Image Tagging with Social Assistance

Yang Yang[†§], Yue Gao[†], Hanwang Zhang[†], Jie Shao[†], Tat-Seng Chua[†]

[†]School of Computing, National University of Singapore
[§]School of Information Technology and Electrical Engineering, The University of Queensland
{dlyyang,kevin.gaoy}@gmail.com; {hanwang,shaojie,chuats}@comp.nus.edu.sg

ABSTRACT

Image tagging, also known as image annotation and image conception detection, has been extensively studied in the literature. However, most existing approaches can hardly achieve satisfactory performance owing to the deficiency and unreliability of the manually-labeled training data. In this paper, we propose a new image tagging scheme, termed *social assisted media tagging* (SAMT), which leverages the abundant user-generated images and the associated tags as the "social assistance" to learn the classifiers. We focus on addressing the following major challenges: (a) the noisy tags associated to the web images; and (b) the desirable robustness of the tagging model. We present a joint image tagging framework which simultaneously refines the erroneous tags of the web images as well as learns the reliable image classifiers. In particular, we devise a novel tag refinement module for identifying and eliminating the noisy tags by substantially exploring and preserving the low-rank nature of the tag matrix and the structured sparse property of the tag errors. We develop a robust image tagging module based on the $\ell_{2,p}$-norm for learning the reliable image classifiers. The correlation of the two modules is well explored within the joint framework to reinforce each other. Extensive experiments on two real-world social image databases illustrate the superiority of the proposed approach as compared to the existing methods.

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Tagging, Robust, Noise, Low Rank, Sparse Coding

ICMR'14, April 01-04, 2014, Glasgow, United Kingdom.

1. INTRODUCTION

In recent years, we have witnessed an explosive growth of the user-generated images on the Web driven by the rapid advance of smart phones, high-speed internet and online sharing sites (e.g., Flickr[1] and Instagram[2]). There is an urgent need for effectively and efficiently searching the image content. The existing commercial image search engines, such as Google[3], provide the image search functionality based on the textual metadata associated to the images. However, the problem with text-based image search is that the textual metadata are usually erroneous, incomplete and inconsistent with the image content.

Many recent research endeavours have been dedicated to learn semantic concepts as intermediate semantic representation to facilitate image retrieval [16, 3, 28, 5, 20, 27]. We refer the task of predicting the presence of semantic concepts in images as image annotation [18, 8] or image tagging [23, 29, 24, 21, 26]. Most of the existing image tagging approaches are developed based on machine learning techniques, where sufficient high-quality training samples are often required in order to address the well-known semantic gap [6] problem and achieve satisfactory performance. The reliability of the training data forms a fundamental basis for most of the components in the learning process, including learning model formulation, model selection, evaluation, etc. However, a long-standing obstacle is that the acquisition of good-quality manually-labelled image data is difficult (even for the domain experts) and expensive.

Over the last decade, the rapid evolution of social media provides us a great opportunity to gather a large number of user-generated images and their associated tags. These valuable resources have enormous potential for handling the deficiency of the training data in image tagging. In this case, one may raise a natural question, i.e., *how can we effectively exploit the social user-tagged images to help supervise the learning of the image classifiers*? We argue that it is non-trivial to involve this kind of user-tagged images in the learning process due to the following two main challenges. On the one hand, the user-generated tags of the web images are inevitably unreliable. In has been reported that many user-generated tags of the images on the public media sharing sites, such as Flickr, are erroneous and only around 50% of tags are actually related to the image content [11]. The underlying reason is that the average users are not well motivated for providing high-quality tags. If we use

[1] http://www.flickr.com/
[2] http://instagram.com/
[3] http://www.google.com/imghp

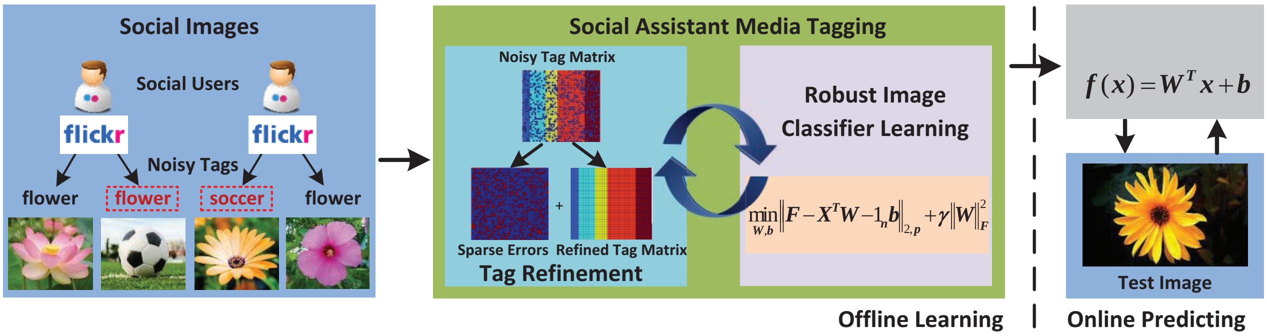


Figure 1: Flowchart of the proposed SAMT framework.

such noisy image tags in the process of learning the image classifiers, the performance of the resulting classifiers will be severely degraded. On the other hand, the conventional learning models are not robust and flexible enough to handle the noisy Web data generated by the social users, which makes it difficult to directly apply them for learning the reliable classifiers and achieving satisfactory performance.

