Learning sparse tag patterns for social image classification

Duan, Ling-Yu; Yuan, Junsong; Li, Qingyong; Luo, Siwei; Lin, Jie

2012

Lin, J., Duan, L.-Y., Yuan, J., Li, Q., & Luo, S. (2012). Learning sparse tag patterns for social image classification. 19th IEEE International Conference on Image Processing (ICIP), 2881-2884.

https://hdl.handle.net/10356/79987

https://doi.org/10.1109/icip.2012.6467501

LEARNING SPARSE TAG PATTERNS FOR SOCIAL IMAGE CLASSIFICATION

Jie Lin*‡ *Ling-Yu Duan*‡ *Junsong Yuan*† *Qingyong Li** *Siwei Luo**

*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
jielinbjtu@gmail.com lingyu@pku.edu.cn jsyuan@ntu.edu.sg {liqy, swluo}@bjtu.edu.cn

ABSTRACT

User-generated tags associated with images from social media (e.g., Flickr) provide valuable textual resources for image classification. However, the noisy and huge tag vocabulary heavily degrades the effectiveness and efficiency of state-of-the-art image classification methods that exploited auxiliary web data. To alleviate the problem, we introduce a Sparse Tag Patterns (STP) model to discover sparsity constrained co-occurrence tag patterns from large scale user contributed tags among social data. To fulfill the compactness and discriminability, we formulate STP as a problem of minimizing a quadratic loss function regularized by the bi-layer l_1 norm. We treat the learned STP as alternative intermediate semantic image feature and verify its superiority within a search-based image classification framework. Experiments on 240K social images associated with millions of tags have demonstrated encouraging performance of the proposed method compared to the state-of-the-art.

Index Terms— Social Data, Sparse Tag Patterns, Image Classification, CBIR

1. INTRODUCTION

With the ever decreasing cost of digital cameras, there are large volumes of images created in our daily life. How to organize and index them automatically remains an important challenge. Previous work [1, 2, 3] tried to train visual classifiers for each concept then use them to label the image test examples. However, such supervised learning methods require large collections of manually labeled training samples and are difficult to scale.

Recently, many studies exploited freely available internet images to facilitate image classification, avoiding the time-consuming human labeling effort[4, 5, 6]. Given a large enough image dataset, it is bound to find very similar images to a query image, even when matching with simple visual features. Torralba et al. [7] verified this observation with a search-based scheme using Euclidean distance of intensity that leads to surprisingly good object recognition results on 80 million tiny images. Furthermore, the text that surrounds internet images (e.g., tags) also provides additional semantic features for image representation. Wang et al. [8] introduced a text-based image feature and demonstrate that it consistently improves performance on object classification problems.

In spite of that, we observe the following issues as leveraging web resources for image classification. On the one hand, due to ambiguous, and incomplete or spurious user-generated tags, the noisy text features can greatly affect the learning performance. On the other hand, the huge tag vocabulary causes large textual feature space and makes it inefficiently to train text classifiers.

In this paper, we introduce a Sparse Tag Patterns (STP) model to alleviate the problem. The intuition for STP is that social images belong to the same semantic concept are mutually complementary in user-generated tags. Our goal is to mine these readable co-occurrence *tag pattern* of each latent concept from large scale user contributed tags among social data. With reasonable assumptions that (1) each tag pattern is relevant to a few mutually complementary tags and (2) user-generated 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 the bi-layer l_1 norm sparsity constraints. Specifically, the learned STP yields compact yet discriminative low-dimensional intermediate semantic image features.

We employ the learned STP as alternative semantic feature and verify its superiority within a search-based image classification framework. Instead of training text classifiers directly [8], we propose to first describe each image as text by aggregating user-generated tags associated with its K-nearest neighbor images from social data via Content Based Image Retrieval (CBIR), then find the most relevant tag patterns for the text using the learned STP. Finally, we use the tag pattern features to train classifiers. Fig. 1 depicts the pipeline of our social image classification system.

We evaluate the proposed method on a large scale dataset with 240K social images and millions of user-generated tags. Experimental results show that the STP model leads to superior image classification results, compared to the state-of-the-art approach [8].

