithm for tag completion by automatic image annotation. We randomly separate each dataset into two collections. One collection consisting of 80% of images is used as training data, and the other collection consisting of 20% of images is used as testing data. We repeat the experiment 20 times. Each run adopts a new separation of the collections. We report the result based on the average over the 20 trials.

To run the proposed algorithm for automatic image annotation, we simply view test images as special cases of partially tagged images, i.e., no tag is observed for test images. We thus apply the proposed algorithm to complete the tag matrix that includes both training and test images. We then rank the tags for test images in the descending order based on their relevance scores in the completed tag matrix, and return the top ranked tags as the annotations for the test images.

We compare the proposed tag matrix completion (TMC) algorithm to the following six state-of-the-art algorithms for automatic image annotation: (i) *Multiple Bernoulli Relevance Models (MBRM)* [50] that models the joint distribution of annotation tags and visual features by a mixture distribution, (ii) *Joint Equal Contribution method (JEC)* [4] that finds appropriate annotation words for a test image by a k nearest neighbor classifier that combines multiple distance measures derived from different visual features, (iii) *Inference Network method (InfNet)* [17] that applies the Bayesian network to model the relationship between visual features and annotation words, (iv) *Large scale max-marin multi-label classification (LM3L)* [8], that overcomes the training bias by incorporating correlation prior, (v) *Tag Propagation method (TagProp)* [38] that propagates the label information from the labeled instances to the unlabeled instances via a weighted nearest neighbor graph, (vi) *social tag relevance by neighbor voting (TagRel)* [53], that explores the tag relevance based on a neighborhood voting approach. The key parameter for TagProp is the number of nearest neighbors used to determine the nearest neighbor graph. We vary this number from 1 to 10, and set it to be 5 because it yields the best empirical performance. For the other baselines, we adopt the same parameter configuration as described in their original reports. For the proposed TMC method, we set $\eta = 1, \mu = 1, \gamma = 1, \lambda = 10$, according to a cross validation procedure. We will discuss the parameter setting in more details in the later part of this section.

Table 2 summarizes the average precision/recall for the first five and ten returned tags for four different datasets. Note that for the Flickr and TinyImg datasets, we only show the results for TagProp, TagRel, and TMC, because the other baseline methods are unable to run over such large datasets. We observe that for all datasets, TMC outperforms the baseline methods significantly in terms of both precision and recall. We also observed that as the number of returned tags increases from five to ten, the precision usually declines while the recall usually improves. This is called precision-recall tradeoff, a phenomenon that is well known in information retrieval.

We now examine a more realistic setup of automatic image annotation in which each training image is partially annotated. This also allows us to test the sensitivity of the proposed method to the number of initially assigned tags. To this end, unlike the previous experiment where all the training images are completely labeled, we vary the number of observed tags for each training image, denoted by n, from 1 to 5 [1] to create the scenario of partially tagged images. We vary the number of predicted tags from 5, 10, 15 to

1. We set the maximum number of observed tags to 5 because the minimum number of tags assigned to an image is 6 except for Corel data. For Corel data we choose 4 as the maximum number of observed tags since the maximum number of annotated tags for the Corel dataset is 5

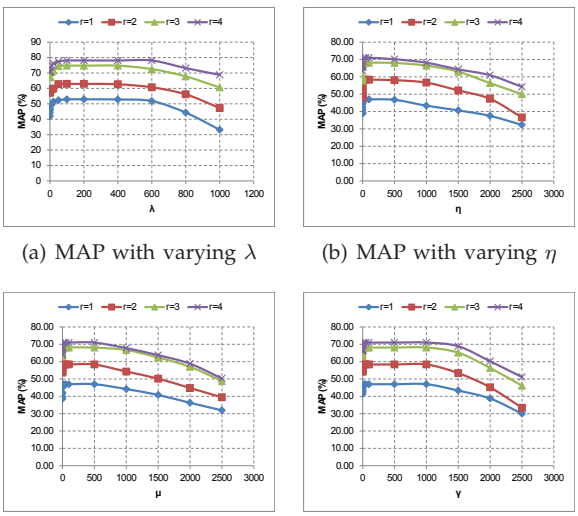
TABLE 2
Average precision and recall for four datasets.

