Tag Suggestion and Localization for Web Videos by Bipartite Graph Matching

Wei-Ta Chu[1], Cheng-Jung Li[1], and Yeh-Kai Chou[2]
[1]National Chung Cheng University, Chiayi, Taiwan
wtchu@cs.ccu.edu.tw, zoneli1987@gmail.com
[2]Industrial Technology Research Institute, Hsinchu, Taiwan
KennyChou@itri.org.tw

ABSTRACT
In this paper, we formulate video tagging as a bipartite graph matching problem. Starting from existing tags that were originally provided by video owners, we conduct keyword-based image search on Flickr. Tags associated with the retrieved images are collected as candidate tags for tag suggestion. Relationships between keyframes extracted from the same video shot and candidate tags are then described as a bipartite graph, and best matching between two disjoint sets is accordingly determined to suggest new tags to this video shot. In constructing the bipartite graph, visual characteristics in terms of the bag of word model and tagging behaviors are jointly considered. Experimental results demonstrate that the proposed features and methodology achieves superior performance over previous approaches.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *indexing methods*. I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - *video analysis*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Tag suggestion, tag localization, bipartite graph matching, video annotation.

1. INTRODUCTION
Currently large amounts of videos are shared on the web, and thus impede the efficiency of large-scale video retrieval and browsing. In the last few years, video annotation or tagging has been widely studied to facilitate keyword-based video retrieval. Many works have been proposed to conduct video annotation based on audiovisual features, temporal information, spatial correlation, context between spatial/temporal information, and even social knowledge implicitly provided by users.

We can classify current video annotation works into two main categories: annotation by concept detection and annotation by social media analysis. As concept detectors flourish in recent years, many researchers tackle with this issue by detecting concepts in video frames, with main consideration on visual features. For example, Li et al. [5] jointly consider spatial correlation, temporal consistency, and temporal dependency of audiovisual features, and formulate video annotation as a sequence multi-labeling problem. Features directly extracted from video content are used to construct classifiers in this kind of work. On the other hand, the idea of social media analysis that exploits user's collective knowledge rather than content itself is recently proposed. Ballan et al. [4] proposed one of the most recent works about using social knowledge in video tagging. Based on existing tags that were originally provided by video owners and were used to describe the whole video, they conduct keyword-based image search on Flickr, retrieve relevant images associated with tags from Flickr, and then rank the retrieved tags to achieve tag suggestion. Instead of describing the whole video, each video shot is suggested a set of new tags (from retrieved tags), and thus tagging results are "localized" into corresponding shots.

Figure 1 illustrates the concept of tag suggestion and localization. The content owner has tagged this video by "iceland," "volcano," and "eruption." These tags may or may not suit every shot in this video. Therefore, the goal of our work is to find more tags, in addition to existing ones, that are appropriately to describe each video shot.

Although a considerable amount of works have been proposed to use concept detectors to annotate videos, performance of such annotation methods is limited due to the notorious semantic gap problem. Therefore, from the state-of-the-art research results, exploring social knowledge to facilitate video annotation seems a more promising approach. We conduct tag suggestion and localization based on the similar idea in [4]. Moreover, motivated by the bipartite graph reinforcement model [6] proposed to image annotation, we model relationship between keyfraems in a video shot and candidate tags as a bipartite graph, and then find best matching to determine the most appropriate tags. Comparing with the work in [4], we further investigate user's tagging behaviors and model tag suggestion as a graph matching problem. Comparing with the work in [6], we describe relationship between keyframes and candidate tags, rather than existing tags and candidate tags. The bag of visual word representation is used to measure visual similarity between images, with the designed adaptive weighting scheme to prioritize different visual words.

The remainder of this paper is organized as follows. Section 2 provides brief literature survey on image and video tagging. Section 3 provides an overview of the proposed framework.

WSM'11, November 28–December 2, 2011, Scottsdale, Arizona, USA.

Section 4 describes graph construction and matching. Experimental results are given in Section 5, followed by the concluding remarks in Section 6.