2. SPARSE TAG PATTERNS

2.1. Problem Formulation

We denote social data as $\Psi = \{(I_1, \mathbf{D}_1), ..., (I_N, \mathbf{D}_N)\}$ consisting of N image-text pairs. Each text $\mathbf{D}_n = [b_{n1}, ..., b_{nM}]^T \in \mathbb{R}^M$ represents user-generated tags of the n^{th} social image I_n, where M is the total number of unique tags appeared among Ψ, and 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 texts by matrix $\mathbf{D} = [\mathbf{D}_1, ..., \mathbf{D}_N]^T = [\mathbf{D}_1^{'}, ..., \mathbf{D}_M^{'}] \in \mathbb{R}^{N\times M}$.

Our goal is to handle noisy user-generated tags for producing more effective and efficient intermediate image representations and boosting image classification results. Intuitively, social images with similar visual content are mutually complementary in user-generated tags. Be analogous to visual pattern [9, 10], we define *tag pattern* as a combination of mutually complementary tags, where each tag pattern represents a *concept*. As shown in Fig. 2, the tag pattern {*rock, park, landscape, red*} expresses the concept *hiking*. Ideally, each tag pattern should contain a few distinct tags in the entire vocabulary.

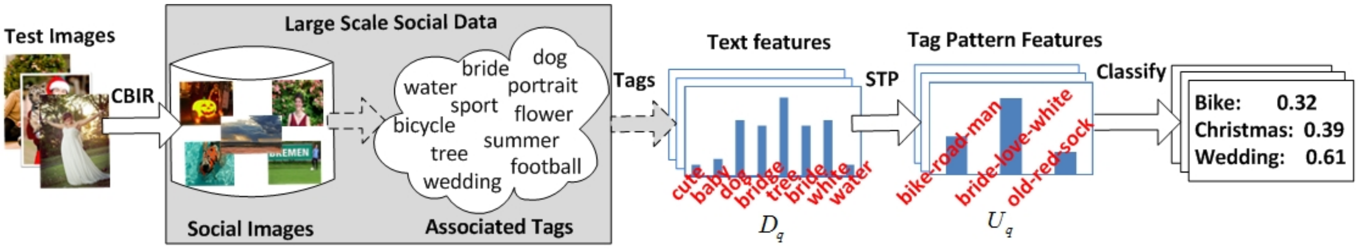

Fig. 1. Pipeline of the search-based image classification framework.

The intersection of any two tag patterns is empty or small, ensuring their discriminative ability. Meanwhile, the text associated with each image should involve as fewer tag patterns as possible, resulting compact semantic representations. Based on these two *sparsity* constraints, we formulate the STP model via the l_1 regularization as in the lasso model [11]:

$$\min_{\mathbf{U},\mathbf{A}} \frac{1}{2} \parallel \mathbf{D} - \mathbf{UA} \parallel_F^2 + \lambda \parallel \mathbf{A} \parallel_1 + \beta \parallel \mathbf{U} \parallel_1, \tag{1}$$

where $\mathbf{U}$ denotes the *text-tag pattern* matrix and is defined as $\mathbf{U} = [\mathbf{U}_1, ..., \mathbf{U}_N]^T = [u_{nz}] \in \mathbb{R}^{N \times Z}; z = 1, ..., Z$. $\mathbf{A}$ denotes the *tag pattern-tag* matrix and is given by $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_Z]^T = [\mathbf{A}_1^{'}, ..., \mathbf{A}_M^{'}] = [a_{zm}] \in \mathbb{R}^{Z \times M}$. The objective is to compute $\mathbf{A}$ such that $\mathbf{UA}$ 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} text. The larger u_{nz} is, the more important role the z^{th} tag pattern plays in representing the n^{th} 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 text-tag pattern relationships.

2.2. Optimization

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

$$\min_{\mathbf{A}} \frac{1}{2} \parallel \mathbf{D} - \mathbf{UA} \parallel_F^2 + \lambda \parallel \mathbf{A} \parallel_1,$$

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

$$\min_{\mathbf{A}_m^{'}} \frac{1}{2} \parallel \mathbf{D}_m^{'} - \mathbf{U}\mathbf{A}_m^{'} \parallel_2^2 + \lambda \parallel \mathbf{A}_m^{'} \parallel_1 . \tag{2}$$

Each subproblem in Eq. 2 is a standard lasso problem, thus, we choose a coordinate descent technique [12] to solve it.

