Image Tag Refinement Towards
Low-Rank, Content-Tag Prior and Error Sparsity

Guangyu Zhu [1], Shuicheng Yan [1], Yi Ma [2,3]

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Department of Computer Engineering, University of Illinois at Urbana-Champaign, USA
[3] Microsoft Research Asia, China
{elezhug, eleyans}@nus.edu.sg, {yima}@uiuc.edu

ABSTRACT

The vast user-provided image tags on the popular photo sharing websites may greatly facilitate image retrieval and management. However, these tags are often imprecise and/or incomplete, resulting in unsatisfactory performances in tag related applications. In this work, the tag refinement problem is formulated as a decomposition of the user-provided tag matrix D into a low-rank refined matrix A and a sparse error matrix E, namely $D = A + E$, targeting the optimality measured by four aspects: 1) *low-rank*: A is of low-rank owing to the semantic correlations among the tags; 2) *content consistency*: if two images are visually similar, their tag vectors (*i.e.*, column vectors of A) should also be similar; 3) *tag correlation*: if two tags co-occur with high frequency in general images, their co-occurrence frequency (described by two row vectors of A) should also be high; and 4) *error sparsity*: the matrix E is sparse since the tag matrix D is sparse and also humans can provide reasonably accurate tags. All these components finally constitute a constrained yet convex optimization problem, and an efficient convergence provable iterative procedure is proposed for the optimization based on accelerated proximal gradient method. Extensive experiments on two benchmark Flickr datasets, with 25K and 270K images respectively, well demonstrate the effectiveness of the proposed tag refinement approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithm, Theory, Experimentation

Keywords

Tag Refinement, Social Images, Content Consistency, Tag Correlation, Low-rank, Error Sparsity

1. INTRODUCTION

With the rapid advance in the technology of digital imaging, there is an explosive growth in the amount of available image data in our daily lives. This trend urgently necessitates the development of effective retrieval technology for large volume of images [1].

Considering the elementary features undertaken in the approaches, a distinction can be generally made between content based image retrieval (CBIR) [2] and text based image retrieval (TBIR) [1]. The features adopted by CBIR are extracted from visual information, *e.g.*, image color, texture and shape of the objects involved in the images. Although CBIR has been extensively studied for more than a decade, there exist three limitations which restrict its practicability [3]. Firstly, the precision of CBIR is usually unsatisfactory because of the semantic gap between low-level visual features and high-level semantic concepts. Secondly, the efficiency of CBIR is usually low due to the high dimensionality of visual features. Thirdly, the query form of CBIR is unnatural for image search owing to the possible absence of appropriate example images. In contrast, TBIR solely adopts the text information to carry through the image indexing and search. Compared with the visual information, text is essentially a kind of representation for image content from the view of human-being concepts and provided with the characteristics in terms of low dimension and easy description. TBIR is a straightforward solution to conquer the disadvantages of CBIR.

The text information used in TBIR can be acquired from image title [4], surrounding text [4, 5] and user tag [3, 6]. Thereinto, user tags are more consistent with semantic concepts and effective to describe image contents. Especially with the prevalence of photo sharing websites such as Flickr and Piscasa, which host vast of digital images with user-provided tags, tag based image retrieval has become potentially popular and practical in extensive applications. Nevertheless, the performance of tag based image retrieval is still far from satisfactory suffering from the inferior quality of image tags. Figure 1 illustrates a representative image for flying eagle downloaded from Flickr and its associated tags. We can observe that only "fly" and "bird" truly describe the visual content of the image, while other tags are imprecise. Meanwhile, some additional tags are missing, such as "sky" and "eagle" that are highly associated with the given image. The reason causing this phenomena is that the current image tagging on the photo sharing websites solely relies on the user inputs, which often prohibits accurate and comprehensive textual description of image visual content. Refining
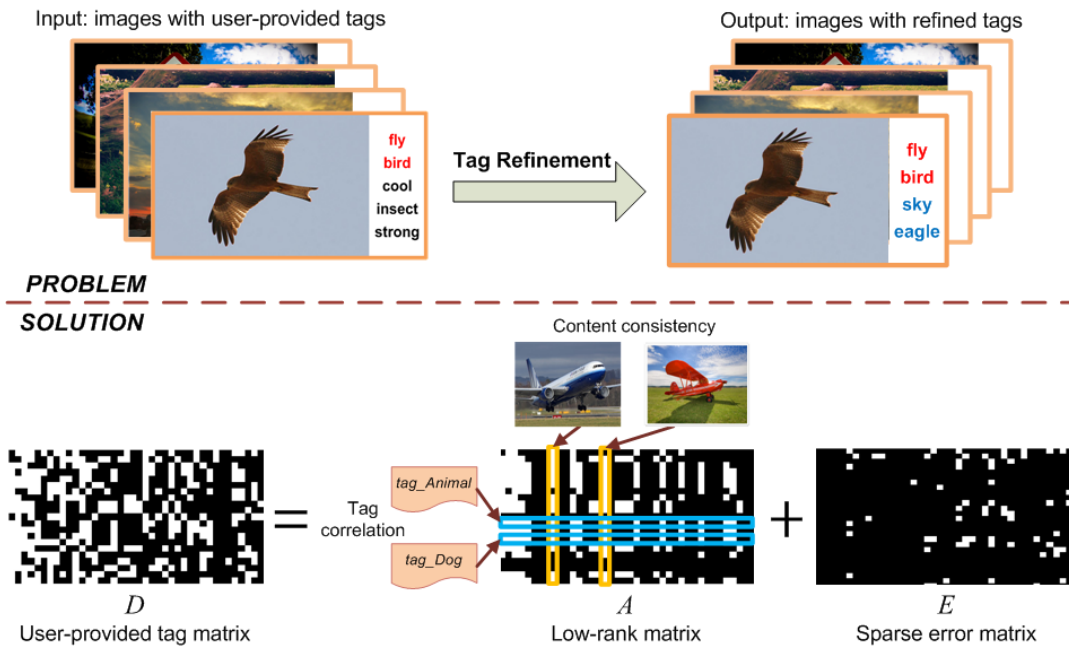
Figure 1: Framework of image tag refinement towards low-rank, content consistency, tag correlation and error sparsity. The column-wise user-provided tag matrix D (Note that D is sub-sampled from a larger real user-provided tag matrix for ease of display), where white grid represents the association of a tag with image and black one represents non-association, is decomposed into a low-rank matrix A (the refined tag matrix and here $\mathrm{rank}(A) = 13$) and a sparse matrix E (tagging error in user-provided tags and sparse error is $\|E\|_0 = 72$ in this illustration) by considering the properties of content consistency and tag correlation.

tags is thus highly desirable for tag based image retrieval and other related applications.

In this paper, to address the aforementioned imprecise and incomplete issues of user-provided image tags, we propose a novel refinement approach aiming to improve the quality of tags. The approach is motivated by the following four observations of image tags from large volume social images.

- **Low-rank**. The existing work on text information processing [7] has demonstrated that the semantic space spanned by text keywords can be approximated by a smaller subset of salient words derived from the original space. As one kind of text information, image tags are consequently subject to such low-rank property.

- **Content consistency**. From large-scale image dataset, we can observe that visually similar images often reflect similar themes and thus are typically annotated with similar tags. Content consistency describes the relationships between content level and semantic level. Being an important prior, this observation has been widely explored in visual category learning [8, 9].