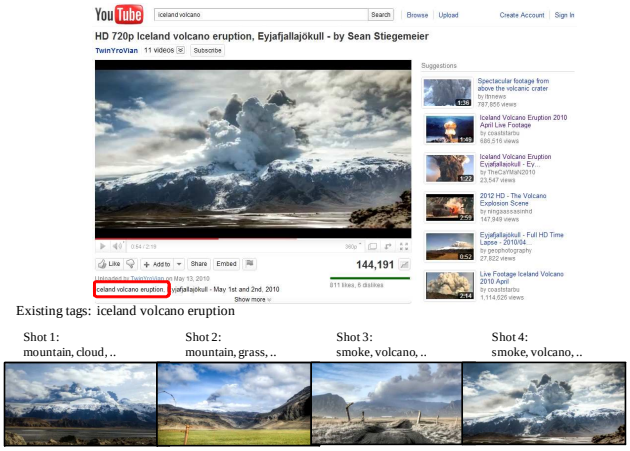

Figure 1. Illustration of tag suggestion and localization.

2. RELATED WORK

2.1 Image Tagging

As the emergence of Web 2.0 and photo sharing websites, image tags have been shown to be important clues to facilitate object recognition and image retrieval. Ames and Naaman [14] investigated the incentives of annotating photos in Flickr and claimed 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 were built to predict semantic concepts in images [15], which handle with the notorious semantic gap problem. Kennedy et al. [16] study performance variations between concept detectors trained by human-annotated data and that trained by data automatically retrieved from the web. They claimed that some concepts would gain much from human efforts. Yan et al. [17] studied manual annotation in a quantitative way and proposed a learning approach to suggest right images or right keywords to reduce annotation time. With a similar purpose, the work in [18] developed a recommendation strategy to support users in photo annotation. In [6], an image is annotated by jointly considering its surrounding text and extended candidate searched 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 [19] used the concept of language models to estimate effectiveness of a tag. From a different perspective, Wu et al. [20] enhance image tagging by learning a more appropriate distance metric.

More recently, Liu et al. proposed a semi-automatic approach that users just need to annotate a small set of representative images, and then the tags are appropriately propagated to related images [21][22]. To make tags more descriptive, Yang et al. [23] 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 [13] model loosely-tagged images and inter-object correlation by a multi-task SVM, and recommend tags for each object instance.

2.2 Video Tagging

Comparing with image tagging, relatively fewer studies have been conducted for video tagging. Ulges et al. [9] proposed one of the first few systems to tag web videos. They constructed statistical models based on global and local visual features, and then estimate the probability of pre-defined tags associated with a video shot. Siersdorfer et al. [10] observe visual redundancy between videos in Youtube, and extensively use the property to recommend tags to videos. Chen et al. [11] considered even richer web information such as news reports, videos and user comments to describe context of a web video. Their extensive work can also be found in [12], in which they verified the effectiveness of tag ranking by extensive experiments. Instead of recommending tags to each individual video shot, Li et al. [5] model video annotation as a sequence multi-labeling problem. They jointly consider spatial and temporal context in consecutive video shots and infer the best labeling sequence in a global optimization manner.

Similar to the idea of loosely-tagged images [13], Ballan et al. [4] suggested and localized tags into video shots, given video tags at the video level. Based on existing user-provided tags, they searched relevant images from Flickr and retrieved the associated tags with candidates. The degree of relevance of a tag to a video shot is estimated by tag co-occurrence frequency. For video tagging, our work has the same goal as [4]. However, we develop a systematic structure to describe relationships between tags, by jointly considering visual similarity, tag co-occurrence, and user's tagging behavior. Furthermore, the proposed unified framework can be extended to both video and image tag suggestion and localization, though temporal localization is conducted for videos and spatial localization is conducted for images.

