Learning to Rank Image Tags With Limited Training Examples

Songhe Feng, Zheyun Feng, and Rong Jin

***Abstract*— With an increasing number of images that are available in social media, image annotation has emerged as an important research topic due to its application in image matching and retrieval. Most studies cast image annotation into a multilabel classification problem. The main shortcoming of this approach is that it requires a large number of training images with clean and complete annotations in order to learn a reliable model for tag prediction. We address this limitation by developing a novel approach that combines the strength of tag ranking with the power of matrix recovery. Instead of having to make a binary decision for each tag, our approach ranks tags in the descending order of their relevance to the given image, significantly simplifying the problem. In addition, the proposed method aggregates the prediction models for different tags into a matrix, and casts tag ranking into a matrix recovery problem. It introduces the matrix trace norm to explicitly control the model complexity, so that a reliable prediction model can be learned for tag ranking even when the tag space is large and the number of training images is limited. Experiments on multiple well-known image data sets demonstrate the effectiveness of the proposed framework for tag ranking compared with the state-of-the-art approaches for image annotation and tag ranking.**

***Index Terms*— Automatic image annotation, tag ranking, matrix recovery, low-rank, trace norm.**

I. Introduction

THE popularity of digital cameras and mobile phone cameras leads to an explosive growth of digital images that are available over the internet. How to accurately retrieve images from enormous collections of digital photos has become an important research topic. Content-based image retrieval (CBIR) addresses this challenge by identifying the matched images based on their visual similarity to a query image [1]. However due to the semantic gap between the low-level visual features used to represent images and the high-level semantic tags used to describe image content, limited performance is achieved by CBIR techniques [1], [2]. To address the limitation of CBIR, many algorithms have been developed for tag based image retrieval (TBIR) that represents images by manually assigned keywords/tags. It allows a user to present his/her information needs by textual information and find the relevant images based on the match between the textual query and the assigned image tags. Recent studies have shown that TBIR is usually more effective than CBIR in identifying the relevant images [3].

Since it is time-consuming to manually label images, various algorithms have been developed for automatic image annotation [4]–[11]. Many studies view image annotation as a multi-label classification problem [12]–[17], where in the simplest case, a binary classification model is built for each tag. The main shortcoming of this approach is that in order to train a reliable model for tag prediction, it requires a large number of training images with clean and complete annotations. In this work, we focus on the tag ranking approach for automatic image annotation [18]–[25]. Instead of having to decide, for each tag, if it should be assigned to a given image, the tag ranking approach ranks tags in the descending order of their relevance to the given image. By avoiding making binary decision for each tag, the tag ranking approach significantly simplifies the problem, leading to a better performance than the traditional classification based approaches for image annotation [25]. In addition, studies have shown that tag ranking approaches are more robust to noisy and missing tags than the classification approaches [24].

Although multiple algorithms have been developed for tag ranking, they tend to perform poorly when the number of training images is limited compared to the number of tags, a scenario often encountered in real world applications [26]. In this work, we address this limitation by casting tag ranking into a matrix recovery problem [27]. The key idea is to aggregate the prediction models for different tags into a matrix. Instead of learning each prediction model independently, we propose to learn all the prediction models simultaneously by exploring the theory of matrix recovery, where a trace norm regularization is introduced to capture the dependence among different tags and to control the model complexity. We shown, both theoretically and empirically, that with the introduction of trace norm regularizer, a reliable prediction model can be learned for tag ranking even when the tag space is large and the number of training images is small. We note that although

Manuscript received April 24, 2014; revised August 29, 2014, November 8, 2014, and January 4, 2015; accepted January 12, 2015. Date of publication January 22, 2015; date of current version February 17, 2015. This work was supported in part by the U.S. Army Research Office under Grant W911NF-11-1-0383, in part by the National Natural Science Foundation of China under Grant 61472028, Grant 61372148, Grant 61300071, Grant 61272352, and Grant 61033013, in part by the Fundamental Research Funds for the Central Universities under Grant 2014JBM035, in part by the Beijing Natural Science Foundation under Grant 4142045, and in part by the Beijing Higher Education Young Elite Teacher Project under Grant YETP0547. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dimitrios Tzovaras.

S. Feng is with the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China, and also with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China (e-mail: shfeng@bjtu.edu.cn).

Z. Feng and R. Jin are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: fengzhey@msu.edu; rongjin@cse.msu.edu).

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/TIP.2015.2395816

the trace norm regularization has been studied extensively for classification [28], [29], this is the first study that exploits trace norm regularization for tag ranking.

The rest of the paper is organized as follows. Section II reviews the related work on automatic image annotation and tag ranking. In Section III, we introduce the formulation details of the proposed framework and describe an efficient algorithm for computing the optimal solution. Experimental results on five different image data sets are reported and analyzed in Section IV. Finally, Section V concludes this work with future directions.

II. Related Work

In this section we review the related work on automatic image annotation and tag ranking. Given the rich literature on both subjects, we only discuss the studies closely related to this work, and refer the readers to [30] and [31] for the detailed surveys of these topics.

A. Automatic Image Annotation

Automatic image annotation aims to find a subset of keywords/tags that describes the visual content of an image. It plays an important role in bridging the semantic gap between low-level features and high-level semantic content of images. Most automatic image annotation algorithms can be classified into three categories (i) generative models that model the joint distribution between tags and visual features, (ii) discriminative models that view image annotation as a classification problem, and (iii) search based approaches. Below, we will briefly review approaches in each category.

Both mixture models and topic models, two well known approaches in generative model, have been successfully applied to automatic image annotation. In [12], a Gaussian mixture model is used to model the dependence between keywords and visual features. In [32]–[34], kernel density estimation is applied to model the distribution of visual features and to estimate the conditional probability of keyword assignments given the visual features. Topic models annotate images as samples from a specific mixture of topics, which each topic is a joint distribution between image features and annotation keywords. Various topic models have been developed for image annotation, including probabilistic latent semantic analysis (pLSA) [35], latent Dirichlet allocation [36], [37] and hierarchical Dirichlet processes [38]. Since a large number of training examples are needed for estimating the joint probability distribution over both features and keywords, the generative models are unable to handle the challenge of large tag space with limited number of training images.

Discriminative models [39], [40] views image annotation as a multi-class classification problem, and learns one binary classification model for either one or multiple tags. A 2D multiresolution hidded Markov model (MHMM) is proposed to model the relationship between tags and visual content [41]. A structured max-margin algorithm is developed in [42] to exploit the dependence among tags. One problem with discriminative approaches for image annotation is imbalanced data distribution because each binary classifier is designed to distinguish image of one class from images of the other classes. It becomes more severe when the number of classes/tags is large [43]. Another limitation of these approaches is that they are unable to capture the correlation among classes, which is known to be important in multi-label learning. To overcome these issues, several algorithms [16], [17], [44] are proposed to harness the keyword correlation as the additional information.

The search based approaches are based on the assumption that visually similar images are more likely to share common keywords [10]. Given a test image $\mathcal{I}$, it first finds out a set of training images that are visually similar to $\mathcal{I}$, and then assigns the tags that are most popular among the similar images. A divide-and-conquer framework is proposed in [45] which identifies the salient terms from textual descriptions of visual neighbours searched from web images. In the Joint Equal Contribution (JEC) model proposed in [4], multiple distance functions are computed with each based on a different set of visual features, and the nearest neighbors are determined by the average distance functions. TagProp [7] predicts keywords by taking a weighted combination of tags assigned to nearest neighbor images. More recently, the sparse coding scheme and its variations are employed in [5], [9], and [14] to facilitate image label propagation. Similar to the classification method, the search based approaches often fail when the number of training examples is limited.

B. Tag Ranking

Tag ranking aims to learn a ranking function that puts relevant tags in front of the irrelevant ones. In the simplest form, it learns a scoring function that assigns larger values to the relevant tags than to those irrelevant ones. In [18], the authors develop a classification framework for tag ranking that computes tag scores for a test image based on the neighbor voting. It was extended in [46] to the case where each image is represented by multiple sets of visual features. Liu et al. [19] utilizes the Kernel Density Estimation (KDE) to calculate relevance scores for different tags, and performs a random-walk to further improve the performance of tag ranking by exploring the correlation between tags. Similarly, Tang et al. [47] proposed a two-stage graph-based relevance propagation approach. In [21], a two-view tag weighting method is proposed to effectively exploit both the correlation among tags and the dependence between visual features and tags. In [26], a max-margin riffled independence model is developed for tag ranking. As mentioned in the introduction section, most of the existing algorithms for tag ranking tend to perform poorly when the tag space is large and the number of training images is limited.

