Social image tagging by mining sparse tag patterns from auxiliary data

Lin, Jie; Yuan, Junsong; Duan, Ling-Yu; Luo, Siwei; Gao, Wen

2012

Lin, J., Yuan, J., Duan, L. -Y., Luo, S., & Gao, W. (2012). Social image tagging by mining sparse tag patterns from auxiliary data. IEEE International Conference on Multimedia and Expo (ICME), 7-12.

<https://hdl.handle.net/10356/99569>

<https://doi.org/10.1109/ICME.2012.170>

© 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/ICME.2012.170>].

Downloaded on 26 Apr 2023 20:43:46 SGT

SOCIAL IMAGE TAGGING BY MINING SPARSE TAG PATTERNS FROM AUXILIARY DATA

Jie Lin^{†‡} Junsong Yuan^{*} Ling-Yu Duan[‡] Siwei Luo[†] Wen Gao[‡]

[†]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China

[‡]The Institute of Digital Media, School of EE&CS, Peking University, Beijing, 100871, China

^{*}School of EEE, Nanyang Technological University, 639798, Singapore

jie.dellinger@gmail.com jsyuan@ntu.edu.sg lingyu@pku.edu.cn swluo@bjtu.edu.cn wgao@pku.edu.cn

ABSTRACT

User-given tags associated with social images from photo-sharing websites (e.g., Flickr) are valuable auxiliary resources for the image tagging task. However, social images often suffer from noisy and incomplete tags, heavily degrading the effectiveness of previous image tagging approaches. To alleviate the problem, we introduce a Sparse Tag Patterns (STP) model to discover noiseless and complementary co-occurrence tag patterns from large scale user contributed tags among auxiliary web data. To fulfill the compactness and discriminability, we formulate the STP model as a problem of minimizing quadratic loss function regularized by bi-layer l_1 norm. We treat the learned STP as a universal knowledge base and verify its superiority within a data-driven image tagging framework. Experimental results over 1 million auxiliary data demonstrate superior performance of the proposed method compared to the state-of-the-art.

Index Terms— Social Image Tagging, Sparse Tag Pattern, Auxiliary Data, CBIR

1. INTRODUCTION

Image tagging (or automatic image annotation) - gives semantically relevant keywords to images - is of great interest as it allows one to retrieve, index, organize and understand large collections of image data. Previous work [1, 2, 3, 4] tried to capture the relationship between image features and keywords. However, such supervised learning methods require huge human-labeled training sets and are difficult to scale. Moreover, it is not applicable to personal photos due to the fixed and limited keyword vocabularies.

Recently, the emerging social media (e.g., Flickr) accommodate enormous auxiliary images augmented with unlimited vocabularies including user-given tags, timestamps, GPS, etc., and benefit a wide variety of potential applications such as annotation [5, 6, 7]. Meanwhile, auxiliary data from the web avoids the time-consuming human labeling effort and prevents scalability issues. With the aid of auxiliary data, Makadia et.al [8] introduced a simple baseline technique for image tagging that treats annotation as a data-driven problem.

Fig. 1. Examples of social images associated with user-given tags from Flickr. Tags in green and red color indicate noisy and incomplete tags, respectively.

That is, it first finds nearest neighbors of a given query image, then the keywords associated with its neighbors are transferred to the query image in a greedy manner. This baseline method [8] outperforms the current state-of-the-art methods [1, 2, 3, 4] on two standard datasets (i.e., Corel5K, IAPR-TC 12) and one large web dataset (i.e., ESP). However, it assumes that auxiliary images have accurate annotations, which is not always true in the context of social media. Fig. 1 depicts an example, where social images often suffer from noisy and incomplete tags, heavily degrading the performance of the baseline method (see Section 5).

In this paper, we introduce a Sparse Tag Patterns (STP) model to alleviate above mentioned problem. The intuition for STP is that social images belong to the same semantic concept are mutually complementary in user-given tags. The goal is to mine the readable co-occurrence *tag pattern* of each latent concept from huge user contributed tags among auxiliary data. With reasonable assumptions that (1) each tag pattern is relevant to a few mutually complementary tags and (2) user-given tags associated with each image is relevant to a few tag patterns, we formulate the STP model as a problem of minimizing a quadratic loss function with bi-layer l_1 norm sparsity constraints. We argue that the learned STP yields compact and discriminative intermediate semantic representations. Specifically, each sparse tag pattern efficiently suppresses noisy user-given tags, and meanwhile promotes the importance of incomplete tags.