Corel	MBRM	JEC	InfNet	LM3L	TagProp	TagRel	TMC
$AP@5(\%)$	24 ± 1.4	27 ± 1.3	17 ± 1.1	32 ± 2.0	33 ± 2.2	33 ± 1.8	43 ± 1.4
$AR@5(\%)$	25 ± 1.6	32 ± 1.2	24 ± 1.4	51 ± 1.8	52 ± 2.6	52 ± 1.6	64 ± 1.2
$AP@10(\%)$	17 ± 1.5	20 ± 1.7	10 ± 1.0	25 ± 1.8	26 ± 1.2	26 ± 1.7	34 ± 1.8
$AR@10(\%)$	28 ± 1.5	34 ± 1.4	26 ± 1.7	53 ± 1.6	54 ± 2.1	54 ± 1.8	66 ± 1.3
Labelme	MBRM	JEC	InfNet	LM3L	TagProp	TagRel	TMC
$AP@5(\%)$	23 ± 2.1	22 ± 1.1	25 ± 1.4	24 ± 1.9	28 ± 2.3	29 ± 2.0	37 ± 1.8
$AR@5(\%)$	24 ± 1.7	24 ± 1.1	28 ± 1.2	28 ± 1.8	35 ± 2.1	35 ± 1.4	47 ± 1.4
$AP@10(\%)$	17 ± 1.8	16 ± 1.2	17 ± 1.3	19 ± 1.7	20 ± 1.8	21 ± 1.7	25 ± 1.5
$AR@10(\%)$	27 ± 1.9	28 ± 1.6	30 ± 1.1	30 ± 1.6	37 ± 1.4	37 ± 1.3	49 ± 1.6
Flickr	MBRM	JEC	InfNet	LM3L	TagProp	TagRel	TMC
$AP@5(\%)$	-	-	-	-	30 ± 2.6	31 ± 1.9	39 ± 1.9
$AR@5(\%)$	-	-	-	-	38 ± 2.1	39 ± 1.8	57 ± 1.8
$AP@10(\%)$	-	-	-	-	22 ± 2.1	23 ± 1.3	31 ± 1.5
$AR@10(\%)$	-	-	-	-	43 ± 1.5	42 ± 1.7	59 ± 1.7
TinyImg	MBRM	JEC	InfNet	LM3L	TagProp	TagRel	TMC
$AP@5(\%)$	-	-	-	-	27 ± 2.1	30 ± 1.8	37 ± 1.5
$AR@5(\%)$	-	-	-	-	31 ± 2.2	37 ± 1.4	52 ± 1.4
$AP@10(\%)$	-	-	-	-	19 ± 1.8	23 ± 1.5	30 ± 1.7
$AR@10(\%)$	-	-	-	-	34 ± 2.1	40 ± 1.3	55 ± 1.7

20, and measure the MAP results for each number of predicted tags, as shown in Table 3. We only include TagProp and TagRel in comparison because the other methods are unable to handle partially annotated training images. It is not surprising to observe that annotation performance of all methods improves with an increasing number of observed annotations. For all the cases, TMC outperforms TagProp and TagRel significantly.

Finally, we examine the sensitivity of the proposed method to the parameter setup. There are four different parameters that need to be determined in the proposed algorithm (i.e., λ, η, μ, and γ). In order to understand how these parameters affect the annotation performance, we conduct four sets of experiments: (i) fixing $\eta = 1, \mu = 1, \gamma = 1$ and varying λ from 0.1 to $1,000$, (ii) fixing $\lambda = 10, \mu = 1, \gamma = 1$ and varying η from 0.1 to $2,500$, (iii) fixing $\lambda = 10, \eta = 1, \gamma = 1$ and varying μ from 0.1 to $2,500$, and (iv) fixing $\lambda = 10, \mu = 1, \eta = 1$ and varying γ from 0.1 to $2,500$. We report the results on the Corel dataset with the number of observed annotations set to 2. The annotation performance, measured in MAP for the four sets of experiments is shown in Figure 2. First, according to the performance with varying λ, we observe that overall the performance is improved as we increase λ, but the performance starts to decline when $\lambda \geq 100$. Second, we observe that the performance of the algorithm is insensitive to these parameters if they fall in certain range ($\lambda < 400$,$\eta < 500$, $\mu < 1,000$, $\gamma < 1,500$), and the performance deteriorates significantly when they are outside the range. Based on the above observations, we set $\lambda = 1$, $\mu = 1$, $\eta = 1$, and $\gamma = 10$ for all the experiments.

We note that these four parameters play different roles in the learning procedure. Parameters μ and γ are used to prevent over-fitting, and to generate a sparse solution. Parameter η controls the constraint on the T by keeping it close to the observations. Parameter λ determines the trade-off between the first two terms in the objective function. Since the larger the λ, the stronger the algorithm is in enforcing the constraint with respect to the tag correlation, it implies that the tag correlation information is important to the tag completion problem.



(a) MAP with varying λ (b) MAP with varying η

(c) MAP with varying μ (d) MAP with varying γ

Fig. 2. Mean Average Precision (MAP) of TMC method on Corel dataset for image annotation with varying λ, η, μ and γ. r is the number of observed annotation tags.