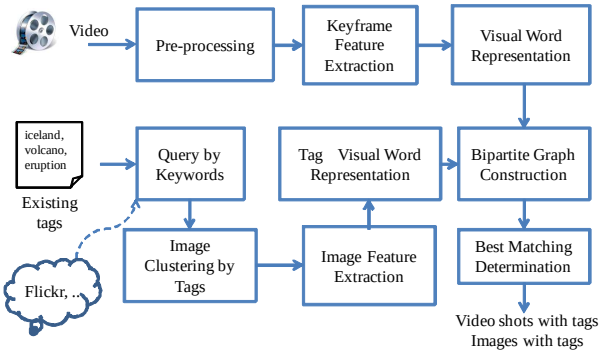

Figure 2. System framework.

3. OVERVIEW OF FRAMEWORK

Figure 2 shows framework of the proposed system. For video data, we first perform shot change detection, and then extract appropriate number of keyframes for each shot based on the global k-means algorithm [1]. The keyframes $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M\}$ are then represented as visual word histograms, in which visual words are derived from clustering SIFT (Scale-Invariant Feature Transform) descriptors [2]. The size of codebook for visual word representation is 50 in this work.

Without loss of generality, we assume that the content owner annotated this video by a single tag t_0. Based on t_0, we search

images on Flickr and retrieve top N images $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$. Each of the retrieved images may be associated with multiple tags, which were provided by corresponding owners, and these tags provide extensive knowledge to facilitate tag expansion and localization. We cluster images of the same tag together. That is, if there are *K* different tags in the retrieved images, *K* image clusters would be formed. Note that an image would be categorized into multiple clusters, because it may have multiple tags. Assume that the set $Y_i = \{\boldsymbol{y}_{i,1}, \boldsymbol{y}_{i,2}, ..., \boldsymbol{y}_{i,L}\}$ denotes the retrieved images associated with the tag t_i, and then we represent the tag t_i by the average visual word histogram of Y_i. With this design, tags are represented as the same way as keyframes.

A bipartite graph is then constructed, in which two disjoint sets of nodes respectively denote keyframes and tags, and each edge between nodes is associated with a weight calculated based on similarity between a pair of keyframe and tag, and tagging behaviors. We apply the Hungarian algorithm [3] to find the best matching between nodes, and determine corresponding tag for each keyframe. Tags associated with keyframes in the same shot are collected to expand annotation for each video shot.

4. BIPARTITE GRAPH CONSTRUCTION AND MATCHING

4.1 Weighting Scheme

Through querying Flickr by the existing tags, we retrieve the top 15 relevant images and collect 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 for describing a specific video shot. With this measurement, we could annotate each video shot with more tags (tag suggestion) that are temporally correlated with visual content (tag localization).

Let $X_j = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ denote keyframes in a video shot, and $T = \{t_1, t_2, ..., t_n\}$ denote the candidate tag pool collected from the retrieved images $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$. One critical issue of utilizing web-based social knowledge is noisy data. Due to user's subjectivity and tagging behaviors, the images that are associated with the tag t_i don't necessarily represent the concept t_i. Robustness of the average visual word histogram that represents t_i is deteriorated by these noisy images. Therefore, in the following we would like to decimate the influence of visual words derived from noisy images by an adaptive weighting scheme.

From the perspective of document analysis, some words play more important roles in presenting main concepts of a document. Based on images associated with the same tag, a tag is viewed as a document constituted by visual words. Different visual words should be prioritized differently so that similarity calculated based on visual word histogram can be estimated well.

Importance of a visual word for a tag t_i depends on two factors:

- A visual word is more important if it frequently appears in the image collection associated with the tag t_i.
- A visual word is more discriminative if it occasionally presents in some images' visual word histograms. If a visual word appears in all retrieved images, it provides less information for distinguishing truth data from noisy data.

According to the factors mentioned above, the term frequency - inverse document frequency strategy can be used to prioritize different visual words. 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. The weight of the *k*th 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 *k*th 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 *k*th bin of the visual word histogram of $\boldsymbol{y}_{i,\ell}$. The first term denotes the normalized occurrence frequency of the *k*th visual word. More frequently this visual word appears in Y_i, larger the first 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.