III. Regularized Tag Ranking

In this section, we first present the proposed framework for tag ranking that is explicitly designed for a large tag space with a limited number of training images. We then discuss a computational algorithm that efficiently solves the related optimization problem.

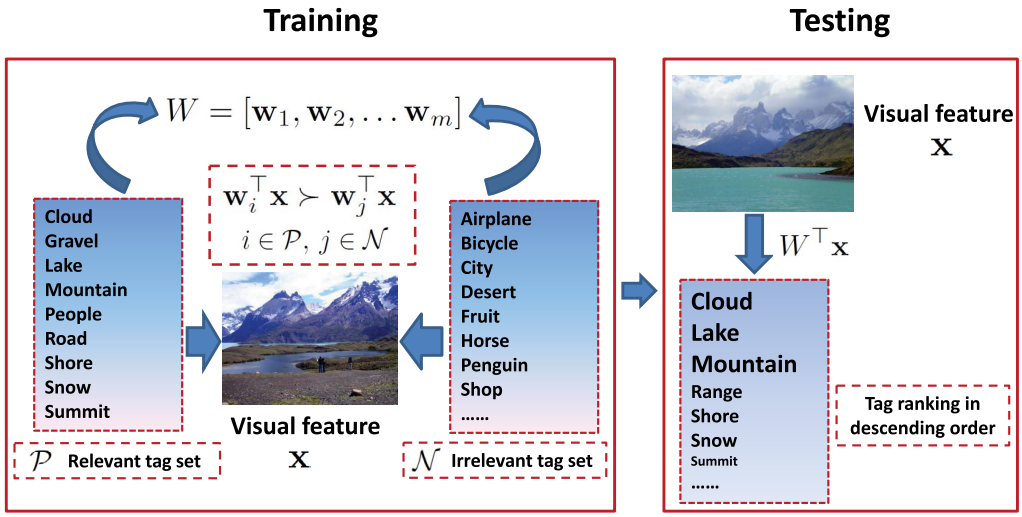


Fig. 1. Schematic illustration of the Proposed Method.

A. A Regularization Framework for Tag Ranking

Let the collection of training images be denoted by $\mathcal{I} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where each image $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d dimensions and n is the number of training examples. Let $\mathcal{T} = \{t_1, t_2, \ldots t_m\}$ be the set of tags used to annotate images. Let $\mathcal{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \{0,1\}^{m \times n}$ represent tag assignments for training images, where $\mathbf{y}_i \in \{0,1\}^m$ represents the tag assignment for the ith image. Here we use $y_{ji} = 1$ to indicate that tag t_j is assigned to image $\mathbf{x}_i$ and zero, otherwise.

In order to learn a tag ranking function, we have to decide in the first place which tags are relevant to a given image, and which ones are not. To this end, we simply assume all the assigned tags are relevant, and the unassigned tags are irrelevant. Although it is arguable that this simple treatment could be problematic for noisy and incomplete tag assignments, it is justified by the empirical study in [24] where tag ranking is shown to be more robust to both noisy and missing tags than the classification approaches. As a result, we would like to learn a ranking function that assign a higher score to tag t_j than to a tag t_k for image $\mathbf{x}_i$ if $y_{ji} = 1$ and $y_{ki} = 0$. More specifically, let $f_i(\mathbf{x})$ be the prediction function for the ith tag, and let $\ell(z)$ be a loss function. Let $\varepsilon_{j,k}(\mathbf{x}, \mathbf{y})$ measure the error in ranking tag t_j and t_k for image $\mathbf{x}$ with respect to the true tag assignments $\mathbf{y}$. It is defined as follows:

$$\varepsilon_{j,k}(\mathbf{x}, \mathbf{y}) = I(y_j \neq y_k)\ell\left(\left(y_j - y_k\right)\left(f_j(\mathbf{x}) - f_k(\mathbf{x})\right)\right) \quad (1)$$

where $I(z)$ is an indicator function that outputs 1 when z is true and 0, otherwise. Using the ranking error $\varepsilon_{j,k}(\mathbf{x}, \mathbf{y})$, we can now define the ranking error for an individual image $\mathbf{x}$ as

$$\varepsilon(\mathbf{x}, \mathbf{y}) = \sum_{j,k=1}^{m} \varepsilon_{j,k}(\mathbf{x}, \mathbf{y})$$

and the overall ranking error for all the training images in collection $\mathcal{I}$ as $\sum_{i=1}^{n} \varepsilon(\mathbf{x}_i, \mathbf{y}_i)$. For the simplicity of computation, we restrict the prediction functions $\{f_i\}_{i=1}^m$ to linear functions, i.e. $f_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x}$. Define $W = [\mathbf{w}_1, \ldots, \mathbf{w}_m] \in \mathbb{R}^{d \times m}$ and the overall loss $f(W)$ as

$$\begin{aligned} f(W) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k=1}^{m} \varepsilon_{j,k}(\mathbf{x}_i, \mathbf{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k=1}^{m} I(y_{ji} \neq y_{ki})\ell\left((y_{ji} - y_{ki})(\mathbf{w}_j^\top \mathbf{x}_i - \mathbf{w}_k^\top \mathbf{x}_i)\right) \end{aligned} \quad (2)$$

Figure 1 illustrates the basic idea of the proposed framework for tag ranking.

A straightforward approach for tag ranking is to search for a matrix W that minimizes the ranking error $f(W)$. This simple approach is problematic and could lead to the overfitting of training data when the number of training images is relatively small and the number of unique tags is large. Like most machine learning algorithms, an appropriate regularization mechanism is needed to control the model complexity and prevent overfitting the training data. In order to effectively capture the correlation among different tags, we follow [48] and assume that the linear prediction functions in W are linearly dependent and consequentially W is a low rank matrix, leading to the following optimization problem

$$\min_{W \in \Omega} f(W) \quad (3)$$

where the domain Ω is defined as

$$\Omega = \{W \in \mathbb{R}^{d \times m}, rank(W) \leq r, |W|_2 \leq s\}$$

Here $|\cdot|_2$ stands for matrix spectral norm. In (3), we restrict the solution W to domain Ω in order to control the complexity of the prediction model. The regularization effect of domain Ω is revealed by the following lemma on generalization error bound.

Theorem 1: Define $\bar{\ell}(W) = \mathrm{E}_{\mathbf{x},\mathbf{y}}[\varepsilon(W, \mathbf{x}, \mathbf{y})]$ and $R = \max_{-sr \leq z \leq sr} \ell(z)$. Assume $\ell(z)$ is L-Lipschitz continuous. Let $\widehat{W}_* \in \Omega$ be the solution in Ω that minimizes $f(W)$. Then, with a probability $1 - n^{-2}$,

we have

$$f(\widehat{W}_*) - \bar{\ell}(\widehat{W}_*) \leq \frac{2L}{n} + R\sqrt{\frac{2}{n}\left[r(1+d+m)\left(\log(18s)+\log n\right)+\log(2n^2)\right]}$$

Proof: We divide our analysis into two step. In the first, we will focus on a fixed solution W, and in the second step, we generalize to any $W \in \Omega$.

1) Step 1: We consider a *fixed* solution $W \in \Omega$. In order to bound $f(W) - \bar{\ell}(W)$, we use the Hoeffiding concentration inequality.