4.2 Experiment (II): Tag based Image Retrieval

Unlike the experiments for image annotation where each dataset is divided into a training set and a testing set, for the experiment of tag-based image retrieval, we include all the images from the dataset except

the queries as the gallery images for retrieval. Similar to the previous experiments, we vary the number of observed tags from 1 to 4. Similar to the previous experiments, we only compare the proposed algorithm to TagProp and TagRel because the other approaches were unable to handle the partially tagged images. Below, we first present the results for queries with single-tag, and then the results for queries consisting of multiple tags.

4.2.1 Results for Single-tag Queries

In this experiment, we restrict ourselves to the queries that consist of a single-tag. Since every tag can be used as a query, we have in total 260 queries for the Corel5k dataset, 495 queries for Labelme dataset, and 1,000 queries for the the Flickr and TinyImage datasets. For generating the initial observed tags, we selected the observed tags by random picking n tags in the annotated tags. If the total number of annotated tags is less than n, we ignore the sample. We adopt a simple rule determining the relevance: an image is relevant if its annotation contains the query. We note that this rule has been used in a number of studies on image retrieval [14], [33], [35]. Besides the TagProp and TagRel methods, we also introduce a reference method that returns a gallery image if its observed tags include the query word. By comparing to the reference method, we will be able to determine the improvement made by the proposed matrix completion method. Table 4 shows the MAP results for the four datasets. We observe that (i) TagProp, TagRel and TMC perform significantly better than the reference method, and (ii) TMC outperforms TagProp and TagRel significantly for all cases. Figure 4 shows examples of single-tag queries and the images returned by different methods.

4.2.2 Experimental results for queries with multiple tags

Similar to the previous experiment, we vary the number of observed tags from 1 to 4 for this experiment. To generate queries with multiple tags, we randomly select 200 images from the Flickr dataset, and use the annotated tags of the randomly selected images as the queries. To determine the relevance of returned images, we manually check for every query the first 20 retrieved images by the proposed method and by the TagProp method. For all the methods in comparison, we follow the method presented in Section 3.3 for calculating the tag-based similarity between the textual query and the completed tags of gallery images. For the TagProp method, we fill in the tag matrix T by applying the label propagation method in TagProp before computing the tag-based similarity. Table 5 shows the MAP scores for the first 5, 10, 15, and 20 images that are returned for each query. For complete comparison, we include two additional baseline methods:

- the *reference* method that computes the similarity between a gallery image and a query based on the occurrence of query tags in the observed annotation of the gallery image, and rank images in the descending order of their similarities;
- the content based image retrieval (*CBIR*) method that represents images by a bag-of-words model using a vocabulary of 10,000 visual words generated by the k-means clustering algorithm, and computes the similarity between a query image and a gallery image using the well-known TF/IDF weighting in text retrieval.

According to Table 5, we first observe a significant difference in MAP scores between CBIR and TBIR (i.e. the reference method, TagProp and TMC), which is consistent with the observations reported in the previous study [23]. Second, we observe that the proposed method TMC outperforms all the baseline methods significantly. Figure 5 shows examples of queries and images returned by the proposed method and the baselines.

4.3 Convergence and Computational Efficiency

We evaluate the computational efficiency by the running time for image annotation. All the algorithms are run on the Intel(R) Core(TM)2 Duo CPU @3.00GHz and 6 GB RAM machine. Table 6 summarizes the running times of both the proposed method and the baseline methods. Note that for the Flickr and TinyImg dataset, we only report the running time for three methods, because the other methods either have memory issue or take more than several days to finish. We observe that although both the TagProp and the TagRel methods are significantly faster than the proposed method for the small dataset, i.e. Corel and Labelme datasets, they show comparable running

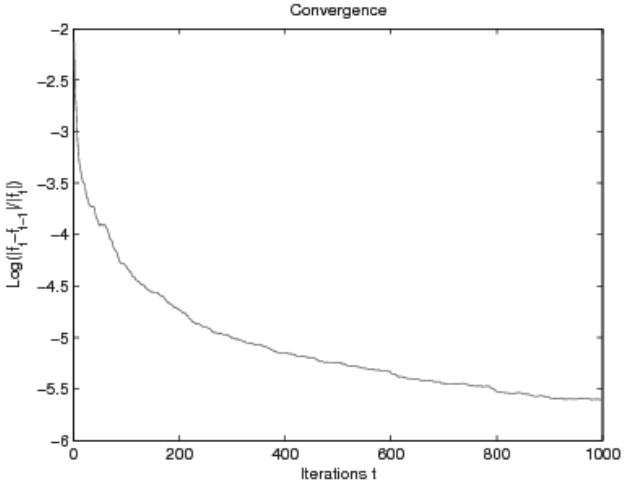


Fig. 3. Convergence of the proposed tag matrix completion method on the Flickr dataset. The x axis is the number of iterations, and y axis is $\log \frac{\|L_{t+1}-L_t\|}{\|L_t\|}$.

