Tag Suggestion and Localization for Images by Bipartite Graph Matching

Wei-Ta Chu[1], Cheng-Jung Li[1], and Jen-Yu Yu[2]
[1]Department of Computer Science and Engineering, National Chung Cheng University, Chiayi, Taiwan
E-mail: wtchu@cs.ccu.edu.tw, zoneli1987@gmail.com
[2]Information and Communication Research Lab, Industrial Technology Research Institute, Hsinchu, Taiwan
E-mail: KevinYu@itri.org.tw

***Abstract*—Given an image that is loosely tagged by a few tags, we would like to accurately localize these tags into appropriate image regions, and at the same time suggest new tags for regions if necessary. In this paper, this task is formulated on a bipartite graph, and is solved by finding the best matching between two disjoint sets of nodes. One set of nodes represents regions segmented from an image, and another set represents a combination of existing tags and new candidate tags retrieved from photo sharing platforms. In graph construction, visual characteristics in the representation of the bag of word model and users' tagging behaviors are jointly considered. By finding the best matching with the Hungarian algorithm, the region-tag correspondence is determined, and tag suggestion and tag localization are accomplished simultaneously. Experimental results show that the proposed unified framework achieves promising image tagging performance.**

I. Introduction

It has been known that tremendous amount of images shared on the web impede efficient retrieval and browsing. In the past few years, image annotation or tagging has been widely studied to facilitate keyword-based retrieval. Many works have been proposed to annotate images based on audiovisual features, temporal information, spatial correlation, context between spatial/temporal information, and even social knowledge implicitly provided by users.

Current annotation works can be categorized into two main groups: annotation by concept detection and annotation by social media analysis. Many researchers develop concept detectors for images, with main consideration on visual features that are directly extracted from visual content. However, performance of concept detectors is limited due to the notorious semantic gap problem. Currently, another alternative is proposed to explore social knowledge in image annotation. In our previous work [2], we conduct tag suggestion and localization based on bipartite graph matching. We modeled relationships between video keyframes and candidate tags as a bipartite graph, and the best matching between two sets of nodes was determined to associate each video shot with appropriate tags. In this paper, we target to push this framework toward image tag suggestion and (spatial) localization.

Fig. 1 illustrates the concept of tag suggestion and localization for images. The content owner has (globally) loosely tagged this image by "mountain" and "tree". These tags may just represent parts of regions of this image, and how existing tags are associated with regions is to be determined. Therefore, the goal of our work is to find more relevant tags from the web that are appropriately to describe all major regions in an image, as well as to spatially localize existing tags into appropriate image regions. The proposed system first segments this image into visually consistent regions. Then, not only existing tags are localized to appropriate regions, but also new tags are suggested to regions that cannot be described by existing tags. In contrast to temporal localization for videos [2], this work spatially localizes tags into image regions.

Contributions of this paper are briefly summarized as follows.

- We introduce the concept of social media analysis to conduct tag suggestion and localization. Users only need to roughly provide a few initial tags, and then more tags would be suggested and localized to appropriate regions.
- A unified framework based on bipartite graph matching is proposed to systematically consider visual characteristics and users' tagging behaviors.

The remainder of this paper is organized as follows. Section II provides literature survey. Section III describes the proposed image tagging framework based on bipartite graph matching. Experimental results are given in Section IV, followed by the concluding remarks in Section V.

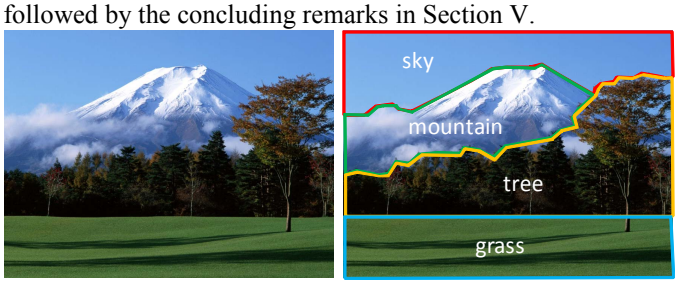

Fig. 1 Illustration of tag suggestion and (spatial) localization for images.