As the learned STP stands for collective intelligence from web data, we treat it as a *universal* knowledge base and verify its superiority within a data-driven image tagging framework.

Instead of transferring tags to the query image directly [8], we propose to first describe it as short text by aggregating user-given tags associated with its K -nearest neighbor images from auxiliary data via Content Based Image Retrieval (CBIR), then find the most relevant tag patterns for the short text using the learned STP. The representative tags associated with the matching tag patterns are then recommended to the query image. Fig. 2 depicts the pipeline of our social image tagging system.

We evaluate the proposed method using more than 1 million auxiliary social images associated with user-given tags. To the best of our knowledge, this is the largest experiments conducted on user-given tags analysis for social image tagging. Experimental results show that the STP model leads to superior image tagging performance, compared to the state-of-the-art approach [8].

The rest of this paper is organized as follows. We briefly review related work in Section 2. Section 3 presents the formulation of the STP model and the optimization algorithm. Section 4 describes the proposed image tagging method. Experimental evaluation is given in Section 5. We draw conclusions in Section 6.

2. RELATED WORK

Studies on image tagging roughly fall into model-based and example-based methods. The *model-based methods* typically learn models from labeled samples of a pre-defined set of concepts. The learned model are then used to annotate new image according to its relevance to the concepts. Cross Media Relevance Models [1], Continuous Relevance Model [2], and Multiple Bernoulli Relevance Model [3] assume different, nonparametric density representations of the joint word-image space. Carneiro et.al [4] proposed Supervised Multi-class Labeling technique to learn class-conditional densities using the training data where each keyword is considered as a class. The *example-based methods* assume that images with similar visual content are annotated by similar tags. For a given image to be annotated, tags associated with its nearest neighbors based on the visual similarity are recommended [8]. Our proposed method belongs to the example-based approach. In contrast to [8], we exploit the collective knowledge of tag patterns embedded in large scale user-provided tags to prevent imprecise tag recommendations.

Besides that, there exist image retagging methods [6] that aim to refine annotation of an image associated with user-given tags. Wang et.al [9] proposed to annotate an input image associated with an initial keyword by leveraging the surrounding text of images crawled from the web. Li et.al [10] measured tag relevance to the image content by using neighbor voting. In this paper, we focus on annotating a query image when its user-given tags are unavailable.

Co-occurrence patterns mining have recently received a lot of attention in object, scene, and action recognition tasks,

Fig. 2. Pipeline of our proposed method.

because they are more discriminative than individual features. In [11, 12], the authors proposed to discover discriminative co-occurrence patterns based on only visual content. While in our work we are more interested in understanding social images in the context of user-provided tags.

3. SPARSE TAG PATTERNS

3.1. Formulation of the STP Model

We denote auxiliary data as $\Psi = \{(I_1, \mathbf{D}_1), \dots, (I_N, \mathbf{D}_N)\}$ consisting of N image-short text pairs, each short text $\mathbf{D}_n = [b_{n1}, \dots, b_{nM}]^T \in \mathbb{R}^M$ represents user-given tags of the n^{th} social image I_n , M is the total number of unique tags appeared among Ψ , each element b_{nm} is a binary value indicating whether the m^{th} tag belongs to I_n ($b_{nm} = 1$) or not ($b_{nm} = 0$). For brevity, we represent the N short texts by matrix $\mathbf{D} = [\mathbf{D}_1, \dots, \mathbf{D}_N]^T = [\mathbf{D}'_1, \dots, \mathbf{D}'_M] \in \mathbb{R}^{N \times M}$.

