Image Tag Completion by Noisy Matrix Recovery

Zheyun Feng[1], Songhe Feng[2], Rong Jin[1], and Anil K. Jain[1]

[1] Michigan State University
[2] Beijing Jiaotong University
{fengzhey,rongjin,jain}@cse.msu.edu, shfeng@bjtu.edu.cn

Abstract. It is now generally recognized that user-provided image tags are incomplete and noisy. In this study, we focus on the problem of *tag completion* that aims to simultaneously enrich the missing tags and remove noisy tags. The novel component of the proposed framework is a noisy matrix recovery algorithm. It assumes that the observed tags are independently sampled from an unknown tag matrix and our goal is to recover the tag matrix based on the sampled tags. We show theoretically that the proposed noisy tag matrix recovery algorithm is able to simultaneously recover the missing tags and de-emphasize the noisy tags even with a limited number of observations. In addition, a graph Laplacian based component is introduced to combine the noisy matrix recovery component with visual features. Our empirical study with multiple benchmark datasets for image tagging shows that the proposed algorithm outperforms state-of-the-art approaches in terms of both effectiveness and efficiency when handling missing and noisy tags.

Keywords: Tag completion, noisy tag matrix recovery, matrix completion, missing/noisy tags, image tagging, image annotation, tag ranking.

1 Introduction

With the ever-growing popularity of digital photography and social media, the number of images with user-provided tags available over the internet has increased dramatically in the last decade. However, many user-provided tags are incomplete or inaccurate in describing the visual content of images [24], making them difficult to be utilized for tasks such as tag based image retrieval and tag recommendation [15,16,27]. This is particularly true for images extracted from social media [16,27], where in most cases, only a few tags are provided for each image and some of them are noisy. In this work, we develop an effective algorithm that can simultaneously recover the missing tags and remove or down weight the noisy tags which are irrelevant to the visual content of images.

We refer to our problem as *tag completion* [7] to distinguish it from previous image tagging work. *Image annotation* [9,11] automatically assigns images with appropriate keywords. As state-of-the-art image annotation approach, search based algorithms [9,11] rely on the quality of tags assigned to training images [9].

Tag recommendation suggests candidate tags to annotators in order to improve the efficiency and quality of the tagging process [14]. It usually identifies missing tags by topic models (e.g. *Latent Dirichlet Allocation* (LDA)) [2,14], but does not address the noisy tag problem. *Tag refinement* applies various techniques (e.g. topic model, tag propagation, sparse training and partial supervision [6,17,25]) to select a subset of user-provided tags based on image features and tag correlation [26]. Although it is able to handle noisy tags, it does not explicitly address the missing tag problem. Unlike most existing studies, the tag completion problem studied in this work simultaneously addresses the challenges of missing and noisy tags [7].

Since the tags of each image can be viewed as a mixture of topics and each topic follows a multinomial distribution over the vocabulary [2,14,25], we use a maximum likelihood component to ensure the learned tag probability matrix to be consistent with the observed tags. To simultaneously address the problem of missing and noisy tags, we assume that the observed tags are sampled independently from a low rank tag matrix; our goal is to recover the tag matrix from the noisy observations. By enforcing the recovered matrix to be low rank, we are able to effectively capture the correlation among different tags, which turns out to be the key in filling out missing tags and down weighting noisy ones [4,19,23]. This is in contrast to the existing studies for tag completion [15,18,24,27] where no principled approach is presented to capture the dependence among tags, which, however in our opinion, is the key issue to the tag completion problem. We refer to the proposed approach as tag completion by noisy matrix recovery, or **TCMR** for short.

We note that although low rank matrix recovery is closely related to topic model that has been applied to many image tag related problems [14,25], it has three novel contributions. First, unlike most existing topic models [1] that need to solve a non-convex optimization problem, the proposed TCMR solves a convex optimization problem and therefore is computationally more efficient. We have shown theoretically that under favorable conditions, the proposed TCMR is guaranteed to recover most of the missing tags even when the user-provided tags are noisy. This is in contrast to most topic models that do not come with theoretical support. Besides, TCMR further improves the performance by effectively exploiting the statistical dependence between image features and tags via a graph Laplacian [26,27], which reduces the impact of incomplete and noisy tags by assigning high weights to tags that are consistent with image features, and low weights to those which are not. Finally, our work is closely related to the theory of matrix completion and recovery [4,5]. Unlike existing studies on matrix completion/recovery, a maximum likelihood estimation is used in this work to recover the underlying low rank tag matrix, adding more complexity to both optimization and analysis.

The paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the noisy matrix recovery and TCMR. Section 4 presents the theoretical support of TCMR. Section 5 summarizes the experimental results, and Section 6 concludes this work with future directions.

2   Related Work

Image Tag Completion. There are only a few studies fitting the category of tag completion with both incomplete and noisy tags. [27] proposes a data-driven framework for tag ranking that optimizes the correlation between visual cues and assigned tags. [16] removes the noisy tags based on the visual and semantic similarities, and expands the observed tags with their synonyms and hypernyms using WordNet. [24] proposes to search for the optimal tag matrix that is consistent with both observed tags and visual similarity. [18] formulates tag completion into a non-negative data factorization problem. [15] exploits sparse learning techniques to reconstruct the tag matrix. None of these studies provides any theoretical guarantee for their approaches. Matrix decomposition is adopted in [3,21,26] to handle both missing and noisy tags. The key limitation of these approaches is that they require a full observed matrix with a small number of errors, making it inappropriate for tag completion.