Theorem 2 (Hoeffding Inequality): Suppose $a_j < b_j$, $j = 1, \ldots, n$, $X_j \in [a_j, b_j]$, $\mathrm{E}[X_j] = 0$, $j = 1, \ldots, n$. Then,

$$\Pr\left\{\sum_{j=1}^{n} X_j \geq t\right\} \leq \exp\left\{-\frac{2t^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right\}$$

To use the Hoeffiding inequality, we define $X_i = \varepsilon(W, \mathbf{x}_i, y_i) - \bar{\ell}(W)$ with $|b_j - a_j| \leq R$. As a result, we have, for a fixed $W \in \Omega$,

$$\Pr\left\{\sum_{i=1}^{n}\left[\varepsilon(W, \mathbf{x}_i, y_i) - \bar{\ell}(W)\right] \geq t\right\} \leq \exp\left(-\frac{2t^2}{nR^2}\right)$$

By setting $\exp(-\frac{2t^2}{nR^2}) = \delta$, we have

$$t = R\sqrt{\frac{n}{2}\log\frac{1}{\delta}}$$

Since $f(W) = \varepsilon(W, \mathbf{x}_i, y_i)$, we also have

$$Pr\{n[f(W) - \bar{\ell}(W)] \geq R\sqrt{\frac{n}{2}\log\frac{1}{\delta}}\} \leq \delta$$

implying that with a probability $1 - \delta$, we have

$$f(W) - \bar{\ell}(W) \leq R\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$$

2) Step 2: We now proceed to consider any $W \in \Omega$. In particular, since

$$f(W_*) - \bar{\ell}(W_*) \leq \sup_{W\in\Omega} f(W) - \bar{\ell}(W)$$

our goal is to bound $\sup_{W\in\Omega} f(W) - \bar{\ell}(W)$, i.e. the bound for $f(W) - \bar{\ell}(W)$ for any $W \in \Omega$. Our approach is based on the theory of covering number. The key idea is to first divide the space Ω into many small cell, and for each solution $W \in \Omega$, we approximate the error $f(W) - \bar{\ell}(W)$ by the error of the center. For the center of each cell, we apply the result of step 1 to bound $f(W) - \bar{\ell}(W)$, and by taking the union bound, we can bound $f(W) - \bar{\ell}(W)$ for all the centers and consequentially bound $f(W) - \bar{\ell}(W)$ for every solution $W \in \Omega$.

To this end, we divide the set Ω into many small cells by using the proper ε-net. It finds the set of centers in Ω, denoted by $\mathcal{N}(\varepsilon, \Omega)$, such that the shortest distance between any $W \in \Omega$ and the set $\mathcal{N}(\varepsilon, \Omega)$ is no larger than ε. Using the result of step 1, for each $W \in \mathcal{N}(\varepsilon, \Omega)$, with a probability $1 - \delta$, we have

$$f(W) - \bar{\ell}(W) \leq R\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$$

By taking the union bound, we have, with a probability $1 - |\mathcal{N}(\varepsilon, \Omega)|\delta$,

$$\max_{W\in\mathcal{N}(\varepsilon,\Omega)} f(W) - \bar{\ell}(W) \leq R\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$$

or if we redefine δ as $\frac{\delta}{|\mathcal{N}(\varepsilon,\Omega)|}$, we have, with a probability $1 - \delta$

$$\max_{W\in\mathcal{N}(\varepsilon,\Omega)} f(W) - \bar{\ell}(W) \leq R\sqrt{\frac{2}{n}\left(\log|\mathcal{N}(\varepsilon, \Omega)| + \log\frac{2}{\delta}\right)}$$

Using the fact [49]

$$|\mathcal{N}(\varepsilon, \Omega)| \leq \left(\frac{18s}{\varepsilon}\right)^{(1+m+d)r}$$

we have, with a probability $1 - \delta$, for any $W \in \mathcal{N}(\varepsilon, \Omega)$,

$$f(W) - \bar{\ell}(W) \leq R\sqrt{\frac{2}{n}\left[r(1+m+d)\left(\log 18s + \log\frac{1}{\varepsilon}\right) + \log\frac{2}{\delta}\right]} \quad (4)$$

To connect the bound for proper ε-net $\mathcal{N}(\varepsilon, \Omega)$ with the bound for Ω, we exploit the property of the loss function $\ell(\cdot)$. Since $\ell(y)$ is L-lipschitz continuous, using the definition of ε-net, we have, for any $W \in \Omega$, there exists $W' \in \mathcal{N}(\varepsilon, \Omega)$ such that

$$|f(W) - f(W')| \leq L|W - W'| \leq L\varepsilon, \quad (5)$$

$$|\bar{\ell}(W) - \bar{\ell}(W')| \leq L|W - W'| \leq L\varepsilon \quad (6)$$

By combining the bounds in (4), (5), (6), we have, with a probability $1 - \delta$, for any $W \in \Omega$,

$$f(W) - \bar{\ell}(W) \leq 2L\varepsilon + R\sqrt{\frac{2}{n}\left[r(1+m+d)\left(\log 18s + \log\frac{1}{\varepsilon}\right) + \log\frac{2}{\delta}\right]}$$

We complete the proof by setting $\varepsilon = 1/n$, $\delta = 1/n^2$, and W to be $\widehat{W}_*$. ∎

As indicated by Theorem 1, to achieve a small generalization error, i.e.

$$\sqrt{2(r(1+m+d)(\log(18s)+\log n) + log(2n^2))/n},$$

n should be at least $r(d+m)(\log r + \log(d+m))$, which is often referred to as sample complexity in learning theory. Since the size of matrix W is dm, the sample complexity $r(d+m)(\log r + \log(d+m))$ is significantly smaller than dm when $r \ll \min(d, m)$, indicating that by restricting the solution to domain Ω, we will be able to avoid the overfitting problem even when the number of tags m is large while the number of training examples is limited.

Directly solving the optimization problem in (3) can be computationally challenging since $rank(W)$ is a non-convex function. Using the fact

$$|W|_* \leq rank(W)|W|_2,$$

where $|\cdot|_*$ stands for the trace norm of matrix, we relax the domain Ω in a convex one as

$$\Omega' = \left\{W \in \mathbb{R}^{d\times m} : |W|_* \leq sr\right\}$$

and consequentially approximate the optimization problem in (3) by

$$\min_{W \in \Omega'} f(W).$$

We can further simplify the above optimization problem by turning the constrained optimization problem into a regularized optimization problem, i.e.

$$\min_{W \in \mathbb{R}^{d \times m}} F(W) := f(W) + \lambda \|W\|_* \tag{7}$$

where $\lambda > 0$ is the regularization parameter used to balance between regularization term and the loss function based on training examples.

B. Optimization

Since both the loss function $f(W)$ and the trace norm $\|W\|_*$ are convex, one popular approach for solving the optimization problem in (7) is gradient descent. For the sake of clarity, we set the loss function in (1) to be a logistic loss. i.e., $\ell(z) = \log\left(1 + e^{-z}\right)$. At each iteration t, given the current solution W_t for (7), we first compute a subgradient of the objective function $F(W)$ at $W = W_t$, denoted by $\nabla F(W_t)$, and then update the solution by

$$W_{t+1} = W_t - \eta_t \nabla F(W_t) \tag{8}$$

where $\eta_t > 0$ is a step size at the t-th iteration. Let $W_t = U_t \Sigma_t V_t^\top$ be the singular value decomposition of W_t. Since $U_t V_t^\top$ is a subgradient of $\|W\|_*$ at $W = W_t$, we have

$$\nabla F(W_t) = \lambda U_t V_t^\top + \frac{1}{n} \sum_{i=1}^{n} \sum_{j,k=1}^{m} \alpha_{jk}^i \mathbf{x}_i \left(\mathbf{e}_j^m - \mathbf{e}_k^m\right)^\top \tag{9}$$

where

$$\alpha_{jk}^i = I(y_{ji} \neq y_{ki}) \ell'((y_{ji} - y_{ki}) \mathbf{x}_i^\top (\mathbf{w}_j - \mathbf{w}_k))$$

and $\mathbf{e}_j^m$ is a vector of m dimensions with all the elements being zero except that its jth entry is 1.

The main computational challenge in implementing the gradient descent approach for optimizing (9) aries from the high cost in computing the singular value decomposition of W_t. It is known [50] that when the objective function is smooth, the gradient method can be accelerated to achieve the optimal convergence rate of $O(T^{-2})$. It was shown recently [51]–[53] that a similar scheme can be applied to accelerate optimization problems where the objective function consists of a smooth part and a trace norm regularization. In this work, we adopt the accelerated proximal gradient (APG) method [54] for solving the optimization problem in (7). Specifically, according to [54], we update the solution W_t by solving the following optimization problem

$$W_t = \arg\min_W \frac{1}{2\eta_t} \|W - W_t'\|_F^2 + \lambda \|W\|_* \tag{10}$$

where

$$W_t' = W_{t-1} - \eta_t \nabla f(W_{t-1})$$

The optimal solution to (10), according to [55], is obtained by first computing the singular value decomposition (SVD) of W_t' and then applying soft-thresholding to the singular values of W_t'. More specifically, the optimal solution is given as

$$W_t = U \Sigma_{\lambda \eta_t} V^\top,$$

where $W_t' = U \Sigma V^\top$ is the SVD of W_t' and $\Sigma_{\lambda\eta_t}$ is a diagonal matrix with its diagonal elements computed as $(\Sigma_{\lambda\eta_t})_{ii} = \max\{0, \Sigma_{ii} - \lambda\eta_t\}$.

