Image Annotation Incorporating Low-Rankness, Tag and Visual Correlation and Inhomogeneous Errors

Yuqing Hou(✉)

Key Laboratory of Machine Perception (MOE), School of EECS,
Peking University, Beijing 100871, China
houyuqing1988@gmail.com

Abstract. Tag-based image retrieval (TBIR) has drawn much attention in recent years due to the explosive amount of digital images and crowdsourcing tags. However, TBIR is still suffering from the incomplete and inaccurate tags provided by users, posing a great challenge for tag-based image management applications. In this work, we propose a novel method for image annotation, incorporating several priors: Low-Rankness, Tag and Visual Correlation and Inhomogeneous Errors. Highly representative CNN feature vectors are adopted to model the tag-visual correlation and narrow the semantic gap. And we extract word vectors for tags to measure similarity between tags in the semantic level, which is more accurate than traditional frequency-based or graph-based methods. We utilize the Accelerated Proximal Gradient (APG) method to solve our model efficiently. Extensive experiments conducted on multiple benchmark datasets demonstrate the effectiveness and robustness of the proposed method.

1 Introduction

The prevalence of social network and digital photography in recent years makes image retrieval an urgent need. Image retrieval methods can be classified into two categories: content-based image retrieval (CBIR) and tag-based image retrieval (TBIR). The performance of CBIR algorithms are limited due to the semantic gap between the low-level visual features used to represent images and the high-level semantic meaning behind images. Tags can represent the semantics of images more precisely than low-level visual features, giving rise to research on TBIR.

However, tags are usually noisy and incomplete due to the arbitrariness of user tagging behaviors, leading to performance degradations of TBIR systems [1]. What's more, manual annotation is laborious, error prone, and subjective, making automatic image annotation an attractive research task.

Many machine learning methods have been developed for image annotation. They can be roughly grouped into three categories: supervised methods, unsupervised methods and semi-supervised methods.

G. Bebis et al. (Eds.): ISVC 2015, Part I, LNCS 9474, pp. 71–81, 2015.
DOI: 10.1007/978-3-319-27857-5_7

Supervised methods use the tagged images to train a dictionary of concept models and formulate image annotation as a supervised learning problem. They annotate images using the likelihood between images and tags. [2] formulates the annotation problem in a probabilistic framework and images are represented as bags of localized feature vectors. [3] learns a two-dimensional Multi-resolution Hidden Markov Model (2D-MHMM) on a fixed-grid segmentation of all category examples. [4] models image annotation procedure as a translation problem between image blobs and tags.

Unsupervised methods, e.g. search based-methods, learn the distribution of images and tags and annotate tags among clusters. Search-based methods always search in the feature space to find the most relevant images to the query image, and transfer tags to it using various tag transfer algorithms [5–7]. JEC [6] demonstrates that simple baseline algorithm can achieve high performance. TagProp [5] applies metric learning in the neighborhood of the feature space to annotate query images.

In recent years, semi-supervised approaches have been proposed in this field [8–10]. Semi-supervised algorithms can exploit the unlabelled information to improve the learning procedure and achieve satisfactory performance. [9] models the annotation task as a matrix completion problem, assuming the low-rankness property of the underlying matrix. [11] combined language model with matrix completion by assuming the independency of tags. Kernel trick and metric learning are exploited in [12] to capture the nonlinear relationships between visual features and semantics of the images. Semi-supervised relational topic model (ssRTM) is exploited to explicitly model the image content and their relations [13].

To utilize the large amount of unlabeled dataset for removing noisy tags and completing the missing ones, we propose a semi-supervised method. We formulate the annotation task as a transduction matrix completion problem, taking the following four priors into consideration:

1. **Low-rankness.** Many methods formulated the image annotation problem in a matrix completion framework by constructing and refining the image-tag matrix [8,10–12]. Existing works have demonstrated that semantic space spanned by tags can be approximated by a much smaller subset of words derived from the original space [14]. As text information, tags are consequently subjected to such subset property [8]. According to the subset property, we assume that the image-tag matrix is a low rank matrix. Thus we can exploit the low rank matrix completion techniques to complete the matrix, thereby annotating the images.