II. Related Works

As Web 2.0 and photo sharing websites emerge, image tags have been shown to be important clues to facilitate object recognition and image retrieval. Ames and Naaman [15] investigate the incentives of annotating photos in Flickr and claim that with tags users can not only easily recall from their own photos, but also make their photos more searchable by other people. To automate the annotation process, various

probabilistic models are built to predict semantic concepts in images [11], which handle with the notorious semantic gap problem. Kennedy et al. [16] study performance variations between concept detectors trained by human-annotated data and concept detectors trained by data automatically retrieved from the web. They claim that some concepts would gain much from human efforts. Yan et al. [17] study manual annotation in a quantitative way and propose a learning approach to suggest right images or right keywords to reduce annotation time. With a similar purpose, the work in [18] develops a recommendation strategy to support users in photo annotation. In [6], an image is annotated by jointly considering its surrounding text and tag candidates retrieved from the web. A bipartite graph is constructed to describe relationship between them, and then a reinforcement algorithm is applied to rank tag candidates. Li et al. [7] take user's tagging behavior into account and evaluate tag relevance to facilitate image ranking or tag ranking. Similarly, Sun and Bhowmick [13] use the concept of language models to estimate effectiveness of a tag. From a different perspective, Wu et al. [12] enhance image tagging by learning a more appropriate distance metric.

More recently, Liu et al. propose a semi-automatic approach. Users just need to annotate a small set of representative images, and then the tags are appropriately propagated to related images [8][9]. Sevil et al. [21] propose a photo tag expansion method. Based on initial tags provided by users, their system retrieves relevant images from Flickr, collects the associated tags, and suggests the most appropriate tags that are associated with the images that have the highest visual similarity to the target image. We adopt a similar idea to this work, but we further consider user's tagging behavior in tag suggestion. Additionally, the proposed systematic framework can be generally employed in image tagging and video tagging.

To make tags more descriptive, Yang et al. [10] associate color, texture, and location properties to existing tags. Their work makes a further step over current image tagging studies. For large amounts of loosely-tagged images (multiple object tags are given loosely at the image level), Shen and Fan [14] model loosely-tagged images and inter-object correlation by a multi-task SVM, and recommend tags for each object instance. For image tagging, our work achieves a similar effect as [14], but the same framework can also be applied to video tagging. Both temporal tag localization and spatial tag localization can be done in the proposed unified framework.

III. TAG SUGGESTION AND LOCALIZATION

A. Overview of Framework

Two perspectives can be considered for the image tag suggestion and localization task. When users share their photos on the web, they just roughly annotate a photo album with a few tags, such as "tour in Paris" and "birthday party in May 19, 2011." The only information for images in the first album is "tour" and "Paris", and thus images are not accurately annotated (or just loosely-tagged [14]). From this perspective, we would suggest more tags to annotate this album, and at the meanwhile appropriate tags are assigned to each image. In this case, tags are suggested and localized from the album level to the image level.

Instead of globally annotating an image with a few tags, another perspective for tag suggestion and localization is to suggest new tags (if necessary) and associate them with appropriate regions in this image. In this case, tags are suggested and spatially localized from the image level to the region level. In this paper, we focus on the second perspective.

The process of tag suggestion and localization is illustrated in Fig. 2. At the beginning, the image is globally tagged with *mountain* and *tree*. Based on these two initial tags, we retrieve relevant images and their associated information, including tags and owner ID, from Flickr. The images that have the same tag t_i, for example, are clustered together, and the average visual word histogram calculated from them is used to characterize the tag t_i. On the other hand, the original image is first segmented into several regions (manual or automatic, see the evaluation section for details). From each region a visual word histogram is extracted to be the region description. A bipartite graph is then constructed, where nodes in one side are image regions and nodes in another side are tags. The best matching is determined by the Hungarian algorithm, and finally the region-tag correspondence is determined. In the following, we describe details of important components in this framework.