Note that the weightings are calculated according to the retrieved images associated with a specific tag, rather than all retrieved images. Therefore, these weightings are adaptively changed for different candidate tags, and thus distances between keyframes and different tags can be appropriately described.

4.2 Tagging Behavior

In addition to weight different visual words based on visual characteristics, we would like to further consider user's tagging behavior to more 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 an image, it implicitly represents consensus of more 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 denotes the number of users utilizing t_i to tag videos, 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 videos 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 videos that simultaneously contains tag 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.

4.3 Graph Construction

To discover relationship between video shots and tags, keyframes extracted from the same shot and candidate tags are respectively viewed as two disjoint sets, and we construct a weighted bipartite graph to describe their relationships. Figure 3 shows an example of such bipartite graph. Based on this graph, best matching between two sets of nodes is accordingly determined.

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

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

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.

4.4 Best Matching

Taking Figure 3 as an example, to provide tag suggestion for the third video shot, a bipartite graph is constructed to include three nodes representing three extracted keyframes and four nodes representing candidate tags. Weights on edges are determined by eqn. (4), and with this graph, the best matching between two sets of nodes are determined as follows.

Given a 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 a keyframe and a tag. Finally, a video shot is annotated by the collection of tags associated with the keyframes from this shot. If n keyframes are extracted from this video shot, n tags would be suggested to annotate this shot.

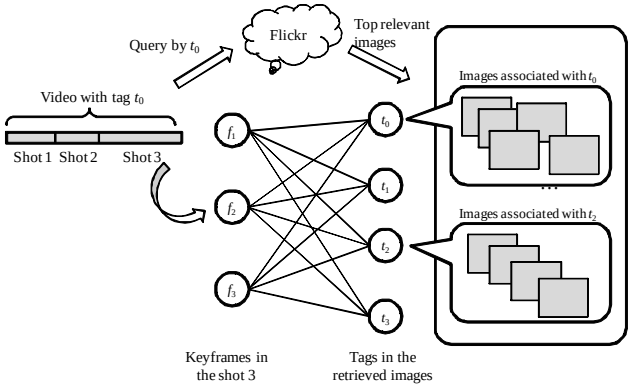

Figure 3. Illustration of a bipartite graph describing relationship between keyframes and tags.

4.5 Extension

The method described in Section 4 can be extended to image tagging. It's often the case that users just give a few tags to a photo album. We can view this album as a video, and segment it into several subsets by time-based clustering or content-based clustering. Photos in the same cluster are like keyframes mentioned above, and we can precede the same process to suggest new tags for each photo cluster. Verification on this extension would be conducted in the future.

5. EXPERIMENTS

5.1 Video Tagging for Youtube Videos – Exp1

We evaluate our system based on a video collection that includes top three rated videos from 15 categories in Youtube. There are totally 45 videos, consisting of 1368 shots and 3176 keyframes. User-provided tags are grabbed by the Youtube API, and there are averagely 11.67 tags for each video after filtering out stop words. The longest video contains 161 video shots, while the shortest video contains only one shot. Through Flickr API, we retrieve images relevant to a query, retrieve associated tags, and obtain user's names from Flickr.

- Performance of Tag Suggestion

In this section we would compare our work with a baseline method, which solely considers frequency of tags in the candidate pool to provide suggestion.

Baseline method. The baseline method suggests tags mainly based on occurrence frequency. Given a set of keyframes extracted from a video shot, e.g. $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, we find three images that are most similar to $\boldsymbol{x}_i$, $1 \le i \le 3$, from the retrieved image set. The tags associated with these images are collected, and the top five tags that most frequently appear in the tag pool are returned for tag suggestion.

Accuracy of tag suggestion for each video shot is defined as

$$accuracy = \frac{|T_C|}{|T_S|}, \quad (6)$$

where T_S denotes the set of suggested tags, and T_C denotes the set of correct tags which are manually evaluated. The notation $|\cdot|$ denotes the number of element in a set. Tagging accuracy values of shots in the same video are averaged, and the accuracy values for the top three videos downloaded from Youtube are averaged to show the final performance.