2. **Tag Correlation.** Tags have high level semantic meanings and often appear correlatively at the semantic level. However, traditional methods treat tags merely as labels, reducing the annotation task as a multi-label classification problem. In recent years, researchers have explored the relation among tags. Graph-based methods calculated the semantic correlation among tags using the WordNet distance [15]. Frequency-based methods [10,16] estimated tag correlations using Jaccard coefficient or co-occurrence in text search results. Jensen-Shannon divergence is introduced in the Flickr Distance to make the algorithm

more precise and reasonable [17]. However, these methods are still imprecise and inefficient. In this work, we utilize the vector representations for tags instead of labels. Word vectors [18], which are seldom used in this field, can present a much higher level semantic meanings than labels, thus we can measure the tag similarity much more easily and precisely.

3. **Tag-Visual Correlation.** Tag-Visual Correlation describes the correlation between the content level and the semantic level. Visually similar images often belong to similar themes and thus are annotated with similar tags. This prior has been widely explored in the image classification field [19,20]. However, there still exists a semantic gap between the content level and the semantic level. Traditional methods usually adopt low level image features, such as color, texture or shape descriptors, to represent the images, which are not so correlated with the semantic level. To narrow the semantic gap and make full use of the correlation property, we utilize high level visual features in our model, such as DeCAF$_6$, which demonstrate much stronger tag-visual correlation than low level visual features.

4. **Inhomogeneous Errors.** Tagging errors come from two aspects: missing tags and noisy tags. Since human-beings are relatively reasonable, we should assume that the tagging results are reasonably accurate. We can observe from the datasets that one image usually has relation with only a few tags, but we have to calculate its association with hundreds or even thousands of tags. For example, images from the MIRFlickr-25K have about 12.7 tags on average [21], but the dataset has $1,386$ unique tags, which means that each image should only be annotated with less than 1 % of all the tags. Hence users are more likely to adding noisy tags than missing noisy tags since there are too many unrelated tags. And the errors are mainly composed of noisy tags rather than missing tags. Thus we should treat these two errors with different strategies. We should put more emphasis on denoising rather than completing, paying more attention to the annotated tags rather than the unannotated ones. In other words, if an image is not originally annotated with a tag, it is more likely that they really have no relation at all.

Existing methods never model these two kinds of errors separately. They simply model the errors as Laplacian noise [8] or Gaussian noise [10]. To our knowledge, our model is the first to model the missing errors and noisy errors separately. The model can further adapt to different datasets according to their noise levels.

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

- We propose a new image annotation model that incorporates four priors: Low-rankness, Tag Correlation, Tag-Visual Correlation, and Inhomogeneous Errors.
- We utilize the word vectors and CNN features for the tag and the visual features, respectively. These high level features can narrow the semantic gap effectively. It is the first time to utilize both the features for image annotation.
- We model tag correlation and tag-visual correlation in different ways according to their semantic levels.

- We model two kinds of errors separately, the model can adapt to different datasets according to the noise level.
- We utilize the APG to solve our model efficiently.

The most related work to our model is LRES [8]. In their work, the authors formulated the image annotation task as a Robust PCA [22] framework, decomposing the original tag matrix into a refined tag matrix and a sparse error matrix. LRES also takes the tag correlation and tag-visual correlation into consideration and achieves good performance. However, our model is different from LRES in several aspects. First, our model measures tag correlation and tag-visual correlation using different models according to their different semantic levels, rather than using the same Graph Laplacian model in LRES. Second, we adopt more representative features such as CNN features and word vectors to narrow the semantic gap. Third, we do not model the error matrix simply as a sparse matrix, since thee errors are inhomogeneous and the distribution varies across different datasets.

2 Our Image Annotation Model

2.1 Low-Rankness

Denote the image collection $I = \{i_1, i_2, \ldots, i_m\}$, where m is the size of the image set. All original tags appearing in the set form a tag set $W = \{w_1, w_2, \ldots, w_n\}$, where n denotes the total number of unique tags. We can construct a binary matrix $\hat{T} \in \{0,1\}^{m \times n}$ whose element $\hat{T}_{i,j}$ indicates the relation between image i_i and tag w_j , i.e. if i_i is annotated with tag w_j, $\hat{T}_{i,j} = 1$, otherwise $\hat{T}_{i,j} = 0$. We use T to represent the refined tag matrix, where $T_{i,j} \in (0,1)$ the confidence score of assigning w_j to i_i. As mentioned above, we want the refined matrix T to be low rank. Since the low-rankness constraint on T is NP-hard to solve, we replace it with the standard relaxation, the trace norm, i.e. sum of singular values : $\|T\|_*$.

