Tag suggestion and localization in user-generated videos based on social knowledge

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo,
Marco Meoni, and Giuseppe Serra
Media Integration and Communication Center (MICC), University of Florence, Italy
{ballan, bertini, delbimbo, meoni, serra}@dsi.unifi.it
http://www.micc.unifi.it

ABSTRACT

Nowadays, almost any web site that provides means for sharing user-generated multimedia content, like Flickr, Facebook, YouTube and Vimeo, has tagging functionalities to let users annotate the material that they want to share. The tags are then used to retrieve the uploaded content, and to ease browsing and exploration of these collections, e.g. using tag clouds. However, while tagging a single image is straightforward, and sites like Flickr and Facebook allow also to tag easily portions of the uploaded photos, tagging a video sequence is more cumbersome, so that users just tend to tag the overall content of a video. Moreover, the tagging process is completely manual, and often users tend to spend as few time as possible to annotate the material, resulting in a sparse annotation of the visual content. A semi-automatic process, that helps the users to tag a video sequence would improve the quality of annotations and thus the overall user experience. While research on image tagging has received a considerable attention in the latest years, there are still very few works that address the problem of automatically assigning tags to videos, locating them temporally within the video sequence. In this paper we present a system for video tag suggestion and temporal localization based on collective knowledge and visual similarity of frames. The algorithm suggests new tags that can be associated to a given keyframe exploiting the tags associated to videos and images uploaded to social sites like YouTube and Flickr and visual features.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Algorithms, Design, Experimentation

Keywords

Tag suggestion, user-generated content, social video retrieval

WSM'10, October 25, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-4503-0173-2/10/10 ...$10.00.

1. INTRODUCTION

In recent years social websites for media sharing have become more and more popular, allowing people to easily upload, share and annotate personal media content with keywords usually referred to as *tags*. These tags provide additional contextual and semantic information by which users can organize and access shared media content. Flickr and YouTube are probably the most popular social image/video sharing web sites. Flickr hosts more than 2 billion images with about 3 millions new uploads per day. YouTube reported in March 2010 more than 2 billion views a day, 24 hours of videos uploaded per minute, and it also estimated that a common user spends on average 15 minutes each day.

Currently the performance of image and video retrieval systems depends mainly on the availability and quality of tags. However existing studies show that tags are few, imprecise, ambiguous and overly personalized [3]. A study on how users tag photos, analysing the kind of tags that are provided, was presented by Sigurbjörnsson and van Zwol [12]; in particular, their analysis of 52 million photos has shown that 64% have few tags (between 1 and 3 tags). For this reason, the authors have presented and compared several tag recommendation strategies, mainly based on tag co-occurrence, to ease the task of tagging the images, proposing the $Vote^+$ algorithm as the more stable and best performing. The proposed approaches do not consider visual content of the images, focusing only on tags and their relations. Another tag recommendation approach has been proposed by Wu *et al.* [16], which learns an optimal combination of tag and visual correlations to generate a ranking function. To address the problem of tag reliability, Li *et al.* [5, 7] have estimated the tag relevance by voting of visually similar images. Tag relevance was used in a tag-based retrieval system, to improve the precision of the queries. Recently, the approach has been extended combining global and local features to better represent visual image content [6]. The problem of tag reliability has been considered also in [4]. In this work Kennedy *et al.* have proposed a method for gathering reliably tagged images; couples of visually similar images are used to select the common tags, as in the ESP game [14]. Liu *et al.* [8] have proposed a tag ranking approach in which the tags of an image are ranked according to their relevance to the content of the image. First is adopted a probabilistic approach to estimate the initial relevance score for each tag, and then the score is refined by using a random walk process over the tags graph.

Most of recent works that tackle the problems related to internet videos have addressed the problem of detect-

VIDEO TAGS: wild, Africa, sky, sunrise, mountain, lake, waterfall, lion, elephant savannah, wildebeest, leopard, zebra, rain, desert, BBC

Figure 1: Example of a YouTube video with its related tags.