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

$$\min_{\mathbf{U}_n} \frac{1}{2} \parallel \mathbf{D}_n - \mathbf{A}^T\mathbf{U}_n \parallel_2^2 + \beta \parallel \mathbf{U}_n \parallel_1, \tag{3}$$

which executes the similar optimization procedure as Eq. 2.

Algorithm 1 Optimization Algorithm for STP

```
Input: $\mathbf{D} \in \mathbb{R}^{N \times M}, Z, \lambda, \beta$,
Initialization: random matrix $\mathbf{U}^0 \in \mathbb{R}^{N \times Z}$
Iterate until convergence of $\mathbf{A}$ and $\mathbf{U}$
    Update $\mathbf{A}$ by solving $M$ lasso problems as in Eq. 2
    Update $\mathbf{U}$ by solving $N$ lasso problems as in Eq. 3
Output: Sparse matrix $\mathbf{A}$ and $\mathbf{U}$
```

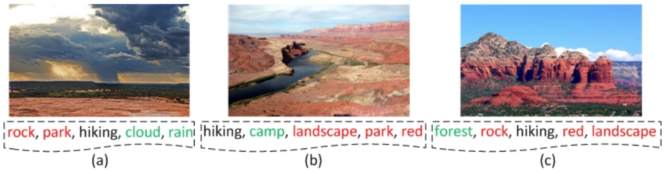
rock, park, hiking, cloud, rain
(a)
hiking, camp, landscape, park, red
(b)
forest, rock, hiking, red, landscape
(c)

Fig. 2. Illustration of mutually complementary tags among social images with similar visual content. Details are seen in text.

3. SOCIAL IMAGE CLASSIFICATION

After obtaining matrix $\mathbf{A}$, given an unlabeled query image I_q, we aim to classify it into one of the pre-defined categories. In this section, we introduce a search-based image classification framework integrated with the learned STP for social image classification (see Fig. 1).

3.1. Generating Text Feature $\mathbf{D}_q$ via CBIR

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

$$d(I_q, I_n) = \sum_{l=0}^{L} \frac{d^l(I_q, I_n)}{L}. \tag{4}$$

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

$$\mathbf{D}_q = \sum_{k=1}^{K} \mathbf{D}_k^q. \tag{5}$$

3.2. Training Classifiers with the Learned STP

We project text feature $\mathbf{D}_q$ into $\mathbf{U}_q = [u_{q1}, ..., u_{qZ}]^T \in \mathbb{R}^Z$ using Eq. 3, where $\mathbf{U}_q$ tends to be sparse. The purpose of this paper is to show that the STP feature $\mathbf{U}_q$, computed from auxiliary social data, is in fact a powerful descriptor. Various classifiers could be applied. We choose SVM classifier for the STP features. The same classifier is used for both the visual features and text features. Specifically, we adopt one-vs-all SVM with a RBF kernel, using 5-fold cross validations for selecting parameters C and γ.

4. EXPERIMENTS

4.1. Datasets

Training/Test Data. We construct our training and 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 and manually labeled 8770 images, ranging from 289 to 750 images for each concept. Then we randomly splitted the dataset into ***Training Set*** and ***Test Set*** evenly. Finally, there are 4659 images for training and 4111 images for testing.

Flickr240K **Social Data**. We randomly crawled 240K social images with associated user-generated tags from Flickr using the most popular tags[1] (including the pre-defined 15 concepts) as query keywords. There were some overlaps between training/test images and social images, so duplicates were removed, resulting a data set with 239,205 images. Many of the user-generated tags were misspelling and meaningless. We removed tags of low-occurrences and then filtered out tags that did not match with the entries in Wikipedia thesaurus. After that, there are a total of 2436 unique tags and 6.2 tags per image on average.

4.2. 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 [13] 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 Flickr240K, we adopt randomized kd-tree forest from VLFeat[2].

Baseline. We compare the proposed STP features to both visual features and text features [8]. To be fair, we trained SVM classifiers with visual features by using equally weighted sum of basic distances. The kernel function that compares two images is thus given by $k(I_q, I_i) = exp(d(I_q, I_n)/\sigma)$. σ is the normalized factor.