Table 1 shows performance of tag suggestion for different Youtube categories. The second, third, and fourth columns respectively show performance for the baseline method, our method jointly considering two tagging behaviors defined as eqn. (2) and (3), and our method considering the tag behaviors defined as eqn. (3) only. We see that our approach has significant performance improvement for all categories. This shows that jointly considering priority of different visual words and tagging behaviors give superior performance over conventional methods solely counting tag frequency. Because video owners provided too noisy tags ("Education") or we could not correctly retrieve relevant images corresponding to scientific terminologies ("Science & Technology"), we obtain much worse performance for these two categories. The work in [4] also evaluated tag suggestion based on Youtube videos, and averagely 0.36 accuracy value was achieved. It's not totally fair to directly compare our work with theirs because the evaluation dataset is not exactly the same. However, we still can see the trend that our approach has apparent superiority. Figure 4 shows two example results of a video from "Film & Animation" and a video from "News & Politics."

Comparing the third column with the fourth column, although we have the same average accuracy, performance variations for different categories reveal interesting observations. For the "Sports" category, names of players often co-occur increases the rank values of some tags that are not truly the players showing in video shots. Therefore, only taking tag co-occurrence into account doesn't achieve the best performance. On the other hand, the top-retrieved videos for the "Education" category are about guns, similar gun names co-occur frequently and truly present content in shots. Relatively fewer distinct users tag images with these gun names, and we obtain worse performance if both factors defined in eqn. (2) and (3) are considered.

- Performance of Tag Localization

To evaluate tag localization, we only examine how existing tags are located in video shots. For this purpose, the denominator in eqn. (6) is the number of existing tags that are located in this shot, while the numerator is the number of correctly located tags. The average localization accuracy is 0.79, which is higher than 0.63 reported in [4]. We have to note again that we don't evaluate exactly the same dataset as in [4] because the top-rated videos keep changing daily in Youtube. Moreover, how to evaluate tag localization is actually not clearly defined in [4].

Figure 4. Example results of tag suggestion.

Table 1. Performance of tag suggestion.

Categories	Baseline	Our (2)+(3)	Our (3)
Autos & Vehicles	0.31	0.83	0.9
Comedy	0.40	0.97	0.97
Education	0.11	**0.34**	**0.42**
Entertainment	0.38	0.94	0.96
Film & Animation	0.38	0.70	0.85
Gaming	0.30	0.59	0.58
Howto & Style	0.24	0.81	0.75
Music	0.23	0.83	0.86
News & Politics	0.26	0.59	0.6
Nonprofits & Activism	0.25	0.68	0.63
People & Blogs	0.45	0.96	0.96
Pets & Animals	0.25	0.73	0.71
Science & Technology	0.18	0.51	0.5
Sports	0.35	**0.83**	**0.68**
Travel & Events	0.44	0.72	0.7
Average	0.30	**0.74**	**0.74**

5.2 Video Tagging for MCG-WEBV – Exp2

The MCG-WEBV web video benchmark [8] is also used to verify our system. The CoreData part of this benchmark includes the "most viewed" videos in months from Dec. 2008 to Nov. 2009 (except for Aug. 2009). There are totally 14,473 videos in this part. Each video is segmented into shots that are then represented by several keyframes. Tags provided by users are associated with each video at the video level. We randomly select three videos from each month, and the shot boundaries, keyframes, and user-provided tags are used to construct the bipartite graph. Table 2 shows the selected videos in our experiments.