Low Rank Matrix Recovery. Low rank matrix recovery has been applied in many applications [4,21], including visual recovery [19,21], multilabel classification [3], tag refinement [26], etc. Since the function of matrix rank is non-convex, a popular approach is to replace it with the nuclear norm, the tightest convex relaxation for matrix rank [4,5,26]. Using the nuclear norm regularization, it is possible to accurately recover a low rank matrix from a small fraction of its entries [5] even if they are corrupted with noise [4,10]. Various algorithms [10,12,21,26] have been developed to solve the related optimization problem. Instead of the ℓ_1-norm loss [10,26], squared loss [23] and max-margin factorization model [19] used in most studies on matrix completion/recovery, a maximum likelihood estimation is used in our work to recover the underlying tag matrix.

3   Image Tag Completion by Noisy Matrix Recovery (TCMR)

In this section, we first describe a noisy matrix recovery framework for tag completion and then discuss how to incorporate visual features into the matrix recovery framework.

3.1   Noisy Matrix Recovery

Let m be the number of unique tags, and $\mathcal{D} = \{\mathbf{d}_1, \cdots, \mathbf{d}_n\}$ be a collection of n tagged images, where $\mathbf{d}_i = (d_{i,1}, \cdots, d_{i,m})$ is the tag vector for the i-th image with $d_{i,j} = 1$ when tag j is assigned to the image and zero, otherwise. For the simplicity of analysis, in this study, we assume that all the images have the same number of assigned tags, denoted by m_*[1].

---

[1] This assumption is only for the convenience of analysis, and does not affect the algorithm. When different number of tags are observed, we apply the weighting technique [22] to handle the variation in the number of tags.

Following the idea of language models [1,2], we assume that all the observed tags in each image are drawn independently from a fixed but unknown multinomial distribution. Let $\mathbf{p}_i = (p_{i,1}, \cdots, p_{i,m})$ be the multinomial distribution used to generate tags in $\mathbf{d}_i$. We use $P = (\mathbf{p}_1, \cdots, \mathbf{p}_n)$ to represent the multinomial distributions for all the images. Our goal is to accurately recover the multinomial distribution P from a limited number of observed tags in $\mathcal{D}$. In general, this is impossible since the number of parameters to be estimated is significantly larger than the number of observed tags. To address this challenge, we follow the key assumption behind most topic models [23,26], *i.e.* tags of any image are sampled from a mixture of a small number of multinomial distributions. A direct implication of this assumption is that matrix P has to be of low rank, the foundation for the theory of low rank matrix recovery [5].

Before presenting our algorithm and analysis, we first introduce the notation that will be used throughout this paper. We use $Q_{*,i}$ to represent the i-th column of matrix Q, $|Q|_F$, $|Q|_{tr}$ and $|Q|_*$ to represent the Frobenius norm, nuclear (trace) norm and spectral norm of matrix Q, respectively. $|Q|_1$ is used to represent the ℓ_1 norm of matrix Q, *i.e.*, $|Q|_1 = \sum_{i,j} |Q_{i,j}|$, and $|\mathbf{v}|_\infty$ is used to represent the infinity norm of vector $\mathbf{v}$, *i.e.*, $|\mathbf{v}|_\infty = \max_i |v_i|$. We also use $\mathbf{e}_i \in \{0,1\}^n$ to represent the i-th canonical basis for $\mathbb{R}^n$, and $\mathbf{1} \in \mathbb{R}^m$ to represent a vector with all its entries being 1.

The proposed approach combines the idea of maximum likelihood estimation, a common approach for topic model, and the theory of low rank matrix recovery. It aims to recover the multinomial probability matrix P by solving the following optimization problem

$$\min_{Q \in \Delta} \quad \mathcal{L}(Q) := \underbrace{-\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{d_{i,j}}{m_*} \log Q_{i,j}}_{:=E_1} + \underbrace{\varepsilon |Q|_{tr}}_{:=E_2}, \qquad (1)$$

where domain $\Delta = \{Q \in (0,1)^{m \times n} : Q_{*,i}^\top \mathbf{1} = 1, i \in [1,n]\}$, and ε is a regularization parameter. We denote by $\hat{Q}$ the optimal solution to (1). Term E_1 in (1) ensures the learned probability matrix $\hat{Q}$ to be consistent with the observed tag matrix, and term E_2 ensures that $\hat{Q}$ is of low rank and therefore all image tags are sampled from a mixture of a small number of multinomial distributions.

3.2 Incorporating Visual Features

The limitation of the noisy matrix recovery method in (1) is that it fails to exploit visual features, an important hint for accurate tag prediction. So we next modify (1) to incorporate visual features.

Let $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ include the visual features of all images, where vector $\mathbf{x}_i \in \mathbb{R}^d$ represents the visual content of the ith image. Let $W = [w_{i,j}]_{n \times n}$ be the pairwise similarity matrix, where $w_{i,j}$ is the visual similarity between images $\mathbf{x}_i$ and $\mathbf{x}_j$, *i.e.*, $w_{i,j} = \exp\left(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma^2\right)$ if $j \in N_k(i)$ or $i \in N_k(j)$, where $N_k(i)$ denotes the index set for the k nearest neighbors of the ith image,

k is empirically set $k = 0.001n$, $d(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and σ is the average distance. We adopt χ distance if $\mathbf{x}_i$ is histogram features and ℓ_2 distance, otherwise. Using matrix W, we can measure the consistency between the estimated tag probability matrix Q and visual similarities by $\sum_{i,j=1}^{n} W_{i,j}|Q_{*,i} - Q_{*,j}|^2 = Tr(Q^\top L Q)$, where $L = \text{diag}(W^\top \mathbf{1}) - W$ is the graph Laplacian. By minimizing $Tr(Q^\top L Q)$, we ensure that the recovered probability matrix Q to be consistent with visual features.