Our goal is to handle noisy and incomplete user-given tags within social images for boosting the accuracy of image tagging (see Section 4). Intuitively, social images with similar visual content are mutually complementary in user-given tags. Fig. 3 depicts an example, user-given tag *red* from image (b) and (c) complements image (a), while *rock* from image (a) and (c) enriches image (b). On the one hand, the mutually complementary tags effectively suppress noisy tags with low co-occurrences among themselves (e.g., the importance of tag *rain* for image (a) is decreased as it did not appear in image (b) and (c)). On the other hand, they also promote the importance of incomplete tags for an image (e.g., *red* for image (a)). We define *tag pattern* as a combination of mutually complementary tags, where each tag pattern represents a *concept*. As shown in Fig. 3, the tag pattern $\{\text{rock, park, landscape, red}\}$ expresses the concept *hiking*. Ideally, each tag pattern should contain a few tags in the entire vocabulary. Thus, the intersection of any two tag patterns is empty or small. Meanwhile, the short text associated with each image should involve as fewer semantic concepts as possible. Based on these two *s-sparsity* constraints, we formulate the STP model via the l_1

Fig. 3. Illustration of mutually complementary tags among social images with similar visual content. See text for details.

regularization as in the lasso model [13]:

$$\min_{\mathbf{U}, \mathbf{A}} \frac{1}{2} \|\mathbf{D} - \mathbf{U}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1 + \beta \|\mathbf{U}\|_1, \quad (1)$$

where $\mathbf{U} = [\mathbf{U}_1, \dots, \mathbf{U}_N]^T = [u_{nz}] \in \mathbb{R}^{N \times Z}$, $z = 1, \dots, Z$ denotes the *short text-tag pattern* matrix. The objective is to compute the *tag pattern-tag* matrix $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_Z]^T = [\mathbf{A}'_1, \dots, \mathbf{A}'_M] = [a_{zm}] \in \mathbb{R}^{Z \times M}$ that $\mathbf{U}\mathbf{A}$ leads to the best reconstruction of $\mathbf{D}$. λ and β denote the positive regularization parameters controlling the density (the number of non-zero entries) of $\mathbf{A}$ and $\mathbf{U}$, respectively. Larger λ (or β) leads to sparser $\mathbf{A}$ (or $\mathbf{U}$). For each row of $\mathbf{A}$, the m^{th} entry a_{zm} denotes the weight of the m^{th} tag in the z^{th} tag pattern. Specifically, the tags with larger weights are more representative in the corresponding tag pattern. For each row of $\mathbf{U}$, the z^{th} entry u_{nz} represents the weight of the z^{th} tag pattern for the n^{th} short text. The larger u_{nz} is, the more important role the z^{th} tag pattern plays in representing the n^{th} short text. With the sparsity constraints on both $\mathbf{A}$ and $\mathbf{U}$, the STP model yields compact yet discriminative representation for both tag pattern-tag and short text-tag pattern relationships.

3.2. Optimization Algorithm

The optimization problem of Eq. 1 is non-convex. But fixing one variable (either $\mathbf{A}$ or $\mathbf{U}$), the objective function with respect to the other is convex. So we alternately minimize Eq. 1 with respect to $\mathbf{A}$ or $\mathbf{U}$. Algorithm 1 summarizes the optimization procedure.

Update A. When $\mathbf{U}$ is fixed, the optimization problem with respect to $\mathbf{A}$:

$$\min_{\mathbf{A}} \frac{1}{2} \|\mathbf{D} - \mathbf{U}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1,$$

can be decomposed into M independent subproblems, each corresponding to one column of $\mathbf{A}$:

$$\min_{\mathbf{A}'_m} \frac{1}{2} \|\mathbf{D}'_m - \mathbf{U}\mathbf{A}'_m\|_2^2 + \lambda \|\mathbf{A}'_m\|_1. \quad (2)$$

Each subproblem in Eq. 2 is a standard lasso problem, thus, we choose a coordinate descent technique [14] to solve it.

Update U. Likewise, the update of $\mathbf{U}$ with $\mathbf{A}$ fixed can be decomposed into N independent subproblems, each corresponding to one row of $\mathbf{U}$:

$$\min_{\mathbf{U}_n} \frac{1}{2} \|\mathbf{D}_n - \mathbf{A}^T \mathbf{U}_n\|_2^2 + \beta \|\mathbf{U}_n\|_1, \quad (3)$$

which executes the similar optimization procedure as Eq. 2.