The final component of the accelerated algorithm is to determine the step size η_t, which could have a significant impact on the convergence of the accelerated algorithm. We follow [54] and apply a simple line search to find an appropriate step size η_t. More specifically, we denote by $p_\eta(W_{t-1})$ the optimal solution to (10) with step size η_t set as η, and by $Q_\eta(p_\eta(W_{t-1}), W_{k-1})$ the optimal value of the objective function in (10). We initialize the step size at iteration t as the one from the last iteration, and perform a simple line search to find the step size such that $F(p_\eta(W_{t-1})) > Q_\eta(p_\eta(W_{t-1}), W_{k-1})$. Algorithm 1 summarizes the key steps of the accelerated algorithm for solving the optimization problem in (7). Figure 2 shows the convergence of the objective function for dataset Corel5K, ESPGame, and IAPRTC-12 (more information about these three datasets can be found in the experimental section).

Algorithm 1 Solving Problem (7) by Accelerated Gradient Algorithm

Input: Training image collection $\mathcal{I} = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$, tag assignments for training images $\mathcal{Y} = \{\mathbf{y}_j \in \{0,1\}^m\}_{j=1}^n$, parameter λ.
Initialize: $\eta_0 = 1$, $\gamma = 2$, $\alpha_1 = 1$, $W_0 = Z_0 = Z_1 \in \mathbb{R}^{d \times m}$
while not converged **do**
1. Set $\bar{\eta} = \eta_{k-1}$
2. While $F\left(p_{\bar{\eta}}(Z_{k-1})\right) > Q_{\bar{\eta}}\left(p_{\bar{\eta}}(Z_{k-1}), Z_{k-1}\right)$, set

$$\bar{\eta} := \frac{\bar{\eta}}{\gamma}$$

3. Set $\eta_k = \bar{\eta}$ and update

$$W_k = p_{\eta_k}(Z_k),$$
$$\alpha_{k+1} = \frac{1 + \sqrt{1 + 4\alpha_k^2}}{2},$$
$$Z_{k+1} = W_k + \left(\frac{\alpha_k - 1}{\alpha_{k+1}}\right)(W_k - W_{k-1}).$$

end while
Output: The optimal solution W_*.

C. Automatic Image Annotation and Tag Ranking

Given the learned matrix W_* and a test image represented by vector $\mathbf{x}_t$, we compute scores for different tags by $\mathbf{y}_t = W_*^\top \mathbf{x}_t$ that indicate the relevance of each tag to the visual content of the test image. The tags are then ranked in the descending order of the relevant scores and only the tags ranked at the top will be used to annotate the test image. Besides image annotation, the learned model can also be used when a subset of tags is provided to the test image and needs to be re-ranked in order to remove the noisy tags.

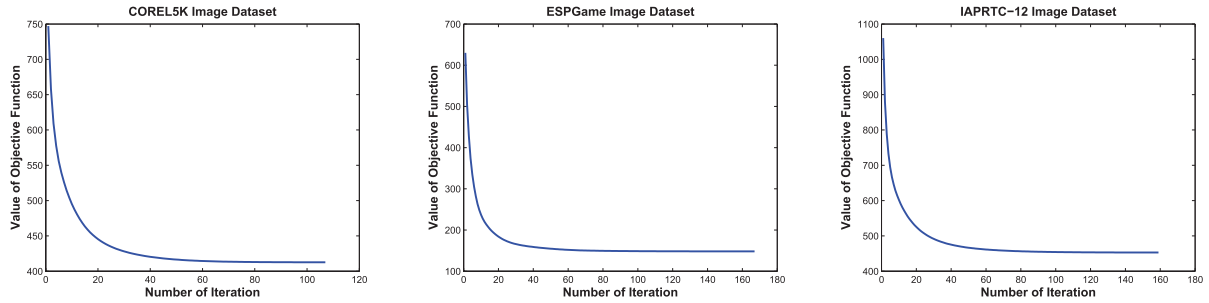


Fig. 2. Convergence of the objective function on image dataset Corel5K, ESPGame, and IAPRTC-12.

TABLE I
STATISTICS FOR THE DATASETS USED IN THE EXPERIMENTS. THE BOTTOM TWO ROWS ARE GIVEN IN THE FORMAT MEAN/MAXIMUM

	Corel5K	ESPGame	IAPRTC-12	Pascal VOC2007	SUNAttribute
No.of images	4,999	20,770	19,627	9,963	14,340
Vocabulary size	260	268	291	399	102
Tags per image	3.4/5	4.69/15	5.72/23	4.2/35	15.5/37
Image per tag	58.6/1,004	363/5,059	386/5,534	53/2,095	2,183/11,878

IV. EXPERIMENTAL RESULTS

In this section, we first describe our experimental setup, including image datasets, feature extraction, and evaluation measures. We then present three sets of experiments to verify the effectiveness of the proposed tag ranking approach, where the first experiment evaluates the performance of image annotation with limited training examples, the second experiment evaluates the performance of image annotation using training images with missing tags, and the last experiment examines the performance of the proposed algorithm for tag ranking. We finally evaluate the sensitivity of the proposed algorithm to parameter λ.

A. Image Datasets

To evaluate the proposed algorithm for image tagging, we conduct extensive experiments on five benchmark datasets for image annotation/taging, including Corel5K, ESPGame, IAPRTC-12, Pascal VOC2007 and SUNAttribute. The first three image datasets are used to evaluate the performance of automatic image annotation, and the last two image datasets are used to evaluate tag ranking since a relevance score is provide for every assigned tag. Table I summarizes the statistics of the image datasets used in our study.

Corel5K: This dataset contains about $5,000$ images that are manually annotated with 1 to 5 keywords. The annotation vocabulary contains 260 keywords. A fixed set of 499 images are used as test and the rest images are used for training.

ESPGame: This dataset is obtained from an online game named ESP. We use a subset of around 20,000 images that are publicly available [11].

IAPRTC-12: This image collection is comprised of $19,627$ images, each accompanied with descriptions in multiple languages that were initially published for cross-lingual retrieval. Nouns are extracted from the textual descriptions to form the keyword assignments to images. We use the annotation results provided in [11].

Pascal VOC2007: This dataset is comprised of $9,963$ images. We use the tags provided in [56] that are collected from 758 workers using Amazon Mechanical Turk. As a result, for each image, we compute the relevance score for each assigned tag based on its votes from different workers. This relevance score will be used to evaluate ranking performance. On average, each image in this dataset is annotated by 4.2 tags from a vocabulary of 399 tags.

SUNAttribute: The SUNAttribute dataset contains 14,340 images and 102 scene attributes spanning from materials, surface properties, lighting, functions and affordances, to spatial envelope properties. Similar to Pascal VOC2007, the annotated tags are collected from a large number of workers using the Amazon Mechanical Turk and therefore the votes from different workers can be used to compute the relevance score for different tags.

For Corel5K, ESPGame and IAPRTC-12 image datasets, a bag-of-words model, based on densely sampled SIFT descriptors, is used to represent the visual content of images [7]. For Pascal VOC2007 dataset, we follow [56] and extract three types of image features: Gist, color histogram, and bag-of-words histograms. For the SUNAttribute dataset, we follow [57] and represent each image using four types of features: Gist, HOG2 $\times$ 2, a self-similarity, and geometric context color histogram. For simplicity, features provided in both Pascal VOC2007 and SUNAttribute datasets are directly combined by merging the feature vectors of each image. We subtract every element of each dimension of features by the mean of all elements in this dimension, and then divide by the standard variation of all elements in this dimension to normalize the features.