By combining the noisy matrix recovery component with the component of visual features, we recover the tag probability matrix Q by solving the following optimization problem

$$\min_{Q \in \Delta} \; -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{d_{i,j}}{m_*} \log Q_{i,j} + \frac{\alpha}{n} Tr(Q^T L Q) + \beta|Q|_{tr}, \tag{2}$$

where both α and β are regularization terms. By minimizing the objective in (2), we are able to simultaneously fill out the missing tags and filter out/down weight the noisy tags.

3.3 Implementation

Incorporation with Irrelevant Tags. Regarding the fact that the initially unobserved tags are with a small probability relevant to the associated image, we also maximize the likelihood of their irrelevance, and the objective in (2) becomes

$$\min_{Q \in \Delta} \; -\sum_{i,j=1}^{n,m} \left[\frac{d_{i,j}}{m_*} \log Q_{i,j} + \frac{1 - d_{i,j}}{m - m_*} \log(1 - Q_{i,j}) \right] + \frac{\alpha}{n} Tr(Q^T L Q) + \beta|Q|_{tr}, \tag{3}$$

where Δ is defined in (1).

Efficient Solution of the Proposed Algorithm. We incorporate several heuristics to improve the computational efficiency. First, we adopt one projection paradigm that has been successfully applied to metric learning [8]. The key idea is to ignore the domain constraint $Q \in \Delta$ during the iteration, and only project the solution Q into Δ at the end of optimization. As a result, we only need to solve an unconstrained optimization problem. Second, we adopt the extended gradient method in [12]. To this end, we rewrite the objective function in (2) as $\mathcal{L}(Q) = f(Q) + \varepsilon|Q|_{tr}$. Given the current solution Q_{k-1}, we update the solution Q_k by solving the following optimization problem

$$\arg\min_{Q} P_{t_k}(Q, Q_{k-1}) = \frac{1}{2}\left| Q - \left(Q_{k-1} - \frac{1}{t_k}\nabla f(Q_{k-1})\right) \right|_F^2 + \frac{\varepsilon}{t_k}|Q|_{tr}. \tag{4}$$

where t_k is the step size for the kth iteration. The detailed algorithm for solving the unconstrained version of (2) can be found in [12].

4 Theoretical Guarantee of TCMR

The following theorem bounds the difference between P and the recovered probability matrix $\hat{Q}$.

Theorem 1. *Let r be the rank of matrix P, and N be the total number of observed tags. Let $\hat{Q}$ be the optimal solution to (1). Assume $N \geq \Omega(n \log(n+m))$, and denote by μ_- and μ_+ the lower and upper bounds for the probabilities in P. Then we have, with a high probability*

$$\frac{1}{n}|\hat{Q} - P|_1 \leq O\left(\frac{rn\theta^2 \log(n+m)}{N}\right), \quad \text{where } \theta^2 := \frac{\mu_+|P\mathbf{1}|_\infty}{n\mu_-^2} \leq \frac{\mu_+^2}{\mu_-^2}. \quad (5)$$

It is clear that the recovery error is $O(rn \log(n + m)/N)$, implying that the tag matrix can be accurately recovered when $N \geq \Omega(rn \log(n + m))$. This is consistent with the standard results in matrix completion [13]. The impact of low rank assumption is analyzed in Section 4.1. We note that unlike standard matrix completion theory where observed entries are sampled uniformly at random from a given matrix, in topic model, each observed tag is sampled from an unknown multinomial distribution. This difference makes the square loss inappropriate for topic model, leading to additional challenges in analyzing the recovery property for topic model.

We now proceed to present a sketch of the proof. More details can be found in the supplementary document. Define matrix M as

$$M := \sum_{i=1}^{n} \left(\frac{1}{m_*}\mathbf{d}_i - \mathbf{p}_i\right) \mathbf{e}_i^\top = \sum_{i=1}^{n} \frac{1}{m_*}\mathbf{d}_i\mathbf{e}_i^\top - P, \quad (6)$$

where $\mathbf{e}_i \in \{0,1\}^n$ is the canonical base for $\mathbb{R}^n$. Since the occurrence of each tag in $\mathbf{d}_i$ is sampled according to the underlying multinomial distribution $\mathbf{p}_i$, it is easy to verify that $\mathbb{E}[M] = 0$.

Before presenting our analysis, we need two supporting lemmas that are important to our analysis.

Lemma 1. *Let $P \in \Delta$ and $Q \in \Delta$ be two probability matrices. We have*

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|P_{i,j} - Q_{i,j}|^2}{Q_{i,j}} \geq \sum_{i=1}^{n}\sum_{j=1}^{m} |P_{i,j} - Q_{i,j}| = |P - Q|_1. \quad (7)$$

Lemma 2. *([13]) Let $Z_1, \cdots, Z_n$ be independent random matrices with dimension $m_1 \times m_2$ that satisfy $\mathbb{E}[Z_i] = 0$ and $|Z_i|_* \leq U$ almost surely for some constant U, and all $i = 1, \cdots, n$. Define*

$$\sigma_Z = \max\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_iZ_i^\top]\right|_*, \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i^\top Z_i]\right|_*\right\}. \quad (8)$$

Then, for all $t > 0$, with a probability $1 - e^{-t}$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}Z_i\right|_* \leq 2\max\left\{\sigma_Z\sqrt{\frac{t + \log(m_1 + m_2)}{n}}, U\frac{t + \log(m_1 + m_2)}{n}\right\}. \quad (9)$$

The following theorem is the key to our analysis. It shows that the estimation error $|P - Q|_1$, measured by ℓ_1 norm, will be small when P can be well approximated by a low rank matrix.