2.2 Tag Correlation Using Word Vectors

To narrow the semantic gap, we extract 300-dimensional word vectors [18] for each tag rather than treating tags merely as labels. Word vectors contain rich semantic information, e.g. semantic similarity. We denote the word vectors as $WV = \{wv_1, wv_2, \ldots, wv_n\}$. Given the completed tag matrix, T^i and T^j are the ith and jth columns of the tag matrix T. Thus we can measure the correlation between tag i and tag j in two ways: (1) similarity between word vectors wv_1 and wv_2, (2) similarity between tag vectors T^i and T^j.

The tag correlation prior can be enforced by solving the following optimization

$$\min_{T} \sum_{i=1}^{n} \sum_{j=1}^{n} \|T^i - T^j\|^2 S_{i,j}, \tag{1}$$

where $\|T^i - T^j\|^2$ measures the similarity between tag vectors T^i and T^j and $S_{i,j}$ measures the similarity between word vectors wv_i and wv_j. The formulation forces tag vectors with large similarities also have large similarity in their corresponding word vectors and vice versa, which essentially embodies the tag correlation prior.

The formulation can be rewritten as $Tr(T^T LT)$, where $L = diag(S^T 1) - S$ is the Graph Laplacian [23]. In our formulation, we define $S_{i,j} = \cos(wv_i, wv_j)$.

2.3 Tag-Visual Correlation Using CNN Features

The tag-visual correlation is not as strong as the correlation between tags owing to the semantic gap. Thus we just formulate the problem in a widely used model [10], which is much more simple and intuitive compared with the Graph Laplacian framework. Denote the image visual features as matrix V, where each visual image is represented as a row vector in $V \in R^{m \times f_v}$. Given the visual feature matrix, we can compute the visual similarity between i_i and i_j as $V_i^T V_j$, where V_i^T and V_j^T are the ith and jth rows of matrix V. Given the completed tag matrix, we can compute the similarity between i_i and i_j basing on the overlap between their corresponding tags, i.e., $T_i^T T_j$, where T_i^T and T_j^T are the ith and jth rows of the tag matrix T [10]. To model the aforementioned tag-visual correlation, we expect $|T_i^T T_j - V_i^T V_j|^2$ to be as small as possible. Thus we can model the tag-visual correlation using the Frobenius norm as $\sum_{i,j}^n |T_i^T T_j - V_i^T V_j|^2 = \|TT^T - VV^T\|_F^2$.

2.4 Inhomogeneous Errors

To model the inhomogeneous errors, we set different weight to the annotated positions and unannotated positions separately: $\lambda_0 \|P_\Omega(\hat{T} - T)\|_F^2 + \lambda_1 \|P_{\Omega^\perp}(\hat{T} - T)\|_F^2$. Ω represents the positions where the images are annotated with tags, P_Ω and $P_{\Omega^\perp}$ are projection operators, λ_0 and λ_1 are positive weighting parameters. λ_0 and λ_1 will change adaptively in different datasets according to their noisy levels. Different from the assumption of sparse errors [8], we model the errors using the Frobenius norm since we observe that large scale noisy datasets tend to be contaminated with dense Gaussian noises rather than Laplacian noises. Experiments on noisy datasets have confirmed our assumption.

2.5 Object Function Formulation: The Four Priors Model

Based on the terms regarding low-rankness, tag correlation, tag-visual correlation and inhomogeneous errors, we formulate the objective function as follows:

$$\begin{aligned} \min_T F(T) = \|T\|_* + \lambda_0 \|P_\Omega(\hat{T} - T)\|_F^2 + \lambda_1 \|P_{\Omega^\perp}(\hat{T} - T)\|_F^2 + \\ \lambda_2 Tr(T^T LT) + \lambda_3 \|TT^T - VV^T\|_F^2. \end{aligned} \tag{2}$$

λ_2 and λ_3 are also weighting parameters.