- **Tag correlation**. Semantic tags associated with images do not appear in isolation, instead often appear correlatively and naturally interact with each other at the semantic level. As another important prior, tag correlation characterizes the relationships within semantic level and is often the preliminary assumption of multi-label and contextual learning algorithms [10, 11].

- **Error sparsity**. With the general knowledge that the human-beings share most of the common concepts in the semantic space, the tagging results for one image

are reasonably accurate to certain level. Moreover, one image usually is labeled with only couple of tags. Such observations lead to the characteristics of error sparsity for image tag matrix.

Figure 1 shows the framework of our problem formulation and solution. Given the user-provided image tag matrix D, to comprehensively characterize the above four factors, we cast the tag refinement task into a convex optimization problem, which simultaneously minimizes the matrix rank and priors as well as error sparsity. Concretely, the nuclear norm, ℓ_1 norm and trace operation are employed to model the properties regarding tag low-rank, error sparsity, content consistency and tag correlation, respectively. The results are the low-rank matrix A which encodes the refined image tags, and the sparse matrix E which represents the tagging errors in user-provided tags. To obtain the results effectively, we also propose an efficient convergence provable iterative procedure to accomplish the optimization.

The novelties and main contributions of this paper are summarized as follows.

- We propose a new tag refinement formulation in form of convex optimization which comprehensively considers the tag characteristics from the points of view of low-rank, error sparsity, content consistency and tag correlation.

- Compared with existing works, the low-rank and error sparsity are firstly integrated into the optimization procedure for image tag refinement. With the assistance of constraints of content consistency and tag correlation, the proposed approach is capable of correcting imprecise tags and enriching the incomplete ones.

- We propose to use an accelerated proximal gradient method to speedup the optimization, which facilitates the proposed approach to be workable on large-scale image datasets.

The rest of the paper is organized as follows. Section 2 reviews the related work on image tag refinement. In Section 3, we introduce the formulation details of our proposed refinement approach. Section 4 describes an efficient iterative procedure for the solution to tag refinement. Experimental results on two Flickr image datasets are reported and analyzed in Section 5. Finally, we conclude the paper with future work discussion in Section 6.

2. RELATED WORK

Many efforts have been devoted to the research on image tag refinement. As a pioneer work, Jin *et al.* [12] utilized WordNet to estimate the semantic correlation among the annotated keywords and remove weakly correlated ones. This method, however, can only achieve limited success as it totally ignores the visual content of the images. In [3, 13], Wang *et al.* proposed a content based approach within the random walk with restart (RWR) framework for image annotation refinement where the visual similarity and word co-occurrence conditioned on visual similarities are both considered. In addition, a similar work was proposed by Jia *et al.* in [14], where the textual similarities of tags and visual similarities of images are fused in a multi-graph reinforcement framework. Liu *et al.* [15] proposed to rank the image tags according to their relevance with respect to the associated images by tag similarity and image similarity in a random walk model. In [16], Xu *et al.* proposed to do tag refinement from topic modeling point of view. A new graphical model named as regularized latent Dirichlet allocation (rLDA) is presented to jointly model the tag similarity and tag relevance. These works are typical based on reranking-and-removing strategy, which focuses on selecting a coherent subset of keywords from the automatically annotated keywords. On the other side, the tags associated with the social images are often imprecise and incomplete. These works are thus not applicable in directly addressing the problems with inferior image tags, which is compelling to further extension on proposed algorithmic framework.

The most related work to our image tag refinement scheme is the improved version of [15] proposed by Liu *et al.* in [17] recently. In this work, the authors formulated image tag refinement as an optimization framework based on the consistency between visual similarity and semantic similarity in social images. An iterative bound optimization algorithm was applied to discover the improved tag assignment. Despite this work has shown encouraging results, it is not scalable to large-scale applications due to its high computation-cost.

3. TAG REFINEMENT TOWARDS LOW-RANK, CONTENT-TAG PRIOR AND ERROR SPARSITY

Denote the image collection $\mathcal{I} = \{x_1, x_2, \ldots, x_n\}$, where n is the size of the image set. All initial tags appearing in the collection form a tag set $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, where m denotes the total number of unique tags. The initial tag membership for the whole image collection can be represented by a binary matrix $D \in \{0, 1\}^{m \times n}$ whose element D_{ij} indicates the presence of tag t_i in image x_j, *i.e.*, $D_{ij} = 1$ if t_i is associated with image x_j, otherwise $D_{ij} = 0$. To represent the final refined results, we define the matrix A whose element $A_{ij} \geq 0$ denotes the confidence score of assigning tag t_i to image x_j, given by tag refinement approach.

For a given tag matrix D, the essential purpose of image tag refinement is to uncover the tag error matrix E, such that,

$$D = A + E, \qquad (1)$$

where matrix A is the ultimate refined tag matrix. As aforementioned as well as introduced later, this decomposition of the matrix D into A and E shall target four properties, which correspond to four items for the objective to optimize. More specifically, we use $T_l(A)$ to characterize the rank of the matrix A, $T_c(A)$ and $T_t(A)$ to measure the image content consistency and tag correlation of the refined tag matrix A, respectively, and $T_e(E)$ to measure the sparsity of the tagging error matrix E. Then the image tag refinement problem is generally formulated as follows:

$$\min_{A,E} \quad T_l(A) + \lambda_1 T_e(E) + \lambda_2 [T_c(A) + T_t(A)],$$
$$\text{subject to} \quad D = A + E . \qquad (2)$$

In the following subsections, we elaborate on how to define these four items.

3.1 Low Rank and Error Sparsity

As we have discussed, the image tags are subject to low-rank property, and the semantically correlated tags in semantic space usually appear in images synchronously. Meanwhile, the tag error matrix is sparse for two reasons: 1) the number of annotated tags is essentially sparse compared with the relatively large tag number, and 2) humans generally share the similar semantic concepts and may provide reasonably accurate tags.

Motivated by the latest research on robust principal component analysis (RPCA) [18], we can formulate this problem as a matrix decomposition problem. In [18], it has been shown that a low-rank matrix A from $D = A + E$ with gross but sparse errors E can be derived by solving the following optimization problem

$$\text{minimize} \quad \|A\|_* + \lambda \|E\|_1 ,$$
$$\text{subject to} \quad D = A + E . \qquad (3)$$

Here, $\| \cdot \|_*$ represents the nuclear norm of a matrix (the sum of its singular values), $\| \cdot \|_1$ is the ℓ_1 norm denoting the sum of the absolute values of matrix entries, and λ is a positive weighting parameter. In our formulation, matrix D is the initial user-provided tags, A and E represent and characterize the low-rank refined tag matrix and sparse error matrix, respectively. Because it is hard to identify the model and control gross amount of the labeling error in the user-provided tags, this optimization strategy is suitable for our problem. Therefore, in our approach, the low-rank and error sparsity are modeled by nuclear norm ($T_l(A) = \|A\|_*$) and ℓ_1 norm ($T_e(E) = \|E\|_1$) according to the definition of Eq. (3).

3.2 Content Consistency for Tag Refinement

Based on the assumption of content consistency that visually similar images often reflect similar themes and thus are typically annotated with similar tags, the content consistency can be formulated as follows.