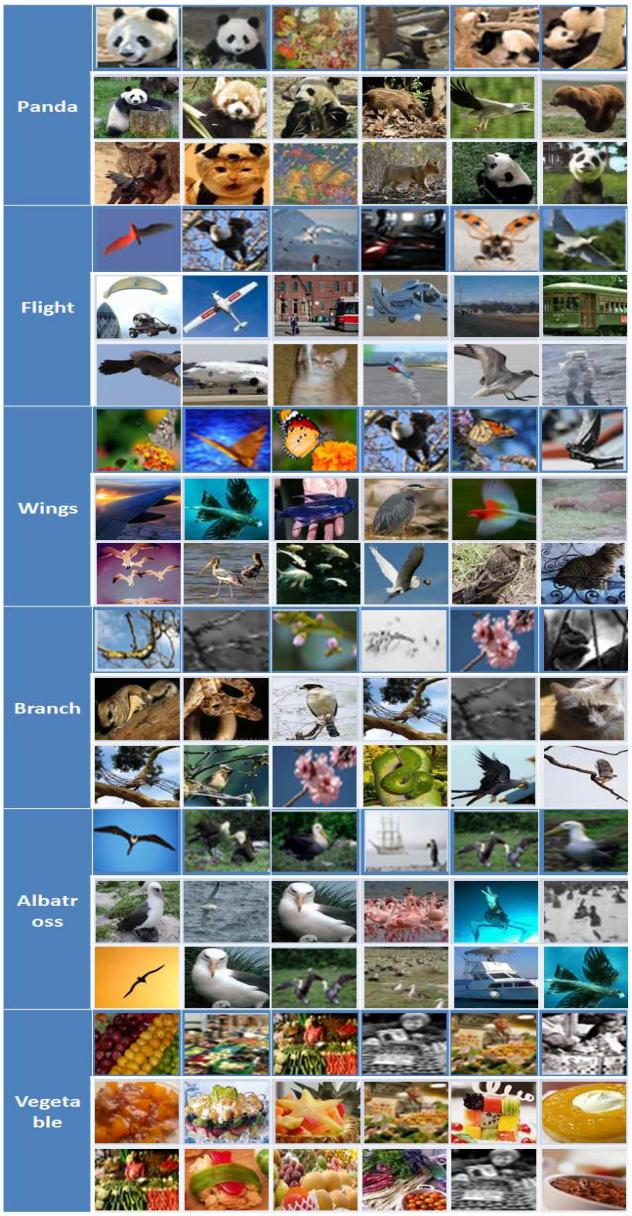


Fig. 4. Illustration of the single-tag based image search. The word on the left is the query, and images on its right are the search results. The images displayed on the three rows are the results returned by the proposed TMC method, the TagProp method, and TagRel method, respectively. The blue outline are the results for the proposed methods, the white lines are the results for the baseline methods.

Fig. 5. Illustration of the similarity retrieval result based on multiple tag querying. Each image on the left is the query image whose annotated tags are used as the query word. The images to its right, from top to bottom, are the results returned by the proposed TMC method, the TagProp method, and the TagRel method, respectively.

time for the two large datasets, i.e., the flickr and Tinyimage datasets. Figure 3 shows how the objective function value is reduced over the iterations. We observe that the proposed algorithm is able to converge within 300 iterations when threshold ε is set to 10^{-5}, and around 1,000 iterations when threshold is $10^{-5.5}$.

5 Conclusions

We have proposed a tag matrix completion method for image tagging and image retrieval. We consider the image-tag relation as a tag matrix, and aim to optimize the tag matrix by minimizing the difference between tag based similarity and visual content based similarity. The proposed method falls into the category of semi-supervised learning in that both tagged images and untagged images are exploited to find the optimal tag matrix. We evaluate the proposed

method for tag completion by performing two sets of experiments, i.e., automatic image annotation and tag based image retrieval. Extensive experimental results on four open benchmark datasets show that the proposed method significantly outperforms several state-of-the-art methods for automatic image annotation. In future work, we plan to exploit computationally more efficient approaches for tag completion based on the theory of compressed sensing and matrix completion.

6 Acknowledgement

This work was supported in part by National Science Foundation (IIS-0643494), US Army Research (W911NF-11-1-0383) and Office of Navy Research (Award N000141210431). Part of Anil Jains research was supported by the WCU (World Class University) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

References

[1] A. B. Goldberg, X. Zhu, B. Recht, J. Xu, and R. D. Nowak. Transduction with matrix completion: Three birds with one stone. *NIPS*, pages 757–765., 2010.

[2] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. *Uncertainty in Artificial Intelligence*, pages 148–155. 1998.