Previous work has been dedicated to tag refinement [22, 17, 30, 11], i.e., the process of improving the quality of user-generated tags associated to the web multimedia data based on multimedia content analysis techniques. They explored the tag-to-image relevance and further refined it with the random walk model. Tang *et al.* [17] proposed to build a robust graph and applied graph-based semi-supervised learning technique to learn a tag ranking model to achieve tag refinement. Liu *et al.* [11] formulated the tag refinement as a tag ranking problem and applied a probabilistic method to rank the associated tags of the images. In [22], Xu *et al.* computed the tag relevance via an extended latent dirichlet allocation model. They added a regularizer to incorporate visual information. Zhu *et al.* [30] simultaneously explored low-rank properties, tag consistency and content consistency to identify a refined tag matrix for web images. While most of the previous work either try to cleanse the noisy user-generated tags or enhance the robustness of the existing approaches, few attempts have been made towards exploring the correlation between these two tasks to make them simultaneously reinforce each other. Besides, under the environment of social media, both the tag matrix and the tag errors may embody certain unique intrinsic properties, which are seldom explored and exploited in previous work.

In this paper, in order to handle the above challenges and problems, we propose a new image tagging framework, termed *social assisted media tagging* (SAMT), which directly takes the noisy social user-tagged images as the training data for learning the reliable image classifiers. The contributions of this paper are summarized as follows:

- We explore the intrinsic nature of the social tags and the embedded tag noise. A novel tag refinement module is proposed based on low rank matrix recovery and sparse group lasso for refining the social tags as well as identifying and eliminating the tag noise.

- We propose a robust image tagging module to further alleviate the influence of tag noise and learn the reliable image classifiers. The robust module utilizes the robust nature of the $\ell_{2,p}$-norm, which extends the applicable range of the $\ell_{2,1}$-norm.

- The correlation between the tag refinement module and the robust tagging module are well investigated. We devise a joint image tagging framework which takes advantage of the reinforcement between the two modules.

- Extensive experiments on two real-world social image datasets illustrate the superiority of the proposed approach as compared to the existing methods.

The reminder of this paper is structured as follows. We elaborate the proposed social assisted media tagging system in Section 2. Section 3 reports the experimental results and analysis on various real-world image datasets. The conclusion is given in Section 4.

2. SOCIAL ASSISTED MEDIA TAGGING

2.1 Preliminary and System Overview

Suppose we are provided with a set of user-tagged images $\mathcal{X} = \{(x_i, y_i)\}|_{i=1}^{n}$ associated to a set of concepts $\mathcal{C} = \{c_j\}|_{j=1}^{m}$. $x_i \in \mathbb{R}^{d \times 1}$ represents the ith image's visual vector in a certain visual space while $y_i \in \{-1, 1\}^{m \times 1}$ is the corresponding label vector. $y_{ij} = 1$ indicates that the ith image is tagged with the jth concept in $\mathcal{C}$ and $y_{ij} = -1$ implies that the image is not associated to the concept c_j. n is the number of the images in $\mathcal{X}$, m is the number of the concepts in $\mathcal{C}$, and d is the dimensionality of the visual space. Without loss of generality, we may assume that a certain number of the images in $\mathcal{X}$ are mislabeled, i.e., the tags of these images are erroneous. In this case, the major goal is to develop a novel image tagging scheme, which is able to not only effectively refine the tags of the noisy user-tagged images but also robustly learn the reliable image classifiers for predicting the tags for unseen images.

Figure 1 illustrates the flowchart of the proposed image tagging framework. Given the concept c, we first collect a set of user-tagged images from the Web. Some of these images are associated to the concept c and some are not. These images are then fed into the proposed social assisted media tagging (SAMT) framework for refining the social tags and learning the reliable image classifiers. The SAMT framework consists of two modules: 1) a tag refinement module based on low rank matrix recovery and sparse group lasso; and 2) a robust image classifier learning module based on $\ell_{2,p}$-based regression. The two modules exert reinforcing effects on each other in order to achieve the better performance. As illustrated, the refined tags are used to "supervise" the learning of the image classifiers while the learning process can help achieve better tag refinement in return. In the online prediction, the learnt classifiers are used to predict whether a given test image should be assigned with the concepts in $\mathcal{C}$ or not.

2.2 Refining User-Generated Tags

As aforementioned, in order to handle the deficiency of training data, we seek help from the user-tagged images. However, in most cases such images are erroneously tagged and thus cannot be directly exploited for learning the image classifiers. Besides, under the circumstance of social media, both the user-generated tags and the tag noise may have certain unique intrinsic properties, which should be explored to guide the tag refinement process.

We observe that there exist many near-duplicate or duplicate images on the Web. These images probably have identical semantics. In other words, the tags of these duplicate images should be the same, which gives us a strong indication that the rank of the corresponding tag matrix should be low enough. On the other hand, the tag errors should also possess certain characteristics. In previous work [30], the sparse property of the tag errors is captured via the ℓ_1-norm, which only controls the magnitude of the erroneous tags and ignores the structural information. One consequence of this method is that these identified erroneous tags tend to spread over all the images, which implicitly indicates that all the images are unreliable. However, according to our observation, only a small proportion of the images are actually mislabeled and only a few tags of these images are erroneously assigned. Therefore, it is more reasonable to assume that the matrix of the tag errors should be sparsely structured, i.e., be sparse both at the sample level and at the element level.