The proposed Four Priors method belongs to transductive learning category, which means it reasons from both labeled and unlabeled data. We can further turn it into a inductive model using traditional machine learning approaches [24].

3 Solving the Four Priors Model

We set $\lambda_0 = 1$ for computational efficiency, and denote the nuclear norm as $g(T)$ and the other terms together as $f(T)$. And $F(T) = g(T) + f(T)$, where $g(\cdot)$ is nonsmooth and $f(\cdot)$ is smooth. We pursuit an effective iterative procedure to solve this optimization based on Accelerated Proximal Gradient method (APG) [25].

Given the following unconstrained problem

$$\min_X F(X) = \mu g(X) + f(X). \tag{3}$$

where $g(\cdot)$ is nonsmooth, $f(\cdot)$ is smooth and its gradient is Lipschitz continuous. To avoid the computation of subgradient, proximal gradient algorithms minimize a sequence of separable quadratic approximations to $F(X)$, denoted as $Q(X, Y)$, formed at specially chosen points Y

$$Q(X, Y) \triangleq \mu g(X) + f(Y) + \langle \nabla f(Y), X - Y \rangle + \frac{L_f}{2} \|X - Y\|^2. \tag{4}$$

Let $M = Y - \frac{1}{L_f} \nabla f(Y)$, we get

$$X = \underset{X}{argmin}\, Q(X, Y) = \underset{X}{argmin} \{\mu g(X) + \frac{L_f}{2} \|X - M\|^2\}. \tag{5}$$

APG set $Y_k = X_k + \frac{b_{k-1}-1}{b_k}(X_k - X_{k-1})$ for a sequence $\{b_k\}$ satisfying $b_{k+1}^2 - b_{k+1} \leq b_k^2$ to get an $O(k^{-2})$ convergence rate. The APG method is described in Algorithm 1.

Algorithm 1. APG Method

```
Require:
while not converged do
    $Y_k = X_k + \frac{b_{k-1}-1}{b_k}(X_k - X_{k-1})$
    $M_k = Y_k - \frac{1}{L_f} \nabla f(Y_k)$
    $X_{k+1} = \underset{X}{argmin}\{\mu g(X) + \frac{L_f}{2} \|X - M_k\|^2\}$
    $b_{k+1} = \frac{1+\sqrt{4b_k^2+1}}{2}$
    $k = k + 1$
end while
Ensure:
```

The main advantage of the APG method is that the minimizer X_{k+1} has a simple or even closed-form solution when the $g(\cdot)$ is ℓ_1 norm or nuclear norm [8].

It is obvious that the APG method naturally fits for the Four Priors model. We estimate the L_f using backtracking method and calculate the $\nabla f(T)$:

$$\nabla f(T) = 2[P_{\Omega}^{*} P_{\Omega}(\hat{T} - T) + \lambda_1 P_{\Omega^{\perp}}^{*} P_{\Omega^{\perp}}(\hat{T} - T) + \lambda_2 LT + \lambda_3 (TT^T - VV^T)T] \quad (6)$$

where P_{Ω}^{*} and $P_{\Omega^{\perp}}^{*}$ are the adjoint operators of P_{Ω} and $P_{\Omega^{\perp}}$, respectively.

Basing on Eqs. (5) and (6) we can obtain the subproblem (Step 4 in Algorithm 1) for our model.

$$T_{k+1} = \underset{T}{argmin} \left\{ \|T\|_* + \frac{L_f}{2} \|T - M_k\|^2 \right\}, \quad (7)$$

where $M_k = T_k + \frac{b_{k-1}-1}{b_k}(T_k - T_{k-1}) - \frac{1}{L_f} \nabla f[T_k + \frac{b_{k-1}-1}{b_k}(T_k - T_{k-1})]$. The solution to (7) is:

$$T_{k+1} = US_{\frac{1}{L_f}}(\Sigma)V^T \quad (8)$$

where $U\Sigma V^T$ is the singular value decomposition (SVD) of M_k and $S_{\tau}(\cdot)$ is the singular value thresholding operator [26].