Algorithm 1 Optimization Algorithm for STP

- 1: **Input:** $\mathbf{D} \in \mathbb{R}^{N \times M}$, Z, λ, β ,
 - 2: **Initialization:** random matrix $\mathbf{U}^0 \in \mathbb{R}^{N \times Z}$
 - 3: **Iterate until convergence of A and U**
 - 4: Update $\mathbf{A}$ by solving M lasso problems as in Eq. 2
 - 5: Update $\mathbf{U}$ by solving N lasso problems as in Eq. 3
 - 6: **Output:** Sparse matrix $\mathbf{A}$ and $\mathbf{U}$
-

3.3. Discussion

In this section, we briefly review the most related methods and point out the difference from our model.

Sparse Latent Semantic Analysis (LSA). LSA [15] aims to derive a mapping that projects input feature space $\mathbb{R}^M$ into a Z -dimensional latent space where $Z \ll M$, which can be achieved by solving the following optimization problem:

$$\min_{\mathbf{U}, \mathbf{A}} \frac{1}{2} \|\mathbf{D} - \mathbf{U}\mathbf{A}\|_F^2 \quad (4)$$

subject to $\mathbf{U}^T \mathbf{U} = \mathbf{I},$

where $\mathbf{U}$ denotes Z uncorrelated latent variables with the orthogonality constraint $\mathbf{U}^T \mathbf{U} = \mathbf{I}$, $\mathbf{I}$ is the identity matrix. Based on Eq. 4, Chen et.al [16] added the sparsity constraint of the projection matrix $\mathbf{A}$:

$$\min_{\mathbf{U}, \mathbf{A}} \frac{1}{2} \|\mathbf{D} - \mathbf{U}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1 \quad (5)$$

subject to $\mathbf{U}^T \mathbf{U} = \mathbf{I}$

Sparse LSA has the ability to automatically select the most relevant tags for each latent dimension.

Sparse Coding. Sparse Coding [17] aims to learn basis functions $\mathbf{A}$ that capture higher-level features in the data. In our notations, Sparse Coding can be modeled as:

$$\min_{\mathbf{U}, \mathbf{A}} \frac{1}{2} \|\mathbf{D} - \mathbf{U}\mathbf{A}\|_F^2 + \lambda \|\mathbf{U}\|_1 \quad (6)$$

subject to $\|\mathbf{A}'_m\|_2^2 \leq c, \quad \forall m = 1, \dots, M,$

where c is constant, $\mathbf{A}$ is called dictionary, each row of $\mathbf{A}$ denotes a basis function; $\mathbf{U}$ are the coefficients. An l_2 -norm constraint on $\mathbf{A}'_m$ is typically applied to avoid trivial solutions. By enforcing the sparsity on $\mathbf{U}$, Sparse Coding selects the most relevant basis functions for each sample.

In contrast with Sparse LSA and Sparse Coding, our model combines them into a unified framework by using the bi-layer l_1 regularization on both $\mathbf{A}$ and $\mathbf{U}$. Another difference is that our model relax the orthogonality constraint $\mathbf{U}^T \mathbf{U} = \mathbf{I}$ and $\|\mathbf{A}'_m\|_2^2 \leq c$ of Eq. 5 and Eq. 6, respectively.

4. SOCIAL IMAGE TAGGING

After obtaining matrix $\mathbf{A}$, given an unlabeled query image I_q , we aim to recommend relevant tags to it. Under the assumption that images with similar visual content are likely to share common tags [8, 10], we employ a data-driven framework integrated with the learned STP for social image tagging (see Fig. 2).

4.1. Generate Short Text $\mathbf{D}_q$ via CBIR

We extract L visual features $f^1, \dots, f^L$ for each image. We denote $d^l(I_q, I_n)$ as normalized distance between I_q and the n^{th} social image on the l^{th} feature, $l = 1, \dots, L$. We use the linear combination with equal weights to fuse distances for various visual features:

$$d(I_q, I_n) = \sum_{l=1}^L \frac{d^l(I_q, I_n)}{L}. \quad (7)$$

We denote the K -nearest neighbor social images of I_q from Ψ as $(I_1^q, \mathbf{D}_1^q), \dots, (I_K^q, \mathbf{D}_K^q)$, ordering by increasing $d(I_q, I_n)$. Then we represent I_q as short text $\mathbf{D}_q$:

$$\mathbf{D}_q = \sum_{k=1}^K \mathbf{D}_k^q. \quad (8)$$

4.2. Tag Recommendation via the Learned STP

We first project $\mathbf{D}_q$ into $\mathbf{U}_q = [u_{q1}, \dots, u_{qZ}]^T \in \mathbb{R}^Z$ using Eq. 3. $\mathbf{U}_q$ tends to be sparse, i.e., there are Z' ($Z' \ll Z$) non-zero entries of $\mathbf{U}_q$, indicating that only Z' tag patterns are relevant to $\mathbf{D}_q$. We choose the corresponding rows of $\mathbf{A}$ according to the Z' non-zero entries and denote them as $\{\mathbf{A}_z^q\}_{z=1}^{Z'}$. Each $\mathbf{A}_z^q = [a_{z1}, \dots, a_{zM}]^T \in \mathbb{R}^M$ is also sparse. We join the tags with non-zero values a_{zm} in each $\mathbf{A}_z^q$ into a union set $\mathbf{W}_q = \{w_{q1}, \dots, w_{qM}\}$. Each entry w_{qm} can be computed as:

$$w_{qm} = \sum_{z=1}^{Z'} u_{qz} a_{zm}. \quad (9)$$

Then tags with larger values w_{qm} are propagated to the query image I_q . Due to the sparsity property on both $\mathbf{U}_q$ and $\mathbf{A}_z^q$, $\mathbf{W}_q$ is tailored to each query image I_q with only a few non-zero entries w_{qm} , producing personalized and precise recommended tags.

5. EXPERIMENTS

5.1. Datasets

Test Data. We construct our test data set using 15 concepts, including *babyshower*, *beach*, *bike*, *birthday*, *camping*, *Christmas*, *concert*, *graduation*, *Halloween*, *hiking*, *skiing*,

soccer, *softball*, *swimming* and *wedding*. The 15 concepts are carefully selected such that they (1) belong to different categories including object, scene, and event, (2) correspond to the most popular tags in Flickr, and (3) have both abstract concepts such as *wedding* and specific concepts such as *bike*. We collected 8770 test images with associated user-given tags from Flickr, ranging from 289 to 750 images for each concept. To evaluate the effectiveness of the image tagging method, we carefully built the ground-truth annotations for each image by manually removing irrelevant tags from its user-given tags. Finally, there are 6.6 tags per test image on average.

Flickr1M Auxiliary Data. We randomly crawled over 1 million social images with associated user-given tags from Flickr using the most popular tags¹ (including the pre-defined 15 concepts) as query keywords. There were some overlaps between test images and social images, so duplicates were removed, resulting a data set with 1,161,499 images and 428,348 unique tags. Many of the unique tags were misspelling and meaningless. We removed low-occurrences tags and then filtered out tags that did not match with the entries in Wikipedia thesaurus. After that, there are a total of 9586 unique tags and 5.8 tags per image on average. We refer to the auxiliary data as **Flickr1M**. Note that the ground-truth annotations of test data are a subset of Flickr1M.

5.2. Experimental Setup

Visual Features. We extract both efficient global and local visual features, including 108-dimensional grid color moments, 320-dimensional Canny edge histogram, and 1000-dimensional Bag of Words (BoW). Specifically, BoW uses Difference of Gaussian as interest point detector and SIFT as descriptor. We randomly sample 10M SIFT descriptors and use Integer k-means clustering for visual codebook construction. To accelerate K-nearest neighbor computation on Flickr1M, we adopt randomized kd-tree forest from VLFeat².

Baseline. We compare the proposed method with the state-of-the-art image annotation algorithm [8]. We choose it as the baseline method because it outperforms previous work [1, 2, 3, 4]. Moreover, it is comparable to our method in a data-driven manner while other approaches [1, 2, 3, 4] are only applicable to small-size data set.