Based on the above analysis, we devise a tag refinement module which simultaneously explores and preserves the endowed low-rank nature of the "genuine" tag matrix as well as the structured sparse property of the tag errors:

$$\begin{aligned} \min_{F,E} \quad & \|F\|_* + \lambda(\|E\|_{2,1} + \|E\|_1), \\ \text{s.t. } & Y = F + E, \end{aligned} \tag{1}$$

where $Y = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^{n\times m}$ is the original tag matrix with tag noise, $F = [f_1, f_2, \dots, f_n]^T \in \mathbb{R}^{n\times m}$ represents the refined tag matrix and $E \in \mathbb{R}^{n\times m}$ is the matrix of the tag errors imposed on F and λ is a trade-off parameter. The constraint $Y = F + E$ indicates that the original tag matrix Y is a combination of the genuine tag matrix F and the error matrix E. $\|\cdot\|_*$ denotes the nuclear norm (i.e., the sum of the singular values of a matrix). It is remarkable that the nuclear norm helps explore the low-rank property of the recovered tag matrix F while the mixed-sparsity term $(\|E\|_{2,1} + \|E\|_1)$ characterizes the structured sparse nature of the tag error E and is expected to induce both sample-wise and element-wise sparsity over E. Here, the $\ell_{2,1}$-norm of E is defined as

$$\|E\|_{2,1} = \sum_{i=1}^{n} \|(E)_i\|_2, \tag{2}$$

where $(E)_i$ denotes the ith row of E, i.e., the tag error of the ith image. We can see that the $\ell_{2,1}$-norm $\|E\|_{2,1}$ is capable of identifying the mislabeled image samples by inducing the sample-wise sparsity over E, while the ℓ_1-norm $\|E\|_1$ further helps to generate the element-wise sparsity within the identified noisy images, thereby pinpointing which tags are actually mislabelled.

With the above tag refinement module, we may cleanse the erroneous tag matrix Y and obtain the refined tag matrix F for the training images. Next, we will detail a robust classifier learning procedure with the refined training data.

2.3 Learning Robust Image Classifier

In this subsection, we focus on how to learn the robust image classifiers with the training images in $\mathcal{X}$ together with the refined tags in F. Under the circumstance of social media, due to the arbitrary and unpredictable tagging behaviors of the social users, it may become much more difficult to conduct a thorough cleansing for the tag noise with the above tag refinement module and some remaining noise may still exist in the tags. If we directly feed the training data into certain traditional model, such as ridge regression [7], the tagging performance will be severely degraded.

In order to further alleviate the influence of the remaining erroneous tags and learn the reliable image classifiers from the refined training data, we expect the learning model to be robust and tolerant to the potential tag noise. Without loss of generality, the target is to learn the following linear classifier:

$$f(x) = W^T x + b, \tag{3}$$

where $x \in \mathbb{R}^{d\times 1}$ denotes the visual feature of an image, $W \in \mathbb{R}^{d\times m}$ is the classification coefficients and $b \in \mathbb{R}^{m\times 1}$ is the bias term.

It has been shown [14] that the $\ell_{2,1}$-norm loss function are endowed with more reliable robustness to noise or outliers than other traditional loss functions, such as hinge loss. Therefore, we may employ the following $\ell_{2,1}$-based learning model for learning the classifiers:

$$\min_{W,b} \left\|F - X^T W - \mathbf{1}_n b^T\right\|_{2,1} + \gamma \|W\|_F^2 \tag{4}$$

where $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d\times n}$ is the visual feature matrix consisting of all the images in $\mathcal{X}$; $\mathbf{1}_n$ is an all-one vector of size $n \times 1$; γ is a trade-off parameter; and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. It is worth noting that the residual of Eq. (4), i.e., $\left\|F - X^T W - \mathbf{1}_n b^T\right\|_{2,1}$, is characterized using the $\ell_{2,1}$-norm loss rather than the squared Frobenius norm. The underlying principle is that we do not expect the error caused by the noisy samples to be squared and over-emphasized, which naturally makes the above model robust to noise.

Further, we may expect to extend the applicable range of the $\ell_{2,1}$-based model in order to adapt to different situations of the tag errors. To this end, we exploit a natural extension of the $\ell_{2,1}$-norm, which is able to flexibly handle different levels of the tag errors:

$$\min_{W,b} \left\|F - X^T W - \mathbf{1}_n b^T\right\|_{2,p} + \gamma \|W\|_F^2, \tag{5}$$

where the $\ell_{2,p}$-norm is defined as follows:

$$\|M\|_{2,p} = \sum_{i=1}^{n} \left\|(M)_i\right\|_2^p \tag{6}$$

where $(M)_i$ is the ith row of the matrix M and $p \in (0, 2]$ is a parameter for controlling the robustness of the model. It is remarkable that both the above $\ell_{2,1}$ model ($p = 1$) and the ridge regression ($p = 2$) are special cases of the $\ell_{2,p}$ model, which further implies its more flexible applicability.

In this way, the model in Eq. (5) provides us a powerful and flexible tool for training the robust image classifiers.

In the next part, we will elaborate a joint image tagging framework which effectively combines the tag refinement module and the classifier learning module by exploring their correlation.