ing near-duplicates [13, 17, 18], because of the fact that a large number of videos uploaded in the video sharing websites have a similar content (search results on YouTube may contain 15 - 27% duplicates [11, 17]) with small variations due to, for example, video editing, compositing and filtering, transcoding or logo/superimposed text insertions. While several methods have been proposed to deal with tags for images [2,4,5,10], the problem of video classification and tagging has been less explored. In [19] Wu *et al.* have proposed a video categorization approach that uses title and tags of a video, the tags of related videos (as selected by YouTube) and of videos of the same user. Liu *et al.* [9] have proposed a method for video topic detection using the "related video" links that YouTube associates to each video to enrich the textual information of a single video. Both these approaches do not consider visual information and depend on the specific features and metadata provided by YouTube. Tags of user-generated videos are typically few and imprecise, and moreover they are usually associated to the entire video and are not located temporally within the sequence. Figure 1 shows a video example from YouTube in which we can see that tags, such as "leopard" or "waterfall", are associated to the entire video. For this reason, the users that searches for a specific tag are forced to watch the whole sequences of the retrieved videos. To the best of our knowledge, few papers address the problem of tag reliability in videos but do not deal with the problem of locating temporally tags within the video sequence. Siersdorfer *et al.* [11] have proposed a tag suggestion and re-ranking approach that exploits near-duplicate and overlapping videos, creating a graph of visually overlapping videos to propagate tags. Zhao *et al.* [20] have exploited an efficient video similarity detection algorithm to retrieve visual near-duplicates of a video, then the tags associated to these videos are re-ranked to suggest new tags. However, these approaches are feasible only for videos that are popular enough to be edited and slightly modified by other users, that have to add also new tags. In [1], Choudhury *et al.* have proposed a method to enrich and rank tags from YouTube videos. First, tags are expanded using contextual information (title and description of the video) and social contexts (e.g. tags of related videos or from playlists that include the video). After they first compute a ranking, based on tag co-occurrence, and then link tags to DBpedia concepts. However the visual information is completely ignored.

In this paper we propose a method for video tag suggestion and temporal localization based on social knowledge. The system exploits the tags associated to user-generated videos and images uploaded to social sites (such as YouTube and Flickr) and their visual similarity, to suggest new tags that can be associated to a particular keyframe at the shot level. The rest of the paper is organized as follow. The proposed method is discussed in details in Section 2; experimental results are presented in Section 3. Conclusions are finally drawn in Section 4.

2. EXPLOITING TAG RELEVANCE FOR VIDEO ANNOTATION

The approach proposed in this paper aims at two goals: to extend the number of tags associated to each video and, at the same time, associate the tags to the relevant shots that compose the video. The first goal is related to the fact that the videos available on media sharing sites like YouTube have relatively few tags (in [20] was observed an average of about only 5 tags per video) that do not allow to annotate thoroughly the content of the whole video. The second goal is related to the fact that tags describe the global content of a video, but they may be associated only to certain shots and not to others. To cope with the large number of videos uploaded daily to media sharing sites, the approach is unsupervised.

Video annotation is performed in two stages; an overview of the approach is shown in Fig. 2. In the first stage the relevance of the video tags is computed for each shot, possibly eliminating tags that are not relevant, then new tags are added to each shot. Each video is segmented into shots, using a fast and simple algorithm that analyzes frame luminance and uses a global threshold to detect transitions and large content changes (in principle this segmentation can be substituted by a simple temporal frame subsampling scheme). From each shot are extracted three keyframes, one from the start, one from the middle and one from the end of the shot, creating a set $K = \{k_1, ..., k_o\}$ of keyframes. The tags $V = \{v_1, ..., v_n\}$ associated to a video are used to select and download from Flickr a set of images $I_{v_i} = \{i_1, ..., i_m\}$, that have been annotated using each tag $v \in V$. Each image $i_j \in I_{v_i}$ has the following set of tags $\{t_1, ..., t_l, v_i\}$. Let $T = \{t_1, t_2, ..., t_k\}$ be the union of all the tags of $I = \{I_{v_1}, ..., I_{v_n}\}$, after that they have been filtered to eliminate stopwords, dates, tags containing numbers, punctuations and symbols. T is considered as the dictionary to be used for the annotation of the video. Since it has been obtained from images tagged by amateurs, such as Flickr users, it is fundamental to evaluate the relevance of the terms that compose the lexicon, to avoid adding incorrect annotations. To this end we have adapted the algorithm for the evaluation of tag relevance in visually similar images, presented in [7], to cope with video shot annotation. Practically, learning tag relevance is based on computing the count of a tag t in the k-nearest neighbours of image i minus the prior frequency of t; this is based on the consideration that the occurrence frequency of t in the visual neighbours of i is related to the importance of t for i. This requires to retrieve efficiently k-nearest visual neighbours for the keyframes and images to be analyzed.