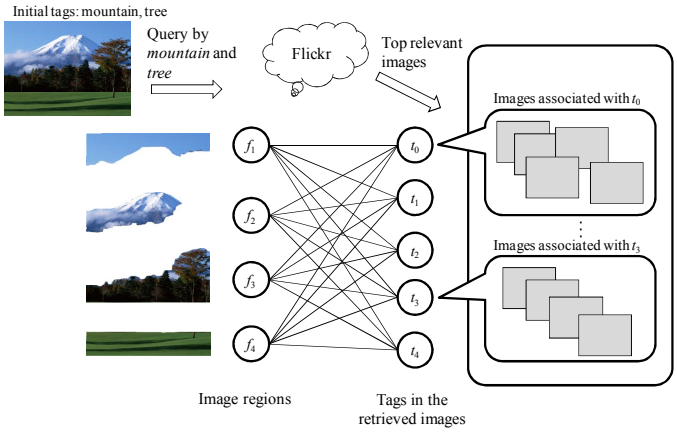

Fig. 2 Illustration of a bipartite graph describing relationships between image regions and tags.

B. Weighting Scheme

Given an image that has been loosely annotated with the tag t_0 (assuming only one existing tag, with loss of generality), we query Flickr by t_0 and retrieve the top 15 relevant images as well as their associated tags as the pool for tag suggestion. With this candidate tag pool, we would like to measure how likely a tag is appropriate to describe a specific image region.

Let $X_j = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ denote the image regions in an image, and $T = \{t_1, t_2, ..., t_n\}$ denote the candidate tag pool which are collected from the whole retrieved images $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$. Due to user's subjectivity and variations of tagging behaviors, the images that are associated with the tag t_i don't necessarily represent the concept t_i. To

decimate the influence of visual words derived from noisy images, an adaptive weighting scheme is designed. In our implementation, visual words are derived from clustering SIFT (Scale-Invariant Feature Transform) descriptors [22] extracted from an independent codebook training dataset.

Importance of a visual word for a tag t_i depends on two factors: 1) A visual word is more important if it frequently appears in the image collection associated with the tag t_i; 2) If a visual word appears in all retrieved images, it provides less information for distinguishing truth data from noisy data.

Let the set $Y_i = \{\boldsymbol{y}_{i,1}, \boldsymbol{y}_{i,2}, ..., \boldsymbol{y}_{i,L}\}$ denote the retrieved images associated with the tag t_i. By combining two factors described above, the weight of the kth visual word is given by

$$w_k = \frac{\sum_{\ell=1}^{L} h_\ell[k]}{\bar{z} + \epsilon} \times \left(1 - \frac{\bar{z}}{L+1}\right), \quad (1)$$

where $\bar{z}$ denotes in Y_i the number of images containing the kth visual word, and the parameter ϵ is set as a small value to avoid zero denominator. The value $h_\ell[k]$ denotes frequency in the kth bin of the visual word histogram of $\boldsymbol{y}_{i,\ell}$. The first term denotes the normalized occurrence frequency of the kth visual word. More frequently this visual word in Y_i, larger this term is. The second term denotes degree of discrimination of this visual word. If this visual word appears in more images in Y_i, less important it is. In [1], we have verified that such weighting scheme effectively reduces the influence of noisy retrieved data.

C. Tagging Behavior

User's tagging behavior is further considered to accurately capture tag properties from a human-centric perspective. Tagging behaviors are classified into two categories. Firstly, if the tag t_i is frequently used to tag images, it implicitly represents consensus of many users, and should be emphasized. Therefore, the first factor $\hat{c}_i$ is defined as

$$\hat{c}_i = \frac{c_i}{max_{1 \le j \le n} c_j}, \quad (2)$$

where c_i is the number of users utilizing t_i to tag images, and n is the number of distinct tags in the candidate tag pool.