Theorem 2. *Let $\hat{Q}$ be the optimal solution to (1). If $\varepsilon \geq |M|_*/\mu_-$, where M is defined in (6), then*

$$|\hat{Q} - P|_1 \leq \min_{Q \in \Delta} \left\{ \frac{1}{\mu_-}|Q - P|_F^2 + 16\varepsilon^2 \mu_+ rank(Q) \right\}. \tag{10}$$

To utilize Theorem 2 for bounding the difference between P and $\hat{Q}$, we need to bound $|M|_*$. The theorem below bounds $|M|_*$ by using Lemma 2.

Theorem 3. *Define γ as*

$$\gamma := \frac{2}{\mu_-} \max \left(\frac{t + \log(m + n)}{m_*}, \sqrt{\max(1, |P\mathbf{1}|_\infty)\frac{t + \log(n + m)}{m_*}} \right). \tag{11}$$

With a probability $1 - e^{-t}$, we have $|M|_* \leq \gamma \mu_-$.

Combining Theorems 2 and 3, we have the following result for recovering the probability matrix P.

Corollary 1. *Set $\varepsilon = \gamma$. With a probability at least $1 - e^{-t}$, we have*

$$|\hat{Q} - P|_1 \leq \min_{Q \in \Delta} \left\{ |Q - P|_F^2/\mu_- + 16\gamma^2 \mu_+ rank(Q) \right\}. \tag{12}$$

Furthermore, let $\hat{P}$ be the best rank-r approximation of P. We have, with a probability $1 - e^{-t}$

$$|\hat{Q} - \hat{P}|_1 \leq |P - \hat{P}|_F^2/\mu_- + 16\gamma^2 \mu_+ r. \tag{13}$$

We now come to the proof of Theorem 1. When the rank of P is r, using Corollary 1, we have, with a high probability, $|\hat{Q} - P|_1 \leq 16\gamma^2 \mu_+ r$. If $|P\mathbf{1}|_\infty \geq 1$ and $m_* \geq O(\log(m + n))$, we have

$$\gamma = O \left(\frac{1}{\mu_-} \sqrt{|P\mathbf{1}|_\infty \frac{\log(n + m)}{m_*}} \right) \tag{14}$$

and therefore, with a high probability, we have

$$\frac{1}{n}|\hat{Q} - P|_1 \leq O \left(\frac{r\log(n + m)}{m_*} \frac{\mu_+|P\mathbf{1}|_\infty}{\mu_-^2} \right) \leq O \left(\frac{rn\log(n + m)}{N} \frac{\mu_+|P\mathbf{1}|_\infty}{n\mu_-^2} \right) \tag{15}$$

where N is the number of observed tags. This immediately implies Theorem 1.

4.1   Impact of Low Rank Assumption on Recovery Error

In order to see the impact of low rank assumption, let us consider the maximum likelihood estimation of multinomial distribution. Since tags for different images are sampled independently, we only need to consider one image at each time. Let $\mathbf{p}$ be the underlying multinomial distribution to be estimated, and let $\mathbf{d}$ be the image tag vector comprised of m_* words sampled from $\mathbf{p}$. We estimate $\mathbf{p}$ by the simple maximum likelihood estimation, *i.e.*,

$$\min_{\mathbf{p} \in [\mu_-, \mu_+]^m : \mathbf{p}^\top \mathbf{1} = 1} -\sum_{i=1}^n d_i \log p_i, \tag{16}$$

where m is the number of unique tags, n is the number of images, μ_- and μ_+ are the lower and upper bounds for the probabilities in matrix $P = (\mathbf{p}_1, \cdots, \mathbf{p}_n)$.

Theorem 4. *Define* $\mathbf{z} = \mathbf{d}/m_* - \mathbf{p}$. *Let* $\hat{\mathbf{q}}$ *be the optimal solution to* (16). *Then*

$$|\mathbf{p} - \hat{\mathbf{q}}|_1 \leq (\mu_+^2/\mu_-^2)|\mathbf{z}|_2^2.$$

Theorem 5. *With a probability* $1 - 2e^{-t}$,

$$|\mathbf{z}|_2 \leq \sqrt{\frac{t + \log m}{\mu_- m_*}}|\mathbf{p}|_2.$$

Following the concentration inequality for vectors in Theorem 5, we bound $|\mathbf{z}|_2$. Then by combining Theorems 4 and 5, we have, with a probability $1 - 2e^{-t}$,

$$|\mathbf{p} - \hat{\mathbf{q}}|_1 \leq \frac{\mu_+^2 |\mathbf{p}|_2^2}{\mu_-^4} \frac{2(t + \log m)}{m_*} \tag{17}$$

By applying the above result to matrix P and taking the union bound, we have, with probability $1 - e^{-t}$,

$$\frac{1}{n}|P - \hat{Q}|_1 \leq \frac{\mu_+^2}{\mu_-^4} \max_{1 \leq i \leq n} |\mathbf{p}_i|_2^2 \frac{2n(t + \log m + \log n)}{|\Omega|}. \tag{18}$$

We now compare the bound in (18) to that in (5). It is easy to verify that $|\mathbf{p}_i|_2^2/\mu^2 \geq m$ for any $\mathbf{p}_i$. Hence, the net effect of the bound in (5) is to replace m with r, which is exactly the impact of low rank assumption.