[3] A. Ioffe. Composite optimization: Second order conditions, value functions and sensityvity. *Analysis and Optimization of Systes*, volume 144 of *Lecture Notes in Control and Information Sciences*, pages 442–451, 1990.

[4] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. *IEEE ECCV*, pages 316–329, 2008.

[5] A. Rakhlin. Applications of empirical processes in learning theory: Algorithmic stability and generalization bounds. *PhD Thesis, MIT*, 2006.

[6] A. Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pages 35–42, 2001.

[7] A. Yavlinsky, E. Schofield, and S. Rger. Automated image annotation using global features and robust nonparametric density estimation. *Proceedings of the International Conference on Image and Video Retrieval*, pages 507–517, 2005.

[8] B. Hariharan, S. V. N. Vishwanathan, and M. Varma. Large Scale Max-Margin Multi-Label Classification with Prior Knowledge about Densely Correlated Labels. *ICML*, 2010.

[9] B.-K. Bao, B. Ni, Y. Mu, and S. Yan. Efficient region-aware large graph construction towards scalable multi-label propagation. *Pattern Recognition*, pages 598–606, 2011.

[10] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. *Proceeding of the 17th International Conference on World Wide Web*, pages 327–336, 2008.

[11] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, pages 157–173, 2008.

[12] C. Cartis, N. I. Gould, and P. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *Optimization Journal*, 2011

[13] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, pages 1–12, 2011.

[14] C. Haruechaiyasak and C. Damrongrat. Improving social tag-based image retrieval with cbir technique. *Proceedings of the role of digital libraries in a time of global change, and 12th International Conference on Asia-Pacific Digital Libraries* pages 212–215, 2010.

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE TPAMI*, 33:2368–2382, 2011.

[16] D. G. Lowe. Object recognition from local scale-invariant features. *IEEE ICCV*, 1999.

[17] D. Metzler and R. Manmatha. An inference network approach to image retrieval. *Proceedings of the Internatinal Conference on Image and Video Retrieval*, pages 42–50, 2004.

[18] E. Akbas and F. T. Y. Vural. Automatic image annotation by ensemble of visual descriptors. *IEEE CVPR*, pages 1–8, 2007.

[19] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. *ACM Multimedia*, pages 348–351, 2004.

[20] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE TPAMI*, 2007.

[21] G. Chen, J. Zhang, F. Wang, C. Zhang, and Y. Gao. Efficient multi-label classification with hypergraph regularization. *IEEE CVPR*, pages 1658–1665, 2009.

[22] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, IEEE ECCV*, pages 1–22, 2004.

[23] G. Wang, D. Hoiem, and D. A. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. *IEEE ICCV*, pages 428–435, 2009.

[24] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. *ACM Multimedia*, pages 461–470, 2010.

[25] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. *Proceedings of the 16th International Conference on World Wide Web*, pages 211–220, 2007.

[26] H. Wang and J. Hu. Multi-label image annotation via maximum consistency. *IEEE ICIP*, pages 2337–2340, 2010.

[27] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE TPAMI*, pages 30:985–1002, 2008.

[28] J. Zhuang and S. C. Hoi. A two-view learning approach for image tag ranking. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 625–634, 2011.

[29] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 2006.

[30] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 2003.

[31] K.-S. Goh, E. Y. Chang, and B. Li. Using one-class and two-class svms for multiclass image annotation. *IEEE TKDE*, pages 1333–1346, 2005.

[32] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. *IEEE CVPR*, pages 3440–3446, 2010.

[33] L. Wu, S. C. Hoi, J. Zhu, R. Jin, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. *ACM Multimedia*, 2009.

[34] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Visual language modeling for image classification. *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 115–124, 2007.

[35] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. *Proceedings of the 18th International Conference on World Wide Web*, pages 361–361, 2009.

[36] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. H. Hoi, and M. Satyanarayanan. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *IEEE TPAMI*, 32(1):30–44, Jan. 2010.

[37] M. E. I. Kipp and G. D. Campbell. Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices. *Annual General Meeting of the American Society for Information Science and Technology*, 2006.

[38] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. *IEEE ICCV*, pages 309–316, 2009.

[39] M. H. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, pages 39–56, 2005.

[40] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, pages 2038–2048, 2007.

[41] M. S. Lew. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, pages 1–19, 2006.