B. Evaluation Measures

Firstly, to evaluate the performance of automatic image annotation, we adopt the Average Precision ($AP@K$) and Average Recall ($AR@K$) as the evaluation metrics, which are

defined as [3]:

$$AP@K = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{N_c(i)}{K} \tag{11}$$

$$AR@K = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{N_c(i)}{N_g(i)} \tag{12}$$

where K is the number of truncated tags, n_t is the number of test images, $N_c(i)$ is the number of correctly annotated tags for the ith test image, $N_g(i)$ is the number of tags assigned to the ith image. Both average precision and recall compares the automatically annotated image tags to the manually assigned ones.

In addition, we use the Normalized Discounted Cumulative Gains at top K ($NDCG@K$) to measure the performance of different tag ranking approaches. It reflects how well a computed ranking agrees with the ideal (ground truth) ranking, with the emphasis on the accuracy of the top ranked items. It is defined as [26]:

$$NDCG@K = \frac{1}{Z}\sum_{i=1}^{K}\frac{2^{rel(i)}-1}{log(1+i)} \tag{13}$$

where K is called truncation level, Z is the normalization constant to make sure the optimal ranking get the NDCG score of 1, and $rel(i)$ is the relevance score for the i-th ranked tag. Finally, for the proposed method, we set $\lambda = 1$ for all experiments except for the last one where we evaluate the impact of parameter λ.

C. Experimental (I): Automatic Image Annotation With Limited Number of Training Images

In the first experiment, we evaluate the annotation performance of the proposed image tagging method with limited training images. To this end, we randomly sample only 10% of images for training and use the remaining 90% for testing. Each experiment is repeated 10 times, each with a different spliting of training and testing data. We report the result based on the average over the trials. The following state-of-the-art approaches for image annotation are used as the baseline approaches in our evaluation:

- *Joint Equal Contribution Method (JEC) [4]:* It finds appropriate annotation words for a test image based on a k nearest neighbor classifier that used a combined distance measure derived from multiple sets of visual features.
- *Tag Propagation Method (TagProp) [7]:* It propagates the tag information from the labeled images to the unlabeled ones via a weighted nearest neighbor graph, where RBF kernel function is used for computing weights between images.
- *Multi-Class SVM Method (SVM) [58]:* It simply implements One-versus-All (OvA) SVM classifier for each tag, and ranks the tags based on the output probability values.
- *Fast Image Tagging Method (FastTag) [59]:* It explores multi-view learning technique for multi-label learning. In particular, it defines two classifiers, one for each view of the data, and introduces a co-regularizer in the objective function to enforce that the predictions based on different views are consistent for most training examples.
- *Efficient Multi-Label Ranking Method (MLR) [23]:* This approach explores the group lasso technique in multi-label ranking to effectively handle the missing class labels. It has been shown to outperform many multi-label learning algorithms [23].

The key parameter for TagProp is the number of nearest neighbors used to determine the nearest neighbor graph. We set it to be 200 as suggested by the original work [7]. For both SVM and MLR methods, linear function instead of RBF kernel function is adopted here for fair comparison. The optimal value for penalty parameter C in both methods is found by cross validation. Note that although FastTag method also adopts linear image feature classifiers, it incorporates non-linearity into the feature space as a preprocessing step.

First, we show the comparison of average precision/recall for the first 5 returned tags for Corel5K dataset[1] and the top 10 returned tags for both ESPGame and IAPRTC-12 datasets in Figure 3. It is not surprising to observe that with increasing number of returned tags, average precision declines while average recall improves. This is also called precision-recall trade-off, a phenomenon that is well known in information retrieval [3]. Second, we observe that our method significantly outperforms two nearest-neighbor based methods (JEC and TagProp) on the given datasets since the performance of nearest-neighbor based methods largely depend on the number of training samples. Specifically, at the truncation level of 4 (AP@4), we see our method yields around 5.6%, 4% and 8.76% improvement over TagProp on Corel5K, ESPGame and IAPRTC-12 dataset, respectively. In addition, the proposed method also outperforms multi-class SVM and FastTag algorithms, two classification based approaches, and MLR, a multi-label ranking approach. We attribute the success of the proposed approach to the special design of the proposed approach that nicely combines the ranking approach with trace norm regularization: it is the ranking approach that allows us to avoid making binary classification decision, and it is the trace norm regularization that makes our approach robust to the limited number of training examples.

To further investigate the advantages of the proposed approach, we evaluate the two components, i.e. ranking loss and trace norm regularization, separately. More specifically, we develop two baseline approaches, one replacing the ranking loss in the proposed framework with classification loss (**C+T**) and the other replacing trace norm with Frobenius norm for regularization (**R+F**). We also include the last baseline (**C+F**) that combines the classification loss with the Frobenius norm regularization. Following the naming convention here, we refer to the proposed approach as **R+T**. Figure 4 shows the prediction results that are based on 10% of images for training. We observe that the proposed framework outperforms the other three baselines, and the classification loss with Frobenius norm yields the worst performance among four approaches.

[1]We only consider the first 5 returned tags in Corel5K image dataset since the maximum tags for each image is 5.

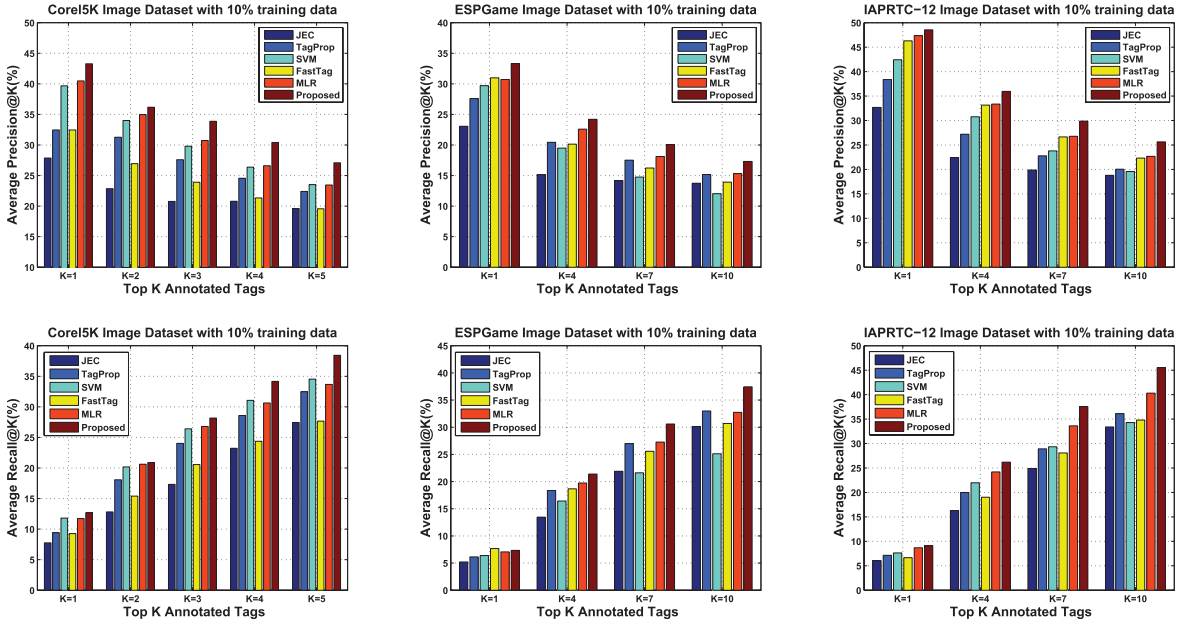


Fig. 3. Average precision and recall for automatic Image annotation on Corel5K, ESPGame and IAPRTC-12 datasets.

