IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON EMOTION-AWARE MOBILE COMPUTING

Received October 10, 2016, accepted October 26, 2016, date of publication November 29, 2016, date of current version May 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2624267

Non-Linear Matrix Completion for Social Image Tagging

XING XU[1], LI HE[2], HUIMIN LU[3], (Member, IEEE), ATSUSHI SHIMADA[4], AND RIN-ICHIRO TANIGUCHI[4]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China
[2]Qualcomm Research and Development Center, San Deigo, CA 92108, USA
[3]Center of Socio-Robotic Synthesis, Kyushu Institute of Technology, Kitakyushu 8048550, Japan
[4]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 8190395, Japan

Corresponding author: X. Xu (xing.xu@uestc.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Project ZYGX2016KYQD114, in part by JSPS KAKENHI under Grant 15F15077, in part by the Leading Initiative for Excellent Young Researcher of Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant 16809746, in part by the State Key Laboratory of Marine Geology, Tongji University, under Grant MGK1608, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, in part by the Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology and in part by the National Natural Science Foundation of China under Project 61602089.

ABSTRACT In this paper, we address the problem of social image tagging using practical vocabulary for mobile users on the social media. On the social media, images usually have an incomplete or noisy set of social tags provided by the mobile users, and we consider this issue as *defective tag assignments*. Previous studies on social image tagging have mostly focused on multi-label classification without considering the defective tags. In these studies, the usage of multi-label classification techniques is expected to synergically exploit the linear relations between the image features and the semantic tags. However, these approaches usually aimed to capture the linear relations from the training data while ignoring the helpful information from the test data. In addition, they failed to incorporate the non-linear associations residing in the visual features as well as in the semantic tags. To overcome these drawbacks, we introduce a novel approach based on non-linear matrix completion for image tagging task with defective tags. Specifically, we first construct the entire feature-tag matrix based on the visual features with non-linear kernel mapping. Then, we present a formal methodology together with an optimization method under the matrix completion framework to jointly complete the tags of training and test images. Experimental evaluations demonstrate that our method shows promising results on image tagging task on two benchmark social image datasets with defective tags, and establishes a baseline for such models in this research domain.

INDEX TERMS Social image tagging, tag completion, defective tag assignment.

I. INTRODUCTION

The last five years have witnessed the surge of social media and mobile computing. More users than ever have been using a wide range of mobile devices to post images to various social media, such as Facebook, Instagram and Google Plus, to record their daily lives. When posting images, the users are usually given the chance to specify a set of tags or labels to describe the semantic content of each image. Proper labels significantly improve the usefulness of the images and facilitate exciting applications such as keyword-based image retrieval/indexing [1]–[3] and large-scale image collection management on the cloud [4]–[6].

However, as revealed in [7] and [8], the user-provided tags tend to be ambiguous, incomplete or even imprecise due to the time-consuming labeling process and the uncertainty of human. These imprecise and incomplete tags, which can be considered as *defective tags*, usually adversely affect the usefulness of the images. One approach for solving this problem is to utilize various sensors found on mobile devices. Qin *et al.* [9] proposed to perform activity recognition on mobile devices using input from various sensors to automatically tagged images when they are taken. However, such methods rely on the appropriate fusion of sensor data and are not always able to extract semantic information embedded in the images. Therefore, we explore a machine learning based approach to complete defective tags for images on social media. The proposed algorithm not only properly completes these defective labels, but

beach, front, palm, people, tree, ***view, wave, sky, water***

flower, red, garden, <u>*orange*</u>*,* <u>*rainbow*</u>***, plant, leaf***

FIGURE 1. Example images with defective tag assignments in social image datasets IAPRTC-12 (left) and MIR Flickr (right): italic tags are provided ground truth tags, underlined tags are incorrect noisy tags, and bold tags are potentially missing tags.

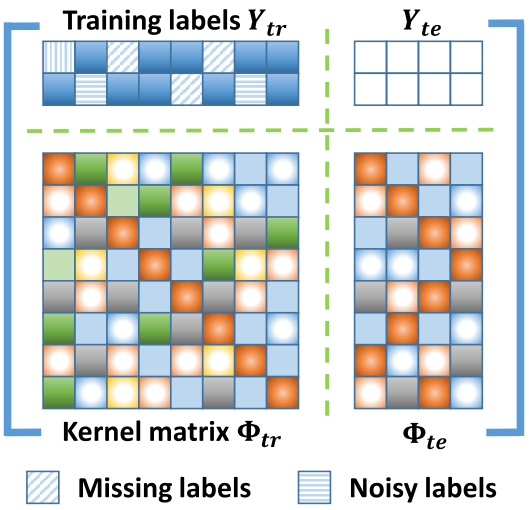


FIGURE 2. Overview of the proposed non-linear matrix completion framework for social image tagging with defective tags. The potentially missing and noisy tags are estimated from the kernel matrix, where the full kernel matrix including both training and test features are utilized during the learning process.

also extracts the semantic information to facilitate further applications.

Given an defective initial tag matrix for the training images, where images are represented by rows and tags by columns, our goal is to complete this matrix by i) recovering the missing tags and ii) removing or reducing the noisy tags that are not relevant to the visual contents of the images. In addition, it is also desired to predict more accurate labels for the untagged test images. The problem setting in this work is different from the traditional image annotation [10]–[13], which assumes that the training images are completely and appropriately labeled. Instead, it is similar to the recent studies of tag refinement [14] and tag completion [15]–[18], which intends to denoise the imprecise labels or enrich the incomplete tags in the training data. Tag refinement utilizes a wide range of techniques (such as tag propagation, sparse training and partial supervision) to choose a subset of user-provided tags based on visual features and tag correlation [19] to handle noisy tags, but it does not explicitly address the problem of missing tags. On the contrary, tag completion treats the missing tags as an independent problem and various algorithms (e.g., data factorization [20], hypergraph model [21]) have been introduced to search for the optimal tagging matrix that is consistent with both observed tags and visual similarities; however, these approaches does not handle the problem of defective tags.