Given two images x_i and x_j associated with tag vectors a_i and a_j, where a_i and a_j are the i-th and j-th column vectors of the refined tag matrix A, we calculate the visual similarities between images. Let $G = \{\mathcal{I}, W^c\}$ be an undirected weighted graph with vertex set $\mathcal{I}$ and similarity matrix $W^c \in \mathrm{IR}^{n \times n}$. Each element of the symmetric matrix W^c measures, for a pair of vertices, its similarity. In our formulation, the similarity matrix W^c is defined based on k-nearest-neighbor (k-NN) graph

$$w_{ij}^c = \begin{cases} \exp(-\|x_i - x_j\|^2/\sigma_c^2) & \text{if } j \in N_{k_c}(i) \text{ or } i \in N_{k_c}(j) , \\ 0 & \text{otherwise} , \end{cases} \tag{4}$$

where $N_{k_c}(\cdot)$ denotes the index set for the k_c nearest neighbors of an image measured by Euclidean distance. In our approach, we empirically set $k_c = 0.001n$ and σ_c as the median value of the entries in W^c. To solve this problem, we define the diagonal matrix P^c as

$$P_{ii}^c = \sum_{j \neq i} w_{ij}^c, \quad \forall\, i . \tag{5}$$

Then, the content consistency among the images can be enforced by solving the following optimization

$$\min_A \sum_{i=1}^n \sum_{j=1}^n \|a_i - a_j\|^2 w_{ij}^c . \tag{6}$$

By following the graph embedding terminologies in [19], this formulation can be regarded as a kind of dimensionality reduction where A encodes the low-dimensional representations of the image set $\mathcal{I}$.

The term $T_c(A)$ in Eq. (2) can then be rewritten as follows based on Eq. (5) and (6)

$$T_c(A) = \mathrm{Tr}[A(P^c - W^c)A^T] , \tag{7}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operation on a matrix.

3.3 Tag Correlation for Tag Refinement

Similar to the formulation for image content consistency, we can construct the regularization item regarding the correlation among image tags. The difference from content consistency is that the k-NN graph for tag correlation is constructed by tag similarity. In this work, we adopt a concurrence based method to estimate the tag similarity, which is analogous to the principle of Google distance [20]. We first estimate the semantic distance $d(t_i, t_j)$ between tags t_i and t_j as

$$d(t_i, t_j) = \frac{\max[\log q(t_i), \log q(t_j)] - \log q(t_i, t_j)}{\log R - \min[\log q(t_i), \log q(t_j)]} , \tag{8}$$

where $q(t_i)$ and $q(t_j)$ are the numbers of images containing tag t_i and tag t_j respectively and $q(t_i, t_j)$ is the number of images containing both tags t_i and t_j. Such numbers can be obtained by performing search "tags only" on Google image website using the tags as queries. In addition, R is the total number of the images in Google image. Then, the semantic similarity between t_i and t_j in matrix W^t is defined as

$$w_{ij}^t = \exp[-d(t_i, t_j)^2/\sigma_t^2] , \tag{9}$$

where σ_t is empirically set as the median value of the entries in $W^t = [w_{ij}^t]$. Similar to the case of content consistency, we calculate the diagonal matrix P^t. The regularization term

of tag correlation for refinement is thus represented as

$$T_t(A) = \mathrm{Tr}[A^T(P^t - W^t)A] . \tag{10}$$

4. OPTIMIZATION VIA ACCELERATED PROXIMAL GRADIENT

Based on the definitions of the terms regarding low-rank, error sparsity, content consistency and tag correlation, we rewrite the objective function in Eq. (2) as follows.

$$\min_{A,E} \quad \|A\|_* + \lambda_1 \|E\|_1 +$$
$$\lambda_2 \{\mathrm{Tr}[A(P^c - W^c)A^T] + \mathrm{Tr}[A^T(P^t - W^t)A]\} ,$$
$$\text{subject to} \quad D = A + E . \tag{11}$$

Note that the regularization terms of content consistency and tag correlation share the same weight λ_2 stemming from the consideration of tradeoff between optimization efficiency and performance accuracy. To solve this problem, we can intuitively employ the traditional Lagrange multiplier method. Unfortunately, this method will be generally computation-intensive. Considering Eq. (11) is convex and often computationally expedient to relax the equality constraint, we here pursuit an effective iterative procedure to solve this optimization based on the accelerated proximal gradient (APG) method with $O(k^{-2})$ convergence rate.

4.1 General Accelerated Proximal Gradient

To better introduce the solution to our problem, in this subsection, we firstly present a brief introduction on accelerated proximal gradient method and then give two basic propositions for the solution.

Given the following unconstrained convex problem

$$\min_{X \in \mathcal{H}} F(X) \doteq \mu g(X) + f(X) , \tag{12}$$

where $\mathcal{H}$ is a real Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$ and a corresponding norm $\| \cdot \|$, $\mu > 0$ is a relaxation parameter. Both $g(X)$ and $f(X)$ are convex and $f(X)$ is further Lipschitz continuous $\|\nabla f(X_1) - \nabla f(X_2)\| \leq L_f \|X_1 - X_2\|$, where L_f is the Lipschitz constant. Instead of directly minimizing $F(X)$, proximal gradient algorithms minimize a sequence of separable quadratic approximations to $F(X)$, denoted as $Q(X, Y)$, formed at specially chosen points Y

$$Q(X, Y) \doteq f(Y) + \langle \nabla f(Y), X - Y \rangle + \frac{L_f}{2}\|X - Y\|^2 + \mu g(X) . \tag{13}$$

Let $G = Y - \frac{1}{L_f}\nabla f(Y)$, then

$$X = \arg\min_X Q(X, Y)$$
$$= \arg\min_X \left\{ \mu g(X) + \frac{L_f}{2}\|X - G\|^2 \right\} . \tag{14}$$

To solve Eq. (12), one may repeatedly set $X_{k+1} = \arg\min_X Q(X, Y_k)$ with Y_k chosen based on $X_0, \ldots, X_k$. The convergence of this iteration depends strongly on the points Y_k at which the approximations $Q(X, Y_k)$ are formed. The natural choice $Y_k = X_k$ can be interpreted as a gradient algorithm and results in a convergence rate no worse than $O(k^{-1})$ [21]. However, the work in [22] has showed that instead setting $Y_k = X_k + \frac{b_{k-1}-1}{b_k}(X_k - X_{k-1})$ for a sequence $\{b_k\}$ satisfying $b_{k+1}^2 - b_{k+1} \leq b_k^2$ can improve the convergence rate to

$O(k^{-2})$. The general proximal gradient method is described in Algorithm 1.

Algorithm 1: General Proximal Gradient Method

while *not converged* **do**

$\quad Y_k \leftarrow X_k + \frac{b_{k-1}-1}{b_k}(X_k - X_{k-1})$.

$\quad G_k \leftarrow Y_k - \frac{1}{L_f}\nabla f(Y_k)$.

$\quad X_{k+1} \leftarrow \arg\min_X \left\{\mu g(X) + \frac{L_f}{2}\|X - G_k\|^2\right\}$.