For all the keyframes in K and Flickr images in I, is computed a 72-dimensional visual feature vector that represent global information of color and texture. The vector is composed by a 48 dimensional color correlogram computed in the HSV color space, 6 color moments computed in the RGB color space and 18 dimensional vector for three Tamura

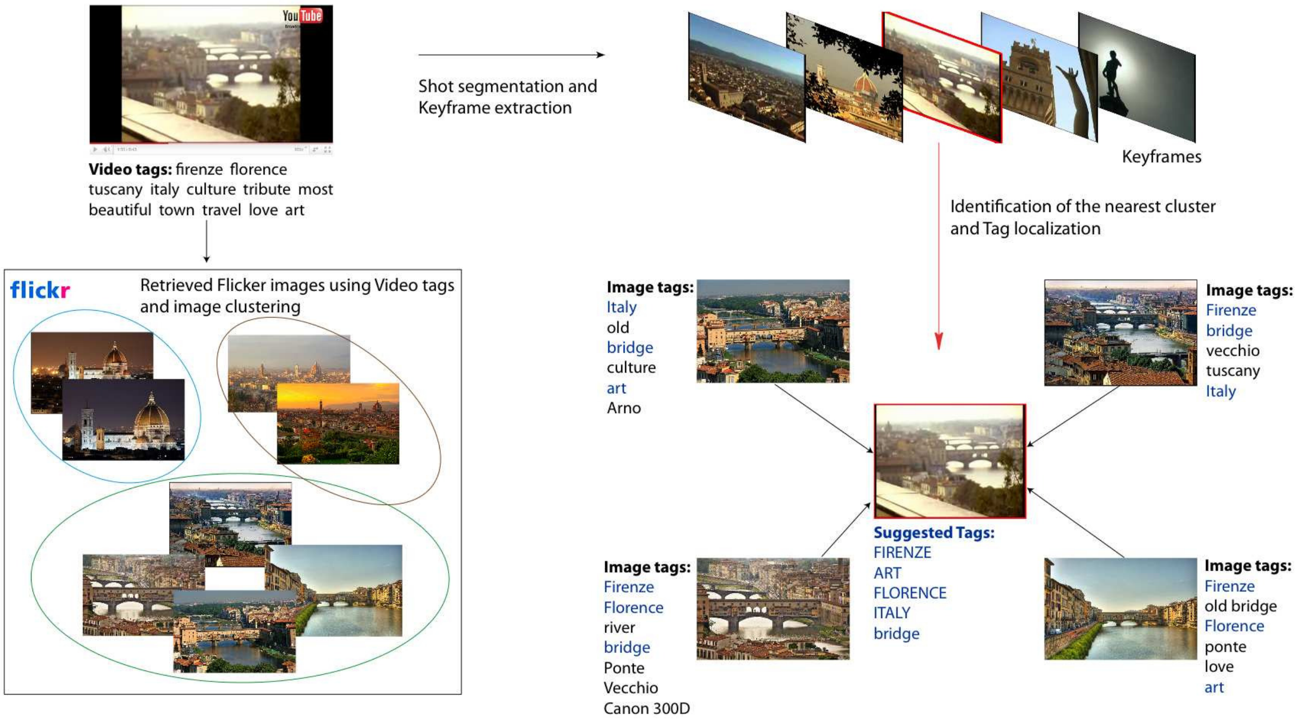
YouTube
Video tags: firenze florence tuscany italy culture tribute most beautiful town travel love art
Shot segmentation and Keyframe extraction
Keyframes
Identification of the nearest cluster and Tag localization
flickr
Retrieved Flicker images using Video tags and image clustering
Image tags: Italy old bridge culture art Arno
Image tags: Firenze bridge vecchio tuscany Italy
Image tags: Firenze Florence river bridge Ponte Vecchio Canon 300D
Image tags: Firenze old bridge Florence ponte love art
Suggested Tags: FIRENZE ART FLORENCE ITALY bridge

Figure 2: Overview of the proposed system.

features that account for texture information (in particular we have used coarseness, contrast and directionality). This combination of features has a low computational cost and has been shown to be effective for scalable image annotation [15]. The images in I are clustered using k-means, because of its convergence speed and empirical success in content-based image analysis and retrieval. Cluster centers are used as an index for approximate nearest neighbor search based on visual similarity of the $k \in K$ keyframes.

For each $k \in K$ keyframe is retrieved the nearest cluster center, and the images belonging to that cluster are selected as neighbors. The set $\{v_1, ..., v_n\}$ is considered as the set of tags of k since in this case, unlike in [7], k has no associated tag.