5   Experiments

5.1   Datasets and Experimental Setup

Four benchmark datasets are used to evaluate our proposed algorithm. ESP Game dataset was collected for a collaborative image labeling task and consists

of images including logos, drawings and personal photos. IAPR TC12 dataset consists of images of actions, landscapes, animals and many other contemporary life, and its tags are extracted from the text captions accompanying each image. Both Mir Flickr and NUS-WIDE datasets [7] include images crawled from Flickr, together with users provided tags. ESP Game and IAPR TC12 are collaboratively human labeled and thus relatively clean, while Mir Flickr and NUS-WIDE are automatically crawled from social media and hence pretty noisy. A bag-of-words model based on densely sampled SIFT descriptors is used to represent the visual content in Mir Flickr, ESP Game and IAPR TC12 datasets[2]. In NUS-WIDE dataset, visual content are represented by six low-level features, including color information, edge distribution and wavelet texture [7].

To evaluate the proposed approach for tag completion, we divide the original tag matrix Y into two parts: the observed tag matrix (*i.e.* training set) D and the left as evaluation ground truth (*i.e.* testing set). We create the observed tag matrix by randomly sampling a subset of tags from D for each image. The number of observed tags m_* is set to 3 for Mir Flickr and 4 for other datasets throughout this section unless it is specified otherwise. To guarantee that the evaluation is meaningful, we ensure that each image has at least one evaluation tag by filtering out images with too few tags and tags associated with only a few images. As a result of this filtering step, Mir Flickr has $5,231$ images with 372 tags, ESP Game has $10,450$ images with 265 tags, IAPR TC12 has $12,985$ images with 291 tags, and NUS-WIDE has $20,968$ images with 420 tags. Detailed statistics about the refined datasets are listed in the supplementary document. All the hyper parameter values used in TCMR, *e.g.* ε, α, β, and the parameter values in the baselines are determined by cross-validation.

Following [15], we evaluate the tag completion accuracy by the *average precision @N* (*AP@N*). It measures the average percentage of the top N recovered tags that are correct. Note that a tag is correctly recovered if it is included in the original tag matrix Y but not observed in D. We also use *average recall* (*AR@N*) to measure the percentage of correct tags that are recovered by a computational algorithm out of all ground truth tags, and *coverage* (*C@N*) to measure the percentage of images with at last one correctly recovered tag. Both the mean and standard deviation of evaluation metrics over 20 experimental trials are reported in this paper.

5.2   Comparison to State-of-the-Art Tag Completion Methods

We first compare our proposed TCMR algorithm[3] to several state-of-the-art tag completion approaches: (1) LRES [26], tag refinement towards low-rank, content-tag prior and error sparsity, (2) TMC [24] that searches for the optimal

---

[2] The features were obtained from
`http://lear.inrialpes.fr/people/guillaumin/data.php`. More detailed description about Mir Flickr, ESP Game and IAPR TC12 can also be found at this site.
[3] The source code can be downloaded from our website
`http://www.cse.msu.edu/~fengzhey/downloads/src/tcmr.zip`

tag matrix consistent with both the observed tags and visual similarity, (3) MC-1 [3] which applies low rank matrix completion to the concatenation of visual features and assigned tags, (4) FastTag [6] that co-regularizes two simple linear mappings in a joint convex loss function, (5) LSR [15] that optimally reconstructs each image and each tag with remaining ones under constraints of sparsity. We also compare the proposed approach with three state-of-the-art image annotation algorithms that are designed for clean tags: (6) TagProp [11], (7)RKML [9], a kernel metric learning algorithm, and (8) vKNN, a nearest neighbor voting algorithm.
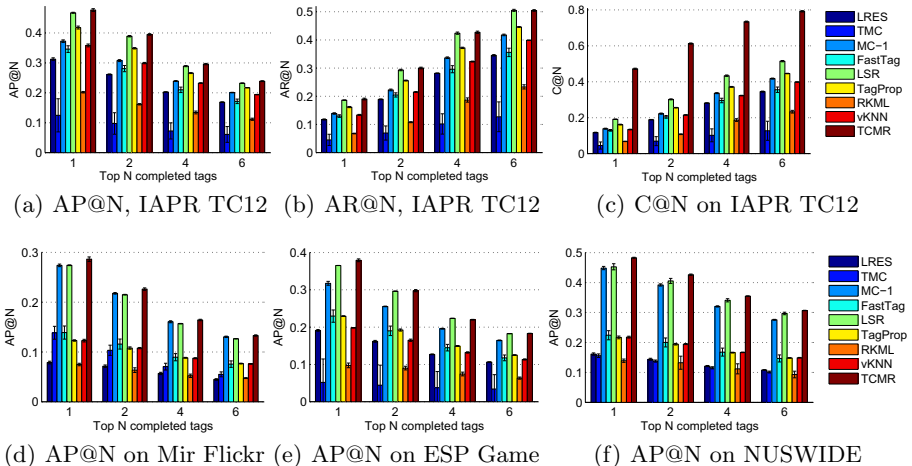

(a) AP@N, IAPR TC12   (b) AR@N, IAPR TC12   (c) C@N on IAPR TC12

(d) AP@N on Mir Flickr (e) AP@N on ESP Game   (f) AP@N on NUSWIDE

Fig. 1. Tag completion performance of TCMR and state-of-the-art baselines

Figs. 1 (a), (b), and (c) show the results for the IAPR TC12 dataset measured by $AP@N$, $AR@N$ and $C@N$, respectively. Figs. 1 (d), (e), and (f) show the results of $AP@N$ for the three remaining datasets; the results of the other two metrics can be found in the supplementary document. We observe that overall, the proposed TCMR and LSR yield significantly better performance than the other approaches in comparison. TCMR performs significantly better than LSR in terms of $C@N$. In particular, TCMR recovers at least one correct tag out of the top six predicted tags for 80% of the images while the other approaches are only able to recover at least one correct tag for less than 50% of the images, indicating that the proposed algorithm is more effective in recovering relevant tags for a wide range of images, an important property for image tag completion algorithm. We also observe that TCMR performs slightly better than LSR in terms of $AP@N$ when the number of predicted tags N is small.