In fact, there are only a few studies [15], [19], [21]–[23] in the area of image annotation with defective tags. Among these works, [15], [19] were based on *matrix completion* (MC) and have achieved promising performance. The observed defective tag matrix is composed of an ideal complete matrix and a sparse noise matrix with the low-rank assumption. Because the matrix rank function is non-convex, a popular approach is to replace it with the nuclear norm so a low rank matrix can be accurately recovered from a small fraction of its entries even if these entries are corrupted with noise. However, a key limitation of these methods is that they are tied to the assumption of linear classification model. They usually aimed to capture the linear relations from the training data while ignoring the helpful information from the test data. In addition, they failed to incorporate the non-linear associations residing in the visual features as well as those in the tags.

To overcome these drawbacks, we impose the non-linear classification via matrix completion for the social image tagging problem with defective tags. As the tags of the training entries may contain multiple missing or noisy tags, the proposed method first constructs the tag-feature matrix that integrates both the information of training labels and visual features of both training and test entries to estimate the labels of test entries. Fig. 2 shows the components of the introduced label-feature matrix. Specifically, the visual features of both training (i.e. $\mathbf{\Phi}_{tr}$) and test (i.e. $\mathbf{\Phi}_{te}$) entries are embedded in the kernel space, where each column in $\mathbf{\Phi}_{tr}$ and $\mathbf{\Phi}_{te}$ represents the non-linear feature vector of each training/test entry. The given training tags (i.e. $\mathbf{Y}_{tr}$) are defective, due to the absence of missing tags and presence of noisy tags. Then, to predict the complete tags for the test entries (i.e. $\mathbf{Y}_{te}$), we cast the social image tagging problem into a finite-dimension optimization problem given the non-linear kernels of visual features and initial training tags.

The main contribution of our work are three-fold: 1) We introduce the non-linear matrix completion approach and apply it to the social image tagging problem with defective tags. The proposed method can effective recover most missing tags despite the noise in user-provided tags. 2) We formulate a non-linear optimization problem which efficiently leverages the kernel matrices of both training and test features along with given defective tags for model learning. 3) We evaluate the proposed method on two benchmark datasets with defective tags. The experimental results show promising taggin results and form a baseline for such models in this research field.

The remainder of this paper is organized as follows. Section II gives an overview of the related work. The details

of our proposed method are presented in Section III. Extensive experimental results are given in Section IV. And finally we conclude this paper with Section V.

II. RELATED WORK

Image tagging has been widely studied during the past decade. Comprehensive literature reviews can be found in [10] and [24]. Generally, previous researches for image annotation can be roughly categorized into three groups: generative models [25]–[27], discriminative models [23], [28] and nearest neighbor (NN) based models [10]–[12], [29], [30]. The generative models define the joint distribution over image features and tags with various probabilistic assumptions, such as latent Dirichlet allocation, mixture of Gaussian, multi-variant Bernoulli, etc. The discriminative models cast image tagging problem into multi-label classification and learn a separate binary classifier for each label via support vector machine [31]. The NN based models generally assume that visually similar images probably share some common tags and propagate the tags from nearest visual neighbors to the test image. In recent years, with the rapidly increasing amount of image data, the NN based models tend to be more preferable because sufficient information of image-similarity and tag-association can be obtained.

Unlike the traditional image annotation problem, recent studies turn to investigate the social image tagging problem, where the problem of noisy and missing tags widely exists. One group of studies developed tag refinement approaches to improve the quality of user-provided tags. Specifically, tag refinement includes tag denoising and completion and has recently become an attractive subject of many ongoing researches [14], [15], [19]–[21], [32]. For example, Lee et al. [33] proposed a scheme to distinguish noisy tags from correct ones by utilizing neighbor voting to learn the relevance of each tag to image. Zhu et al. [19] cast the problem of tag refinement as decomposing the initial tagging matrix into a low-rank refined tagging matrix and a sparse error matrix. Liu et al. [21] built a hyper-graph model with semantic unities for tag clustering and refinement. Lin et al. [34] developed an image tag completion framework via image-specific and tag-specific linear sparse reconstructions. Among the latest works, the well-known MC philosophy is applied to the problem of defective tag assignments. Wu et al. [15] proposed to complete the optimal matrix while preserving the consistence between both observed tags and visual similarities. Liu et al. [20] utilized the non-negative matrix factorization algorithm to perform tag completion, embedding various contextual information that is available, such as within-image and cross-image relations.

Although a few of these methods (i.e. [15], [20], [21]) demonstrated their capability to accomplish both tag denosing and completion, their performance still needs further improvement, especially for the methods based on MC, which only considered the linear relations while ignoring the non-linear cues residing in the images with defective tag assignments. In our work, we impose the non-linear classification via matrix completion for the image tagging problem with defective tag assignments to capture the non-linearity by utilizing both the training and test data.