$\quad b_{k+1} \leftarrow \frac{1+\sqrt{4b_k^2+1}}{2}, k \leftarrow k+1$.

end

The main motivation for forming the separable quadratic approximation in Algorithm 1 is that in many cases of interest, the minimizer X_{k+1} has a simple or even closed-form expression. Before presenting the details, we first introduce a soft-thresholding operators. For $x \in \mathbb{R}$ and $\varepsilon > 0$, the soft-thresholding operation is defined as

$$\mathcal{S}_\varepsilon[x] \doteq \begin{cases} x - \varepsilon & \text{if } x > \varepsilon, \\ x + \varepsilon & \text{if } x < -\varepsilon, \\ 0 & \text{otherwise}. \end{cases} \quad (15)$$

By extending this operator to vector and matrix, we then have [23]

PROPOSITION 1. *If $\mathcal{H}$ is an Euclidean space endowed with the Frobenius norm $\|\cdot\|_F$ and $g(\cdot)$ is ℓ_1 norm, then X_{k+1} is given by soft-thresholding the entries of G_k as*

$$X_{k+1} = \mathcal{S}_\varepsilon[G_k] = \arg\min_X \varepsilon\|X\|_1 + \frac{1}{2}\|X - G_k\|_F^2. \quad (16)$$

PROPOSITION 2. *if $\mathcal{H}$ is an Euclidean space endowed with the Frobenius norm and $g(\cdot)$ is the matrix nuclear norm, then X_{k+1} is given by soft-thresholding the singular values as*

$$X_{k+1} = U\mathcal{S}_\varepsilon(\Sigma)V^T = \arg\min_X \varepsilon\|X\|_* + \frac{1}{2}\|X - G_k\|_F^2, \quad (17)$$

where $U\Sigma V^T$ is the singular value decomposition (SVD) of G_k.

Based on Eq. (11) and (12), we can obtain the objective function $F(X)$ for our relaxed problem as

$$\begin{aligned} F(X) \doteq{}& \mu\|A\|_* + \mu\lambda_1\|E\|_1 \\ &+ \frac{\mu\lambda_2}{2}\{\text{Tr}[A(P^c - W^c)A^T] + \text{Tr}[A^T(P^t - W^t)A]\} \\ &+ \frac{1}{2}\|D - A - E\|_F^2, \end{aligned} \quad (18)$$

then $g(X)$ and $f(X)$ are respectively defined as

$$\begin{aligned} g(X) ={}& \mu\|A\|_* + \mu\lambda_1\|E\|_1, \\ f(X) ={}& \frac{\mu\lambda_2}{2}\{\text{Tr}[A(P^c - W^c)A^T] + \text{Tr}[A(P^t - W^t)A^T]\} \\ &+ \frac{1}{2}\|D - A - E\|_F^2, \end{aligned} \quad (19)$$

where $X = \begin{pmatrix} A \\ E \end{pmatrix}$ and the term $\frac{1}{2}\|D - A - E\|_F^2$ penalizes violations for the equality constraint of Eq. (11). In the next subsections, we will show that $f(X)$ in Eq. (19) satisfies Lipschitz continuous condition and present the iterative solution of A and E based on the definitions in Eq. (19).

4.2 Proof of Lipschitz Continuity

Observing terms in Eq. (11), we can have that $\|A\|_*$ and $\|E\|_1$ are convex [22]. Moreover, because $P^c - W^c$ and $P^t - W^t$ in the regularization terms regarding content consistency and tag correlation are the Laplacian matrices which are positive semidefinite, the terms $\text{Tr}[A(P^c - W^c)A^T]$ and $\text{Tr}[A^T(P^t - W^t)A]$ are therefore convex with respect to A. Consequently, the linear combination of convex terms, namely the objective function for our problem, is also convex.

Based on the definition in Eq. (19) and let $R^c = P^c - W^c$ and $R^t = P^t - W^t$, we can calculate $\nabla f(X)$ as

$$\begin{aligned} \nabla f(A) &= \mu\lambda_2(AR^c + R^t A) + A + E - D, \\ \nabla f(E) &= A + E - D. \end{aligned} \quad (20)$$

Given $X_1 = (A_1\ E_1)'$ and $X_2 = (A_2\ E_2)'$, we have

$$\begin{aligned} &\|\nabla f(X_1) - \nabla f(X_2)\|_F \\ &= \left\| \begin{array}{c} \mu\lambda_2(\triangle AR^c + R^t\triangle A) + \triangle A + \triangle E \\ \triangle A + \triangle E \end{array} \right\|_F, \end{aligned} \quad (21)$$

where $\triangle A = A_1 - A_2$ and $\triangle E = E_1 - E_2$. Then we have

$$\begin{aligned} &\|\nabla f(X_1) - \nabla f(X_2)\|_F^2 \\ &\leq [4\sigma_{\max}^2(\mu\lambda_2 R^c) + 4\sigma_{\max}^2(\mu\lambda_2 R^t) + 6]\|\triangle A\|_F^2 + 6\|\triangle E\|_F^2 \\ &\leq L_f^2 \left\| \begin{array}{c} A_1 - A_2 \\ E_1 - E_2 \end{array} \right\|_F^2, \end{aligned} \quad (22)$$

where $\sigma_{\max}(\cdot)$ represents the maximum singular value of a matrix. Therefore, the Lipschitz constant is

$$L_f = \sqrt{4\sigma_{\max}^2(\mu\lambda_2 R^c) + 4\sigma_{\max}^2(\mu\lambda_2 R^t) + 6}. \quad (23)$$

As the objective function is convex, we are able to use iterative optimization similar to Algorithm 1 to achieve the globally optimal solution [22].

4.3 Optimizing Low-Rank Refined Matrix

From Eq. (13), (14) and (20), we can infer the optimal solution of X as following

$$\begin{aligned} X_{k+1} &= \arg\min_X Q(X, Y_k) \\ &= \arg\min_X f(Y_k) + \langle\nabla f(Y_k), X - Y_k\rangle \\ &\quad + \frac{L_f}{2}\|X - Y_k\|_F^2 + \mu g(X), \end{aligned} \quad (24)$$

where $X = \begin{pmatrix} A \\ E \end{pmatrix}$ and $Y = \begin{pmatrix} Y^A \\ Y^E \end{pmatrix}$. When solving A_{k+1}, E is assumed to be constant and set as E_k. The solution of A_{k+1} is thus simplified as

$$\begin{aligned} A_{k+1} &= \arg\min_A f(Y_k) + \frac{L_f}{2}\|A - Y_k^A\|_F^2 + \mu g(A) \\ &\quad + \langle\mu\lambda_2(Y_k^A R^c + R^t Y_k^A) + Y_k^A + Y_k^E - D, A - Y_k^A\rangle. \end{aligned} \quad (25)$$

Let $O_k^A = \mu\lambda_2(Y_k^A R^c + R^t Y_k^A) + Y_k^A + Y_k^E - D$, we have

$$\begin{aligned} A_{k+1} &= \arg\min_A \frac{L_f}{2}\|A - Y_k^A + \frac{1}{L_f}O_k^A\|_F^2 + \mu\|A\|_* \\ &\quad + \mu\lambda_1\|E_k\|_1 + f(Y_k) - \frac{1}{2L_f}\|O_k^A\|_F^2. \end{aligned} \quad (26)$$

In Eq. (26), the term $\mu\lambda_1\|E_k\|_1 + f(Y_k) - \frac{1}{2L_f}\|O_k^A\|_F^2$ is

constant for the optimization of A. Therefore,

$$A_{k+1} = \arg\min_A \frac{\mu}{L_f}\|A\|_* + \frac{1}{2}\|A - (Y_k^A - \frac{1}{L_f}O_k^A)\|_F^2 \ , \tag{27}$$

and $G_k^A = Y_k^A - \frac{1}{L_f}O_k^A$. According to Proposition 2 (Eq. (17)), this optimization can be solved by singular value thresholding algorithm as

$$A_{k+1} = U\mathcal{S}_{\frac{\mu}{L_f}}[S]V^T \ , \tag{28}$$

where USV^T is the SVD of G_k^A.