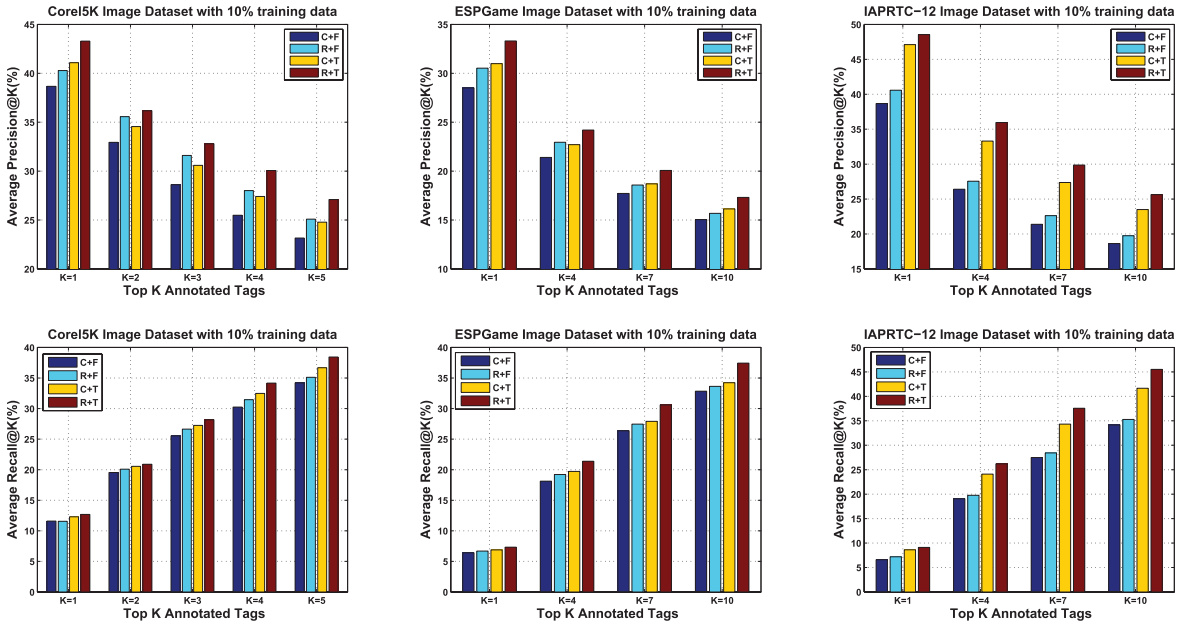


Fig. 4. Evaluation of different loss functions and matrix regularizers for automatic image annotation.

Both observations indicate that the combination of ranking loss with trace norm regularization is important when the number of training images is limited.

Finally, for the completeness of our experiment, we evaluate the performance of automatic image annotation by varying the number of training samples from 10% to 90%. Figure 5 summarizes the performance of $AP@5$ for three different datasets. We observe that annotation performance of all methods improves with increasing numbers of training images. We also observe that the improvement made by the proposed algorithm over the baseline methods reduces as the number of training images increase.

D. Experiment (II): Automatic Image Annotation With Incomplete Image Tags

In this experiment, we examine the performance of the proposed method when training image are partially annotated. To this end, similar to [24], we randomly select only 20%, 40%, and 60% of the assigned tags for training images. This setting allows us to test the sensitivity of the proposed method to the missing tags. Since the maximum number of annotated tags for the Corel5K dataset is 5, we only conduct the experiments on ESPGame and IAPRTC-12 datasets, where the maximum number of assigned tags are 15 and 23, respectively. The results of average precision

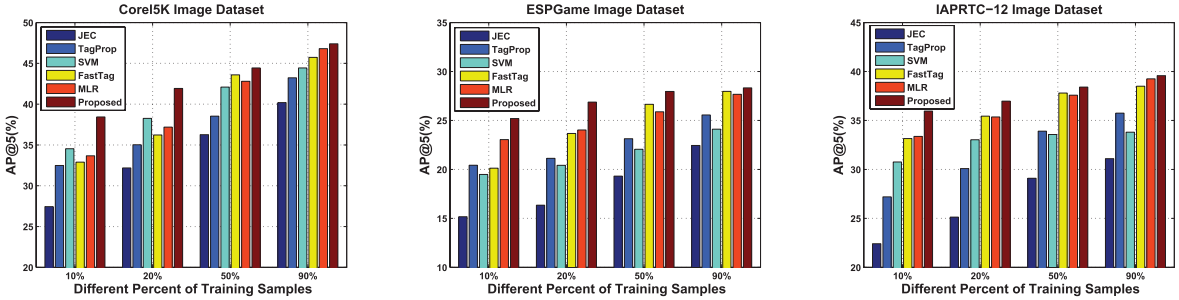


Fig. 5. Average precision at rank 5 ($AP@5$) with varied numbers of training images for datasets Corel5K, ESPGame, and IAPRTC-12.

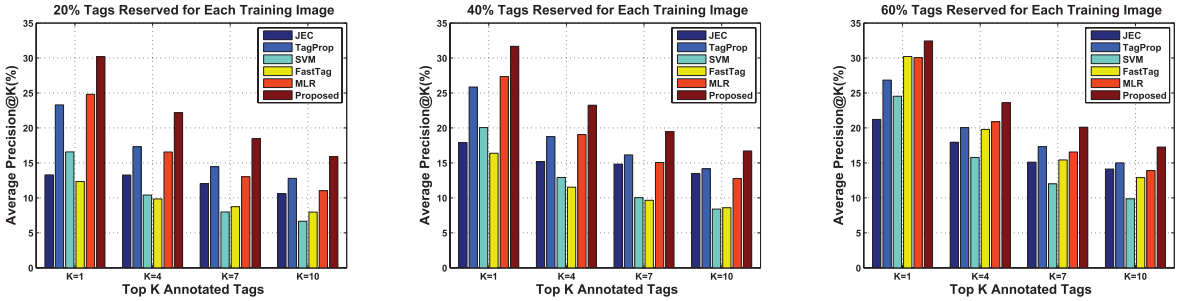


Fig. 6. Performance of automatic image annotation on the ESPGame dataset with incomplete image tags, where the number of observed tags is varied from 20%, 40% to 60%.

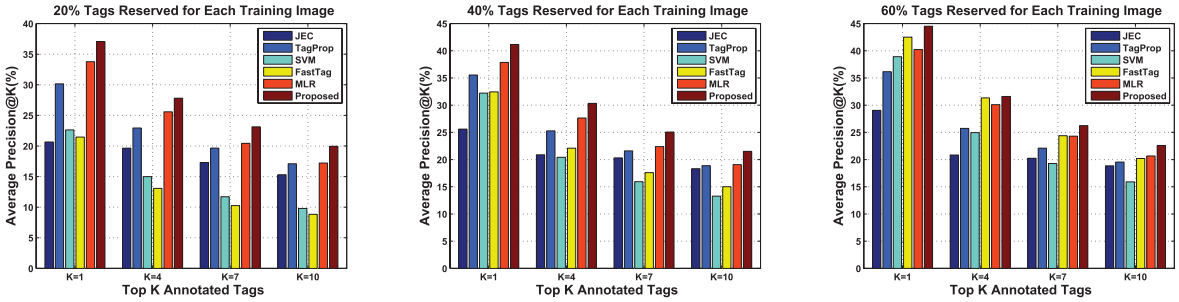


Fig. 7. Performance of automatic image annotation on the IAPRTC-12 dataset with incomplete image tags, where the number of observed tags is varied from 20%, 40% to 60%.

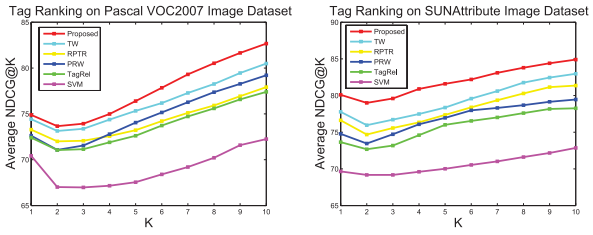


Fig. 8. Performance of tag ranking, measured by NDCG, for dataset Pascal VOC2007 and SUNAttribute.

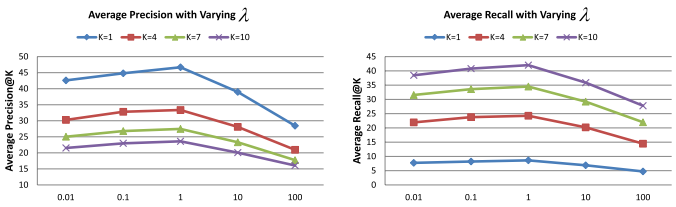


Fig. 9. Average precision and recall of the proposed method for the IAPRTC-12 dataset with varied λ.

for both datasets are reported in Figure 6 and Figure 7, respectively.

It is not surprising to observe that annotation performance of all methods drops as the number of observed annotations decreases, indicating that the missing annotations could greatly affect the annotation performance. On the other hand, compared to the baseline methods, the proposed method is more resilient to the missing tags: on the ESPGame dataset, it only experiences a 1.41% drop in average precision when the number of observed tags decreases from 60% to 20%, while the other five baseline methods suffer from 4% to 8% loss for $AP@4$. This result indicates that the proposed method is more effective in handling missing tags. Figure 10 provides examples of annotations generated by different approach for the ESPGame and IAPRTC-12 datasets when only 20% of the assigned tags are observed for each training image.