In this experiment we evaluate our method developed based on the eqn. (4) (jointly considering two tagging behaviors). Table 3 shows the average accuracy values for the baseline method, our tag suggestion method, and our tag localization method. These results have consistent trends to that in Section 5.1. These reveal that our system has consistent performance no matter the evaluated dataset covers a wide or a short temporal range of videos. Figure 5 shows performance variations of tag suggestion and localization in different time periods. Accuracy values for the three videos in the same month are averaged. From Figure 5 we see it is not necessary that our method performs better when the baseline method has better performance. For example, in May 2009, lower accuracy is achieved by the baseline method, but our method achieves very high accuracy. Because the baseline method solely considers tag frequency, such performance variation highlights the importance of analyzing tagging behavior. We can view user's tagging behaviors as orthogonal information to tag occurrence frequency.

Table 2. Information of the selected evaluation videos from MCG-WEBV.

VideoID	Video title	Tags
2008-12 3135481	All-New 2010 BMW Z4 Roadster	2010, BMW, Z4, Roadster
3107302	President Bush Attacked By Shoes	Bush, Shoes, MSNBC
3107350	ein Iraqi wirft Bush Schuhe / MamThrows Shoes At Bush	Bush, Schuh, Iraq, schuhe, schoes
2009-01 3138158	Google Latitude	Google, Latitude
3147786	Land of the Lost - Superbowl TV Spot	Danny McBride, Anna Friel
3139850	So cute, does anyone know what is this animal called?	Cute, animal, amazing, beautiful
2009-02 3138631	3138631 Lim Ding Wen	iPhone
3147696	Nokia 5630 XpressMusic	Nokia, 5630, XpressMusic, music
3147839	february gmc truck cold start	truck, gmc, chev, chevy, cold, start, crank, pump, pedal, prime, gas, 400, sb, v8, carb, winter
2009-03 3248576	Tinchy Stryder Ft. N -Dubz - Number 1	Tinchy, Stryder, N-Dubz, Dappy, Number
3248772	Japanese National Robot HRP 4C In technology of Honda Motor 01	Japanese, National, Robot, HRP, 4C, 01
3248821	Mimiron's Flying Mount	wotlk, mount, mimiron, ulduar, world of warcraft, raid
2009-04 3251306	American Idol scream heard round the world!	AI Gokey, cat, American Idol, Danny
3251189	Fat Kid and TNT	Fat Kid, and, TNT, comedy, ac, dc, funny
3251142	Denise Richards 7th Inning Stretch	MLB, Denise Richards, Wrigley Field, 7th inning stretch
2009-05 3279010	Kimi Raikkonen crash! Rally della Marca 2009	raikkonen, crash, kimi, rally, marca, 2009
3279064	BMW X5 on the beach	BMW, X5, on the beach
3279141	Monkeys	Monkeys
2009-06 3280543	The All New XJ	New XJ, Jaguar
3280775	Firefox TV	Firefox, TV
3280880	Zombieland Trailer HD	zombieland, zombie, movie, trailer, hollywood.com
2009-07 3281746	Ferrari 458 Italia teaser	Ferrari, 458, F458, Italia
3281883	News Anchor Fail	News, Anchor, Fail, Failblog, Blog, Funny, Videos, Comedy
3282073	little girl goes fishing	little, girl, fishing
2009-09 3283247	Samoa Tsunami 2009	Samoa, Tsunami, 2009
3283283	LEXUS LF-A	LEXUS, LF-A, CAR
3284312	What Dogs Are Really Thinking	Dogs, talking, translation, dog, chihuahua, talk, funny, funniest, puppy
2009-10 3284716	Bugatti Veyron Lake Crash!	Bugatti, veyron, crash
3285523	Thierry Henry handball controversy	France, Ireland, handball, Thierry, henry, maradona, world cup
3285943	Opera Mobile 10 beta	Opera, mobile 10, beta, software, nokia, smartphones, symbian, s60, browsing
2009-11 3286629	Insane Canadian Fisherman	Insane, Canadian, Fisherman, funny, moments, today, break
3287783	Kobe Hits From Behind the Backboard	Nba, amazing, highlights, Kobe Bryant, Los Angeles Lakers
3287864	Dharni from Singapore - Beatbox Battle TV	Beatbox, Dharni, Asia

Table 3. Performance of tag suggestion and localization for selected MCG-WEBV videos.

