12-2011

# Tag-based social image search with visual-text joint hypergraph learning

Yue GAO
*Tsinghua University*

Meng WANG
*National University of Singapore*

Huanboo LUAN
*National University of Singapore*

Jialie SHEN
*Singapore Management University*, jlshen@smu.edu.sg

Shuicheng YAN
*National University of Singapore*

***See next page for additional authors***

**DOI:** https://doi.org/10.1145/2072298.2072054

Published in MM '11: Proceedings of the 2011 ACM Multimedia Conference & Co-Located Workshops: November 28 - December 1, 2011, Scottsdale, AZ, USA, pp. 1517-1520.
http://doi.org/10.1145/2072298.2072054

# Tag-Based Social Image Search with Visual-Text Joint Hypergraph Learning

Yue Gao [†], Meng Wang [♯], Huanbo Luan [‡], Jialie Shen [◇], Shuicheng Yan [‡], Dacheng Tao [§]
[†] TNLIST, Dept. of Automation, Tsinghua University, Beijing, China
[♯] Hefei University of Technology, Hefei, China
[‡] National University of Singapore, Singapore
[◇] School of Information Systems, Singapore Management University, Singapore
[§] Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia
kevin.gaoy@gmail.com, {mengwang,luanhb}@comp.nus.edu.sg
jlshen@smu.edu.sg, eleyans@nus.edu.sg, dacheng.tao@gmail.com

## ABSTRACT

Tag-based social image search has attracted great interest and how to order the search results based on relevance level is a research problem. Visual content of images and tags have both been investigated. However, existing methods usually employ tags and visual content separately or sequentially to learn the image relevance. This paper proposes a tag-based image search with visual-text joint hypergraph learning. We simultaneously investigate the bag-of-words and bag-of-visual-words representations of images and accomplish the relevance estimation with a hypergraph learning approach. Each textual or visual word generates a hyperedge in the constructed hypergraph. We conduct experiments with a real-world data set and experimental results demonstrate the effectiveness of our approach.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Content Analysis and Index; H.4 [**Information Systems Applications**]: Miscellaneous

## Keywords

Tag-based image search, hypergraph learning, visual-text

## 1. INTRODUCTION

In recent years, the amount of social media grows in an explosive way due to the fast development of multimedia and network technology, such as Flickr and Youtube. The development of efficient search for these media corpus becomes highly desired. On these websites, users are allowed to not only upload multimedia data but also annotate their content with tags. Therefore, many social media entities are associated with user-provided tags. By indexing media data with these tags, tag-based search becomes a solution for users in interesting data on social media websites.

However, the performance of existing tag-based search methods is usually not satisfactory. This can mainly be attributed to the following two facts. First, user-provided tags are usually noisy. Second and more important, there lacks a good ranking strategy to order the multimedia entities that contain a query tag. For example, currently Flickr provides two ranking options for tag-based image search, one is time-based ranking and the other is interestingness-based ranking. However, they both rank images according to measures that are not related to relevance and thus in many cases the search results are not good enough in terms of relevance.

Several research efforts have been dedicated to developing relevance-based ranking for social media search. Given a query tag, the task is to estimate the relevance levels of the images that contain the tag. The visual content of images and tags have both been explored. However, existing methods usually use the two information sources separately or sequentially. For example, the method in [10, 19] mainly works as follows. First, an initial relevance score of each image is learned according to the similarity between the query tag and the image's tag set. A graph-based learning is then performed to refine the relevance scores based on the pairwise visual similarities of images. Therefore, tag and visual information is actually used in the first and second steps, respectively. This is due to the fact that visual features and tags are different characteristics and are not easy to be integrated.

In this work, we propose a visual-text joint hypergraph learning approach to simultaneously explore the two information sources. Each image can be represented by bag-of-words and bag-of-visual-words, which are generated from the image's tags and visual content, respectively. A hypergraph is constructed to model the relationship of all images, in which each vertex denotes an image and a hyperedge is a visual or textual word (i.e., tag), and a hyperedge connects to multiple vertices. We define the weight of each edge based on the visual similarities of images belonging to the edge. The relevance scores of images are learned based on the hypergraph and then we can order the images with their relevance scores in descending order. Experiments on a Flickr data set demonstrates the superiority of our approach over the state-of-the-art methods.