Parameters and Evaluation. We choose parameter Z from a set of values $\{30, 50, 80, 100, 200, 400\}$. Regularization parameters λ and β are adjusted in the interval $[0.01, 1]$ and $[0.01, 1]$, respectively. In subsequent experimental results, we select the optimal parameters $Z = 50$, $\lambda = 0.5$, and $\beta = 0.1$. Due to limited space, the detail effects of STP parameters is included in our future work. For evaluation, we use Average Precision (AP) and Average Recall (AR) at top t recommended tags as performance metrics.

¹<http://www.flickr.com/photos/tags/>

²<http://www.vlfeat.org>

(a) Average Precision

(b) Average Recall

Fig. 4. Quantitative comparison between our method and the state-of-the-art Baseline [8] approach.

5.3. Experimental Results

Sparse Tag Patterns. In this section, we qualitatively show that the sparse tag pattern-tag relationship learned by the STP model using Flickr1M data set. We vary the value of regularization parameter λ so that each tag pattern has at least 20 tags with non-zero values a_{zm} . The top tags for some randomly selected sparse tag patterns are listed in Table 1. We observed that (1) the learned tag patterns cover diverse concepts not restricted to the pre-defined 15 concepts in this paper (e.g., TP 2). Thus, the learned STP can be treated as an universal knowledge base and generalized to broad concepts, (2) the representative tags of each tag pattern are noiseless and complementary to each other, (3) tag pattern can join synonyms tags together (e.g., *motorcycle*, *motorbike* and *moto* in TP 42) and split polysemy tags away (e.g., *park* in TP 15 and TP 30). However, there also exist few meaningless sparse tag patterns (e.g., TP 38).

Quantitative Comparison. Fig. 4 (a) and (b) report AP and AR at top t recommended tags when we set $K = 10$, respectively. We have the following observations (1) the proposed method outperforms the Baseline method on both AP and AR consistently, (2) as shown in Fig. 4 (a), the proposed method achieves significant better AP compared to the Baseline method when t is small, especially for AP@1 (26.6% vs. 14.9%). It verifies that the STP model promotes the importance of lower ranked relevant tags or incomplete tags that not appeared in $\mathbf{D}_q$, (3) as shown in Fig. 4 (b), the proposed method achieves significant better AR compared to the Base-

Fig. 5. Effect of varied K on Average Precision.

	Ground-truth	Baseline	Our method
	bicycle cycle biking	toy holga canal biking urban	biking street urban city people
	football soccer	sport running waterfront soccer football	football soccer sport green ball
	water pool swimming dog	nature beach caribbean reef snorkeling	water pool animal nature swimming
	bride groom dress forest tree bridge	portrait wedding napa park hike	park tree girl wedding family

Fig. 6. Qualitative comparison of some social image tagging results. Besides the ground-truth annotations, we list top 5 recommended tags for each image. Tags in green color are irrelevant to the corresponding image, while tags in red color are relevant to the corresponding image but not appeared in the ground-truth annotations.

line method when t is large, especially for AR@15 (24.1% vs. 18.8%). It validates that the STP model efficiently filters out high frequency but low co-occurrence noisy tags.

Effect of Varied K . Fig. 5 reports the average precision of the proposed method at top t recommended tags with varied K ranging from 10 to 50. The performance improves consistently for $K < 30$ and slows down (or even drops down) for $K > 30$. If K is too large, lots of noisy tags may be included as there exist many irrelevant images among the nearest neighbors. However, if K is too small, some relevant tags may not appear. In both situations, the inferred $\mathbf{U}_q$ may be inaccurate and degrade the performance.

Qualitative Comparison. Fig. 6 examines some social image tagging results. we observe that the proposed method recommends more accurate tags compared to the Baseline approach, especially for the incomplete tags (in red color). Note that the ground-truth annotations for each test image are generated by manually removing irrelevant tags from user-given

Table 1. Sparse Tag Patterns. “TP” denotes tag pattern.