[42] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based

image tagging. *IEEE TPAMI*, 33:1281–1294, 2011.
[43] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *IEEE ECCV*, pages 97–112, 2002.
[44] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 195–206, 2008.
[45] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, pages 249–258, 2008.
[46] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 197–206, 2011.
[47] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. *ACM Multimedia*, pages 977–986, 2006.
[48] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. *ACM Multimedia*, pages 892–899, 2004.
[49] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. *IEEE CVPR*, pages 2072–2078, 2006.
[50] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. *IEEE CVPR* pages 1002–1009, 2004.
[51] S. Shalev-Shwartz and N. Srebro. Svm optimization: inverse dependence on training set size. *ICML*, pages 928–935, 2008.
[52] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *NIPS*, 2003.
[53] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, pages 1310–1322, 2009.
[54] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. *IEEE CVPR*, pages 1483–1490, 2006.
[55] Y. Guo and S. Gu. Multi-label classification using conditional dependency networks. *IJCAI*, pages 1300–1305, 2011.
[56] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 494–501, 2007.
[57] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. *IEEE CVPR*, pages 506–513, 2004.
[58] Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Textual Query of Personal Photos Facilitated by Large-Scale Web Data. *IEEE TPAMI*, 33:1022–1036, 2011.
[59] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. *AAAI*, pages 421–426, 2006.
[60] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, pages 97–103, 2009.
[61] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE TPAMI*, 33:1991–2001, 2011.

Lei Wu received Ph.D. in Dept. of Electronic Engineering and Information Science and B.S. degree in Special Class for Gifted Young (SCGY) from University of Science and Technology of China. His research interests include distance metric learning, multimedia retrieval, and object recognition. Dr. Wu has also filed four U.S. patents, one of the patents received Microsoft Patent Award in 2009. Dr. Lei Wu received Microsoft Fellowship 2007, the President Special Scholarship of Chinese Academy of Science in 2010, which is the highest honor granted by Chinese Academy of Science. Dr. Lei Wu's PhD Thesis is honored Outstanding Doctoral Thesis by Chinese Academy of Science in 2011. Dr. Lei Wu has served as a technical editor for International Journal of Digital Content Technology of its Application, Lead Guest Editor at an SI in Advances in Multimedia, program committee member at AAAI 2012, ACM Multimedia 2012, ACM CIKM 2012, IJCAI 2011, IEEE ICIP 2011, ACM SIGMAP 2011-2012, etc. He also served as a technical reviewer for multiple international conferences and journals.

Rong Jin focuses his research on statistical machine learning and its application to information retrieval. He has worked on a variety of machine learning algorithms and their application to information retrieval, including retrieval models, collaborative filtering, cross lingual information retrieval, document clustering, and video/image retrieval. He has published over eighty conference and journal articles on related topics. Dr. Jin holds a Ph.D. in Computer Science from Carnegie Mellon University. He received the NSF Career Award in 2006.

Anil K. Jain is a university distinguished professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1991-1994). The holder of six patents in the area of fingerprints, he is the author of a number of books, including Handbook of Fingerprint Recognition (2009), Handbook of Biometrics (2007), Handbook of Multibiometrics (2006), Handbook of Face Recognition (2005), BIOMETRICS: Personal Identification in Networked Society (1999), and Algorithms for Clustering Data (1988). He served as a member of the Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. Dr. Jain received the 1996 IEEE TRANSACTIONS ONNEURALNETWORKS Outstanding Paper Award and the Pattern Recognition Society best paper awards in 1987, 1991, and 2005. He is a fellow of the AAAS, ACM, IAPR, and SPIE. He has received Fulbright, Guggenheim, Alexander von Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu awards. ISI has designated him a highly cited researcher. According to Citeseer, his book Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice-Hall, 1988) is ranked $\#93$ in most cited articles in computer science.

TABLE 3
Performance of automatic image annotation with varying numbers of observed tags on four datasets. r is the number of observed tags.