III. PROPOSED METHOD

A. PROBLEM FORMULATION

Let $\mathcal{O}_{tr} = \{o_1, \ldots, o_{N_{tr}}\}$ be a collection of N_{tr} training entries. Here $o_i = \{\mathbf{x}_i, \mathbf{Y}_i\}|_{i=1}^{N_{tr}}$ contains the visual feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and the tag vector $\mathbf{Y}_i \in \mathbb{R}^m$, $\mathbf{Y}_i \subseteq \mathcal{Y}$ is a set of tags, where $\mathcal{Y} = \{y_1, \ldots, y_m\}$ is a vocabulary of m tags and $\mathbf{Y}_{i,k}$ is set to one if tag k is assigned to image i and zero otherwise. For all the N_{tr} training entries, we denote the visual feature matrix $\mathbf{X}_{tr} = [\mathbf{x}_1, .., \mathbf{x}_{N_{tr}}] \in \mathbb{R}^{d \times N_{tr}}$ and the tag matrix $\mathbf{Y}_{tr} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_{tr}}] \in \mathbb{R}^{m \times N_{tr}}$. Similarly, we also have a set of N_{te} testing entries $\mathcal{O}_{te} = \{o_1, \ldots, o_{N_{te}}\}$, where $o_j = \{\mathbf{x}_j, \mathbf{Y}_j\}|_{j=1}^{N_{te}}$. For all the N_{te} testing entries, we denote the visual feature matrix $\mathbf{X}_{te} = [\mathbf{x}_1, .., \mathbf{x}_{N_{te}}] \in \mathbb{R}^{d \times N_{te}}$ and the tag matrix $\mathbf{Y}_{te} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_{te}}] \in \mathbb{R}^{m \times N_{te}}$. Note that for $\mathcal{O}_{te}$, $\mathbf{X}_{te}$ is known while $\mathbf{Y}_{te}$ is unknown. Our target is to learn an annotation model from the labeled training entries in $\mathcal{O}_{tr}$ and then apply the learned model to predict the tag matrix $\mathbf{Y}_{te}$ for the test entries in $\mathcal{O}_{tr}$. In the supervised setting, linear classifiers are introduced to learn mapping between the visual feature space of $\mathcal{X}$ and the label space $\mathcal{Y}$. The above process can be formulated by minimizing the loss between the output space and the projection of the input space:

$$[\mathbf{Y}_{tr} \quad \mathbf{Y}_{te}] = \mathbf{W}^\top [\mathbf{X}_{tr} \quad \mathbf{X}_{te}], \tag{1}$$

with $\mathbf{W} \in \mathbb{R}^{d \times m}$ being the classifiers parameters. Inspired by [35], we cast predicting $\mathbf{Y}_{te}$ as a Matrix Completion (MC) problem. Specifically, we concatenate the tag matrices and visual feature matrices of training and test entries into a joint tag-feature matrix $\mathbf{M}$, which is denoted as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{Y}_{tr} & \mathbf{Y}_{te} \\ \mathbf{X}_{tr} & \mathbf{X}_{te} \end{bmatrix} \in \mathbb{R}^{(m+d) \times (N_{tr}+N_{te})}. \tag{2}$$

If the assumption of linearity hold, the joint matrix $\mathbf{M}$ should be rank deficient. Thus, the classification process can be considered as filling the unknown tags $\mathbf{Y}_{te}$ of test entries. As mentioned in [35], the classifiers in Eq. 2 impose a linear dependency between the rows of matrix, indicating that the intra-relation in the visual feature space and the label space are not explicitly incorporated. In addition, the unknown tags of $\mathbf{Y}_{te}$ are directly estimated based on remaining part in $\mathbf{M}$. In practice, there are several challenges in the image auto-annotation problem that are not well explored in previous studies [15], [19], [34]. Firstly, there may exist errors and partial knowledge in the given tags $\mathbf{Y}_{tr}$ of training entries; in another words, the given tags may be incomplete and noisy. Secondly, the given tag matrix $\mathbf{Y}_{tr}$ of training entries are fixed during the process of matrix completion for M, indicating that the matrix completion can only be applied to test entries rather than the training entries. Thirdly, the matrix completion framework as introduced in Eq. 2 assumes the linear mapping between the visual feature space and label

space; this assumption may prevent it from capturing the non-linearity inside the two spaces. To overcome the difficulties, we present a formal methodology that integrate the classification capability of the matrix completion framework with the representation power of non-linear kernels for the visual features and achieve more accurate tag completion results for the testing entries. Simultaneously, the completion process can also be applied to the tags of training entries to better refine these tags.

B. DESIGN OF NON-LINEAR MATRIX COMPLETION

In the non-linear case, we assume the visual feature is mapped into a h-dimensional space through a feature mapping $\boldsymbol{\phi}$ (e.g., RBF kernel mapping). Specifically, for the i-th training entry o_i, its nonlinear feature mapping is denoted as $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$. Similarly, we have Φ_{tr} and Φ_{te} which are analogous to $\mathbf{X}_{tr}$ and $\mathbf{X}_{te}$. Furthermore, we denote the kernel matrices as $\mathbf{K}_{00} = \Phi_{tr}^\top \Phi_{tr} \in \mathbb{R}^{N_{tr}\times N_{tr}}$, $\mathbf{K}_{01} = \Phi_{tr}^\top \Phi_{te} \in \mathbb{R}^{N_{tr}\times N_{te}}$ and $\mathbf{K}_{11} = \Phi_{te}^\top \Phi_{te} \in \mathbb{R}^{N_{te}\times N_{te}}$. The new tag-feature matrix is define as:

$$\hat{\mathbf{Z}} = \begin{bmatrix} \mathbf{Y}_{tr} & \mathbf{Y}_{te} \\ \Phi_{tr} & \Phi_{te} \end{bmatrix} \in \mathbb{R}^{(m+N_{tr})\times(N_{tr}+N_{te})}. \tag{3}$$

Here we still assume the linearity between the tags and the new features after kernel mapping and we seek for a low-rank approximation of the matrix $\hat{Z}$ as:

$$\begin{aligned} \mathbf{Z}^* &= arg\min_{\mathbf{Z}} \ rank(\mathbf{Z}), \\ &s.t.\ \mathbf{B}(\mathbf{Z} - \hat{\mathbf{Z}}) = 0, \end{aligned} \tag{4}$$

where $\mathbf{B}$ plays as an a binary mask over the new tag-feature matrix of alls features and training tags, ensuring that the visual feature matrix of Φ_{te} be utilized at the training time. Directly minimizing the rank of $\hat{\mathbf{Z}}$ in Eq 4 is NP-hard and intractable. Previous matrix completion works such as [19] and [20] use the nuclear norm $\|\cdot\|_*$ to compute the sum of the singular values to formulate the tightest convex envelope for the rank of $\hat{\mathbf{Z}}$. Then the minimization problem in Eq. 4 can be approximated as

$$\begin{aligned} \mathbf{Z}^* &= arg\min_{\mathbf{Z}} \ \|\mathbf{Z}\|_*, \\ &s.t.\ \mathbf{B}(\mathbf{Z} - \hat{\mathbf{Z}}) = \mathbf{0}. \end{aligned} \tag{5}$$