2.4 A Joint Framework

So far, we have developed two independent modules to undertake the tasks of refining the erroneous tag matrix as well as learning the robust image classifiers, respectively. In order to perform the effective image tagging, a straightforward two-step approach can be used, i.e., first refine the erroneous tag matrix and then feed it together with the training images into the robust model for learning the image classifiers. One limitation of this approach is that the intrinsic correlation between these two modules are not well explored to reinforce the performance of each other.

In this subsection, we propose a novel joint image tagging framework, termed *social assisted media tagging* (SAMT), which simultaneously conducts the tag refinement and classifier learning by exploring their intrinsic correlations. The fundamental design principle of the SAMT framework lies in that the tag refinement module and classifier learning module should form a mutually-reinforcing learning loop, in which the refined tags of the Web images should be well explored to better supervise the learning of the image classifiers, while the classifier learning process is supposed to guide a better tag refinement in return. Based on the above analysis, we formulate the image tagging under the noisy circumstance as follows:

$$\begin{aligned}\min_{F,E,W,b}\ & \|F\|_* + \lambda\left(\|E\|_{2,1} + \|E\|_1\right)\\ & + \mu\left(\left\|F - X^TW - \mathbf{1}_nb^T\right\|_{2,p} + \gamma\|W\|_F^2\right),\\ \text{s.t. } & Y = F + E,\end{aligned} \tag{7}$$

where μ is a trade-off parameter which controls the balance between the two modules. The objectives of the two modules are simultaneously achieved under the joint framework with their correlation substantially explored. Next, we will present an effective solution for optimizing the SAMT model.

2.5 Optimization

Note that Eq. (7) is difficult to solve because it is nonconvex w.r.t. all the variables at the same time, and the non-smooth property of the weighted loss function makes it non-trivial to optimize the problem as a whole. To address the above challenges, we devise an effective algorithm to optimize the model.

First, we introduce an alternative problem as below:

$$\begin{aligned}\min_{F,E,W,b}\ & \|F\|_* + \lambda\left(\|E\|_{2,1} + \|E\|_1\right)\\ & + \mu\left(Tr\left((F-X^TW-\mathbf{1}_nb^T)^TD(F-X^TW-\mathbf{1}_nb^T)\right)\right.\\ & \left.+\gamma\|W\|_F^2\right),\\ \text{s.t. } & Y = F + E,\end{aligned} \tag{8}$$

where $Tr(\cdot)$ is the trace of a matrix. D is a diagonal matrix with its ith diagonal element defined as

$$D_{ii} = \frac{1}{\frac{2}{p}\|(F - X^TW - \mathbf{1}_nb^T)_i\|_2^{2-p}}, \tag{9}$$

where $(F-X^TW-\mathbf{1}_nb^T)_i$ is the ith row of $(F-X^TW-\mathbf{1}_nb^T)$. We further rewrite the problem in Eq. (8) as follows:

$$\begin{aligned}\min_{J,F,E,W,b}\ & \|J\|_* + \lambda\left(\|E\|_{2,1} + \|E\|_1\right)\\ & + \mu\left(Tr\left((F-X^TW-\mathbf{1}_nb^T)^TD(F-X^TW-\mathbf{1}_nb^T)\right)\right.\\ & \left.+\gamma\|W\|_F^2\right),\\ \text{s.t. } & Y = F + E \wedge J = F.\end{aligned} \tag{10}$$

Note that D is actually dependent on F, W and b, which makes the above problem difficult to solve. To handle this problem, we design an iterative algorithm, which updates D in each iteration with the F, W, b of the previous iteration. In this way, D is disconnected with F, W, b, which makes the problem in Eq. (10) solvable via exact or inexact Augmented Lagrange Multiplier (ALM) method [10]:

$$\begin{aligned}\min_{J,F,E,W,b,G_1,G_2}\ & \|J\|_* + \lambda\left(\|E\|_{2,1} + \|E\|_1\right)\\ & + \mu\left(Tr\left((F-X^TW-\mathbf{1}_nb^T)^TD(F-X^TW-\mathbf{1}_nb^T)\right)\right.\\ & \left.+\gamma\|W\|_F^2\right) + Tr\left(G_1^T(Y - F - E) + G_2^T(F - J)\right)\\ & + \frac{\beta}{2}\left(\|Y - F - E\|_F^2 + \|F - J\|_F^2\right),\end{aligned} \tag{11}$$

where G_1 and G_2 are the Lagrange multipliers and $\beta > 0$ is a trade-off parameter.

Update W and b by fixing others. When updating W and b by keeping others fixed, we obtain the following sub-problem:

$$\min_{W,b} Tr((F-X^TW-\mathbf{1}_nb^T)^TD(F-X^TW-\mathbf{1}_nb^T))+\gamma\|W\|_F^2, \tag{12}$$

By setting the derivative of the above objective function w.r.t. b to zero, we have

$$b^T = \frac{\mathbf{1}_n^TD}{\mathbf{1}_n^TD\mathbf{1}_n}(F - X^TW). \tag{13}$$

By substituting Eq. (13) into Eq. (12) we arrive at

$$\min_W Tr\left((HF-HX^TW)^TD(HF-HX^TW)\right) + \gamma\|W\|_F^2, \tag{14}$$

where $H = I_n - \frac{\mathbf{1}_n\mathbf{1}_n^TD}{\mathbf{1}_n^TD\mathbf{1}_n}$. I_n is the $n \times n$ identity matrix. Then, setting the derivative of the above objective function w.r.t. W to zero, we have

$$W = (XH^TDHX^T + \gamma I_n)^{-1}XH^TDHF. \tag{15}$$

Update J by fixing others. When we fix all others except J, we have the following sub-problem:

$$\min_J \frac{1}{\gamma}\|J\|_* + \frac{1}{2}\left\|J - (F + \frac{G_2}{\beta})\right\|_F^2, \qquad (16)$$

which can be solved by the singular value thresholding algorithm [1].