4.4 Optimizing Sparse Error Matrix

Similar to the deduction procedure of A, the optimal solution of E with fixed A set as A_{k+1} is inferred as

$$\begin{aligned}
E_{k+1} &= \arg\min_E Q(E, Y_k) \\
&= \arg\min_E f(Y_k) + \langle Y_k^A + Y_k^E - D, E - Y_k^E\rangle \quad (29) \\
&\quad + \frac{L_f}{2}\|E - Y_k^E\|_F^2 + \mu g(E) \ .
\end{aligned}$$

Let $O_k^E = Y_k^A + Y_k^E - D$, we have

$$\begin{aligned}
E_{k+1} = \arg\min_E &\frac{L_f}{2}\|E - Y_k^E + \frac{1}{L_f}O_k^E\|_F^2 + \mu\|A_{k+1}\|_* \\
&+ \mu\lambda_1\|E\|_1 + f(Y_k) - \frac{1}{2L_f}\|O_k^E\|_F^2 \ .
\end{aligned} \tag{30}$$

In Eq. (30), the term $\mu\|A_{k+1}\|_* + f(Y_k) - \frac{1}{2L_f}\|O_k^E\|_F^2$ is constant for the optimization of E. Therefore,

$$E_{k+1} = \arg\min_E \frac{\mu\lambda_1}{L_f}\|E\|_1 + \frac{1}{2}\|E - (Y_k^E - \frac{1}{L_f}O_k^E)\|_F^2 \ . \tag{31}$$

Let $G_k^E = Y_k^E - \frac{1}{L_f}O_k^E$, then the solution E_{k+1} can be obtained according to Proposition 1 (Eq. (16))

$$E_{k+1} = \mathcal{S}_{\frac{\mu\lambda_1}{L_f}}[G_k^E] \ . \tag{32}$$

As we empirically set an upper bound for μ in the optimization, the formulation (and solution) is robust to grossly corrupted data, *e.g.*, Gaussian error.

4.5 Implementation Issues

Based on the aforementioned analysis and deduction, we summarize the procedure for tag refinement via accelerated proximal gradient method in Algorithm 2. To speedup the convergence rate, we vary μ in the procedure starting from a large initial value μ_0 and decreasing it geometrically with each iteration until it reaches the floor $\bar{\mu}$, rather than applying the proximal gradient algorithm directly to Eq. (18). We observe that this greatly reduces the number of iterations and therefore, the number of SVD computations. The general proof of convergence of Algorithm 2 can follow the general case in [22]. A "grid-search" strategy [26] can be employed to set λ_1 and λ_2. We set $\lambda_1 \in \{2^{-6}, 2^{-4}, \ldots, 2^4\}$ and $\lambda_2 \in \{2^{-10}, 2^{-8}, \ldots, 2^0\}$. Various pairs of (λ_1, λ_2) values were tried and the one with the best performance was picked. In the experiments, μ_0 was empirically initialized as 1.0 and $\delta = 10^{-9}$. The decreasing rate η was set as 0.9.

As one tag unassociated with an image in the user-provided tag matrix does not simply imply that it is definitely irrelevant while perhaps because of missing tags. Therefore, it is natural to re-initialize D by a simple tag propagation step before tag refinement. The user-provided tags for each image are propagated based on its k nearest neighbors measured by visual features. Given image x_i and its tag vector p_i where element $p_{ij} = 1$ means t_j is associated with x_i whereas $p_{ij} = 0$ if t_j is not associated, its k-NN set of images is denoted as $M_i = \{(x_i^l, p_i^l), l = 1, \ldots, k\}$ where p_i^l is the tag vector of image x_i^l. The propagated tag vector v_i for image x_i is defined as

$$v_i = \sum_{l=0}^k \frac{\exp(-\|x_i - x_i^l\|^2/\sigma^2)}{c}p_i^l \ , \tag{33}$$

where $c = \sum_{l=0}^k \exp(-\|x_i - x_i^l\|^2/\sigma^2)$ is a normalization constant. The parameter σ is set as the same as σ_c in Eq. (4) and k is empirically set as $0.1\overline{n}$ where $\overline{n}$ is the average number of images for each tag. Note that $x_i^0 = x_i$ is the image x_i itself.

The low-rank matrix A is the final refinement result, where each entry produces the confidence score for a tag associated with an image. To identify the ultimate image tags, we rank the tags of each image based on their confidence scores and then retain the top m (*e.g.*, $m = 5$) ones to be the refined tags associated with an given image.

Algorithm 2: Tag Refinement via Accelerated Proximal Gradient Method

Input: User-provided tag matrix $D \in \mathbb{R}^{m \times n}$, matrix $R^c \in \mathbb{R}^{n \times n}$ and $R^t \in \mathbb{R}^{m \times m}$ for content consistency and tag correlation, and weighting parameters λ_1, λ_2.

$A_0, A_{-1} \leftarrow 0$; $E_0, E_{-1} \leftarrow 0$; $b_0, b_{-1} \leftarrow 1$; $\bar{\mu} \leftarrow \delta\mu_0$.

while *not converged* **do**

$\quad Y_k^A \leftarrow A_k + \frac{b_{k-1}-1}{b_k}(A_k - A_{k-1})$,

$\quad Y_k^E \leftarrow E_k + \frac{b_{k-1}-1}{b_k}(E_k - E_{k-1})$.

$\quad G_k^A \leftarrow Y_k^A - \frac{1}{L_f}[\mu\lambda_2(Y_k^A R^c + R^t Y_k^A) + Y_k^A + Y_k^E - D]$,

$\quad (U, S, V) \leftarrow \text{svd}(G_k^A), A_{k+1} = U\mathcal{S}_{\frac{\mu_k}{L_f}}[S]V^T$.

$\quad G_k^E \leftarrow Y_k^E - \frac{1}{L_f}(Y_k^A + Y_k^E - D), E_{k+1} = \mathcal{S}_{\frac{\mu_k\lambda_1}{L_f}}[G_k^E]$.

$\quad b_{k+1} \leftarrow \frac{1+\sqrt{4b_k^2+1}}{2}, \mu_{k+1} \leftarrow \max(\eta\mu_k, \bar{\mu})$.

$\quad k \leftarrow k + 1$.

end

Output: $A \leftarrow A_k$, $E \leftarrow E_k$.