4 Experimental Evaluation

4.1 Datasets and Experimental Setup

The proposed algorithm is denoted as Four Priors and is evaluated on two well known benchmark datasets: MIRFlickr-25K , Corel5K and Labelme. MIRFlickr-25K is collected from Flickr. Compared to the Corel5K, tags in Labelme and MIRFlickr-25K are rather noisy and many of them are misspelled or meaningless words. Hence, a pre-processing is performed. We match each tag with entries in a Wikipedia thesaurus and only retain the tags in accordance with Wikipedia. We use the pre-trained word and phrase vectors [18] to extract tag vectors from the tags in these two datasets. To narrow the semantic gap, we utilized DeCAF [27] to extract the DeCAF$_6$ features, which have high level semantic meanings (Table 1).

We compare the proposed Four Priors model with the state-of-the-art methods, including matrix completion-based model LRES [8], TCMR [11], RKML [12], search-based algorithms (i.e. JEC [6], TagProp [5], and TagRelevance

Table 1. Statistics of 3 datasets

Statistics	Corel5K	Labelme	MIRFlickr-25K
No. of images	4,918	2,900	25,000
Vocabulary size	260	495	1,386
Tags per Image (mean/max)	3.4/5	10.5/48	12.7/76
Images per Tag (mean/max)	65.3/1,120	67.1/379	416.5/76,890

[7]), mixture models (i.e. CMRM [28] and MBRM [29]), tag recommendation approaches (i.e. Vote+ [30] and Folk [31]), co-regularized learning model FastTag [32] and Bayesian network model InfNet [33]. Note that the parameters of adopted baselines are also carefully tuned on the validation set of Corel5K with corresponding proposed tuning strategy.

We measure all the algorithms in terms of *average precision*@N (i.e. $AP@N$), *average recall*@N (i.e. $AR@N$) and *coverage*@N (i.e. $C@N$). In the top N completed tags, *precision*@N is to measure the ratio of correct tags in the top N competed tags and *recall*@N is to measure the ratio of missing ground-truth tags, both averaged over all test images. *Coverage*@N is to measure the ratio of test images with at least one correctly completed tag.

4.2 Evaluation of Tag Completion on Corel5*K*

We adopt the tuning strategy used in [10] to set $\lambda_1 = 0.6, \lambda_2 = 1$, and $\lambda_3 = 0.8$. Table 2 demonstrates the performance comparisons. Due to the space limit, we only report results when $N = 2, 3, 5, 10$.

4.3 Evaluation of Tag Completion on MIRFlickr-25*K* and Labelme

We tuned $\lambda_1 = 0.2, \lambda_2 = 1.0$, and $\lambda_3 = 0.5$ using cross validation on MIRFlickr-25K. The two datasets use the same parameters since they are both noisy. Note that as the datasets become large or noisy, the semantic gap expands, leading to the decrease of λ_3. And λ_1 varies according to different noisy level.

Table 2. Performance comparison on Corel5K dataset

	Corel5K											
	N = 2			N = 3			N = 5			N = 10		
	AP	AR	C	AP	AR	C	AP	AR	C	AP	AR	C
Four Priors	**0.58**	**0.42**	**0.52**	**0.50**	**0.50**	**0.62**	**0.42**	**0.60**	**0.66**	0.37	**0.65**	**0.92**
LRES [8]	0.56	0.39	0.47	0.48	0.48	0.57	0.41	0.53	0.62	0.37	0.62	0.85
TCMR [11]	0.57	0.39	0.49	0.48	0.47	0.58	0.44	0.55	0.64	**0.38**	0.62	0.88
RKML [12]	0.29	0.21	0.24	0.25	0.24	0.29	0.23	0.25	0.34	0.19	0.29	0.67
JEC [6]	0.36	0.34	0.39	0.31	0.40	0.47	0.27	0.32	0.59	0.20	0.33	0.76
TagProp [5]	0.46	0.40	0.50	0.38	0.48	0.57	0.33	0.51	0.63	0.26	0.54	0.86
TagRel [7]	0.43	0.41	0.48	0.37	0.47	0.57	0.31	0.50	0.60	0.26	0.53	0.90
CMRM [28]	0.29	0.20	0.23	0.24	0.24	0.27	0.21	0.25	0.35	0.16	0.27	0.63
MBRM [29]	0.35	0.29	0.35	0.28	0.34	0.42	0.24	0.24	0.39	0.17	0.28	0.70
FastTag [32]	0.54	0.31	0.45	0.46	0.44	0.51	0.40	0.52	0.63	0.36	0.63	0.82
Vote+ [30]	0.41	0.34	0.40	0.35	0.40	0.48	0.29	0.35	0.56	0.24	0.37	0.81
Folk [31]	0.29	0.29	0.34	0.22	0.34	0.41	0.20	0.24	0.41	0.18	0.30	0.61
InfNet [33]	0.26	0.19	0.24	0.20	0.22	0.29	0.17	0.24	0.30	0.12	0.19	0.64