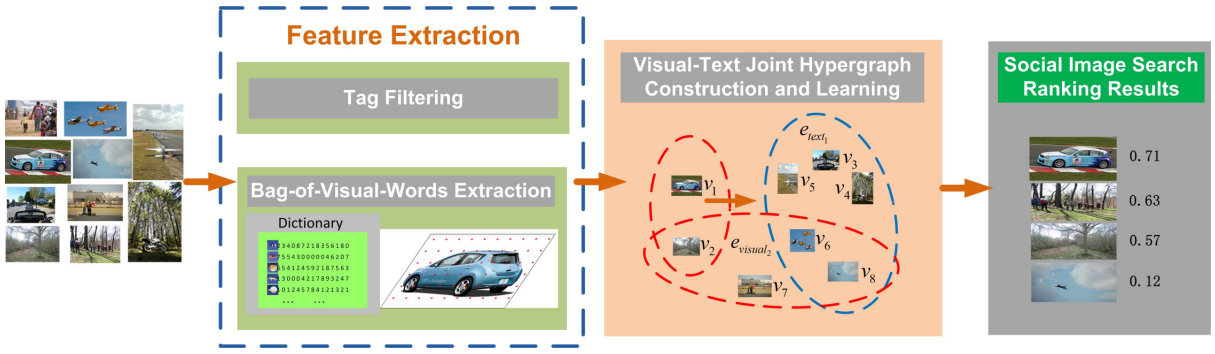


**Figure 1: The flowchart of the proposed method.**

The rest of the paper is organized as follows. Section 2 briefly reviews related work. In Section 3, we introduce the tag-based image search with visual-text joint hypergraph learning approach. Experiments on a Flickr data set are provided in Section 4. Finally, we conclude the paper in Section 5.

## 2. RELATED WORK

Multimedia analysis and retrieval [5,6,14,15,17,18,20,21] has attracted a lot of attention recently, and extensive research efforts have been dedicated to topics related to web image search in the past years [3, 4, 16]. Different from these web images, social images can be indexed with user-contributed tags and tag-based image search is an effective approach for social image search. However, user-provided tags are usually noisy. Therefore, several research has been conducted towards improving search performance by tag refine or tag relevance learning [2, 9, 11–13, 19, 24]. Liu et al. [**?**] proposed a relevance-based ranking method for social image search. It first learns relevance scores based on images' tags and then refines the scores by exploring images' visual content. In [19], a diverse relevance ranking scheme was proposed to re-rank images by exploring the content of images and their associated tags. The first component of their approach estimates the relevance scores of images and it is actually the same with [10]. However, these methods usually use visual and tag information separately or sequentially, whereas our approach integrates them in a hypergraph learning scheme such that they can be simultaneously investigated. Experiments will demonstrate the superiority of our approach.

Hypergraph has been widely investigated in information retrieval and pattern recognition tasks [1, 3, 7, 23] for its capability of capturing high-order relationship of samples. A probabilistic hypergraph matching method was proposed in [22] to match two feature sets. In the transductive learning framework for image retrieval [8], each image was represented by a vertex in a probabilistic hypergraph, and the image retrieval was formulated as a hypergraph ranking task.

Considering the capability of hypergraph in high-order relationship mining and unified modeling (the hyperedge can be generated based on different information sources), our work employs the hypergraph learning method for a tag-based image search with joint visual-text information.

## 3. VISUAL-TEXT JOINT HYPERGRAPH LEARNING

We introduce the proposed tag-based image search with visual-text joint hypergraph learning in this section. Figure 1 demonstrates the schematic illustration.

### 3.1 Feature extraction

By feature extraction, we aim to generate the bag-of-words and the bag-of-visual-words representations for each social image.