---

5. EXPERIMENTS

To systematically demonstrate the effectiveness of our proposed tag refinement algorithm, we performed thorough experiments on two large volume image datasets: MIRFlickr-25K [6] and NUS-WIDE-270K [24]. The MIRFlickr-25K and NUS-WIDE-270K datasets are both collected from Flickr website. The MIRFlickr-25K dataset contains $25,000$ images with $1,386$ tags. The second dataset, NUS-WIDE-270K, comprises a total of $269,648$ images with $5,018$ unique tags. Note that the tags in the above two collections are rather noisy and many of them are misspelling or meaningless words. Hence, a pre-processing was performed to filter out these tags. We matched each tag with entries in a Wikipedia thesaurus and only the tags with coordinates in Wikipedia were retained. Moreover, to avoid sample insufficiency issue in optimization, we further removed those tags

Table 1: Average performances of different algorithms for tag refinement on MIRFlickr-25K

	UT	RWR	TRVSC	LR_ES	LR_ES_CC	LR_ES_TC	LR_ES_CC_TC
F-score	0.221	0.338	0.412	0.423	0.465	0.463	0.477

whose occurrence numbers are below 50. Consequently, 205 and 521 unique tags were obtained in total for MIRFlickr-25K and NUS-WIDE-270K, respectively.

To calculate the visual similarities between images, each image was extracted a 428-dimensional feature vector as the content representation, including 225-dimensional block-wise color moment features generated from 5-by-5 fixed partition on image, 128-dimensional wavelet texture features and 75-dimensional edge distribution histogram features [15, 17]. To evaluate the performance of tag refinement, we evaluated the performance on 18 tags in MIRFlickr-25K and 81 tags in NUS-WIDE-270K where the ground-truth annotations of these tags have been provided. The F-score, which was widely used as evaluation metric of tag refinement [13, 15, 16, 17], was calculated to measure the refinement results for each tag and average them as the final evaluation.

In the experiments, the datasets of MIRFlickr-25K and NUS-WIDE-270K were both employed to evaluate the performance of tag refinement. In addition, MIRFlickr-25K was also utilized for the systematic evaluation of refinement performance against the noise rate, namely to study at least what percentage of tags are required to be annotated such that the tags are reasonably refinable, which is an important index for tag/annotation refinement, yet to our best knowledge, none of previous research effort has ever been devoted to this study.

5.1 Evaluation of Tag Refinement on MIRFlickr-25K

In this subsection, we demonstrate the convergence property of the optimization process. Two types of results are reported in terms of the value of objective function Eq. (18) and the refinement performance (F-score) against iteration times. Figure 2 shows the detailed results. From Figure 2(a), we can observe that the objective function converges after about 30 iterations. For this experiment, the relaxation parameter μ was fixed as 1.0 to avoid the affection from varying μ in Algorithm 2. Figure 2(b) shows the refinement performance against iteration times. It is shown that the F-score increases steadily as the iteration proceeds and finally reaches a satisfactory result.
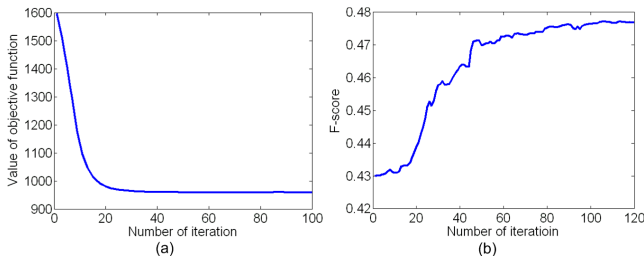

Figure 2: Convergence properties of the optimization process. (a) Convergence of the objective function value vs. iteration times with fixed $\mu = 1.0$; (b) Convergence of the performance F-score vs. iteration times.

To study the tag refinement performance, three algorithms were employed as the baselines:

- User tagging (UT): *i.e.*, the original user-provided tags.

- Random walk with restarts (RWR): the tag refinement algorithm based on random walk with restarts proposed in [13].

- Tag refinement based on visual and semantic consistency (TRVSC): the tag refinement algorithm based on visual and semantic consistency proposed in [17].

The RWR and TRVSC are the two state-of-the-art algorithms for tag refinement task. In addition, we compared the performance of our approach with different combinations of regularization terms: 1) low-rank and error sparsity only (LR_ES), 2) low-rank, error sparsity and content consistency (LR_ES_CC), 3) low-rank, error sparsity and tag correlation (LR_ES_TC), and 4) all regularization terms (LR_ES_CC_TC). Figure 3 shows the detailed performances of image refinement for individual tags between our proposed approach and the baselines. From Figure 3, we can observe that the performance of tag refinement from our approach is much superior over those from user-provided tags, RWR and TRVSC algorithms. Table 1 further lists the average performances measured by F-score for different tag refinement schemes.

The experiments were performed in Matlab platform on a server machine with dual quad-core 3.0GHz Intel Xenon processors and 32GB RAM. Our proposed approach (here refers to LR_ES_CC_TC, note that the four versions of our proposed approach with different combinations of regularization terms are quite similar in computational cost) only cost 1 hour while, in contrast, RWR algorithm cost nearly 22 hours and TRVSC consumed more than 9 hours. Such results can be explained by analyzing the details of different algorithms. For RWR algorithm, it has to calculate large amount of visual similarity matrices conditioned on all the tags of each image against all the other images, which is highly computation-intensive. TRVSC algorithm equally needs to perform an intensive optimization during the refinement procedure, of which the algorithmic complexity is of $O(mn^2)$, m and n are the total numbers of tags and images, respectively. Nevertheless, for our approach, the key computational bottleneck is the SVD required by each iteration, and the computational cost is lower compared with RWR and TRVSC. In addition, for the singular value thresholding step at each iteration, it is observed that full SVD computation is not always necessary, especially in the first few iterations when μ_k is quite large. A partial SVD [25] can be used to speed up the optimization procedure, and thus the computational cost can be further alleviated significantly.

Figure 4 further shows the tag refinement results for some exemplary images produced by our proposed approach. It can be seen that our proposed approach can effectively correct and enrich the imprecise and incomplete image tags. For example, in Figure 4(c), only the tag "architecture" is related with the image content while other tags are annotated by users with their personal intentions. After tag refinement, the weakly related tags are removed and the close-related tags, *e.g.*, "building" and "house", are enriched by considering the content consistency and tag correlation. Moreover,
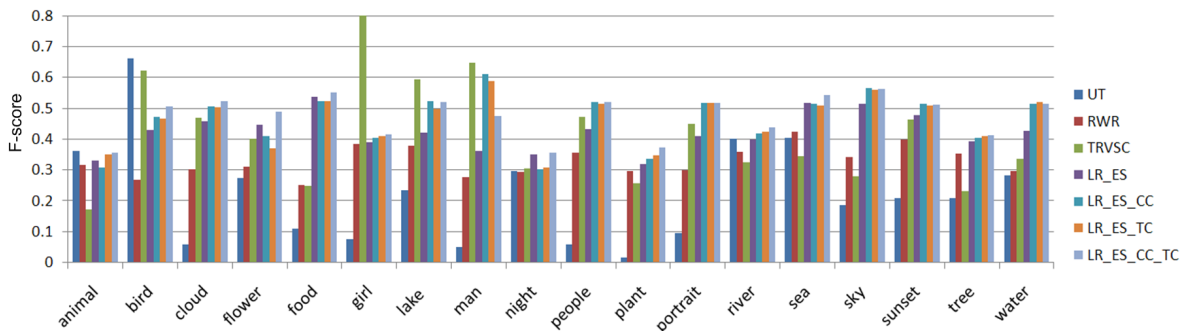
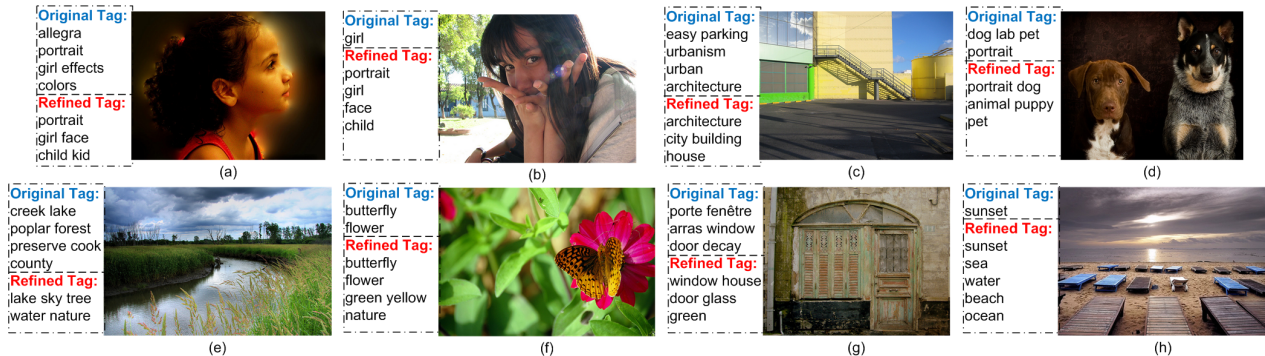**Figure 3: Detailed performances for individual tags of different approaches on MIRFlickr-25K dataset.**