Furthermore, inspired by [35], we also impose the decomposition $\mathbf{Z} = \mathbf{L}\mathbf{Q}^\top$, where $\mathbf{L} \in \mathbb{R}^{(m+N_{tr})\times r}$ and $\mathbf{Q} \in \mathbb{R}^{(N_{tr}+N_{te})\times r}$, and we also restraint that $\mathbf{Q}$ be orthogonal to avoid many identical local minima in the objective function. Finally, the optimization problem in our model is formulated as:

$$\begin{aligned} (\mathbf{L}^*, \mathbf{Q}^*) &= arg\min_{\mathbf{L},\mathbf{Q}} \ \|\mathbf{L}\|_F^2 + \|\mathbf{Q}\|_F^2, \\ &s.t.\ \mathbf{B}(\mathbf{L}\mathbf{Q}^\top - \mathbf{Z}) = 0, \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}. \end{aligned} \tag{6}$$

C. OPTIMIZATION ALGORITHM

For simplicity, we rewrite $\mathbf{L}$ and $\mathbf{Q}$ in the form of tag-feature matrices such as $\mathbf{M}$ and $\hat{\mathbf{Z}}$. In particular, $\mathbf{L} = [\mathbf{L}_0; \mathbf{L}_1]$, where $\mathbf{L}_0 \in \mathbb{R}^{m\times r}$ and $\mathbf{L}_1 \in \mathbb{R}^{N_{tr}\times r}$. Similarly, $\mathbf{Q} = [\mathbf{Q}_0; \mathbf{Q}_1]$, where $\mathbf{Q}_0 \in \mathbb{R}^{N_{tr}\times r}$ and $\mathbf{Q}_1 \in \mathbb{R}^{N_{te}\times r}$. In practice, since given tags and the extracted visual features may be noisy, the original constrain $\mathbf{B}(\mathbf{L}\mathbf{Q}^\top - \mathbf{Z}) = \mathbf{0}$ in Eq. 6 is too rigorous. Instead, we make a more appropriate relaxation for the original problem by measuring how close the predictions are to the available observations. Here the observations includes the known matrices in tag-feature matrix $\hat{\mathbf{Z}}$. Hence, the objective function can be approximated as:

$$\begin{aligned} &F(\mathbf{L}_{tr}, \mathbf{L}_{te}, \mathbf{Q}_{tr}, \mathbf{Q}_{te}) \\ &\quad = \|\mathbf{Y}_{tr} - \mathbf{L}_0\mathbf{Q}_0^\top\|_F^2 + \|\Phi_{tr} - \mathbf{L}_1\mathbf{Q}_0^\top\|_F^2 \\ &\qquad + \|\mathbf{Y}_{te} - \mathbf{L}_1\mathbf{Q}_1^\top\|_F^2 + \lambda(\|\mathbf{L}_0\|_F^2 + \|\mathbf{L}_1\|_F^2), \end{aligned} \tag{7}$$

where λ is a regularization parameter of the ℓ_2 regularization terms of $\|\mathbf{L}_0\|_F^2$ and $\|\mathbf{L}_1\|_F^2$. In Eq. 7, directly computing the $\mathbf{L}_1$ is infeasible since the features Φ_{tr} and Φ_{te} are in kernel space and cannot be explicitly computed. We adopt the scheme proposed in [35] to iteratively solve the sub-problems that are derived with respect to $\mathbf{L}_1$ and the other three variables successively.

The detailed optimization steps are as follows:

1) UPDATING $\mathbf{L}_1$

We first take the derivative of Eq. 7 with respect to $\mathbf{L}_1$ as

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{L}_1} &= -2\Phi_0\mathbf{Q}_0 + 2\mathbf{L}_1\mathbf{Q}_0^\top - 2\Phi_1\mathbf{Q}_1 \\ &\quad + 2\mathbf{L}_1\mathbf{Q}_1^\top\mathbf{Q}_1 + 2\lambda\mathbf{L}_1. \end{aligned} \tag{8}$$

Let Eq. 8 to be zero, we obtain the close-form solution for L_1 as

$$\mathbf{L}_1 = \frac{1}{\lambda+1}(\Phi_0\mathbf{Q}_0 + \Phi_1\mathbf{Q}_1). \tag{9}$$

2) UPDATING $\mathbf{L}_0$

We then replace the solution of $\mathbf{L}_1$ in Eq. 7 and we obtain a new objective function as

$$\begin{aligned} F(\mathbf{L}_0, \mathbf{Q}_1, \mathbf{Q}_0) &= -2Tr(\mathbf{Y}_{tr}^\top\mathbf{L}_0\mathbf{Q}_0^\top) + Tr(\mathbf{L}_0^\top\mathbf{L}_0\mathbf{Q}_0^\top\mathbf{Q}_0) \\ &\quad + \lambda Tr(\mathbf{L}_0^\top\mathbf{L}_0) - \frac{1}{\lambda+1}Tr(\mathbf{Q}^\top\mathbf{K}\mathbf{Q}), \end{aligned} \tag{10}$$

where $\mathbf{K} = [K_{ij}]_{ij} \in [0, 1]$ contains the kernel matrices of visual features of training and test entries. After we the derivative of $\mathbf{L}_0$ for Eq. 10, we obtain the close-from solution for $\mathbf{L}_0$ as

$$\mathbf{L}_0 = \mathbf{Y}_{tr}\mathbf{Q}_0(\mathbf{Q}_0^\top\mathbf{Q}_0 + \lambda\mathbf{I})^{-1}. \tag{11}$$

3) UPDATING $\mathbf{Q}_1$ AND $\mathbf{Q}_0$

By substituting the solution of $\mathbf{L}_0$ in Eq. 11 to Eq. 10, we have

$$\begin{aligned} F(\mathbf{Q}_1, \mathbf{Q}_0) &= -Tr(\mathbf{Q}_0^\top\mathbf{Y}_{tr}^\top\mathbf{Y}_{tr}\mathbf{Q}_0(\mathbf{Q}_0^\top\mathbf{Q}_0 + \lambda\mathbf{I})^{-1}) \\ &\quad - \frac{1}{\lambda+1}Tr(\mathbf{Q}^\top\mathbf{K}\mathbf{Q}). \end{aligned} \tag{12}$$

Eq. 12 can be solved with the classical interior-point algorithm, given the gradient values of $\mathbf{Q}_1$ and $\mathbf{Q}_0$. The

TABLE 1. General statistics of the two datasets.