Corel	MAP@5			MAP@10			MAP@15			MAP@20		
	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC
r=1	47.95	47.84	49.48	41.85	41.83	45.01	36.60	37.01	39.24	30.98	31.14	33.52
r=2	48.99	49.98	59.36	43.70	44.71	53.00	37.58	38.14	47.48	32.50	33.39	41.13
r=3	57.52	58.27	70.60	52.16	52.48	64.17	46.54	47.38	59.02	41.34	42.31	52.93
r=4	60.23	62.25	74.99	55.07	56.53	67.73	50.07	52.25	62.13	44.41	46.69	55.99
Labelme	MAP@5			MAP@10			MAP@15			MAP@20		
	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC
r=1	50.12	50.33	52.18	43.78	43.81	47.33	41.87	41.67	42.47	35.63	35.12	37.87
r=2	51.56	52.33	62.26	45.22	46.27	55.67	42.21	43.34	50.37	37.83	37.91	44.76
r=3	59.33	61.53	72.28	54.73	54.89	66.36	51.33	52.65	62.87	46.77	45.28	55.72
r=4	62.65	65.78	76.82	57.91	58.18	69.59	56.41	57.27	65.58	50.36	51.27	57.99
r=5	67.14	68.96	80.03	63.36	63.76	72.14	59.18	59.73	68.34	53.12	54.16	60.03
Flickr	MAP@5			MAP@10			MAP@15			MAP@20		
	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC
r=1	61.65	62.09	71.26	61.18	61.54	71.23	60.55	61.16	69.13	57.35	58.01	70.02
r=2	72.33	73.27	78.91	71.83	71.93	78.86	71.54	71.97	78.55	69.10	69.91	76.76
r=3	76.48	77.75	83.83	76.36	77.17	83.59	75.47	76.52	81.37	71.77	72.28	83.72
r=4	78.96	79.98	86.78	78.82	79.08	86.61	77.35	77.87	84.91	76.41	76.27	80.99
n=5	80.87	81.16	88.53	80.45	80.76	88.52	79.99	80.57	86.39	76.74	77.16	86.03
TinyImg	MAP@5			MAP@10			MAP@15			MAP@20		
	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC	TagProp	TagRel	TMC
r=1	50.11	51.15	61.16	48.32	48.11	58.13	43.32	42.13	49.21	40.23	40.33	45.11
r=2	61.23	62.43	65.21	58.25	58.23	61.58	50.54	50.42	56.34	47.57	47.46	52.25
r=3	65.54	66.22	69.43	63.36	64.43	64.62	58.67	58.15	61.53	53.86	54.75	58.64
r=4	67.65	68.34	71.76	64.76	65.77	67.23	60.22	61.64	65.55	56.35	57.23	61.21
r=5	69.66	70.52	73.54	66.77	67.56	69.74	62.37	62.67	67.37	60.31	61.44	64.61

TABLE 4
Mean Average Precision (MAP) for TBIR using single-tag queries. The number of observed annotated tags varies from 1 to 4.

Corel	MAP@5				MAP@10			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	79.48	85.61	86.12	89.77	77.09	84.08	85.12	89.54
r=2	84.73	89.41	90.08	95.88	84.89	88.61	89.91	95.51
r=3	86.94	91.53	91.98	97.12	87.03	90.21	91.17	97.64
r=4	87.83	93.11	94.21	98.85	88.09	92.28	92.86	98.12
	MAP@15				MAP@20			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	76.59	81.69	82.28	86.85	74.39	78.30	79.18	85.36
r=2	82.27	84.80	85.51	92.54	79.28	82.30	82.81	89.63
r=3	83.72	87.56	88.09	94.74	81.58	85.60	86.31	91.77
r=4	85.90	89.65	90.19	95.39	83.48	86.64	87.09	92.17
Labelme	MAP@5				MAP@10			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	71.12	87.11	87.21	90.71	67.23	86.53	87.23	91.51
r=2	76.33	90.23	91.18	96.17	69.43	88.14	91.24	92.33
r=3	77.42	91.13	91.68	97.31	73.52	91.63	92.55	93.25
r=4	80.56	91.42	92.01	98.19	76.67	92.66	92.83	95.63
	MAP@15				MAP@20			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	66.44	83.22	83.21	88.22	63.33	75.44	76.78	83.54
r=2	68.57	86.32	86.32	92.53	64.24	76.23	76.75	84.23
r=3	69.86	88.43	88.46	93.75	66.53	77.46	77.46	85.48
r=4	71.54	91.52	90.74	94.46	67.37	79.36	79.47	87.82
Flickr	MAP@5				MAP@10			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	73.93	77.78	78.81	84.83	71.30	75.03	76.16	83.59
r=2	79.69	82.52	83.38	90.99	77.36	80.24	82.07	90.61
r=3	81.95	85.46	86.51	92.84	78.90	83.18	83.88	92.12
r=4	84.35	86.46	87.79	94.04	81.35	83.77	84.14	93.57
	MAP@15				MAP@20			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	65.74	68.44	69.01	79.67	51.97	54.68	55.14	71.81
r=2	70.24	73.85	74.24	85.13	57.16	59.98	69.62	77.51
r=3	72.61	77.23	78.16	89.14	59.20	63.21	64.17	81.38
r=4	75.57	78.28	79.35	90.00	62.10	65.11	66.62	82.06
TinyImg	MAP@5				MAP@10			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	59.32	75.42	76.32	82.23	52.32	72.34	72.23	80.42
r=2	64.65	77.24	77.53	84.45	56.34	74.53	74.89	81.53
r=3	66.64	79.64	78.53	87.56	58.53	76.26	77.12	83.67
r=4	67.24	82.17	81.77	90.26	62.25	79.74	79.99	96.25
	MAP@15				MAP@20			
	Ref.	TagProp	TagRel	TMC	Ref.	TagProp	TagRel	TMC
r=1	47.31	68.34	68.77	78.53	44.32	61.32	61.12	72.13
r=2	49.63	70.53	71.87	79.42	49.56	63.65	62.42	74.42
r=3	54.84	72.52	72.79	80.47	51.77	65.24	64.55	77.52
r=4	58.23	74.34	75.98	85.85	53.36	67.77	67.63	80.77

TABLE 5
Mean Average Precision (MAP) for TBIR using multiple tag queries. r stands for the number of observed tags.