Table 3. Performance comparison on Labelme dataset

	Labelme											
	N = 2			N = 3			N = 5			N = 10		
	AP	AR	C	AP	AR	C	AP	AR	C	AP	AR	C
Four Priors	**0.53**	**0.36**	**0.44**	**0.50**	**0.39**	**0.53**	**0.46**	**0.50**	**0.65**	**0.35**	**0.62**	**0.79**
LRES [8]	0.42	0.32	0.39	0.40	0.36	0.50	0.35	0.45	0.55	0.27	0.56	0.69
TCMR [11]	0.44	0.32	0.42	0.41	0.36	0.51	0.37	0.45	0.60	0.29	0.55	0.75
RKML [12]	0.21	0.14	0.20	0.20	0.16	0.21	0.19	0.20	0.23	0.14	0.22	0.28
JEC [6]	0.33	0.29	0.31	0.30	0.32	0.37	0.27	0.38	0.45	0.20	0.48	0.58
TagProp [5]	0.39	0.31	0.36	0.35	0.37	0.45	0.33	0.45	0.52	0.25	0.56	0.64
TagRel [7]	0.43	0.32	0.36	0.37	0.35	0.44	0.34	0.45	0.51	0.11	0.55	0.62
CMRM [28]	0.20	0.14	0.18	0.18	0.15	0.20	0.18	0.19	0.25	0.12	0.22	0.29
MBRM [29]	0.23	0.14	0.18	0.21	0.16	0.21	0.18	0.20	0.25	0.12	0.27	0.37
FastTag [32]	0.43	0.34	0.40	0.48	0.36	0.44	0.37	0.44	0.53	0.28	0.57	0.70
Vote+ [30]	0.32	0.28	0.32	0.31	0.30	0.38	0.28	0.38	0.47	0.20	0.50	0.60
Folk [31]	0.25	0.24	0.30	0.19	0.30	0.36	0.17	0.20	0.39	0.14	0.45	0.51
InfNet [33]	0.22	0.19	0.20	0.16	0.20	0.24	0.14	0.24	0.26	0.09	0.16	0.49

Table 4. Performance comparison on MIRFlickr-25K dataset

	MIRFlickr-25K											
	N = 2			N = 3			N = 5			N = 10		
	AP	AR	C	AP	AR	C	AP	AR	C	AP	AR	C
Four Priors	**0.55**	**0.39**	**0.46**	**0.50**	**0.43**	**0.55**	**0.40**	**0.47**	**0.61**	**0.31**	**0.61**	**0.80**
LRES [8]	0.43	0.35	0.40	0.40	0.39	0.53	0.32	0.40	0.57	0.26	0.45	0.73
TCMR [11]	0.45	0.35	0.44	0.43	0.38	0.54	0.35	0.41	0.60	0.28	0.48	0.77
RKML [12]	0.21	0.15	0.15	0.23	0.22	0.25	0.13	0.23	0.31	0.13	0.22	0.55
JEC [6]	0.33	0.30	0.32	0.31	0.38	0.45	0.25	0.34	0.55	0.19	0.35	0.66
TagProp [5]	0.39	0.35	0.39	0.36	0.42	0.51	0.28	0.37	0.59	0.20	0.41	0.73
TagRel [7]	0.42	0.34	0.37	0.37	**0.43**	0.52	0.30	0.37	0.57	0.20	0.40	0.78
CMRM [28]	0.20	0.15	0.16	0.18	0.21	0.24	0.13	0.18	0.30	0.11	0.20	0.50
MBRM [29]	0.22	0.16	0.18	0.17	0.30	0.35	0.13	0.18	0.33	0.10	0.22	0.55
FastTag [32]	0.43	0.35	0.38	0.39	0.48	0.51	0.30	0.41	0.57	0.27	0.42	0.75
Vote+ [30]	0.34	0.29	0.33	0.28	0.33	0.40	0.23	0.33	0.52	0.21	0.37	0.70
Folk [31]	-	-	-	-	-	-	-	-	-	-	-	-
InfNet [33]	-	-	-	-	-	-	-	-	-	-	-	-