Dataset	Images	Tags	Tags per image	Images per tag
IAPRTC-12	17665, 1962	291	5.7, 1, 23	34, 153, 4999
MIRFlickr	9359, 9335	457	4.55, 1, 45	145, 50, 1483

optimization details for $\mathbf{Q}_1$ and $\mathbf{Q}_0$ are similar as in [35]. Once the optimal solution of $\mathbf{Q}_1$ and $\mathbf{Q}_0$ are obtained, we can recursively compute $\mathbf{L}_0$ with Eq. 11 and $\mathbf{L}_1$ with Eq. 9.

D. TAG PREDICTION FOR OUT-OF-SAMPLE

Through the iterative updating step depicted above, we can obtain the optimal solutions for the model parameters $\{\mathbf{L}_0, \mathbf{L}_1, \mathbf{Q}_0, \mathbf{Q}_1\}$. Under the matrix completion framework, we can predict the unknown tags of the test entries via learned $\mathbf{L}_0$ and $\mathbf{Q}_1$ as:

$$\mathbf{Y}_{te} = \mathbf{L}_0\mathbf{Q}_1^{\top}. \tag{13}$$

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

1) DATASETS

We use two datasets for benchmarking, namely IAPRTC-12 [36] and MIRFlickr [37] and compare the performance of our method with previous approaches. The IAPRTC-12 dataset was introduced for cross-lingual retrieval where each image has a description. We only keep the nouns and treat them as annotations. The MIRFlickr dataset has been introduced to evaluate keyword-based image retrieval methods. The dataset contains images that were downloaded from Flickr website. The tags for each images are extracted from the user-assigned tags as well as the EXIF fields. Note that these two datasets are very challenging due to the large vocabulary and the significant diversity among visual content. Statistics of the datasets are shown in Table 1, where the counts of images are in the format "training, test" in column 2, and the counts are in the format "mean, minimum, maximum" in column 4 and 5, respectively. It is worth noting that a large portion of tags has a frequency less than the mean tag frequency for both dataset and the median tag frequency is also far less than the mean frequency. This justifies the assumption of defective tags that we have previously made in Section I.

2) FEATURES

To make fair performance comparison with previous methods, we use similar features in [11], where a combination of local and global features were used for both datasets. Local features include SIFT and hue descriptors obtained from multi-scale grid and Harris-Laplacian key points. Global features include histograms in RGB, HSV and LAB color spaces as well as Gist features. We adopt the scheme proposed in [38] to obtain the nonlinear representation for each type of feature, and transform all the features by the nonlinear kernel mapping. For Gist feature, we use the random Fourier feature mapping that approximates the Gaussian kernel. All the other descriptors above are histograms, and for them we extract the kernel mapping of term-wise square root as in [38]. In our experiments, we reduce each kernel-mapped feature to 500 dimensions and concatenate them into a 4,000-dimensional vector.

3) DEFECTIVE TAG ASSIGNMENTS

To simulate the condition of defective tags, we follow the setting in the previous works [22], [25]. Specifically, we consider two cases: incomplete case and noisy case. For the incomplete case, partial tags are randomly detected from the given tags for each images. The deletion process complies the principle $\min(1, \lceil N \times (1-\gamma)\rceil)$ ensuring that each image has at least one tag. For the noisy case, tags other than the given tags are randomly added to each image. This process follows the principle $N + \min(1, \lceil N \times ratio\rceil)$, ensuring that each image is corrupted by at least one noisy tag. Here N represents the number of originally tagged tags for an image, $\lceil\cdot\rceil$ is the ceiling function that returns the largest integer smaller than the given value, *ratio* represents the degree of incompletion or noise. In the experiments, we choose $\gamma = \{30\%, 50\%, 70\%, 90\%\}$, where the larger the ratio, the higher the degree of incompletion or noise.

4) COMPARED METHODS AND EVALUATION METRICS

For baselines in performance evaluation, we adopt widely used image auto-annotation methods (JEC [10] and TagProp [11]), tag recommendation approaches (Vote+ [29]) and recently proposed unified tag refinement frameworks of denoising and completion (LSR [34] and TMC [15]). Following [25], [34], the experimental results are evaluated using three standard measures: *average precision@N* (i.e. AP@N), *average recall@N* (AR@N) and *coverage@N* (C@N). In the top N predicted tags, AP@N measures the ratio of correctly predicted tags and AR@N measures the percentage of correct tags that are recovered out of all ground-truth tags. Both AP@N and AR@N are averaged over all test images. Besides, Coverage@N measures the ratio of test images with at least one correct tags. For all the three criteria, a larger value indicates better performance.

B. SENSITIVITY TO INCOMPLETION

For tag prediction results on IAPRTC-12 and MIRFlickr datasets, we measure all the algorithms in terms of AP@N, AR@N and C@N, with the ratio of incompletion γ varying from [30%, 50%, 70%, 90%]. Fig. 3 and Fig. 4 show the results for the two datasets measured by AP@N, AR@N and C@N, respectively. From the experimental results we can draw the following conclusions. 1) The proposed method outperform the image auto-annotation, tag refinement baselines, providing a demonstration for their effectiveness. 2) In general, both LSR and the proposed method significantly outperform all the other approaches. In addition, the proposed method performs competitively with the state-of-the-art approach LSR. 3) The proposed method recovers relevant tag more effectively for a large range of images