Ground Truth	**building door frame sky street window**	**adult child front house square woman**	**fog mountain roof stripe train**	**grandstand lawn roof round spectator stadium**	**hill landscape mountain rock woman**	**bike car cycling sky cyclist frame helmet jersey rack landscape roof short**	**boat green hill jacket lake orange mountain people range shore life sky summit**
JEC	*table front man house roof wall woman boy* **building** *chair*	*tourist* **woman child front** *wall classroom people room table building*	*front child classroom man sky table board car* **mountain** *wall*	*grass bush hill house* **lawn** *man people player short slope*	*sky* **landscape hill rock** *tree lake man* **mountain** *bay cliff*	**cyclist jersey helmet sky** *side road* **short bike** *sand* **cycling**	**mountain** *sea* **sky shore** *cloud man house* **lake** *rock tourist*
TagProp	*front wall table man rail house woman level* **building** *child*	*people wall tourist* **front woman** *table man side round classroom*	*man sky* **mountain** *front classroom wall child table tourist house*	*man short sky tree* **lawn** *house grass bush rock people*	*sky* **rock** *man* **landscape mountain** *jeep sea* **hill** *tree palm*	**cyclist sky helmet jersey short** *highway road side meadow* **bike**	**sea shore sky mountain lake** *cloud man house* **boat** *woman*
SVM	*rail level front house* **building** *table roof wall man* **sky**	*people side* **front** *wall tree* **woman** *tourist building sky* **house**	**mountain** *sky front man cloud wall child classroom table car*	*man short woman sky tree house grass wall bush people*	*jeep sky* **landscape mountain rock** *lake road* **hill** *man tree*	**sky** *side landscape highway* **car** *tree* **short** *road* **bike** *people*	**mountain lake** *sea* **sky** *cloud* **boat** *fountain* **hill** *house man*
FastTag	**sky** *front man stripe level tree people house rack pond*	*sky* **front** *man people rail* **house** *frame tree bedside centre*	*sky gate front man people* **train** *tree penguin portrait carpet*	*sky people front man tree* **lawn** *dirt short house tussock*	*sky front man tree house tussock kid people* **mountain** *formation*	**sky** *front tree man highway people* **short rack** *trouser pinnacle*	**sky** *man front lagoon tree* **shore** *fountain garden* **mountain** *cloth*
MLR	*house wall room cobblestone* **door** *front palm lamp tower bed*	**house** *tree landscape people jacket* **square** *building sweater shelf dog*	**mountain** *house palm flower landscape wall gate orange sky* **train**	*people house* **lawn round** *view green field building stand* **stadium**	**mountain** *grass* **landscape** *house shrub* **rock** *wall tussock slope snow*	**cyclist sky car** *tree* **jersey helmet cycling short** *sign sand*	**range shore** *field* **sky** *lagoon bay* **lake** *river tourist road*
Proposed	*wall* **building** *front* **window** *house* **door street** *room column balcony*	**front** *man* **house** *wall people* **woman child** *tourist room building*	**mountain** *sky* **train** *front cloud tourist door* **roof** *window wall*	**stadium lawn** *slope house* **grandstand** *road field tree player people*	**mountain landscape** *sky* **rock** *middle* **hill** *desert lake man cliff*	**cyclist jersey sky short** *road* **cycling helmet bike** *pole* **car**	**mountain sky lake range** *cloud* **shore summit** *sea stone* **hill**

Ground Truth	**front group meadow stone tourist wall**	**boat man ocean sea water wave**	**balcony building sky fountain lamp palm square street**	**cloud leave palm plant sea sky**	**bed bedcover bedside curtain lamp room table wall**	**bush cactus man slope tree woman**	**face girl hair smile white woman**
JEC	*slope jersey man tree cyclist helmet cycling* **wall** *sky* **front**	**man** *green people grass picture photo family blue woman black*	**sky** *front table child grey tree wall classroom house man*	*man round tee shirt wall woman building child classroom jacket*	*front* **room** *fence gravel child cloud* **curtain** *door floor man*	**man** *forest* **grass bush** *cliff middle adult fence hill leave*	*website red* **woman** *black art box brown building church colors*
TagProp	*man jersey slope* **front** *tree* **wall** *mountain bush* **meadow** *people*	*airplane arrow red* **man** *blue sky* **water** *maga ine word* **ocean**	*front* **sky** *group people table man wall tree room house*	*man bench blanket wall table woman front child classroom* **sky**	*front child people* **wall room table** *sky fence man house*	**man** *hill grass forest* **tree** *pant sky* **bush** *rock middle*	*man cd red* **white** *black circle green logo* **face** *people*
SVM	*man* **front meadow wall** *cycling cyclist forest helmet tree bike*	**water man sea** *sand tree word blue car gray mountain*	*front group* **sky** *wall tree people table man house* **building**	*bench* **sky** *man wall people woman* **sea** *tree building tourist*	*front child* **wall room** *man people tourist fence gravel* **table**	*hill pant* **man** *sky* **tree** *grass* **bush** *forest meadow cliff*	*man logo red black cd* **face white** *circle green pink*
FastTag	*sky man tree bone people* **front** *shelter building pant house*	**water** *red sky* **ocean** *tree smile beak nose green cloud*	**sky** *front tree harbour man people house* **building** *pot* **fountain**	**sky** *jetty front man tree house corridor edge lagoon flagpole*	*sky tree table cloth flagpole* **wall** *man ravine* **room** *neck*	*sky* **man** *front* **tree** *people house grass bicycle lawn leave*	**hair face** *man* **smile** *nose* **woman girl** *circle glasses teeth*
MLR	*building trunk dog man* **tourist** *photo sky* **meadow** *sand paving*	**water ocean** *tree nose blue fire white roof cloud wing*	***building sky*** *house tree tower* ***palm street*** *people sign man*	**sky** *woman house tree* **sea cloud** *salt lake orange tower*	**room bed wall** *tree house* **curtain** *night side painting* **bedside**	*grass* **tree** *leave sand rock dog* **bush** *house cliff trail*	**white** *pink square purple* **face** *bald nose man word blue*
Proposed	**front** *man* **wall** *people* **tourist** *helmet* **group** *woman photo* **meadow**	**ocean sea** *sky* **water** *sand blue red* **boat man** *tree*	**building sky lamp** *tree* **street** *front people tower* **palm** *grey*	**sky cloud** *man* **sea palm** *tree lake grass house shore*	**wall room curtain** *front* **bed** *painting* **bedside** *fence blanket* **table**	**tree bush man** *grass forest* **slope** *middle rock jungle path*	**white** *man* **hair** *black red* **face** *blue* **woman girl** *hat*

Fig. 10. Examples of test images from both the ESPGame and IAPRTC-12 datasets with top 10 annotations generated by different methods. The correct tags are highlighted by bold font whereas the incorrect ones are highlighted by italic font.

These examples further confirm the advantage of using the proposed approach for automatic image annotation when training images are equipped with incomplete tags.

E. Experimental (III): Tag Ranking

In this subsection, we evaluate the proposed algorithm for tag ranking. Given an image and a list of associated tags, the goal of tag ranking is to rank the tags according to their relevance to the image content. Both the Pascal VOC2007 and SUNAttribute datasets are used in this experiment since a relevance score is provided for each assigned tag. We randomly select 10% of images from each dataset for training, and use the remaining 90% for testing. We repeat the experiment 10 times and repeat the averaged NDCG.

Using the votes collected from different workers, according to the settings in [26], we create three levels of relevance score for each assigned tag: Most Relevant (score 4), Relevant (score 3) and Less Relevant (score 2). To make the problem challenging enough, for each image, we add three randomly sampled irrelevant tags (score 1) to the tag list. As a rsult, each tag list is comprised of labels with four relevance levels, ranging from the irrelevant category to the most relevant one.

The following algorithms are used as the baselines in the evaluation of tag ranking. The first baseline uses the classification scores output from the one-vs-all SVM with

linear function to rank tags. The second baseline, named TagRel [18], is based on the neighbor voting strategy for tag ranking, and the neighbor number is empirically set to 100. The third baseline, abbreviated as PRW [19], combines the probabilistic tag ranking approach with a random walk-based tag ranking approach, and we use the same parameter settings suggested by the origin work. The fourth baseline, named RPTR [47], is a relevance propagation tag ranking approach which combines both tag graph and image graph. The last baseline, which is known as TW [21], is a two-view tag weighting method that combines the local information both in tag space and visual space, and the trade-off hyper-parameters used in the algorithm is adopted as suggested by the origin work.