Update F by fixing others. When we update F with all other variables fixed, the problem reduces to

$$\begin{aligned}\min_F \; & Tr\left((F-\hat{F})^T D(F-\hat{F})\right) + tr\left((G_2^T - G_1^T)F\right) \\ & + \frac{\beta}{2}\left(\|Y-F-E\|_F^2 + \|F-J\|_F^2\right),\end{aligned} \qquad (17)$$

where $\hat{F} = X^T W + \mathbf{1}_n b^T$. Then, we can update F with the following closed-form solutions:

$$F = (2\mu D + 2\beta I)^{-1}\left(2\mu D\hat{F} + G_1 - G_2 + \beta(J+Y-E)\right). \qquad (18)$$

Update E by fixing others. When updating E, we need to solve the following sub-problem:

$$\min_E \frac{\lambda}{\beta}\left(\|E\|_{2,1} + \|E\|_1\right) + \frac{1}{2}\left\|E - (Y - F + \frac{G_1}{\beta})\right\|_F^2, \quad (19)$$

which can be regarded as a sparse group lasso model [25, 4] and can be efficiently solved using the SLEP optimization toolbox [12].

Finally, we summarize the algorithm for solving the problem in Algorithm 1. It can be similarly proven [15] that by iteratively solving the problem in Eq. (8) we can converge to the optima of the problem in Eq. (7).

Algorithm 1 An iterative algorithm for solving the problem in Eq. (7).

```
Input: The Web image data set $X$ and the corresponding noisy tag matrix $Y$;
Output: The refined tag matrix $F$ for the training image set, the image classifier coefficients $W$ and $b$;
Initialize $J, F, E, W, b, G_1, G_2$;
Initialize $\beta = 10^{-6}, \beta_{max} = 10^{10}, \rho = 1.1$;
repeat
    Update $D$ according to Eq. (9);
    Update $W$ according to Eq. (15);
    Update $b$ according to Eq. (13);
    Update $J$ by applying the singular value thresholding operator on Eq. (16);
    Update $F$ according to Eq. (18);
    Update $E$ by solving the sparse group lasso problem in Eq. (19) with the SLEP toolbox;
    Update the multipliers $G_1$ and $G_2$:
        $\begin{cases} G_1 \leftarrow G_1 + \beta(Y - F - E) \\ G_2 \leftarrow G_2 + \beta(F - J) \end{cases}$
    Update $\beta \leftarrow \min(\rho\beta, \beta_{max})$;
until convergence
return $F, W, b$;
```

3. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed social assisted media tagging framework.

3.1 Data and Experimental Settings

We utilized two real-world social image data sets, namely NUS-WIDE-SCENE and NUS-WIDE [2], for evaluation. NUS-WIDE contains $269,648$ Flickr images manually tagged with 81 tags, while NUS-WIDE-SCENE comprises $34,928$ Flickr images with the ground-truth of 33 tags. The images in both data sets are associated with $5,018$ user-generated tags, which are quite noisy. We used the tags with ground-truth as the evaluated tags. For each evaluated tag, we randomly selected $\{10, 20, 30, 40, 50\}$ positive samples to form the training set.

We extracted four types of visual features for image representation, including a 9D edge feature, a 512D color feature, a 512D texture feature [9] and a $4,096$D bag-of-visual-word feature based on Scale Invariant Feature Transform (SIFT) descriptors [13]. For the color, texture, and SIFT descriptors, we used the locality-constrained linear sparse coding (LLC) [19] method with max-pooling to generate the visual representations. Further, we divided each image into 2×3 grids, and for each grid and the whole image, we extracted all the four features. Thus, we have a $35,903$D ($5,129 \times 7$) feature for each image. Finally, we used PCA to reduce the dimensionality to $1,024$.

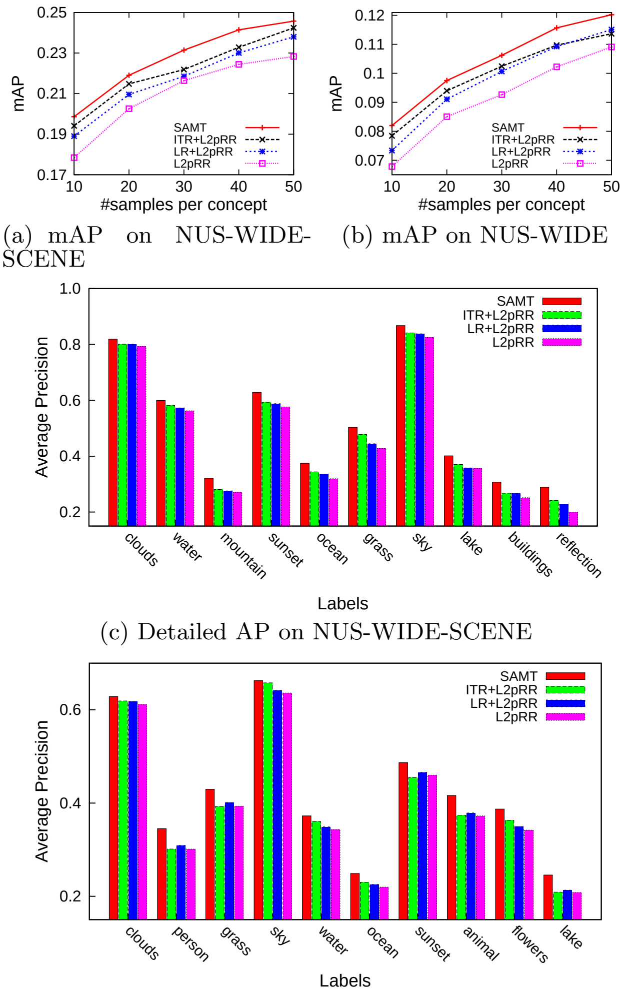
We compared the proposed SAMT framework with the following state-of-the-art approaches and baselines:

- **ITR [30]**, which utilizes an low-rank model and content consistency to refine the tags of the web images. It does not learn any classifiers for future use.
- **L2pRR [14]**, which extends the original $\ell_{2,1}$-norm model to its more flexible variant $\ell_{2,p}$-norm model as described in Section 2.4. It directly takes the noisy tags as input for learning.
- **ITR+L2pRR [30, 14]**, which first uses the ITR method to cleanse the tags and then learns the classifiers with the robust ridge regression (L2pRR) model [14].
- **SAMTL2**, which is a variant of the SAMT approach except that its loss function is based on ℓ_2-norm instead of $\ell_{2,p}$-norm.
- **LR+L2pRR**, which first uses the tag refinement proposed in this work (referred as LR) and then learns the classifiers with the L2pRR model.
- **SocialTagging**, which refers to the process of the social users generating the tags for the web images.

We test all the balance parameters in all the evaluated methods in $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$ for fair comparison. For all the methods that are based on $\ell_{2,p}$-norm, we set p in the range of $\{0.5, 1.0, 1.5\}$. We used the average precision (AP) of each individual tag and the mean average precision (mAP) of all the evaluated tags as the evaluation metrics. We repeated each experiment five times and reported the average results.

3.2 Comparison in Image Tagging

Figure 2 reports the comparison of image tagging performance between the proposed SAMT framework and the state-of-the-art methods and baselines, including ITA+L2pRR [30], LR+L2pRR [14] and L2pRR [14]. Figure 2(a) and 2(b) report the mAP performance of all the comparing methods



(a) mAP on NUS-WIDE-SCENE

(b) mAP on NUS-WIDE

(c) Detailed AP on NUS-WIDE-SCENE

(d) Detailed AP on NUS-WIDE

Figure 2: Comparison of different image tagging approaches: (a) mAP on NUS-WIDE-SCENE; (b) mAP on NUS-WIDE; (c) Detailed AP of representative tags on NUS-WIDE-SCENE; and (d) Detailed AP of representative tags on NUS-WIDE.

over all the evaluated tags with different numbers of positive samples per tag on NUS-WIDE-SCENE and NUS-WIDE, respectively. Due to space limit, different from Figure 2(a) and 2(b) with all the evaluated tags, we only report several representative tags of NUS-WIDE-SCENE and NUS-WIDE in Figure 2(c) and 2(d), respectively. From these results, we derive the following observations and analysis from these results:

- The proposed SAMT approach consistently achieves the best performance amongst all the comparing approaches w.r.t. different numbers of training samples on the two data sets. The superiority of the proposed SAMT approaches over the other approaches mainly attributes to the substantial exploration of the properties of the genuine tag matrix and the error, as well as the correlation of the refinement module and the classifier learning module.

- L2pRR always achieves the worst performance w.r.t. different numbers of training samples on the two data sets. We understand that the original tags of the web images are noisy and this phenomenon indicates that the tag refinement strategies exploited in different methods really help alleviate the influence of the errors and improve the tagging performance. It is worth mentioning that the tag refinement strategy of the SAMT approach is the best one because all the other methods fail to capture the intrinsic correlation of the tag refinement and learning modules.

- When we increase the number of the training images, even though more potential tag noise may be brought in the learning process, the performance of all the comparing methods keeps improving. This phenomenon shows that the number of training images is a significant factor in achieving better performance.