Table 1 summarizes the running time of all algorithms in comparison. We observe that although TCMR is not as efficient as several baselines, it is more efficient than LSR which yields similar performance as TCMR in multiple cases. The high computational cost of LSR is due to the fact that it has to train a different model for each instance, which does not scale well to large datasets.

Table 1. Running time (seconds) for tag completion baselines. All algorithms are run in Matlab on an AMD 4-core @2.7GHz and 64GB RAM machine.

	LRES	TMC	MC-1	FastTag	LSR	TagProp	RKML	vKNN	TCMR
MirFlickr	5.6e2	4.7e3	8.6e2	1.4e3	6.2e3	2.5e2	3.0e2	2.1e2	1.3e3
ESP Game	3.4e2	5.8e3	1.0e3	8.6e2	1.3e4	6.7e2	1.3e3	4.3e2	5.9e3
IAPR TC12	5.2e2	1.2e4	1.7e3	1.6e3	1.6e4	1.1e3	1.5e3	1.0e3	9.4e3
NUS-WIDE	6.8e3	2.9e4	1.8e3	2.6e3	2.8e4	1.5e3	3.8e3	1.2e3	1.9e4

Evaluation of Noisy Matrix Recovery. The key component of the proposed approach is a noisy matrix recovery framework. To independently evaluate the effectiveness of noisy matrix recovery component proposed in this work, we compare it (TCMR0) to several baseline approaches for matrix completion that do not take into account visual features: (1) Freq, which assigns the most frequent tags to all the images, (2) LSA [20], Latent Semantic Analysis, (3) tKNN, majority voting among the nearest neighbors in the tag space, (4) LDA [2], (5) LRES0 [26], a version of LRES algorithm without using visual features, and (6) pLSA, probabilistic LSA.
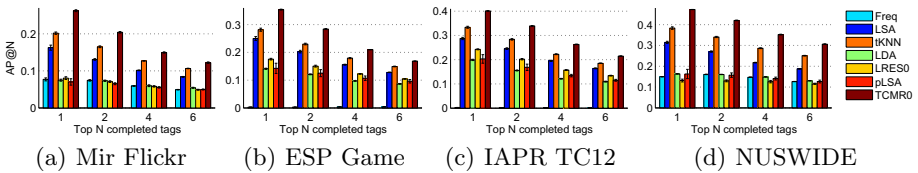

(a) Mir Flickr   (b) ESP Game   (c) IAPR TC12   (d) NUSWIDE

Fig. 2. *AP@N* for different topic models and matrix completion algorithms

Fig. 2 compares the tag completion performance without using visual features. We observe that the proposed noisy matrix recovery algorithm performs significantly better than the other baseline methods, implying that it can successfully capture the important dependency among tags. We also observe that a simple tKNN algorithm works better than the topical models (LSA, LDA and pLSA), suggesting that directly applying a topical model may not be appropriate for the tag completion problem.

From Figs. 1 and 2, we observe that TMC and RKML perform much worse than the other algorithms in comparison, while LSA and tKNN perform quite good. Accordingly, we exclude TMC and RKML, and include LSA and tKNN in the following evaluation cases.

Sensitivity to the Number of Observed Tags. We also examine the sensitivity of the proposed TCMR to the number of initially observed tags by comparing it to the baseline algorithms on IAPR TC12 and NUS-WIDE datasets. To make a meaningful evaluation, we only keep images with 6 or more tags for IAPR TC12 dataset, and images with 9 or more tags for NUS-WIDE dataset. As before, we

divide the tags into testing and training sets, and randomly sample m_* tags for each image from the training tag set to create the partially observed tag matrix, where the number of sampled tags m_* is varied. We evaluate the tag completion performance on the testing tag sets.
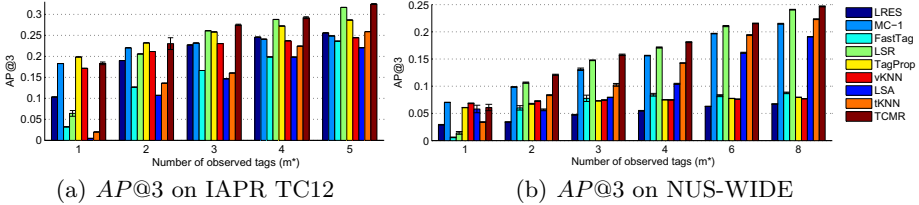

Fig. 3. Tag completion performance (AP@3) with varied number of observed tags

Fig. 3 shows the influence of the number of partially observed tags to the final tag completion performance measured by $AP@3$; results of the metric $AR@5$ are reported in the supplementary document. We observe that the performance of all algorithms improves with increasing number of observed tags. We also observe that when the number of observed tags is 3 or larger, TCMR and LSR perform significantly better than the other baseline approaches. When the number of observed tags is small (*i.e.* 1 or 2), TCMR performs significantly better than LSR, indicating that the proposed algorithm is noticeably effective even when the number of observed tags is small.