Tables 3 and 4 demonstrate the performance comparisons. Note that Folk and InfNet is unable to run on the large dataset MIRFlickr-25K. Besides, search-based baselines (JEC, TagProp, and TagRel) cost a lot of time to run on the dataset.

4.4 Observations on Experimental Results

We observe that: (1) Generally algorithms achieve better performance on Corel5K, since tags in MIRFlickr-25K are more noisy. (2) Matrix completion-based methods, such as Four Priors, LRES and TCMR, usually achieve the best performances. (3) Four Priors shows increasing advantage to LRES as the data become more and more noisy, justifying our assumption and model of the noises. (4) Four Priors nearly outperforms all the other algorithms in all cases. (5) Performance on MIRFlickr-25K in some sense provides an evidence for the robustness of Four Priors.

5 Conclusions and Future Work

We have proposed an effective method for image annotation. The model takes four priors into consideration: Low-Rankness, Tag Correlation, Tag-Visual Correlation and Inhomogeneous Errors. This is the first work to model inhomogeneous errors in the image annotation field. We utilize word vectors to calculate tag correlation and CNN features to measure tag-visual correlation. It achieves the state-of-the-art performance in extensive experiments conducted on benchmark datasets for image annotation.

References

1. Ntalianis, K., Tsapatsoulis, N., Doulamis, A., Matsatsinis, N.: Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution. Multimedia Tools Appl. **69**, 397–421 (2014)
2. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 394–410 (2007)
3. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1075–1088 (2003)
4. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
5. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
6. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
7. Li, X., Snoek, C.G., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Trans. Multimedia **11**, 1310–1322 (2009)
8. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: ACM MM (2010)
9. Goldberg, A., Recht, B., Xu, J., Nowak, R., Zhu, X.: Transduction with matrix completion: three birds with one stone. In: NIPS (2010)

10. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 716–727 (2013)
11. Feng, Z., Feng, S., Jin, R., Jain, A.K.: Image tag completion by noisy matrix recovery. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 424–438. Springer, Heidelberg (2014)
12. Feng, Z., Jin, R., Jain, A.: Large-scale image annotation by efficient and robust kernel metric learning. In: ICCV (2013)
13. Niu, Z., Hua, G., Gao, X., Tian, Q.: Semi-supervised relational topic model for weakly annotated image recognition in social media. In: CVPR (2014)
14. Zhao, R., Grosky, W.I.: Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE Trans. Multimedia **4**, 189–200 (2002)
15. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: ACM MM (2005)
16. Cilibrasi, R.L., Vitanyi, P.: The google similarity distance. IEEE Trans. Knowl. Data Eng. **19**, 370–383 (2007)
17. Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr distance. In: ACM MM (2008)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
19. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1958–1970 (2008)
20. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In: CVPR (2006)
21. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: MIR 2008: Proceedings of the 2008 ACM ICMI (2008)
22. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**, 11 (2011)
23. Chung, F.R.: Spectral Graph Theory. American Mathematical Society, Providence (1997)
24. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: UAI (1998)
25. Toh, K.C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pac. J. Optimiz. **6**, 615–640 (2010)
26. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
27. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
28. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR (2003)
29. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: CVPR (2004)
30. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: ACM WWW (2008)
31. Lee, S., De Neve, W., Plataniotis, K.N., Ro, Y.M.: Map-based image tag recommendation using a visual folksonomy. Pattern Recogn. Lett. **31**, 976–982 (2010)
32. Chen, M., Zheng, A., Weinberger, K.: Fast image tagging. In: ICML (2013)
33. Metzler, D., Manmatha, R.: An inference network approach to image retrieval. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 42–50. Springer, Heidelberg (2004)