TP 2	airplane, flying, airport, plane, aircraft, flight, museum, air, jet, aviation, aeroplane, military
TP 5	halloween, death, costume, party, pumpkin, dead, cemetery, scary, skull, grave, ghost, orange
TP 15	animal, bird, zoo, wildlife, nature, lion, bear, park , panda, monkey, tiger, wild, elephant, fish
TP 16	christmas, light, red, tree, holiday, color, green, blue, white, decoration, xmas, orange, winter
TP 22	snow, winter, skiing, mountain, ice, cold, switzerland, lake, landscape, white, storm, hill, district
TP 24	water, pool, swimming, reflection, blue, summer, lake, hotel, beach, sun, wet, river, green, rain
TP 29	food, picnic, restaurant, dinner, wine, kitchen, bbq, friend, fruit, family, party, cooking, beer, bottle
TP 30	park , baseball, softball, woman, city, event, rebels, game, fun, league, people, fair, group, man
TP 38	day, angel, life, from, love, world, hand, big, labor, time, weekend, you, one, fun, long, god, place
TP 42	bike, bicycle, motorcycle , racing, cycling, road, motorbike , ride, park, gear, street, moto , wheel

tags, many recommended tags that relevant to the test image but not appeared in the ground-truth are improperly ignored.

6. CONCLUSIONS

To attack the problem of social image tagging, we propose an STP model to exploit sparsity constrained co-occurrence tag patterns from huge user-given tags of social images, yielding compact and discriminative intermediate semantic representations. The learned STP can be regarded as a universal knowledge base and integrated into a data-driven framework for boosting the accuracy of social image tagging. Experimental results verify our ideas and show that the proposed method outperforms the state-of-the-art approach. In future work, we aim to systematically (1) study the effect of parameters of the STP model and (2) compare our model to other related methods such as Sparse Coding, Sparse LSA, etc.

7. ACKNOWLEDGEMENT

This work was supported by National Basic Research Program of China (2009CB320902), in part by Chinese National Nature Science Foundation (60902057) and Nanyang Assistant Professorship SUG M4080134.

8. REFERENCES

- [1] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *ACM SIGIR*, 2003, pp. 119–126.
- [2] Victor Lavrenko, R. Manmatha, and Jiwoon Jeon, “A model for learning the semantics of pictures,” in *NIPS*, 2004.
- [3] S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *CVPR*, 2004, pp. 1002–1009.
- [4] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” *IEEE Transactions on PAMI*, vol. 29, pp. 394–410, 2007.
- [5] Borkur Sigurbjornsson and Roelof van Zwol, “Flickr tag recommendation based on collective knowledge,” in *WWW*, 2008, pp. 327–336.
- [6] Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang, “Image retagging,” in *ACM Multimedia*, 2010, pp. 491–500.
- [7] Junsong Yuan, Jiebo Luo, and Ying Wu, “Mining compositional features from gps and visual cues for event recognition in photo collections,” *IEEE Transactions on Multimedia*, vol. 12, pp. 705–716, 2010.
- [8] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar, “Baselines for image annotation,” *Int. J. Comput. Vision*, vol. 90, pp. 88–105, 2010.
- [9] Xin-Jing Wang, Lei Zhang, Feng Jing, and Wei-Ying Ma, “Anosearch: Image auto-annotation by search,” in *CVPR*, 2006, pp. 1483–1490.
- [10] Xirong Li, Cees G. M. Snoek, and Marcel Worring, “Learning social tag relevance by neighbor voting,” *IEEE Transactions on Multimedia*, vol. 11, pp. 1310–1322, 2009.
- [11] Junsong Yuan, Ming Yang, and Ying Wu, “Mining discriminative co-occurrence patterns for visual recognition,” in *CVPR*, 2011, pp. 2777–2784.
- [12] Shen-Fu Tsai, Liangliang Cao, Feng Tang, and Thomas S. Huang, “Object pattern: a new model for album event recognition,” in *ACM Multimedia*, 2011.
- [13] Robert Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.
- [14] H. Hofling et. al. J. Friedman, T. Hastie, “Pathwise coordinate optimization,” in *ANN APPL STAT*, 2007.
- [15] S. Deerwester, S.T. Dumais, T. Landauer, G.W. Furnas, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society of Information Science*, vol. 41, pp. 391–407, 1990.
- [16] X. Chen, Y. Qi, B. Bai, Q. Li, and J. Carbonell, “Sparse latent semantic analysis,” in *SDM*, 2011.
- [17] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng, “Efficient sparse coding algorithms,” in *NIPS*, 2007, pp. 801–808.