However, following this simplistic approach does not yield good results for video annotation: in fact the video tags may be associated only with certain keyframes, i.e. some v may be related only to a certain shot and not to another; considering all the $v \in V$ for each shot would simply result in a re-ranked list of the same tags. To solve this problem we have adopted the following approach: a tag v is kept in the list T_k of the tags of k, and thus its relevance is computed, only if it is present among the tags of the visual neighbourhood. In case that a relevant tag is incorrectly eliminated in this phase, it may be recovered during the following stage of annotation. Also the WordNet synonyms of all the v that are kept after this filtering, are added to extend the list T_k associated to each shot. To compute the relevance of synonyms a new set of images is downloaded from Flickr but, to cope with the fact that synonyms introduce a semantic drift, we heuristically chose to download a number of images that is one third of that used for the $\{v_1, ..., v_n\}$ tags.

To add new tags to each shot we compute a set of candidate tags for each shot from the dictionary T. $\forall t \in T$ is computed its tag relevance and resulting rank position $rank_t$. For each tag associated with a keyframe ($t_k \in T_k$), as obtained from the previous step, is computed the co-occurrence with all the $t \in T$, creating a tag candidate list C of the tags that have a co-occurrence value that is above the average. $\forall c \in C$ is computed a suggestion score $score(c, T_k)$, according to the $Vote^+$ algorithm. Finally, for each candidate tag of each keyframe k is computed a score according to the suggestion score proposed in [7]:

$$score(c, k) = score(c, T_k) \cdot \frac{\lambda}{\lambda + (rank_c - 1)}$$

This score is used to order the tags to be added to the shot, and the five most relevant tags are then used to annotate the shot.

The union of all the tags added to all the shots in this second step, is used to annotate the video at the global level.

3. EXPERIMENTAL RESULTS

We have evaluated the performance of our proposed approach using a dataset designed to represent the variety of content on YouTube. The dataset was created by choosing 4 YouTube videos selected from each of the 14 categories used by YouTube, to cover the most of different types of videos. The number of detected shots is 1135, resulting in 3405 keyframes analyzed. All the videos in the dataset had been previously tagged by YouTube users. The number of tags per video varies from a minimum of 3 to a maximum of 26. The videos and the related tags were collected through the YouTube API[1].

[1] YouTube APIs and Tools
http://code.google.com/apis/youtube

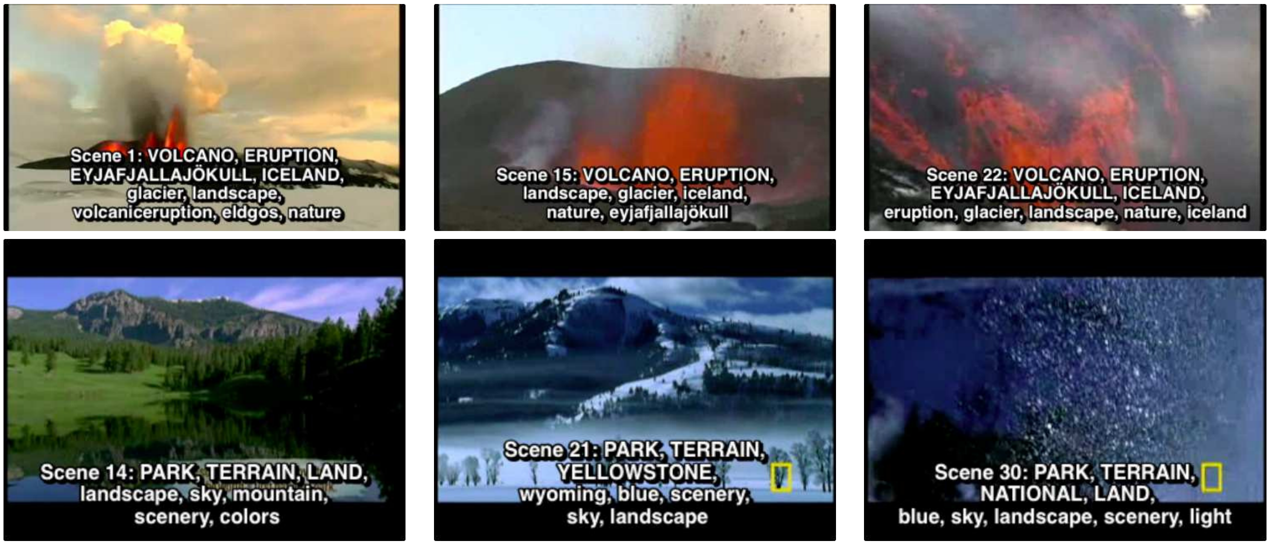
Scene 1: VOLCANO, ERUPTION, EYJAFJALLAJÖKULL, ICELAND, glacier, landscape, volcaniceruption, eldgos, nature
Scene 15: VOLCANO, ERUPTION, landscape, glacier, iceland, nature, eyjafjallajökull
Scene 22: VOLCANO, ERUPTION, EYJAFJALLAJÖKULL, ICELAND, eruption, glacier, landscape, nature, iceland
Scene 14: PARK, TERRAIN, LAND, landscape, sky, mountain, scenery, colors
Scene 21: PARK, TERRAIN, YELLOWSTONE, wyoming, blue, scenery, sky, landscape
Scene 30: PARK, TERRAIN, NATIONAL, LAND, blue, sky, landscape, scenery, light