	Baseline	Tag suggestion	Tag localization

Average accuracy	0.28	0.73	0.74

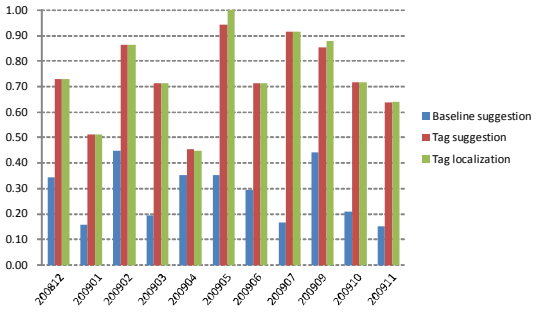

Figure 5. Performance of tag suggestion and localization in different time periods.

6. CONCLUSION

To accomplish tag suggestion and location for web videos, we search relevant images from user-shared photo collections based on existing tags, and then model relationship between tags associated with retrieved images and keyframes of the original video 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 with the proposed method we can well capture association between keyframes and tags, and achieve significantly better performance in tag suggestion. Relationship between video shots and more social knowledge would be investigated to more enhance the performance in the future.

Acknowledgement

The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 100-2221-E-194-061 and NSC 99-2221-E-194-036.

7. REFERENCES

[1] Likas, A., Vlassis, N., and Verbeek, J.J.. The global k-means clustering algorithm. Pattern Recognition, vol. 36, pp. 451-461, 2003.

[2] Lowe, D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.

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

[4] Ballan, L., Bertini, M., Del Bimbio, A., Meoni, M., and Serra, G. 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] Li, Y., Tian, Y., Duan, L.-Y., Yang, J., Huang, T., and Gao, W. Sequence multi-labeling: a unified video annotation scheme with spatial and temporal context. IEEE Transactions on Multimedia, vol. 12, no. 8, pp. 814-828, 2010.

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

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

[8] Cao, J., Zhang, Y.D., Song, Y.C., Chen, Z.N., Zhang, X., and Li, J.T. MCG-WEBV: A benchmark dataset for web video analysis. Technical Report, ICT-MCG-09-001, Institute of Computing Technology, May. 2009.

[9] Ulges, A., Schulze, C., Keysers, D., and Breuel, T. Content-based video tagging for online video portals. In Proc. of MUSCLE/ImageCLEF Workshop, 2007.

[10] Siersdorfer, S., Pedro, J.S., and Sanderson, M. Automatic video tagging using content redundancy. In Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 395-402, 2009.

[11] Chen, Z., Cao, J., Song, Y., Guo, J., Zhang, Y., and Li, J. Context-oriented web video tag recommendation. In Proc. of International Conference on World Wide Web, pp. 1079-1080, 2010.

[12] Chen, Z., Cao, J., Xia, T., Song, Y., Zhang, Y., and Li, J. Web video retagging. Multimedia Tools and Applications, 2011.

[13] Shen, Y., and Fan, J. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In Proc. of ACM Multimedia, pp. 5-14, 2010.

[14] Ames, M., and Naaman, M. 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.

[15] Zhou, N., Cheung, W.K., Qiu, G., and Xue, X. 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.

[16] Kennedy, L., Chang, S.-F., and Kozintsev, I. 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] Yan, R., Natsev, A., and Campbell, M. 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] Sigurbjornsson, B., and van Zwol, R. Flickr tag recommendation based on collective knowledge. In Proc. of ACM International Conference on World Wide Web, pp. 327-336, 2008.

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

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

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

[22] Liu, D., Yan, S., Hua, X.-S., and Zhang, H.-J. Image retagging using collaborative tag propagation. IEEE Transactions on Multimedia, 2011.

[23] Yang, K., Hua, X.-S., Wang, M., and Zhang, H.-J. Tag tagging: towards more descriptive keywords of image content. IEEE Transactions on Multimedia, 2011.