Secondly, for the existing tag t_0, the tag t_i is more important if more images were simultaneously tagged with t_0 and t_i. This idea was also adopted in [4] and [7]. Given the candidate tag pool, we count the number of images that are simultaneously tagged with t_0 and t_i. According to this count, tags in the candidate pool are sorted in descending order. Let $\{r_1, r_2, ..., r_n\}$ denote ranks of candidate tags, i.e., $r_i = 1$ if t_i is the first top-ranked tag, and $r_i = 2$ if t_i is the second top-ranked tag. The second factor $\hat{r}_i$ for tagging behaviors is defined as

$$\hat{r}_i = \frac{\lambda}{\lambda + (r_i - 1)}, \quad (3)$$

where λ is a positive value to avoid zero denominator.

D. Graph Construction and Matching

To discover relationships between image regions and candidate tags, they are respectively viewed as two disjoint sets, and we construct a weighted bipartite graph to describe their relationships. Based on this graph, the best matching between two sets of nodes is accordingly determined.

Weight on each edge is defined as the weighted similarity between image regions and tags. With the weighting scheme and factors of tagging behaviors described above, the similarity value between the region $\boldsymbol{x}_i$ and the tag t_j is calculated as weighted histogram intersection:

$$S(\boldsymbol{x}_i, t_j) = \hat{r}_j \times \hat{c}_j \times \left(\sum_{k=1}^{K} w_k \times \min(h_i[k], h_j[k])\right), \quad (4)$$

where K is the number of visual words, and w_k is the weight for kth visual word. Note that the tag t_j is represented by the average visual word histogram of the retrieved images tagged with t_j. That is,

$$h_j[k] = \frac{1}{L} \sum_{\ell=1}^{L} h_\ell[k], \quad (5)$$

where L is the number of retrieved images tagged with t_j.

Given a weighted bipartite graph $G = (V, E)$, where $V = (X, T)$, a matching is a set of pairwise non-adjacent edges, in which each edge connects one node in X and one node in T, and no two edges share a common node. A maximum weighted matching is a matching that contains the largest possible edges and the sum of edge weights is maximal. This problem is well studied, and can be solved by the Hungarian algorithm [3]. By this algorithm, the determined matching describes the best association between image regions and tags.

IV. Experiments

We evaluate image tagging based on the MSRC dataset v1 [19], which includes 240 images of 9 object classes. The image resolution is 320×213. Ground truth of image segmentation is provided for each image, and each image belongs to one of the object classes. With these ground truths we avoid the intractable image segmentation problem and focus on evaluating the proposed tagging method. Fig. 3 shows some image samples from the *animal* and *bicycle* object classes. We see that various objects would appear in images of the same object class. If multiple regions in an image correspond to the same object, these regions jointly form a node in the left side of the bipartite graph. For example, in the leftmost image of the first row in Fig. 3, the three regions in blue are all cows, and thus they jointly represent a node in the bipartite graph. Therefore, conceptually a node in the left side of Fig. 2 represents "a set of image regions with the same semantic concept."

Images belonging to the same object class, e.g. *animal*, are viewed as in the same album. We manually annotate each image in the album with a few tags. For example, for each image in the animal album, we tag it with *sheep*, *horse*, *grass*, and *cow*, which are viewed as the initial tags. In evaluation, we eliminate the *face* class from the MSRC dataset v1 because most images in this class simply consist of a face object. Table I shows the manually-defined initial tags for images in each object class.