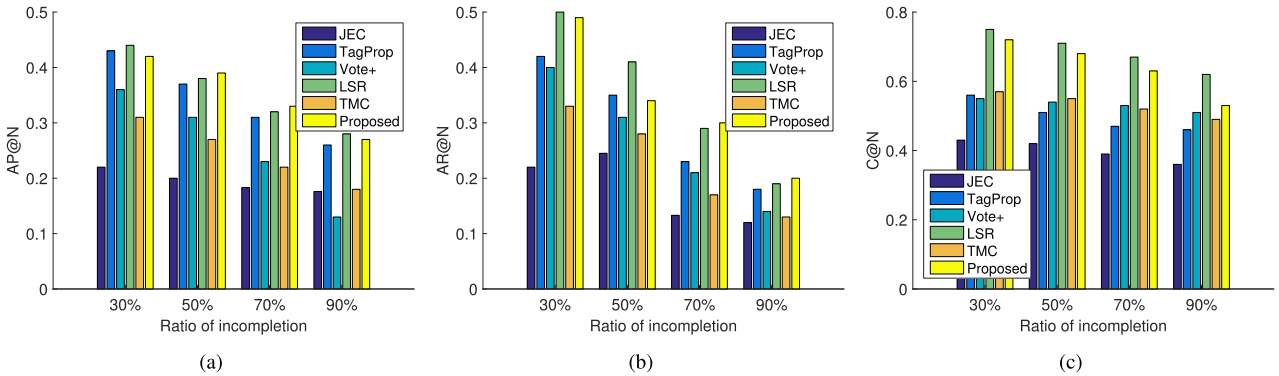


FIGURE 3. Tag prediction performance of the proposed method and other baselines with different ratio of incompletion on IAPRTC-12 dataset. (a) AP@N. (b) AR@N. (c) C@N.

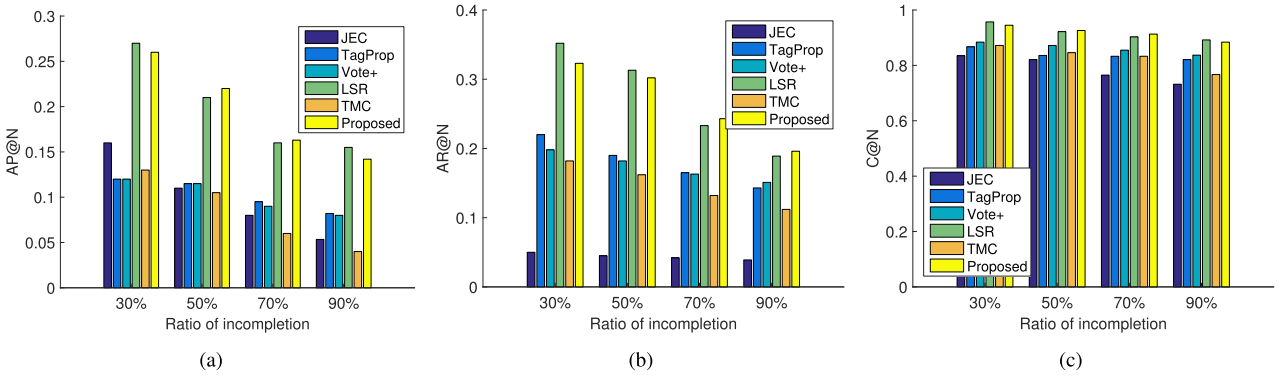


FIGURE 4. Tag prediction performance of the proposed method and other baselines with different ratio of incompletion on MIRFlickr dataset. (a) AP@N. (b) AR@N. (c) C@N.

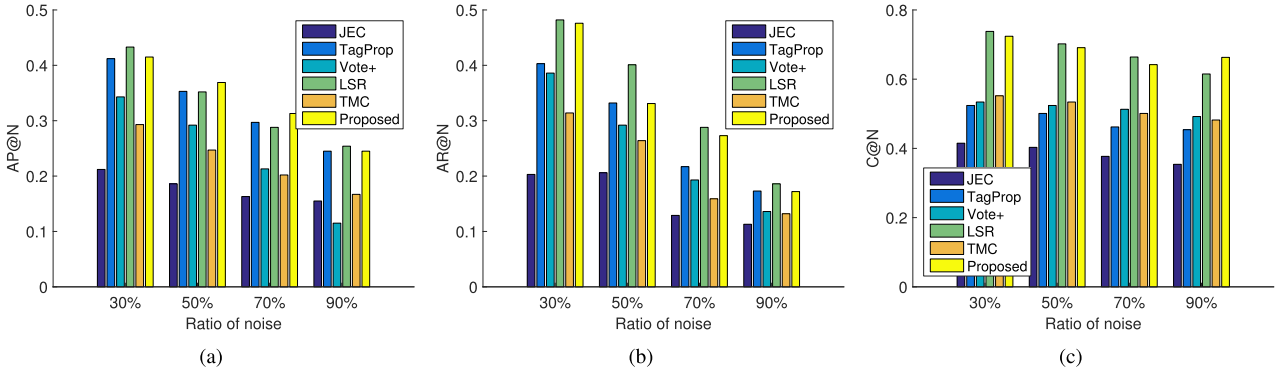


FIGURE 5. Tag prediction performance of the proposed method and other baselines with different ratio of noise on IAPRTC-12 dataset. (a) AP@N. (b) AR@N. (c) C@N.

with different ratio of incompletion. This is an important capability of tag completion methods. We can also see that the proposed method slightly outperforms LSR in terms of AR@N with larger γ when the degree of incompletion is serious (i.e. 70%, 90%).

C. SENSITIVITY TO NOISE

We conduct experiments with noisy observed tags to evaluate the sensitivity to noise. Fig. 5 and Fig. 6 shows the tag prediction performance for different algorithms with different ratio of noise on the two datasets, respectively. Not surprisingly, the performance of all methods degrades with the increasing amount of noise. This is rather expected because with severe noise, certain images do not have accurate observed tags for training the model, which is especially true for IAPRTC-12 where the original tags are already noisy. We can also see that the proposed method generally performs on par with LSR on three measures with different degree of noise.