Sensitivity to Noise. To evaluate the sensitivity to noise, we conduct experiments with noisy observed tags on datasets IAPR TC12 and NUS-WIDE. To generate noisy tags, we replace some of the sampled tags with the incorrect ones that are chosen uniformly at random from the vocabulary. The percentage of noisy tags among the total observed ones in the whole gallery is varied from 0 to 0.9. To ensure there are a sufficient number of noisy tags as well as sufficient number of images, we set m_*, the number of sampled tags, to be 8 for NUS-WIDE dataset and to be 4 for IAPR TC12 dataset in this experiment.

Fig. 4 shows the tag completion performance for different algorithms using noisy observed tags. It is not surprising to observe that the performance of all algorithms in comparison degrades with increasing amounts of noise. We also
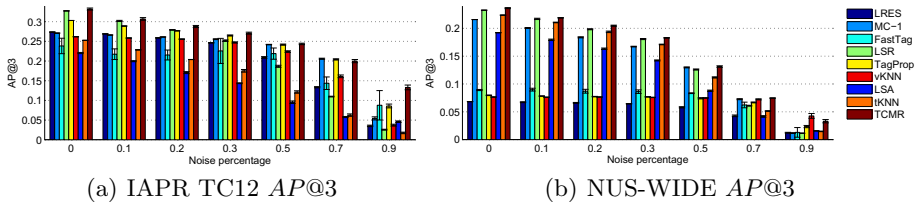

Fig. 4. Comparison of tag completion performance (AP@3)using noisy observed tags

Table 2. Examples of tag completion results generated by some baseline algorithms and the proposed TCMR. The observed tags in red italic font are noisy tags, and others are randomly sampled from the ground truth tags. The completed tags are ranked based on the recovered scores in descending order, and the correct ones are highlighted in blue bold font.

Ground truth	building, front, group, people, palm, lawn, tree, square, statue	boy, cap, hair, house, power, pole, roof, sky, shirt, sweater, terrain, tree	bank, bush, helmet, jacket, life, people, river, rock, tree	balcony, door, entrance, car, flag, front, lamp, house, sky, window	bed, brick, curtain, leg, man, short, sweater, wall, woman
Observed tags	lawn, people, square, *cloud*	cap, terrain, sky, *meadow*	life, river, tree, *llama*	balcony, car, window, *water*	curtain, wall, *floor*, *team*
LSER	**people**, bike, wall, cloud, **square**, **tree**, house, **lawn**, **palm**	**terrain**, **sky**, **hair**, **sweater**, **roof**, mountain, wall, meadow, **cap**, trouser	**tree**, **river**, **life**, **helmet**, **rock**, woman, llama, **jacket**, gravel, **people**	**entrance**, **car**, **front**, **balcony**, water, **window**, building, people, harbour, **sky**	**woman**, **wall**, table, room, hand, **curtain**, floor, team, person, front
MC-1	**people**, **square**, cloud, **lawn**, **tree**, sky, **building**, **front**, **hair**	**sky**, meadow, **terrain**, **cap**, wall, mountain, man, **house**, woman, **hair**	**tree**, **river**, **life**, man, llama, wall, **people**, front, mountain, sky	**window**, **car**, **balcony**, water, man, **front**, building, wall, **house**, woman	**wall**, **curtain**, floor, team, window, room, **man**, table, front, **bed**
FastTag	**tree**, tourist, footpath, shirt, river, **group**, woman, tile, **people**	wall, **boy**, desk, meadow, mountain, girl, **hair**, tee-shirt, plane, fence	**life**, mountain, **people**, front, tourist, railing, **river**, llama, **tree**, wall	building, **front**, **house**, **car**, grey, **window**, rail, **balcony**, street, photo	**wall**, room, table, window, **bed**, **curtain**, hand, night, cup, towel
LSR	sky, **square**, **building**, **people**, **tree**, house, **lawn**, street, cloud	**house**, **sky**, hill, **boy**, grey, jacket, **tree**, **terrain**, cloud, landscape	**bank**, **jacket**, **river**, **helmet**, **bush**, tourist, boat, mountain, **tree**, **people**	**front**, building, **house**, wall, **sky**, cliff, **door**, **window**, street, man	**wall**, room, window, front, uniform, **bed**, table, jersey, **short**, round
TagProp	**people**, **tree**, **square**, house, **front**, wall, tourist, man, woman	wall, woman, man, **sky**, front, **sweater**, **hair**, mountain, table, desert	**people**, **tree**, woman, front, man, **rock**, wall, **river**, sky, mountain	wall, **front**, man, building, woman, table, people, **house**, **sky**, **entrance**	front, **woman**, **wall**, table, **man**, house, room, people, tree, window
vKNN	**tree**, wall, house, **people**, sky, woman, bike, **front**, **square**	**sweater**, desert, **sky**, landscape, **terrain**, **hair**, mountain, wall, cloud, front	**people**, **tree**, **helmet**, front, **river**, **bush**, woman, **life**, sky, man	**front**, building, people, **house**, **entrance**, **sky**, wall, **balcony**, tree, **window**	room, **woman**, table, front, house, **wall**, **man**, chair, window, child
LSA	**people**, cloud, **square**, roof, **group**, meadow, **building**, tower, landscape	**sky**, meadow, cloud, **hair**, **roof**, road, short, **tree**, woman, **boy**	**tree**, **bush**, lake, palm, meadow, **river**, tourist, slope, building, grass	**car**, **window**, street, **house**, building, room, **lamp**, **front**, bed, bush	**wall**, room, table, **bed**, window, hair, girl, wood, boy, **curtain**
tKNN	**people**, **square**, cloud, **lawn**, sky, **tree**, mountain, street, **building**	**sky**, meadow, **terrain**, **cap**, people, cloud, hill, mountain, road, **tree**	**tree**, **river**, **life**, **bush**, house, sky, building, man, **people**, **bank**	**window**, **car**, **balcony**, wall, **house**, **front**, building, bed, room, curtain	**wall**, floor, **curtain**, room, **bed**, front, window, girl, team, **brick**
TCMR	**people**, **square**, **lawn**, sky, **building**, **tree**, cloud, street, **palm**	**sky**, **terrain**, **cap**, **boy**, hill, **house**, **hair**, landscape, **sweater**, cloud	**tree**, **river**, **life**, boat, **jacket**, **bank**, llama, **helmet**, **rock**, mountain	**car**, **window**, **balcony**, **door**, building, wall, **front**, **house**, water, **sky**	**wall**, floor, **curtain**, **bed**, **brick**, room, window, front, table, team