For bag-of-words representation, we simply generate it by selecting several informative tags. From our dataset, which will be introduced in the next section, we have 12,921 unique tags. We first perform a filtering with the help of Wikipedia, and those tags that do not have coordinate in Wikipedia are removed (they are usually misspelling or meaningless words). Then we select 2,000 tags that are with the highest TF-IDF values. With the 2,000 tags, we generate the bag-of-words representation for each image.

For bag-of-visual-words representation, we perform Difference-of-Gaussian (DoG) method to detect keypoints in each image and then employ 128D SIFT descriptor. A 1,000-D codebook is built by grouping the keypoint features with hierarchical K-means and in this way a 1000-D bag-of-visual-words representation is generated for each image.

### 3.2 Hypergraph construction

A hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ is composed by the vertex set $\mathcal{V}$, the hyperedge set $\mathcal{E}$, and the hyperedge weight $w$ . Each hyperedge $e_i$ is given a weight $w(e_i)$. The hypergrpah $G$ can be denoted by a $|\mathcal{V}| \times |\mathcal{E}|$ incidence matrix $\mathbf{H}$ with entries:

$$h(v,e) = \begin{cases} 1 & if\ v \in e \\ 0 & if\ v \notin e \end{cases} \quad (1)$$

For a vertex $v_i \in \mathcal{V}$, the vertex degree is estimated by:

$$d(v_i) = \sum_{e \in \mathcal{E}} \omega(e)\, h(v_i, e). \quad (2)$$

For a hyperedge $e_i \in \mathcal{E}$, the edge degree is estimated by

$$\delta(e_i) = \sum_{v \in \mathcal{V}} h(v, e_i). \quad (3)$$

We let $\mathbf{D}_v$ and $\mathbf{D}_e$ denote the diagonal matrices of the vertex degrees and the hyperedge degrees, respectively, and we let $\mathbf{W}$ denote the diagonal matrix of the hyperedge weights.

In our approach, we regard each image in the database as

a vertex in the visual-text hypergraph $G = (\mathcal{V}, \mathcal{E}, w)$. For example, assuming there are totally $n$ images in the database, the generated hypergraph $G = (\mathcal{V}, \mathcal{E}, w)$ thus contains $n$ vertices. Taken both the visual content and the text information into consideration, there are two types of hyperedges generated from the visual content and the tags respectively.

For visual content-based hyperedge, each visual word is selected as a hyperedge, and the images that contain the same visual word are connected by the hyperedge. These visual-content based hyperedge is denoted by $\mathcal{E}_{visual}$. Analogously, for tag-based hyperedge, each tag is selected as a hyperedge, and the images containing the same tag are connected by the hyperedge. These tag-based hyperedges is denoted by $\mathcal{E}_{text}$. Let $n_{visual}$ and $n_{text}$ denote the number of the two types of hyperedges, there are totally $n_E = n_{visual} + n_{text}$ hyperedges.

We let $\mathbf{D}_v$ and $\mathbf{D}_e$ denote the diagonal matrices of the vertex degrees and the hyperedge degrees respectively, and the incidence matrix $\mathbf{H}$ is constructed using Equation (1). The weight of a hyperedge $w$ is estimated based on the similarity of images connected by the hyperedge, i.e.,

$$w(e_i) = \sum_{I_a, I_b \in e_i} \exp\left(-\frac{\|I_a - I_b\|^2}{\sigma^2}\right), \tag{4}$$

## 3.3 Hypergraph Learning

For hypergraph learning, the Normalized Laplacian method proposed in [23] is employed, and it is formulated as a regularization framework:

$$\arg\min_f \{\lambda R_{emp}(f) + \Omega(f)\}, \tag{5}$$

where $f$ is the classification function to be learned, $\Omega(f)$ is a regularizer on the hypergraph, $R_{emp}(f)$ is empirical loss, and $\lambda > 0$ is a weighting parameter. The regularizer on the hypergraph is defined as

$$\begin{aligned}\Omega(f) &= \tfrac{1}{2}\sum_{e\in\mathcal{E}}\sum_{u,v\in\mathcal{V}} \frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 \\ &= \tfrac{1}{2}\sum_{e\in\mathcal{E}_{visual}}\sum_{u,v\in\mathcal{V}} \frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 \\ &+ \tfrac{1}{2}\sum_{e\in\mathcal{E}_{text}}\sum_{u,v\in\mathcal{V}} \frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2\end{aligned} \tag{6}$$