Actually, LSR reconstructs the ideal tag matrix from image- and tag-specific views jointly, where the image-

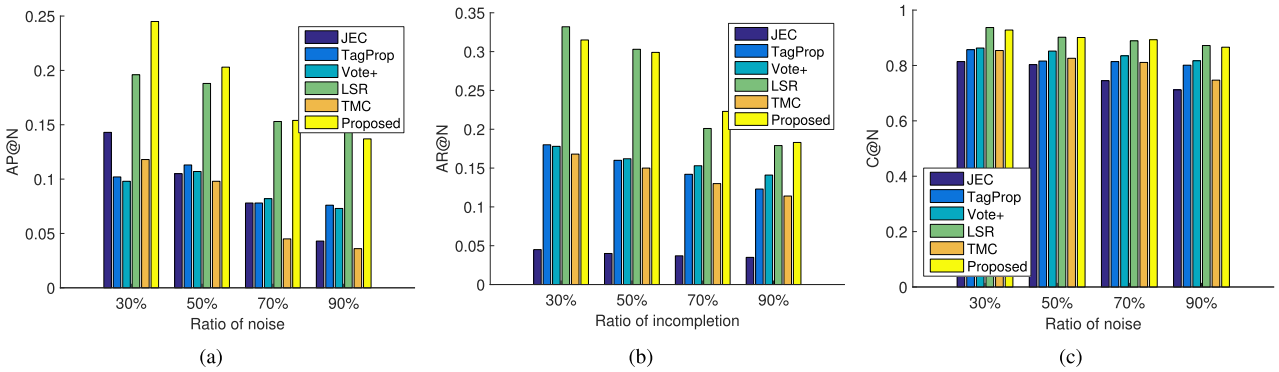


FIGURE 6. Tag prediction performance of the proposed method and other baselines with different ratio of noise on MIRFlickr dataset. (a) AP@N. (b) AR@N. (c) C@N.

FIGURE 7. Samples of tag completion results obtained by the proposed method and other two baselines (JEC and LSR) on the two datasets.

image, image-tag, tag-tag associations are explicitly incorporated. In contrast, the proposed method formulates a matrix that consists of both visual features and tags of both training and testing data, and further completes the matrix under a nonlinear matrix completion framework. Since these two method use different methodologies, they may have different performance on various conditions. When the degree of noise is large, the potential associations of image-tag and tag-tag are hard to capture, thus affecting the completion results of LSR and the proposed method.

D. EVALUATION WITH COMPLETELY LABELED TRAINING SET

Similar as the settings in LSR, we further conduct experiments on the two datasets with completely labeled training sets to investigate its performance. Specifically, we preserve the originally given tags for the training set and consider them as the complete tags. All the algorithms are then applied to the same test set. Table 2 shows the overall tag prediction results in terms of AP@N, AR@N and C@N. The best results obtained by the proposed method and the baselines are highlighted in bold font. Specifically, N is selected according to the mean number of tags for each datasets, i.e., $N = 6$ for IAPRTC-12 and $N = 4$ for MIRFlickr. It can be concluded from Table 2 that 1) all methods consistently performs better

TABLE 2. Experimental results on the two datasets with the completely labeled training sets, in terms of in terms of AP@N, AR@N and C@N.

	IAPRTC-12			MIRFlickr		
Method	AP@N	AR@N	C@N	AP@N	AR@N	C@N
JEC	0.263	0.278	0.859	0.158	0.053	0.828
TagProp	0.453	0.332	0.864	0.135	0.233	0.877
Vote+	0.389	0.267	0.833	0.118	0.203	0.881
LSR	**0.472**	0.322	0.872	0.232	**0.351**	**0.956**
TMC	0.356	0.323	0.763	0.134	0.184	0.876
Proposed	0.468	**0.326**	**0.882**	**0.258**	0.344	0.953

with completely labeled training set on the two datasets, indicating the importance of the quality of training data; 2) the proposed method and LSR performs significantly better than the other baselines, further demonstrating their effectiveness; 3) the proposed method performs competitively with LSR, thus it establishes another baseline for the annotation model with defective tags.

Fig. 7 shows qualitative samples of image tagging results obtained by the proposed method and two baselines JEC and LSR. on the two datasets. The first row contains the raw images, and the second row of "Groundtruth" represents the user-provided tags. The last three rows contain the top five tags predicted by the three methods. It can be observed that 1) the user-provided tags may be defective, for example, some proper tags are missing and some noisy tags are assigned for these images. 2) the predictions obtained by

JEC are usually inaccurate, since JEC ignores the issue of defective tag assignments on the training images, thus it fails to efficiently capture the relationships of different tags. For example, JEC prone to prediction frequent tags such as *nikon*, *d50*, and explore, which are usually noisy to describe the image content. 3) Compared with JEC, the LSR and the proposed methods achieve much better tagging results as they fully consider the issue of defectiveness. Specifically, the proposed method can recall more appropriate tags than those of LSR for the first two images such as the tags people, *cloud*, *grass*; while LSR gives some unrelated tags such as front, *road*, *snow*. Moreover, LSR seems to be inferior to tackle the noisy tags (e.g. *etsy*, *d50*) as it is initially designed for tag completion with incomplete tags. The proposed method is able to improve LSR for removing noisy tags, as the nonlinear matrix completion framework ensures the proposed method to jointly explore the relationships between visual features and tags and effectively distinguish the importance of different tags to each image.

V. CONCLUSION

In this paper we proposed an effective method for social image tagging with defective tags using nonlinear matrix completion. The proposed method first constructs the tag-feature matrix of both training and test data and then formulates a formal methodology together with an optimization method under the matrix completion framework to jointly complete the tags of training and test images. Extensive experiments conducted on two social image datasets with defective tags assignments verified that the proposed method achieved competitive performance compared with the state-of-the-art approaches.

REFERENCES

[1] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.

[2] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, "Image annotation by large-scale content-based image retrieval," in *Proc. 14th ACM Int. Conf. Multimedia*, 2006, pp. 607–610.