Figure 3: Example of tag suggestion and localization: the top keyframe is part of a video in the "Travels" category, the bottom one is part of a video in "People and Blogs" category. The upper cap tags (e.g. *PARK*, *TERRAIN*, *LAND*, *VOLCANO*, *ERUPTION*) are part of the set of tags associated with the whole video that have been localized in this shot, while the lower cap tags (e.g. *landscape, sky, mountain, scenery, colors, glacier, iceland, nature, eyjafjallajökull*) have been both suggested and localized in this shot.

For each YouTube tag our system downloads the first 15 Flickr images ranked according to the "relevance" criterion provided by the Flickr API[2]. Furthermore, in the WordNet query expansion experiments the system downloads 5 additional Flickr images for each WordNet synonym.

To ease the task of manually evaluating the results of tag suggestion and localization, the system outputs its results in MPEG-7 and SRT subtitle format[3], that can be shown while playing the video. The subtitles contain both original and suggested tags, to facilitate the manual checking of results.

To evaluate the performance of our approach we use accuracy, computed as the proportion of true positives against the total number of true and false positives. Fig. 3 shows an example of the output of the system obtained on two videos of two different categories (respectively "Travel & Events" and "People and Blogs"): the tags written in capital letters (e.g. *PARK*, *TERRAIN*, *LAND*, *VOLCANO*, *ERUPTION*) are those that are part of the original set of YouTube tags associated with the video. The tags in lower caps are those suggested for this shot (e.g. *landscape, sky, mountain, scenery, colors, glacier, iceland, nature, eyjafjallajökull*[4]). In the experiments we evaluated the performance of our system in terms of:

- **Shot level tag localization: STL.** Evaluation of the performance of the tag localization at shot level. This measure shows the accuracy of YouTube video tags localization in the correct shots.

- **Shot level tag suggestion and localization: STSL.** Evaluation of the performance of the tag localization at shot level for both user-generated and suggested tags.

- **STSL with WordNet query expansion: STLS-WN.** Performance measure of STSL with WordNet synset expansion of the YouTube tags that have been kept at the end of the localization process.

The results are reported in Table 1. The overall performance of the system is encouraging. We can notice that average accuracy of tag localization is analogue to the image-based existing methods [7]. As expected, user-generated tag localization accuracy is always higher than the accuracy of tag suggestion and localization. The WordNet synset expansion method can outperform the basic suggestion method in the majority of the categories, improving the selection of Flickr similar images through the added semantic information. The cases in which the use of synonyms reduces the performance are due to differences in the context of the meaning of the terms of the WordNet synset w.r.t. the context of the video (e.g. *dog* and *hot dog*).

From the results we can also see that some categories are more tractable with our approach than the others. In the 10 best performing categories usually visual content is more closely related the tags used and to the category itself, or certain tags can be applied to many shots of the same video (e.g. the names of football teams in sports videos). Tags associated with scenes and landscapes obtain good results (e.g. *airplanes* and *boats*, *waterfall* and *volcano*), because the global features used to measure the similarity of frames with the Flickr photos are able to capture the overall content of the setting. This motivates the good results for "Cars & Vehicles" and "Travel & Events" categories. Also concepts related to topics of interest for social communities achieve a good performance; this is probably due to the fact that in this case users usually provide more accurately tagged videos and images, as shown for the videos belonging to the "People & Blogs" category. Instead, categories such as "News & Politics" or "Comedy" gather content that is typically tagged using tags that are more related to the feelings

[2] Flickr API Documentation:
http://www.flickr.com/services/api/
[3] SubRip format: http://en.wikipedia.org/wiki/SubRip
[4] The name of the Icelandic volcano filmed in this video.

and political views of the user than that of the visual content of videos. Other categories such as "Gaming" ,"Entertainment", "Howto & Style" gather too heterogeneous and different content to be correctly discriminated by the features described in Sect. 2.