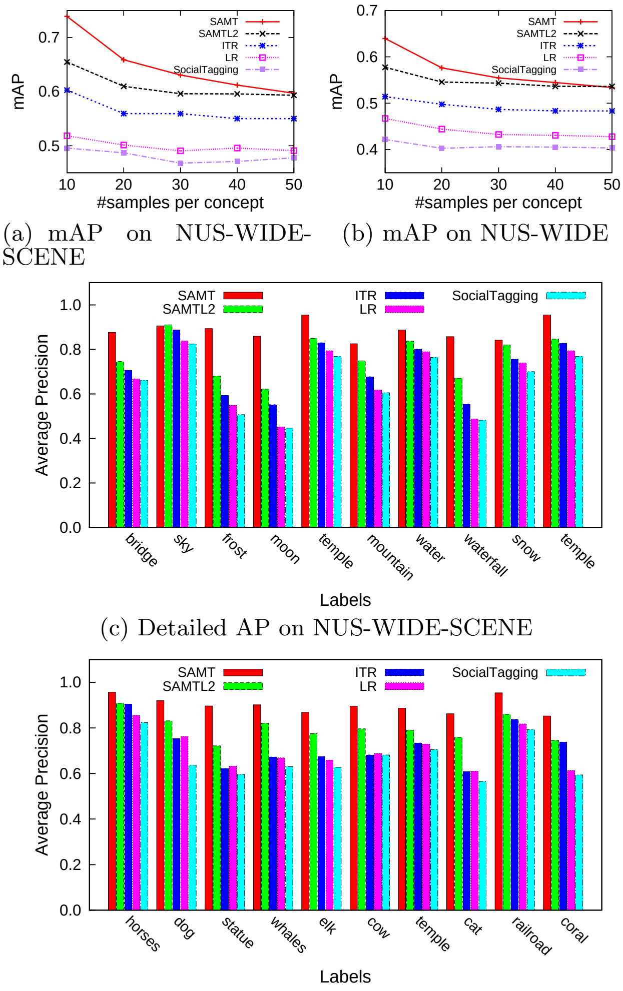
3.3 Effects of Tag Refinement

In this subsection, we provide a quantitative evaluation on the capability of the proposal SAMT approach in cleansing the tag noise by comparing to the state-of-the-art tag refinement methods and baselines, including ITA, LR, SAMTL2 and SocialTagging. In particular, for different tag refinement methods, we examine the quality of the refined tags of the training data by comparing them to the ground-truth. As we can see, Figure 3(a) and 3(b) report the mAP performance with different numbers of positive samples per tag on NUS-WIDE-SCENE and NUS-WIDE, respectively. Again, we report the AP of several representative tags of NUS-WIDE-SCENE and NUS-WIDE in Figure 3(c) and 3(d), respectively. We have the following conclusions from these results:

- SocialTagging consistently performs the worst under different settings on the two data sets. In particular, the performance of mAP achieves only around 40% on NUS-WIDE and 50% on NUS-WIDE-SCENE, which clearly indicates that the original tags of the web images contain quite a proportion of noise. Among all the tag refinement methods, our proposed SAMT and SAMTL2 always outperform the other methods due to the in-depth investigation of the low-rank property of the tag matrix and the structured sparse property of the tag errors, as well as the correlation of the refinement module and the classifier learning module.

- The SAMT approach consistently achieves better performance than its non-robust version, namely SAMTL2. This further demonstrates that the robust $\ell_{2,p}$-based ridge regression actually exerts more positive reinforcement on the tag refinement module than the traditional ridge regression does. The underlying reason is because the robust ridge regression model is able to alleviate more influence caused by the remaining tag errors in each iteration, resulting in a better tag refinement process and a more reliable learning of the image classifiers.

- As shown in Figure 3(a) and 3(b), as the number of the training images increases, the performance of all the comparing methods for refining the tags of the training data shows descending trends. The reason is that when we use more training data, more potential noise may be introduced into the learning process, which



(a) mAP on NUS-WIDE-SCENE

(b) mAP on NUS-WIDE

(c) Detailed AP on NUS-WIDE-SCENE

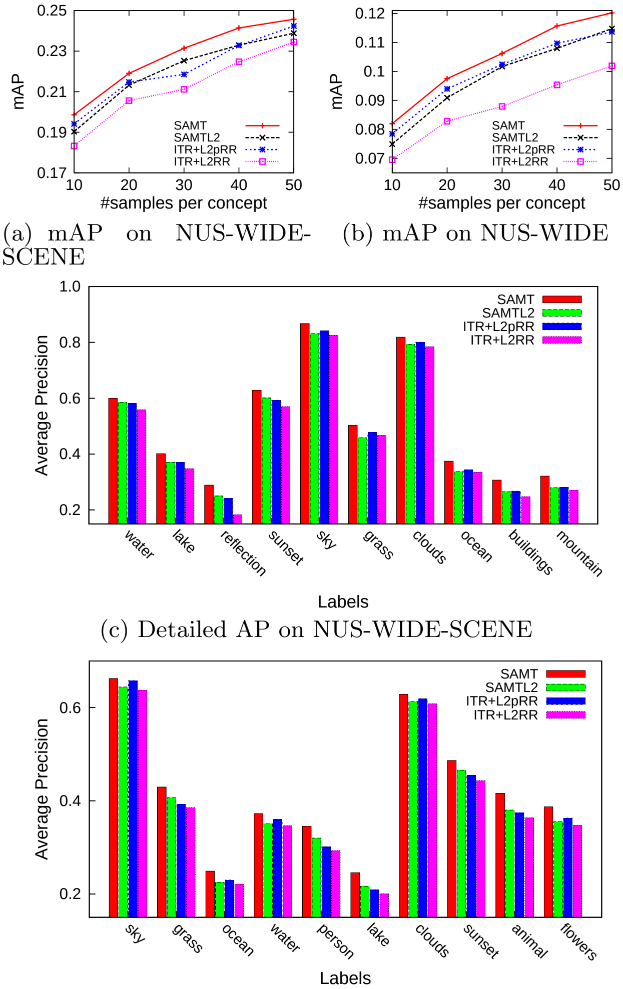
(d) Detailed AP on NUS-WIDE

Figure 3: Effects of tag refinement of different comparing methods on the training data: (a) mAP on NUS-WIDE-SCENE; (b) mAP on NUS-WIDE; (c) Detailed AP of representative tags on NUS-WIDE-SCENE; and (d) Detailed AP of representative tags on NUS-WIDE.

exerts much negative influence on the tag refinement performance.