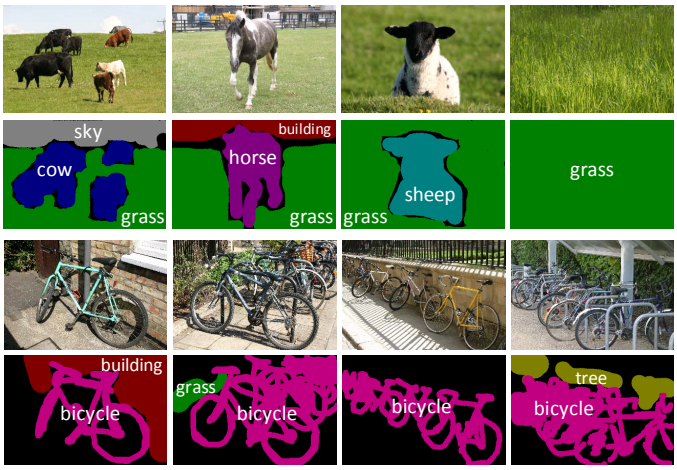

Fig. 3 Image samples from the "animal" and "bicycle" object classes.

TABLE I. OBJECT CLASSES AND THE INITIAL TAGS WE PROVIDED FOR MSRC DATASET V1.

Object class	Initial tags
Animal	cow, horse, sheep, grass
Tree	tree, sky, grass
Building	building, sky, grass
Aeroplane	aeroplane, sky, grass, building
Cow	cow, grass, water
Car	car, building
Bicycle	bicycle, grass, building

In the proposed framework, we consider weighted visual word histogram intersection to evaluate visual similarity between tags. However, nodes in the left side are image regions, and we often cannot extract enough number of SIFT descriptors to build reliable visual words from smooth regions. Therefore, we further consider HSV (hue-saturation-value) color histogram intersection in evaluating tag similarity:

$$S(\boldsymbol{x}_i, t_j) = \hat{r}_j \times \hat{c}_j \times \Big(\alpha \times \big(\textstyle\sum_{k=1}^{K} w_k \times \min(h_i[k], h_j[k])\big) + (1-\alpha) \times \big(\textstyle\sum_{k'=1}^{K'} \min(h'_i[k'], h'_j[k'])\big)\Big), \quad (6)$$

where $h'[k']$ is the histogram value of the k'th bin of the HSV color histogram, in which hue, saturation, value are quantized into 8, 4, and 4 bins, respectively. The value α is set as 0.6.

- Performance of Tag Suggestion

We compare our work with a state-of-the-art image annotation system, i.e. Alipr [20], which predicts 15 most probable tags for each uploaded image. For an image constituted by k regions, k tags will be suggested by our system, and only k most tags predicted by Alipr are considered for fair comparison. Each suggested tag is evaluated manually. Accuracy of tag suggestion for an image is calculated as k_c/k, where k_c is the number of correctly suggested tags.

Although directly utilizing the segmentation ground truths provided by the MSRC dataset largely reduces the efforts for image tagging, it is often not the case in real-world usage. To evaluate how image segmentation errors affect the performance of image tagging, we employ a graph-based image segmentation method [5] to automatically find image segments, and accordingly conduct tag suggestion and localization by the proposed method.

Furthermore, determining whether a tag is "correctly" suggested to annotate an image region is not always easy. If we evaluate tagging results in a broad sense, multiple tags may be used to annotate an image region. For example, it may also make sense if an image region containing *grass* is tagged with *green* or *plant*, in addition to the definitely correct answer *grass*. In this experiment, we evaluate average accuracy of tag suggestion under both tight and loose conditions. In the former condition, only the definitely correct answer is counted in evaluation, while in the latter condition broader ranges of tags can be counted if they make sense (determined manually). Another reason for allowing loose tag suggestion in evaluation is that Alipr often suggest generic tags. Allowing the loose condition makes performance comparison fairer.

Table II shows the average suggestion accuracy for each object class, based on true segmentation results (GT) or automatic segmentation results (Auto), under tight or loose conditions. From the tight condition (left part), our approach achieves much better performance than Alipr, because further information such as initial tags and web-retrieved data provides more clues for tag suggestion. This result seems straightforward, but meanwhile it reveals that utilizing web-based context in tag suggestion is promising. Comparing results based on true segmentation and automatic segmentation, it is not surprising that the ones based on automatic segmentation have worse performance due to segmentation errors. However, even with segmentation errors, our approach still achieves promising performance.