YouTube category	**STL**	**STSL**	**STSL-WN**
Cars & Vehicles	0.74	0.42	0.49
Comedy	0.47	0.23	0.25
Education	0.63	0.43	0.51
Entertainment	0.39	0.18	0.13
Film & Animation	0.79	0.52	0.65
Gaming	0.57	0.11	0.10
Howto & Style	0.47	0.21	0.18
Music	0.62	0.39	0.51
News & Politics	0.59	0.44	0.50
People & Blogs	0.86	0.58	0.66
Pets & Animals	0.44	0.40	0.15
Science & Technology	0.62	0.55	0.35
Sport	0.91	0.21	0.38
Travel & Events	0.72	0.31	0.21
Average	0.63	0.35	0.36

Table 1: Tag localization and tag suggestion accuracy at video and shot level, with or without WordNet query expansion. VTS: tag suggestion at video level. STL: tag localization at shot level. STSL: tag suggestion and localization at shot level. STSL-WN: tag suggestion and localization at shot level with WordNet query expansion.

4. CONCLUSIONS

In this work we have presented a method for video annotation based on social knowledge. The tags provided by the users that upload the videos are localized within the shots and new tags are added. The preliminary results are encouraging for almost all the categories of videos uploaded to YouTube. Our future work will deal with the improvement of the features used to evaluate the visual similarity of keyframes: in fact some of the errors of tag suggestion are due to the use of global features that do not allow to discriminate certain tags (e.g. names of players and teams in sport videos). Other work will deal with exploitation of semantic relations between tags and the use of other sources of social knowledge to improve semantic relatedness of the suggested tags.

Acknowledgments. This work is partially supported by the EU IM3I Project (Contract FP7-222267).

5. REFERENCES

[1] S. Choudhury, J. Breslin, and A. Passant. Enrichment and ranking of the YouTube tag space and integration with the linked data cloud. In *Proc. of International Semantic Web Conference (ISWC)*, 2009.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. of ICCV*, 2009.

[3] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label? Predicting the performance of search-based automatic image classifiers. In *Proc. of ACM MIR*, 2006.

[4] L. S. Kennedy, M. Slaney, and K. Weinberger. Reliable tags using image similarity. In *Proc. of ACM MM Workshop on Web-Scale Multimedia Corpus*, Beijing, China, 2009.

[5] X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proc. of ACM MIR*, 2008.

[6] X. Li, C. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proc. of ACM CIVR*, 2010.

[7] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.

[8] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. of International World Wide Web Conference (WWW)*, 2009.

[9] Y. Liu and N. Yu. Dual linkage refinement for YouTube video topic discovery. In *Proc. of IEEE ICME*, 2010.

[10] S. G. Sevil, O. Kucuktunc, P. Duygulu, and F. Can. Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools and Applications*, 49(1):81–99, 2009.

[11] S. Siersdorfer, J. San Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *Proc. of ACM SIGIR*, pages 395–402, New York, NY, USA, 2009.

[12] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of International World Wide Web Conference (WWW)*, 2008.

[13] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *Proc. of ACM Multimedia*, pages 145–154, 2009.

[14] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of ACM Conference on Human Factors in Computing Systems*, 2004.

[15] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Scalable search-based image annotation of personal images. In *Proc. of ACM MIR*, pages 269–278, New York, NY, USA, 2006.

[16] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *Proc. of International World Wide Web Conference (WWW)*, 2009.

[17] X. Wu, A. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *Proc. of ACM Multimedia*, pages 218–227, 2007.

[18] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan. Real-time near-duplicate elimination for web video search with content and context. *IEEE Transactions on Multimedia*, 11(2):196–207, 2009.

[19] X. Wu, W.-L. Zhao, and C.-W. Ngo. Towards Google challenge: Combining contextual and social information for web video categorization. In *Proc. of ACM Multimedia*, 2009.

[20] W. Zhao, X. Wu, and C. Ngo. On the annotation of web videos by efficient near-duplicate search. *IEEE Transactions on Multimedia*, to appear in 2010.