Figure 4: Exemplar tag refinement results from our proposed approach on MIRFlicker-25K dataset.

the enrichment capacity of our proposed algorithm can be particularly seen from Figure 4(b) and (h), where only one tag is associated with each image as original annotations, and the incomplete tags are finally enriched by our proposed refinement strategy reasonably.

5.2 Evaluation of Tag Refinement on NUS-WIDE-270K

NUS-WIDE-270K is one of the most representative large-scale image datasets. The evaluation on this dataset can not only verify the algorithmic performance, but also check the efficiency of our proposed approach on large-scale dataset. In this evaluation, we only compared the performance of our proposed tag refinement strategy in different optimization configurations, as the existing RWR and TRVSC algorithms are hard to be implemented for large-scale datasets due to their computation-intensive insights discussed above. The refinement duration on NUS-WIDE-270K is about 6 hours for our proposed approach (LR_ES_CC_TC). This demonstrates that the APG based iterative procedure is efficient for optimization. Figure 5 shows the detailed tag refinement results for individual tags and Table 2 summaries the average F-score for all tags. The baseline algorithm UT is evaluated based on user-provided initial tagging results.

Figure 6 displays some exemplary images with their initial tags and the refined tags. Such examples demonstrate the capacity of our proposed approach on large-scale image dataset. From Figure 6(b), (c) and (d), we can see that the imprecise tags (*e.g.*, girl, excellence and sigma) are removed by tag refinement. In Figure 6(g) and (h), the missed tags (*e.g.*, sky, night, people, and portrait) are enriched. From these results, we particularly observe that the proposed tag

refinement approach is capable of automatic image annotation. There are about 25K images without any initial tags in NUS-WIDE-270K data collection. Figure 6(a) and (f) illustrate two unlabeled images with refined tags. We can see that the added tags are reasonable regarding the image contents.

5.3 Evaluation of Refinement Performance against Noise Rate

In this subsection, we systematically evaluate the tag refinement capability of our proposed approach at different tag noise rate and aim to uncover the breakdown point, where the performance of tag refinement shall decrease significantly. This breakdown point is valuable for guiding users to provide proper amount of tags, such that reasonably accurate tags can be recovered by tag refinement algorithms. However, to our best knowledge, no research effort has ever been devoted to this study. As our proposed approach is relatively more efficient and more effective, the experiments were performed based on our proposed tag refinement approach. The MIRFlickr-25K dataset and its tags with ground truth were utilized for these experiments.

Denote the binary tag membership matrix from the ground truth as D_g. We simulate the tag noises of different rates as follows. Assume that D_g is with N_1 entries of value 1 and N_0 entries of value 0, and the corresponding user-provided tag matrix is denoted D_u. Based on D_g and D_u, we can calculate two statistics $TF_{1\to0}$ and $FT_{0\to1}$. $TF_{1\to0}$ is the number of entries with value 1 in D_g whereas with value 0 in D_u, and $FT_{0\to1}$ represents the number of entries with value 0 in D_g whereas with value 1 in D_u. Then two statistics can be used to compute the ratio of two types of errors in human tagging. For a given rate ($0 \leq \alpha \leq 1$) of noise,

Table 2: Average performances of different approaches for tag refinement on NUS-WIDE-270K

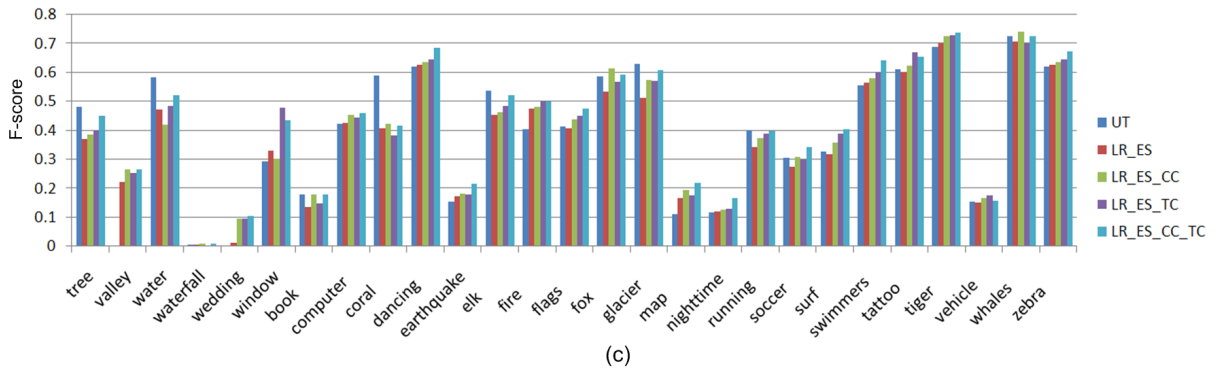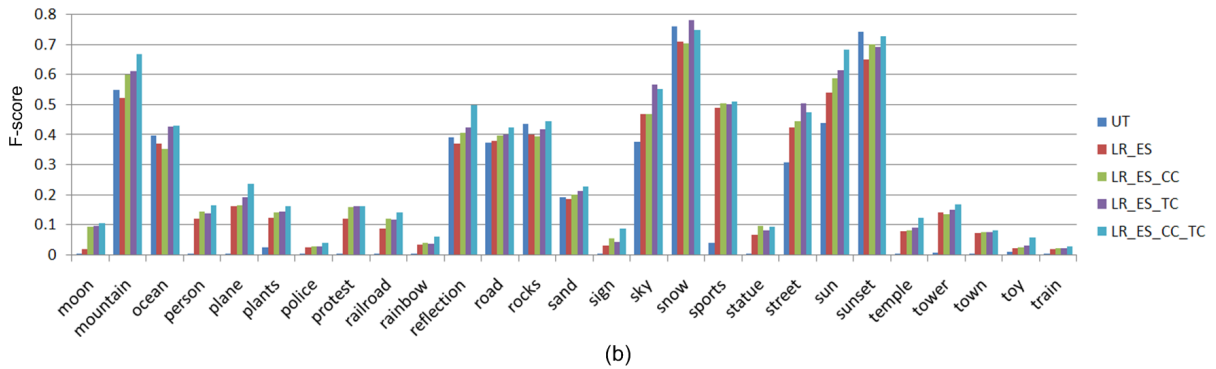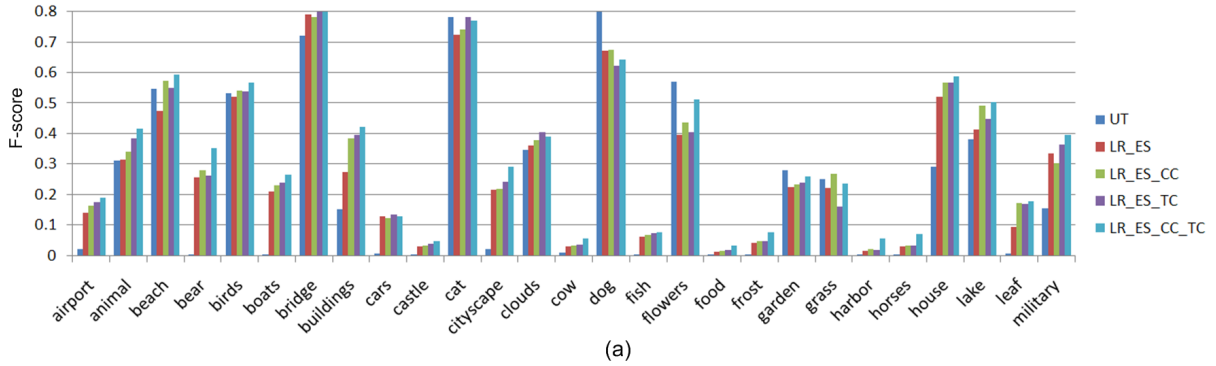|	UT	RWR	TRVSC	LR_ES	LR_ES_CC	LR_ES_TC	LR_ES_CC_TC
F-score	0.269	-	-	0.299	0.321	0.330	0.353