Parameters and Evaluation. Regularization parameters λ and β in Eq. 1 are adjusted in the interval $[0.01, 1]$ and $[0.01, 1]$, respectively. Due to limited space, we set the optimal parameters $\lambda = 0.5$, and $\beta = 0.1$ in subsequent experimental results. We choose parameters $Z \in \{30, 50, 100, 150, 200, 400\}$ and $K \in \{10, 20, 30, 40, 50\}$ for evaluating the influences of varied Z and K, respectively. We use standard mean Average Precision (mAP) for comprehensive classification evaluation.

4.3. Experimental Results

Sparse Tag Patterns. We qualitatively show the sparse tag pattern-tag relationship learned by the STP model using Flickr240K 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 observe that the learned STP covers diverse semantic concepts. Fig. 3 depicts the STP feature for an test image, indicating the sparse text-tag pattern relationship. Note that the largest tag pattern in Fig. 3 (i.e., TP 41 in Table 1) precisely describes the semantic meaning of the test image.

Table 1. Sparse Tag Patterns. "TP" denotes tag pattern.

TP 1	water, pool, swimming, reflection, blue, summer
TP 7	halloween, death, costume, party, pumpkin, dead
TP 15	snow, winter, skiing, mountain, ice, cold
TP 19	wedding, bride, love, groom, couple, dress
TP 40	football, soccer, sport, ball, game, stadium
TP 41	graduate, school, student, college, class, high
TP 42	war, prisoner, flag, freedom, people, peace

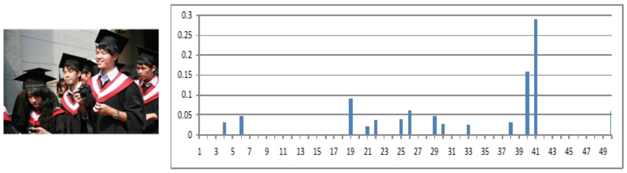

Fig. 3. The right column shows the STP feature corresponds to the left column test image.

Quantitative Comparison. Table 2 shows the performance with different types of features for each category. We observe that the STP features and text features significantly outperform visual features for all categories except *skiing*, *softball* and *wedding*, meanwhile, the STP features also achieve comparable or better performance than text features. We argue that the STP features effectively preserve the latent-semantic meaning of corresponding text features, and efficiently reduces the high-dimensional tag features to a compact representation (from 2436 to 50 on Flickr240K dataset in our experiments).

Effect of Varied Z. In this section, we study the effect of varying number of sparse tag patterns. As shown in Fig. 4, the performance increases when $Z \leq 50$, then reduces gradually from 50 to 400. If Z is too small, the relevant concepts are merged into same tag patterns (e.g., the concepts "soccer" and "softball" appear in tag pattern "sport"). However, if Z is too large, abstract concepts may be split into scattered tag patterns (e.g., the concept "Christmas" contains tag patterns such as "lights, holiday, decorations", "tree, xmas, red", "snow, white, green", "gifts, presents", etc).

Effect of Varied K. Fig. 5 reports the performance of the STP features with varied K ranging from 10 to 50. We observe that the performance improves consistently as K increases. 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. The STP features seems less sensitive to noisy tags.

Robustness to training sample noise. In this section, we inspect the robustness property of the STP features. We extend the Training Set to ***Noisy Training Set*** by adding junk images into each category in proportion, leading to a new training dataset with 9076 images in total. Table 3 shows the performance of the STP features ($K \in 10, 30, 50$) and visual features trained with or without noisy training set. We observe that the performance of visual classifier de-

[1] http://www.flickr.com/photos/tags/

[2] http://www.vlfeat.org

Table 2. Comparison in terms of mAP with different features for each category of our test dataset, as well as mAP for all categories in last column ($Z = 50$, $K = 50$). The best performance in each panel is indicated in bold.

	babyshower	beach	bike	birthday	camping	Christmas	concert	graduation
Visual Feature	0.362	0.405	0.352	0.090	0.165	0.318	0.472	0.281
Text Feature	0.434	**0.684**	0.462	**0.129**	**0.205**	0.479	**0.640**	0.285
STP Feature	**0.475**	0.600	**0.533**	0.083	0.193	**0.489**	0.606	**0.294**
	Halloween	hiking	skiing	soccer	softball	swimming	wedding	all
Visual Feature	0.274	0.375	**0.509**	0.308	**0.464**	0.502	**0.372**	0.359
Text Feature	**0.305**	0.563	0.491	0.338	0.396	0.542	0.312	0.435
STP Feature	0.302	**0.576**	0.504	**0.357**	0.328	**0.589**	0.312	**0.436**