Corel	MAP@5					MAP@10				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	86.32	67.33	88.32	88.33	93.53	78.23	63.21	84.32	84.23	90.44
r=2	90.42	67.33	92.53	91.46	95.32	82.42	63.21	85.57	84.44	91.25
r=3	92.22	67.33	94.66	93.87	98.16	84.53	63.21	87.76	86.64	92.74
r=4	93.48	67.33	95.47	95.42	99.74	86.23	63.21	90.87	89.57	94.32
	MAP@15					MAP@20				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	74.32	58.43	79.43	79.12	85.32	70.25	50.35	71.42	71.45	78.32
r=2	76.56	58.43	80.67	81.45	87.45	72.64	50.35	73.66	72.63	81.42
r=3	78.67	58.43	83.36	83.57	89.63	74.38	50.35	74.47	74.26	82.46
r=4	81.64	58.43	85.76	85.32	91.42	75.84	50.35	75.73	75.11	84.37
Labelme	MAP@5					MAP@10				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	83.12	65.44	86.33	86.32	92.23	79.34	61.24	83.45	82.23	89.42
r=2	83.32	65.44	87.43	87.53	94.42	80.42	61.24	84.65	84.42	91.54
r=3	85.53	65.44	88.54	89.43	96.53	81.45	61.24	85.76	86.64	92.56
r=4	96.64	65.44	91.64	91.45	98.56	83.65	61.24	87.87	87.53	94.77
	MAP@15					MAP@20				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	75.23	55.43	80.32	81.44	86.32	71.23	41.34	75.32	75.32	79.33
r=2	77.45	55.43	81.44	81.65	87.45	72.45	41.34	76.45	77.45	80.23
r=3	79.67	55.43	82.54	82.53	88.64	73.64	41.34	76.67	78.23	82.45
r=4	81.75	55.43	83.52	84.24	90.23	74.43	41.34	78.86	79.53	84.13
Flickr	MAP@5					MAP@10				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	81.03	61.89	86.71	87.27	91.28	75.88	52.43	81.62	82.19	86.75
r=2	85.00	61.89	89.79	90.19	95.20	80.05	52.43	83.61	84.34	91.61
r=3	87.16	61.89	91.31	91.98	97.53	83.04	52.43	85.48	85.17	93.28
r=4	88.49	61.89	92.97	93.31	99.53	84.84	52.43	88.21	89.98	95.23
	MAP@15					MAP@20				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	72.06	41.18	76.72	77.27	82.77	66.35	31.82	70.71	71.11	77.05
r=2	76.39	41.18	78.77	79.35	86.30	71.23	31.82	74.75	75.28	82.41
r=3	78.40	41.18	80.47	81.17	88.82	72.98	31.82	73.95	74.45	83.80
r=4	80.90	41.18	82.47	83.37	90.57	76.65	31.82	77.39	78.25	87.37
TinyImg	MAP@5					MAP@10				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	67.34	57.43	79.34	80.32	89.34	61.42	50.35	73.53	74.34	84.34
r=2	69.13	57.43	81.41	81.53	91.53	63.57	50.35	74.23	76.54	86.42
r=3	71.54	57.43	83.45	83.17	93.37	64.84	50.35	76.34	77.65	90.45
r=4	73.64	57.43	85.76	84.24	95.54	65.75	50.35	79.64	79.46	93.24
	MAP@15					MAP@20				
	Ref.	CBIR	TagProp	TagRel	TMC	Ref.	CBIR	TagProp	TagRel	TMC
r=1	61.34	39.32	69.34	70.32	78.34	56.32	33.56	62.34	64.23	71.32
r=2	63.23	39.32	70.54	71.25	80.54	57.45	33.56	64.23	65.34	74.36
r=3	65.43	39.32	72.26	73.56	83.65	59.54	33.56	67.45	66.56	76.54
r=4	67.31	39.32	74.67	75.23	85.45	60.34	33.56	68.56	68.34	78.36

TABLE 6
Running time (seconds) of different methods.

TIME	MBRM	JEC	InfNet	LM3L	TagProp	TagRel	TMC
Corel	34.18	23.19	33.27	46.79	2.63	4.28	31.89
Labelme	27.33	18.81	22.37	34.41	1.98	2.28	28.18
Flickr	-	-	-	-	8,770	7,740	10,630
TinyImg	-	-	-	-	3,560	2,710	4,413