(a)

(b)

(c)

Figure 5: Detailed performances for individual tags of different approaches on NUS-WIDE-270K dataset.

Figure 6: Examples of tag refinement results by our approach on NUS-WIDE-270K dataset.
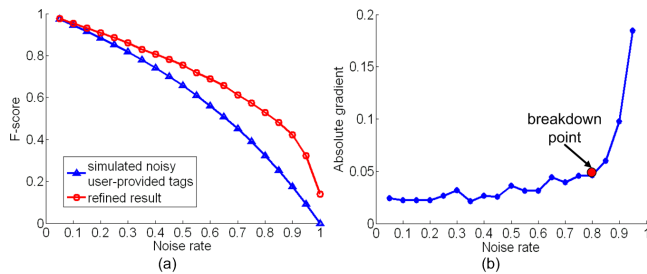
Figure 7: Detailed tag refinement results with different tag noise rate by our proposed algorithm on MIRFlickr-25K dataset. (a) Tag refinement performance against different noise rate compared with simulated noisy user-provided tags; (b) Absolute degradation gradient of the tag refinement performance against different noise rate.

we first randomly select $N_{1 \to 0} = \alpha N_1$ entries of which the values are changed from 1 to 0. Then, $N_{0 \to 1} = \frac{FT_{0 \to 1}}{TF_{1 \to 0}} N_{1 \to 0}$ entries with value 0 are randomly selected and adjusted as 1.

Our proposed tag refinement approach is then performed on the noise-added tag matrix. Figure 7(a) shows the tag refinement performance against different noise rate. From these results, we can observe that the tag refinement generally improves the tag quality by comparing with noise-added tag matrix. By increasing the noise rate, the tag refinement performance degrades gradually, and this phenomena is further detailed by calculating the curve of the absolute values of the degradation gradients as shown in Figure 7(b). The degradation gradient g_i is defined as $g_i = r_i - r_{i+1}$ where r_i and r_{i+1} are the refinement performance (F-score value) against i-th and $(i+1)$-th configurations of noise rates. From Figure 7(b), we can conclude that our proposed approach is robust to the noise when increasing the noise rate from 0.1 to 0.8. When α is increased higher than 0.8, the degradation gradient is significantly increased, which is right at the breakdown point as marked in Figure 7(b). This breakdown point means that, given that the ratio of two types of tagging errors is fixed, at least about 20% tags need be correctly tagged if we expect to achieve reasonably good performance in tag refinement task.

6. CONCLUSIONS AND FUTURE WORK

Motivated by the fact that the existing user-provided image tags in public photo sharing websites are imprecise and incomplete, we proposed an efficient iterative approach for image tag refinement by pursuing the low-rank, content consistency, tag correlation and error sparsity. Extensive experiments on large-scale image datasets, 25K and 270K respectively, well demonstrated the effectiveness and efficiency of our proposed algorithm.

Our future work shall focus on two directions. First, more effective visual features shall be integrated in the current framework for measuring the image content consistency. Second, more efficient iterative optimization procedure shall be further exploited, such that the proposed algorithm can work on even larger volume image corpus.

7. ACKNOWLEDGMENT

8. REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

[2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 2000.

[3] C. Wang, F. Jing, L. Zhang, and H. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 2008.

[4] X. Li, L. Chen, L. Zhang, W. Ma, and F. Lin. Image annotation by large-scale content-based image retrieval. *ACM MM*, 2006.

[5] X. Rui, M. Li, Z. Li, W. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. *ACM MM*, 2007.

[6] M. Huiskes and M. Lew. The mir flickr retrieval evaluation. *ACM MIR*, 2008.

[7] R. Zhao and W. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *TMM*, 2002.

[8] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *TPAMI*, 2008.

[9] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: discriminative nearest neighbor classification for visual category recognition. *CVPR*, 2006.

[10] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. *AAAI*, 2006.

[11] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, 2009.

[12] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence and wordnet. *ACM MM*, 2005.

[13] C. Wang, F. Jing, L. Zhang, and H. Zhang. Content-based image annotation refinement. *CVPR*, 2007.

[14] J. Jia, N. Yu, X. Rui, and M. Li. Multi-graph similarity reinforcement for image annotation refinement. *ICIP*, 2008.

[15] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang. Tag ranking. *WWW* , 2009.

[16] H. Xu, J. Wang, X. Hua, and S. Li. Tag refinement by regularized lda. *ACM MM*, 2009.

[17] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *ACM MM*, 2010.

[18] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, (submitted). http://watt.csl.illinois.edu/~perceive/matrix-rank/Files/RobustPCA.pdf

[19] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extension: a general framework for dimensionality reduction. *TPAMI*, 2007.

[20] R. Cilibrasi and P. Vitany. The google similarity distance. *TKDE*, 2007.

[21] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2009.

[22] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *UIUC Technical Report UILU-ENG-09-2214*, 2009.

[23] J. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *In preprint*.

[24] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. *CIVR*, 2009.

[25] R. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *Aarhus University Technical Report DAIMI-PB-357*, 1998.

[26] C. Hsu, C. Chang, and C Lin. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.