Figure 8 reports NDCG values for the proposed algorithm and the four baseline methods on datasets Pascal VOC2007 and SUNAttribute. We can see that the proposed method significantly outperforms most baselines on both datasets. When evaluating with respect to the first five ranked tags (i.e. $NDCG@5$), we see our method yields about 9% improvement over SVM and 3.5% to 4% improvement over TagRel, RPTR and PRW on Pascal VOC2007 dataset. Furthermore, although our method only achieves around 2% improvement over TW, it is much more scalable than TW due to the fact that TW is essentially a transductive learning manner, which is not suitable for unseen test images. Similar improvements are observed on the SUNAttribute dataset. The experimental results prove that the proposed method is effective for tag ranking especially when the training samples are limited.

F. Experiment (IV): Sensitivity to Parameter λ

In this experiment, we examine the sensitivity of the proposed method to parameter λ using the dataset IAPRTC-12. In general, a larger λ will lead to a higher regularization capacity, and as a sequence, a larger bias and a smaller variance for the final solution. In order to understand how the parameter affects the annotation performance, we conduct the experiment by varying λ from 0.01 to 100 and measure average precision and recall for the learned annotation model, as shown in Fig. 9. We observe that the proposed method yields the best performance when λ is around 1.

V. Conclusion

In this work, we have proposed a novel tag ranking scheme for automatic image annotation. The proposed scheme casts the tag ranking problem into a matrix recovery problem and introduces trace norm regularization to control the model complexity. Extensive experiments on image annotation and tag ranking have demonstrated that the proposed method significantly outperforms several state-of-the-art methods for image annotation especially when the number of training images is limited and when many of the assigned image tags are missing. In the future, we plan to apply the proposed framework to the image annotation problem when image tags are acquired by crowdsouring that tend to be noisy and incomplete.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, 2008, Art. ID 5.

[2] J. Wu, H. Shen, Y. Li, Z.-B. Xiao, M.-Y. Lu, and C.-L. Wang, "Learning a hybrid similarity measure for image retrieval," *Pattern Recognit.*, vol. 46, no. 11, pp. 2927–2939, 2013.

[3] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.

[4] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for image annotation," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 88–105, 2010.

[5] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 1–16, 2011.

[6] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 223–232.

[7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 309–316.

[8] W. Liu and D. Tao, "Multiview Hessian regularization for image annotation," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2676–2687, Jul. 2013.

[9] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas, "Automatic image annotation using group sparsity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3312–3319.

[10] Y. Verma and C. V. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.

[11] Z. Feng, R. Jin, and A. Jain, "Large-scale image annotation by efficient and robust kernel metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1609–1616.

[12] 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.

[13] C. Yang, M. Dong, and F. Fotouhi, "Region based image annotation through multiple-instance learning," in *Proc. 13th ACM Int. Conf. Multimedia*, 2005, pp. 435–438.

[14] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1643–1650.

[15] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[16] H. Wang, H. Huang, and C. Ding, "Image annotation using multi-label correlated Green's function," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2029–2034.

[17] X. Cai, F. Nie, W. Cai, and H. Huang, "New graph structured sparsity model for multi-label image annotations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 801–808.

[18] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

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

[20] Z. Wang, J. Feng, C. Zhang, and S. Yan, "Learning to rank tags," in *Proc. ACM Int. Conf. CIVR*, 2010, pp. 42–49.

[21] J. Zhuang and S. C. H. Hoi, "A two-view learning approach for image tag ranking," in *Proc. 4th ACM Int. Conf. WSDM*, 2011, pp. 625–634.

[22] J. Liu, Y. Zhang, Z. Li, and H. Lu, "Correlation consistency constrained probabilistic matrix factorization for social tag refinement," *Neurocomputing*, vol. 119, pp. 3–9, Nov. 2013.

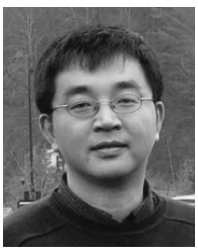
[23] S. S. Bucak, P. K. Mallapragada, R. Jin, and A. K. Jain, "Efficient multi-label ranking for multi-class learning: Application to object recognition," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2098–2105.

[24] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2801–2808.

[25] Z. Li, J. Liu, C. Xu, and H. Lu, "MLRank: Multi-correlation learning to rank for image annotation," *Pattern Recognit.*, vol. 46, no. 10, pp. 2700–2710, 2013.
[26] T. Lan and G. Mori, "A max-margin riffled independence model for image tag ranking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3103–3110.
[27] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[28] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 281–288.
[29] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 676–683.
[30] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," *ACM Comput. Surv.*, vol. 44, no. 4, pp. 1–28, 2012.
[31] D. Liu, X.-S. Hua, and H.-J. Zhang, "Content-based tag processing for internet social images," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 723–738, 2011.
[32] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. ACM Int. Conf. SIGIR*, 2003, pp. 119–126.
[33] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA, USA: MIT Press, 2003.
[34] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun./Jul. 2004, pp. II-1002–II-1009.
[35] F. Monay and D. Gatica-Perez, "PLSA-based image auto-annotation: Constraining the latent space," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 348–351.
[36] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Mar. 2003.
[37] D. Putthividhya, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent Dirichlet allocation for image annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3408–3415.
[38] O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical Dirichlet process model," in *Proc. 9th Int. Workshop Multimedia Data Mining*, 2008, pp. 1–7.
[39] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," in *Proc. Electron. Imag.*, 2004, pp. 330–338.
[40] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.
[41] 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.
[42] J. Fan, Y. Shen, C. Yang, and N. Zhou, "Structured max-margin learning for inter-related classifier training and multilabel image annotation," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 837–854, Mar. 2011.
[43] Q. Mao, I. W.-H. Tsang, and S. Gao, "Objective-guided image annotation," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1585–1597, Apr. 2013.
[44] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1719–1726.
[45] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.
[46] X. Li, C. G. M. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *Proc. ACM Int. Conf. CIVR*, 2010, pp. 10–17.
[47] M. Li, J. Tang, H. Li, and C. Zhao, "Tag ranking by propagating relevance over tag and image graphs," in *Proc. ACM 4th Int. Conf. Internet Multimedia Comput. Service*, 2012, pp. 153–156.
[48] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 461–470.
[49] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Berlin, Germany: Springer-Verlag, 2011.
[50] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372–376, 1983.
[51] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[52] Y. Nesterov, "Gradient methods for minimizing composite objective function," Dept. Center Oper. Res. Econometrics, Univ. Catholique Louvain, Louvain-la-Neuve, Belgium, Tech. Rep. 2007/76, 2007.
[53] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, no. 3, pp. 615–640, 2010.
[54] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.
[55] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[56] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, 2012.
[57] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2751–2758.
[58] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[59] M. Chen, A. Zheng, and K. Weinberge, "Fast image tagging," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1274–1282.

Songhe Feng received the Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2009. He is currently an Associate Professor with the School of Computer and Information Technology, Beijing Jiaotong University. He was a Visiting Scholar with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA, from 2013 to 2014. His research interests include computer vision and machine learning.

Zheyun Feng is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA. Her research interests include computer vision, machine learning, data mining, pattern recognition, and information retrieval. She has worked on image tagging with machine learning techniques, including distance metric learning, matrix completion, and statistical models. She received the M.E. degree in image processing from Telecom ParisTech, Paris, France, and the B.S. degree in electrical engineering from Nanjing University, Nanjing, China.



Rong Jin is currently a Professor with the Department of Computer and Science Engineering, Michigan State University, East Lansing, MI, USA. He has been involved in statistical machine learning and its application to information retrieval. He has extensive research experience in a variety of machine learning algorithms, such as conditional exponential models, support vector machine, boosting, and optimization for different applications. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *ACM Transactions on Knowledge Discovery from Data*. He received the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, USA, in 2003. He was a recipient of the NSF CAREER Award in 2006, and the best paper award from the Conference of Learning Theory in 2012.