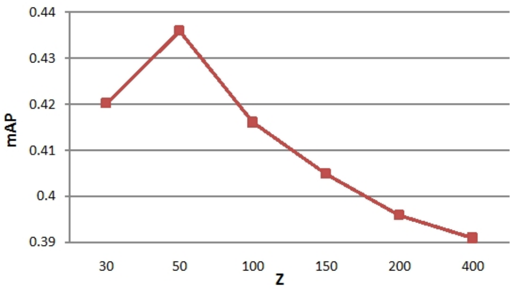

Fig. 4. Comparison in terms of mAP of STP Classifier with varied STP number Z on Flickr240K when set $K = 50$.

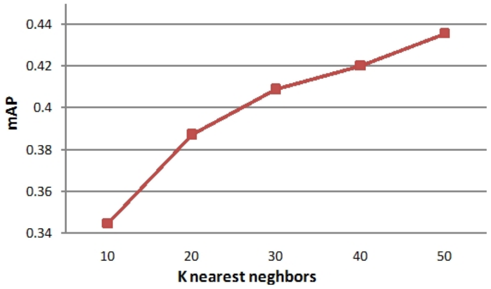

Fig. 5. Comparison in terms of mAP of STP Classifier with varied nearest neighbors K on Flickr240K when set $Z = 50$.

Table 3. Comparison in terms of mAP of STP and Visual Classifiers with/without Noisy Training Set

	Without Noisy	With Noisy	
Visual Feature	0.359	0.331	↓ 7.8%
STP Feature_K10	0.345	0.345	→
STP Feature_K30	0.409	0.399	↓ 2.45%
STP Feature_K50	0.436	0.431	↓ 1.15%

clines faster than that of the STP classifiers. The STP features are more robust to training sample noise as they are stable to appearance changes.

5. CONCLUSIONS

To attack the problem of noisy and unlimited user-generated tag vocabulary, we propose a STP model to exploit sparsity constrained co-occurrence tag patterns from large scale user-generated tags of social images, yielding compact yet discriminative intermediate semantic descriptors. The learned STP can be regarded as an alternative image features and integrated into a search-based image classification framework for boosting the accuracy of social image classification. Experimental results verify our ideas and show that the proposed method outperforms the state-of-the-art approach.

6. ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research Program of China (2009CB320902), in part by grants from the National Science Foundation of China (60902057 and 61121002) and Nanyang Assistant Professorship SUG M4080134.

7. REFERENCES

[1] Li. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in *CVPR*, 2005.

[2] Li-Jia Li and Li Fei-Fei, “What, where and who? classifying events by scene and object recognition,” in *ICCV*, 2007.

[3] Chong Wang, David Blei, and Li Fei-fei, “Simultaneous image classification and annotation,” in *CVPR*, 2009.

[4] A. Quattoni, M. Collins, and T. Darrell, “Learning visual representations using images with captions,” in *CVPR*, 2007.

[5] M. Guillaumin, J. Verbeek, and C. Schmid, “Multimodal semi-supervised learning for image classification,” in *CVPR*, 2010.

[6] A. Bergamo and L. Torresani, “Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach,” in *NIPS*, 2010.

[7] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.

[8] Gang Wang, Derek Hoiem, and David Forsyth, “Building text features for object image classifications,” in *CVPR*, 2009.

[9] Junsong Yuan, Ming Yang, and Ying Wu, “Mining discriminative co-occurrence patterns for visual recognition,” in *CVPR*, 2011.

[10] Shen-Fu Tsai, Liangliang Cao, and et. al., “Object pattern: a new model for album event recognition,” in *ACM Multimedia*, 2011.

[11] Robert Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, 1996.

[12] H. Hofling et. al. J. Friedman, T. Hastie, “Pathwise coordinate optimization,” in *ANN APPL STAT*, 2007.

[13] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, 2004.