observe that LSR seems to be significantly sensitive to the noise in the observed tags than the proposed TCMR algorithm. In particular, we find that TCMR outperforms LSR significantly when the percentage of noisy tags is large. The contrast is particularly obvious for the IAPR TC12 dataset, where LSR starts to perform worse than several other baselines when the noise level is above 50%. Besides, all algorithms reduce their performance dramatically as the noise level increases from 70% to 90%. This is not surprising because at the 90% noise level, a number of images do not have accurate observed tags for training the model, especially for the NUS-WIDE dataset whose originally assigned tags are pretty noisy. However, the proposed TCMR algorithm is overwhelmingly better in this case, especially on IAPR TC12, indicating that it is more powerful in recovering expected tags from severely noisy tagged images. Table 2 shows the tag completion results of exemplar images by different algorithms, where both partially true and noisy tags are observed.

6   Conclusions

We have proposed a robust yet efficient image tag completion algorithm (TCMR), which is capable of simultaneously fill in the missing tags and remove/down weight the noisy tags. TCMR introduces a noisy matrix recovery procedure that captures the underlying dependency by enforcing the recovered matrix to be of low rank. Besides, a graph Laplacian based on the image visual features is also applied to ensure the recovered tag matrix is consistent with the visual content. Experiments over four different scaled image datasets demonstrate the effectiveness and efficiency of the proposed TCMR algorithm by comparing it to state-of-the-art tag completion approaches. In the future, we plan to improve the tag completion performance by incorporating the visual features a more effectively, and adopting more efficient nuclear norm optimization procedure.

References

1. Blei, D.M.: Probabilistic topic models. Communications of the ACM (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research (2003)
3. Cabral, R.S., la Torre, F.D.D., Costeira, J.P., Bernardino, A.: Matrix completion for multi-label image classification. In: NIPS, pp. 190–198 (2011)
4. Candès, E.J., Plan, Y.: Matrix completion with noise. Proceedings of the IEEE 98(6), 925–936 (2010)
5. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9(6), 717–772 (2009)
6. Chen, M., Zheng, A., Weinberger, K.Q.: Fast image tagging. In: ICML (2013)
7. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: CIVR (2009)

8. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML, pp. 209–216 (2007)
9. Feng, Z., Jin, R., Jain, A.K.: Large-scale image annotation by efficient and robust kernel metric learning. In: ICCV (2013)
10. Ganesh, A., Wright, J., Li, X., Candès, E.J., Ma, Y.: Dense error correction for low-rank matrices via principal component pursuit. In: ISIT (2010)
11. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image annotation. In: ICCV (2009)
12. Ji, S., Ye, J.: An accelerated gradient method for trace norm minimization. In: ICML, pp. 457–464. ACM (2009)
13. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics (2011)
14. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: ACM Conference on Recommender Systems, pp. 61–68. ACM (2009)
15. Lin, Z., Ding, G., Hu, M., Wang, J., Ye, X.: Image tag completion via image-specific and tag-specific linear sparse reconstructions. In: CVPR (2013)
16. Liu, D., Hua, X.S., Wang, M., Zhang, H.J.: Image retagging. In: Proceedings of the International Conference on Multimedia, pp. 491–500. ACM (2010)
17. Liu, D., Yan, S., Hua, X.S., Zhang, H.J.: Image retagging using collaborative tag propagation. IEEE Transactions on Multimedia 13(4), 702–712 (2011)
18. Liu, X., Yan, S., Chua, T.S., Jin, H.: Image label completion by pursuing contextual decomposability. TOMCCAP, 21:1–21:20 (2012)
19. Loeff, N., Farhadi, A.: Scene discovery by matrix factorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 451–464. Springer, Heidelberg (2008)
20. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: ACM International Conference on Multimedia, pp. 275–278. ACM (2003)
21. Mu, Y., Dong, J., Yuan, X., Yan, S.: Accelerated low-rank visual recovery by random projection. In: CVPR, pp. 2609–2616. IEEE Computer Society (2011)
22. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. JMLR 13, 1665–1697 (2012)
23. Tian, Q., Aggarwal, C., Qi, G.J., Ji, H., Huang, T.S.: Exploring context and content links in social media: A latent space method. PAMI 34(5), 850–862 (2012)
24. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. PAMI 35(3) (2013)
25. Xu, H., Wang, J., Hua, X.S., Li, S.: Tag refinement by regularized lda. In: ACM International Conference on Multimedia, pp. 573–576. ACM (2009)
26. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: International Conference on Multimedia. ACM (2010)
27. Zhuang, J., Hoi, S.C.H.: A two-view learning approach for image tag ranking. In: WSDM, pp. 625–634 (2011)