Let $\mathbf{\Theta} = \mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}}$, and $\mathbf{\Delta} = \mathbf{I} - \mathbf{\Theta}$, the normalized cost function can be written as

$$\Omega(f) = f^T\mathbf{\Delta}f. \tag{7}$$

Here $\mathbf{\Delta}$ is a positive semi-definite matrix, and it is usually called hypergraph Laplacian.

The transductive inference is formulated with a regularization on hypergraphs $\arg\min_f \{\lambda R_{emp}(f) + \Omega(f)\}$, and the loss term is defined as follows:

$$\|f - y\|^2 = \sum_{u\in\mathcal{V}} (f(u) - y(u))^2, \tag{8}$$

where $y$ is the label vector. Assuming the number of all images in the database is $n$, and the $i$-th image is selected as the query image. Denote by $y$ an $n \times 1$ vector, where all elements of $y$ are 0 except its $i$-th value is 1. Then the learning task for social image search is to minimize the sum of the two terms:

$$\Phi(f) = f^T\mathbf{\Delta}f + \lambda\|f - y\|^2, \tag{9}$$

where $\lambda > 0$ is the regularization parameter. Differentiating $\Phi(f)$ with respect to $f$, we can obtain:

$$f = \left(\mathbf{I} + \frac{1}{\lambda}\mathbf{\Delta}\right)^{-1} y. \tag{10}$$

After obtaining the relevance score vector $f$, we can rank the images that contain the query tag with the scores in descending order.

## 3.4 Analysis of computational cost

According to the process introduced above, it can be analyzed that the computational cost of hypergraph learning scales as $O(n^3)$, where $n$ is the number of images in the hypergrpah learning procedure. But in fact we can solve Eq. (10) can be solved with an iterative process, which can reduce the computational cost to $O(n^2)$.

# 4. EXPERIMENTS

## 4.1 Experimental Settings

To evaluate the proposed tag-based image search method with visual-text joint hypergraph learning, we conduct experiments on the dataset in [19]. The data are collected from Flickr with the search results of 52 tags: *airshow, apple, beach, bird, car, cow, dolphin, eagle, flower, fruit, jaguar, jellyfish, lion, owl, panda, starfish, triumphal, turtle, watch, waterfall, wolf, chopper, fighter, flame, hairstyle, horse, motorcycle, rabbit, shark, snowman, sport, wildlife, aquarium, basin, bmw, chicken, decoration, forest, furniture, glacier, hockey, matrix, Olympics, palace, rainbow, rice, sailboat, seagull, spider, swimmer, telephone*, and *weapon*. Time-based ranking is used in the data collection process and thus we can estimate the search relevance of time-based ranking. Mean Average Precision (MAP) is adopted as our performance evaluation metric.

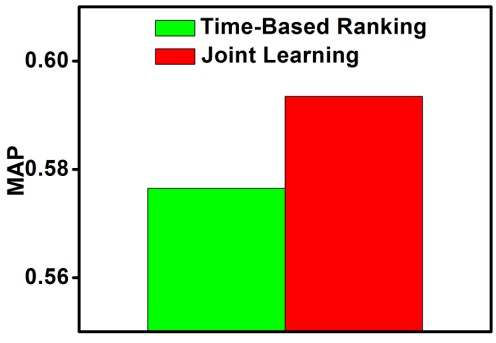


**Figure 2: The MAP comparison of the two compared methods.**

## 4.2 Experimental Results

We compare the following methods in the experiment:

- Time-Based Ranking. Time-based ranking orders the images according to their uploading time.
- Tag-Based Image Search with Visual-Text Joint Hypergraph Learningčň i.e., the proposed method. We denote it as "Joint Learning".