The right part of Table II shows average suggestion accuracy under the loose condition. In this case the Alipr system also achieves satisfactory performance (0.52 average accuracy when true segmentation results are used), and our approach keeps working very well even if it is based on error-prone automatic segmentation results.

- Performance of Tag Localization

Tags are spatially localized to each image region by the proposed approach. In many cases, more than one concept can be used to tag an image region. For example, the image region consisting of a *sheep* should be perfectly tagged as *sheep*. However, we cannot say it is not correct at all if this region is tagged with *mammal* or *lamb*. Instead of making a hard decision, we evaluate the tag for each region as one of three categories: good, neutral, and bad. For the above example, *sheep* is a good tag for this region, while *mammal* is a neutral tag. Localization results are also evaluated manually.

Fig. 4 shows average percentages of different types of tags for each object class. Most tag localization results are either good or bad. Overall, averagely 56% of localization results are deemed as "good" results, and averagely 62% of localization results are deemed as either "good" or "neutral". We obtain worse performance for the *aeroplane* and *bicycle* classes. The images retrieved by the *aeroplane* tag often simultaneously consist of an *aeroplane* object and a (very)

TABLE II. AVERAGE TAG SUGGESTION ACCURACY, BASED ON MANUAL AND AUTOMATIC SEGMENTATION RESULTS, UNDER TIGHT AND LOOSE CONDITIONS.

Object class	Tight				Loose			
	Alipr(GT)	**Alipr(Auto)**	**Our(GT)**	**Our(Auto)**	**Alipr(GT)**	**Alipr(Auto)**	**Our(GT)**	**Our(Auto)**
Animal	0.33	0.31	0.88	0.76	0.81	0.69	0.96	0.93
Tree	0.45	0.33	0.70	0.74	0.56	0.53	0.85	0.96
Building	0.36	0.44	0.90	0.79	0.51	0.49	0.93	0.89
Aeroplane	0.49	0.31	0.67	0.97	0.57	0.43	0.71	0.98
Cow	0.13	0.06	0.78	0.74	0.63	0.48	0.94	0.92
Car	0.05	0.01	0.94	0.49	0.28	0.02	0.94	0.59
Bicycle	0.17	0.05	0.96	0.75	0.31	0.09	0.96	0.85
Overall	0.28	0.22	0.83	0.75	0.52	0.39	0.90	0.87

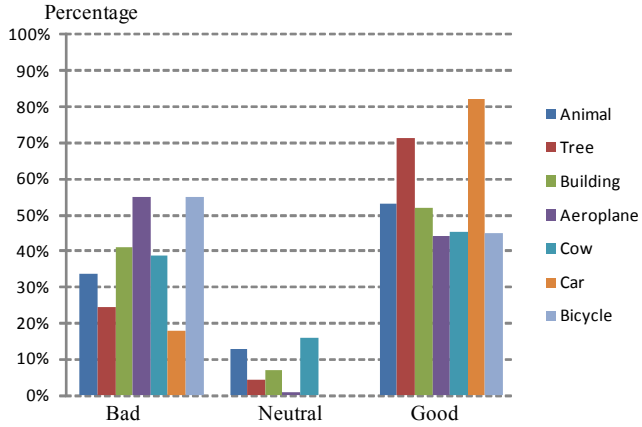

Fig. 4 The average percentage of different types of tags.

large-area *sky* object, which makes aeroplane images similar to sky images. As a consequence, we observe in this class an *aeroplance* object is often mis-tagged as a *sky* object. In constructing the bipartite graph, we evaluate visual similarity between tags by comparing an image region with the whole images of the same tag, because we don't have accurate image segmentation results for web-retrieved images. The imbalance of visual similarity evaluation may degrade localization performance, especially when the main semantic object does not occupy a large area.