3.4 Effects of Robustness

In this subsection, we evaluate the robustness of the $\ell_{2,p}$-norm for alleviating the influence of the noisy user-generated tags. We report the mAP and detailed AP of representative individual tags on the test data in the two data sets. We report the results of the comparison between the proposed SAMT approach, its non-robust variant SAMTL2, the ITA+L2pRR and its non-robust version ITA+L2RR in Figure 4. Figure 4(a) and 4(b) report mAP performance with different numbers of positive samples per tag on NUS-WIDE-SCENE and NUS-WIDE, respectively. We illustrate the AP of several representative tags of NUS-WIDE-SCENE and NUS-WIDE in Figure 4(c) and 4(d), respectively. From these results, we can see that the proposed SAMT approach consistently achieves better performance than its SAMTL2 variant; while ITR+L2pRR always outperforms its non-robust version ITR+L2RR. Besides, even though the correlation between the tag refinement module and the learning module, the ITA+L2pRR still performs better than the SAMTL2 approach. These observations clearly shows that the robust ridge regression is indeed crucial for both eliminating the tag noise and the learning of the reliable image classifiers. While the $\ell_{2,p}$-based ridge regression takes advantages of the robust property of the $\ell_{2,p}$-norm, which effectively reinforces the tag refinement module than the traditional ridge regression does. Further, the more refined tags help the $\ell_{2,p}$-based ridge regression to learn the more reliable classifiers, leading to a virtuous circle for the whole learning process.



(a) mAP on NUS-WIDE-SCENE

(b) mAP on NUS-WIDE

(c) Detailed AP on NUS-WIDE-SCENE

(d) Detailed AP on NUS-WIDE

Figure 4: Effects of robustness on different datasets: (a) mAP on NUS-WIDE-SCENE; (b) mAP on NUS-WIDE; (c) Detailed AP of representative tags on NUS-WIDE-SCENE; and (d) Detailed AP of representative tags on NUS-WIDE.

4. CONCLUSIONS

In this paper, we investigated the problem of facilitating image tagging with “social assistance”. A long-standing obstacle for most of the existing image tagging approaches to achieve satisfactory performance is the scarcity of the high-quality training data. In order to handle this problem, we presented a new image tagging scheme, termed *social assisted media tagging* (SAMT), which uses the abundant web images associated with plentiful yet erroneous user-generated tags. We mainly addressed the following challenges, i.e., noisy tags associated to web images and the desirable robustness of the tagging model. In particular, we proposed a novel tag refinement approach based on low rank matrix recovery and sparse group lasso for identifying and eliminating tag noise and a robust tagging model based on $\ell_{2,p}$-norm for further alleviating the influence of noise and learning reliable image classifiers. Further, we designed a unified model which explores the in-depth reinforcement between the two components. Extensive experiments on two real-world social image databases illustrate the superiority of the proposed approach as compared to existing methods. In future, we will improve the scalability of the proposed framework to handle big multimedia data.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

5. REFERENCES

[1] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[2] T.-S. 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. In *CIVR*, 2009.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5, 2008.

[4] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint*, 2010.

[5] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *TIP*, 22(1):363–376, 2013.

[6] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *TMM*, 9(5):958–966, 2007.

[7] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *TCYB*, (99):1–1, 2013.

[9] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[10] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint*, 2010.

[11] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang. Tag ranking. In *WWW*, pages 351–360, 2009.

[12] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *NIPS*, 23:1813–1821, 2010.

[15] F. Nie, H. Huang, and C. H. Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *AAAI*, 2012.

[16] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.

[17] J. Tang, S. Yan, R. Hong, G. Qi, and T. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM MM*, pages 223–232, 2009.

[18] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua. Semantic-gap-oriented active learning for multilabel image annotation. *TIP*, 21(4):2354–2360, 2012.

[19] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[20] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu. Multimodal graph-based reranking for web image search. *TIP*, 21(11):4649–4661, 2012.

[21] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys (CSUR)*, 44(4):25, 2012.

[22] H. Xu, J. Wang, X. Hua, and S. Li. Tag refinement by regularized lda. In *ACM MM*, pages 573–576, 2009.

[23] Y. Yang, Z. Huang, H. T. Shen, and X. Zhou. Mining multi-tag association for image tagging. *WWW*, 14(2):133–156, 2011.

[24] Y. Yang, Z. Huang, Y. Yang, J. Liu, H. T. Shen, and J. Luo. Local image tagging via graph regularized joint group sparsity. *PR*, 46(5):1358–1368, 2013.

[25] Y. Yang, Y. Yang, Z. Huang, H. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, pages 881–888, 2011.

[26] Y. Yang, Y. Yang, and H. T. Shen. Effective transfer tagging from image to video. *TOMCCAP*, 9(2):14, 2013.

[27] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua. Visual query suggestion: Towards capturing user intent in internet image search. *TOMCCAP*, 6(3):13, 2010.

[28] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM MM*, pages 33–42. ACM, 2013.

[29] X. Zhang, Z. Huang, H. T. Shen, Y. Yang, and Z. Li. Automatic tagging by exploring tag information capability and correlation. *WWW*, 15(3):233–256, 2012.

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