[3] Z. Fu, X. Wu, C. Guan, X. Sun, and K. Ren, "Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2706–2716, Dec. 2016.

[4] D. Tao, L. Jin, W. Liu, and X. Li, "Hessian regularized support vector machines for mobile image annotation on the cloud," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 833–844, Jun. 2013.

[5] Z. Yang, S. I. Kamata, and A. Ahrary, "NIR: Content based image retrieval on cloud computing," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 3. Nov. 2009, pp. 556–559.

[6] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 340–352, Feb. 2016.

[7] B. Sigurbjornsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 327–336

[8] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Corrections to 'exploiting web images for semantic video indexing via robust sample-specific loss,'" *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 246–256, Feb. 2015.

[9] C. Qin, X. Bao, R. R. Choudhury, and S. Nelakuditi, "TagSense: Leveraging smartphones for automatic image tagging," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 61–74, Jan. 2014.

[10] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 316–329.

[11] 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. (ICCV)*, Oct. 2009, pp. 309–316.

[12] Y. Verma and C. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7574. 2012, pp. 836–849.

[13] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma, "ARISTA—Image search to annotation on billions of web photos," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2987–2994.

[14] H. Xu, J. Wang, X. Hua, and S. Li, "Tag refinement by regularized LDA," in *Proc. 17th Int. Conf. Multimedia*, 2009, pp. 573–576.

[15] 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.

[16] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *Proc. IEEE CVPR*, Jun. 2011, pp. 2801–2808.

[17] X. Lou and F. A. Hamprecht, "Structured learning from partial annotations," in *Proc. ICML*, 2012, pp. 1519–1526.

[18] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *Proc. ICML*, 2013, pp. 1–9.

[19] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. 18th Int. Conf. Multimedia*, 2010, pp. 461–470.

[20] X. Liu, S. Yan, T. Chua, and H. Jin, "Image label completion by pursuing contextual decomposability," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 8, no. 2, p. 21, 2012.

[21] Y. Liu, F. Wu, Y. Zhang, J. Shao, and Y. Zhuang, "Tag clustering and refinement on semantic unity graph," in *Proc. 11th IEEE Int. Conf. Data Mining*, Dec. 2011, pp. 417–426.

[22] X. Xu, A. Shimada, and R. Taniguchi, "Tag completion with defective tag assignments via image-tag re-weighting," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[23] Y. Verma and C. V. Jawahar, "Exploring SVM for image annotation in presence of confusing labels," in *Proc. BMVC*, 2013, pp. 25.1–25.11.

[24] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *CoRR*, 2015. [Online]. Available: http://dblp.uni-trier.de/rec/bibtex/journals/csur/LiUBBSB16

[25] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. CVPR*, vol. 2. Jul. 2004, pp. 1002–1009.

[26] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo, "A revisit of generative model for automatic image annotation using Markov random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1153–1160.

[27] X. Xu, A. Shimada, and R. Taniguchi, "Latent topic model for image annotation by modeling topic correlation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Jose, CA, USA, Jul. 2013, pp. 1–6.

[28] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu, "Correlative multi-label multi-instance image annotation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 651–658.

[29] 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.

[30] X. Xu, A. Shimada, H. Nagahara, and R. Taniguchi, "Learning multi-task local metrics for image annotation," *Multimedia Tools Appl.*, vol. 75, no. 4, pp. 2203–2231, 2016.

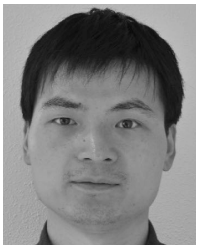
[31] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for support vector regression," *Neural Netw.*, vol. 67, pp. 140–150, Jul. 2015.

[32] X. Xu, A. Shimada, and R. Taniguchi, "Exploring image specific structured loss for image annotation with incomplete labelling," in *Proc. 12th Asian Conf. Comput. Vis.*, 2014, pp. 704–719.

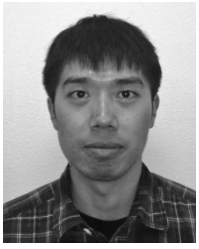
[33] S. Lee, W. D. Neve, and Y. M. Ro, "Image tag refinement along the 'what' dimension using tag categorization and neighbor voting," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 48–53.

[34] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1618–1625.

[35] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5240–5248.
[36] M. Grubinger, "Analysis and evaluation of visual information systems performance," Ph.D. dissertation, School Comput. Sci. Math., Victoria Univ., Melbourne, VIC, Australia, 2007.
[37] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. ACM MIR*, 2008, pp. 39–43.
[38] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.



XING XU received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan, in 2015. His current research interests include multimedia information retrieval and machine learning.



LI HE received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, China, in 2006 and 2009, respectively, and the M.S. degree in electrical engineering from the University of Nebraska-Lincoln, Lincoln, NE, USA, in 2011. He is currently an Engineer with the Qualcomm Research and Development Center, San Diego. His current research interests include object detection/tracking and machine learning.

HUIMIN LU (M'10) received the B.S. degree in electronics information science and technology from Yangzhou University in 2008, the M.S. degrees in electrical engineering from the Kyushu Institute of Technology and Yangzhou University in 2011, the Ph.D. degree from the Kyushu Institute of Technology in 2014. In 2013, he became a JSPS Research Fellow. His current research interests include computer vision, artificial intelligence, and deep-sea information processing. He is a member of the ACM.

ATSUSHI SHIMADA received the M.E. and D.E. degrees from Kyushu University in 2004 and 2007, respectively. Since 2007, he has been an Assistant Professor with the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include image processing, pattern recognition, and neural networks.

RIN-ICHIRO TANIGUCHI received the B.E., M.E., and D.Eng. degrees from Kyushu University in 1978, 1980, and 1986, respectively. Since 1996, he has been a Professor with the Graduate School of Information Science and Electrical Engineering, Kyushu University, where he directs several projects including multiview image analysis and software architecture for cooperative distributed vision systems. His current research interests include computer vision, image processing, and parallel and distributed computation of vision-related applications.