In our "Joint Learning" method, the parameter $\lambda$ is simply set to 0.9, and the parameter $\sigma$ in Eq. (4) is simply set to the median value of the pairwise distances of all images. For each query tag, we only randomly select 2000 images that do not contain the tag as negative samples.

Figure 2 demonstrates the results. From the results we can see that, time-based ranking performs worse than the "Joint" method. Among the 52 tags, there are 21 queries where the improvement of search performance is more than 10%. The MAP measures of the two compared methods are 0.576 and 0.593, respectively. Figure 3 demonstrate the top 10 search results of the two compared ranking methods of an example query "Forest".

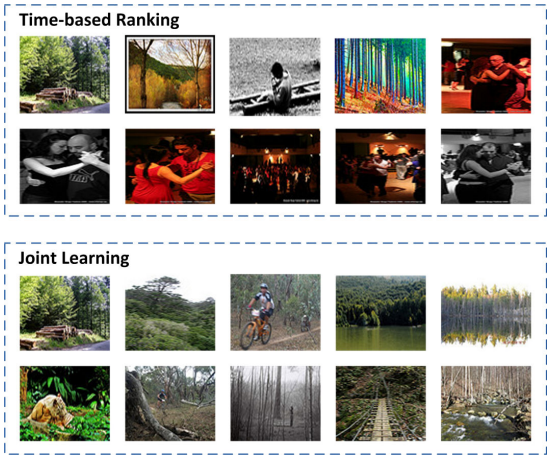


**Figure 3: The top 10 results of the two ranking methods of an example query "Forest".**

# 5. CONCLUSION

In this work, we propose a tag-based image search with visual-text joint hypergraph learning. In this method, we simultaneously investigate the visual and text information of images with a hypergraph learning approach. Hyperedges are generated based on visual and textual words. We conduct experiments on a Flickr data set and compare the proposed approach with existing methods. Experimental results show that our method can achieve better performance.

# 6. REFERENCES


[1] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the ACM International Conference on Multimedia*, 2010.

[2] L. Chen, D. Xu, W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[3] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. *ACM Trans. Graph*, 28, 2009.

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 409–416, 2011.

[5] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, and N. Zhang. Less is more: Efficient 3d object retrieval with query view selection. *IEEE Transactions on Multimedia*, 11, 2011.

[6] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Ranking model adaptation for domain-specific search. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[7] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Video object segmentation by hypergraph cut. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–6, Miami, USA, 2009.

[8] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2010.

[9] X. Li, C. G. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010.

[10] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the International Conference on World Wide Web*, 2009.

[11] D. Liu, M. Wang, X.-S. Hua, and H.-J. Zhang. Semi-automatic tagging of photo albums via exemplar selection and tag inference. *IEEE Transaction on Multimedia*, 13:82–91, 2011.

[12] D. Liu, M. Wang, L. Yang, X.-S. Hua, and H. Zhang. Tag quality improvement for social images. In *Proceeding of IEEE International Conference on Multimedia*, pages 350–353, 2009.

[13] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang. Image tagging via collaborative tag propagation. *IEEE Transaction on Multimedia*, 13:702–712, 2011.

[14] J. Shen, J. Shepherd, B. Cui, and K.-L. Tan. A novel framework for efficient automated singer identification in large music databases. *ACM Transactions on Information Systems*, 27, 2009.

[15] J. Shen, D. Tao, and X. Li. Modality mixture projections for semantic video event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18:1587–1596, 2008.

[16] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[17] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2:10–31, 2011.

[18] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11:465–476, 2009.

[19] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 12:829–842, 2010.

[20] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 10:2761–2773, 2010.

[21] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10:437–446, 2008.

[22] R. Zass and A. Shashua. Probabilistic graph and hypergraph matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Alaska, USA, 2008.

[23] D. Zhou, J. Huang, and B. Schokopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Proceedings of Advances in Neural Information Processing Systems 19*, pages 1601–1608, 2007.

[24] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the ACM International Conference on Multimedia*, 2010.