V. CONCLUSION

To accomplish tag suggestion and location for images, we search relevant images from user-shared photo collections based on existing tags, and then model relationships between candidate tags and image regions as a bipartite graph. Tag suggestion is then transformed into a bipartite graph matching problem. In constructing the bipartite graph, priority of different visual words (visual similarity) and frequency of tags utilized by users (tagging behavior) are jointly considered. The experimental results demonstrate that the proposed method can well capture association between image regions and tags, and achieve promising performance in tag suggestion and localization.

In the future, relationship between images and more social knowledge would be investigated to enhance tagging performance. The influence of noisy data, and the imbalance similarity evaluation in the image tagging process, will be studied more.

ACKNOWLEDGMENT

The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 101-2221-E-194-055-MY2.

REFERENCES

[1] W.-T. Chu and C.-J. Li, "Somebody helps me: travel video scene detection using web-based context." Neurocomputing, vol. 95, pp. 60-71, 2012.

[2] W.-T. Chu, C.-J. Li, and Y.-K. Chou, "Tag suggestion and localization for web videos by bipartite graph matching." Proceeding of International ACM Workshop on Social Media, pp. 35-40, 2011.

[3] R. Diestel, Graph Theory. Heidelberg, Springer, 2005.

[4] L. Ballan, M. Bertini, A. Del Bimbio, M. Meoni, and G. Serra, "Tag suggestion and localization in user-generated videos based on social knowledge." in Proc. of ACM SIGMM Workshop on Social media, pp. 3-8, 2010.

[5] P. Felzenszwalb and D. Huttenlocher. "Efficient graph-based image segmentation." International Journal on Computer Vision, vol. 59, no. 2, pp. 167-181, 2004.

[6] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu, "Bipartite graph reinforcement model for web image annotation." in Proc. of ACM Multimedia, pp. 585-594, 2007.

[7] X. Li, C.G.M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting." IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1310-1322, 2009.

[8] D. Liu, M. Wang, X.-S. Hua, and H.-J. Zhang, "Semi-automatic tagging of photo albums via exemplar selection and tag inference." IEEE Transactions on Multimedia, vol. 13, no. 1, pp. 82-91, 2011.

[9] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation." IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 702-712, 2011.

[10] K. Yang, X.-S. Hua, M. Wang, and H.-J. Zhang, "Tag tagging: towards more descriptive keywords of image content." IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 662-673, 2011.

[11] N. Zhou, W.K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 7, pp. 1281-1294, 2011.

[12] L. Wu, S.C.H. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging." in Proc. of ACM Multimedia, pp. 135-144, 2009.

[13] A. Sun and S.S. Bhowmick, "Image tag clarity: in search of visual-representative tags for social images." in Proc. of ACM Workshop on Social Media, 2009.

[14] Y. Shen and J. Fan, "Leveraging loosely-tagged images and inter-object correlations for tag recommendation." in Proc. of ACM Multimedia, pp. 5-14, 2010.
[15] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media." in Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 971-980, 2007.
[16] L. Kennedy, S.-F. Chang, and I. Kozintsev, "To search or to label? Predicting the performance of search-based automatic image classifiers." in Proc. of ACM Workshop on Multimedia Information Retrieval, pp. 249-258, 2006.
[17] R. Yan, A. Natsev, and M. Campbell, "A learning-based hybrid tagging and browsing approach for efficient manual image annotation." in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.
[18] B. Sigurbjornsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge." in Proc. of ACM International Conference on World Wide Web, pp. 327-336, 2008.
[19] MSRC object image database, http://research.microsoft.com/en-us/projects/objectclassrecognition/
[20] Alipr, http://alipr.com
[21] S.G. Sevil, O. Kucuktunc, P. Duygulu, and F. Can, "Automatic tag expansion using visual similarity for photo sharing websites." Multimedia Tools and Applications, vol. 49, no. 1, 2010.
[22] D. Lowe, “Distinctive image features from scale-invariant keypoints.” International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.