Enriching and Localizing Semantic Tags in Internet Videos

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra
Media Integration and Communication Center, Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy
{ballan, bertini, delbimbo, serra}@dsi.unifi.it

ABSTRACT

Tagging of multimedia content is becoming more and more widespread as web 2.0 sites, like Flickr and Facebook for images, YouTube and Vimeo for videos, have popularized tagging functionalities among their users. These user-generated tags are used to retrieve multimedia content, and to ease browsing and exploration of media collections, e.g. using tag clouds. However, not all media are equally tagged by users: using the current browsers is easy to tag a single photo, and even tagging a part of a photo, like a face, has become common in sites like Flickr and Facebook; on the other hand tagging a video sequence is more complicated and time consuming, so that users just tend to tag the overall content of a video. In this paper we present a system for automatic video annotation that increases the number of tags originally provided by users, and localizes them temporally, associating tags to shots. This approach exploits collective knowledge embedded in tags and Wikipedia, and visual similarity of keyframes and images uploaded to social sites like YouTube and Flickr.

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 refinement, tag relevance learning, internet videos, social video retrieval

1. INTRODUCTION

In the last years, social media repositories such as Flickr and Youtube have become more and more popular, allowing users to upload and share media content[1], annotating it with personal keywords called *tags*. Tags provide contextual and semantic information which can be used to organize and facilitate media content search and access. The performance of social image and video retrieval systems depends mainly on the availability and quality of tags. However, these are often imprecise, ambiguous and overly personalized [2], and also very few (typically one-three tags per image) [13].

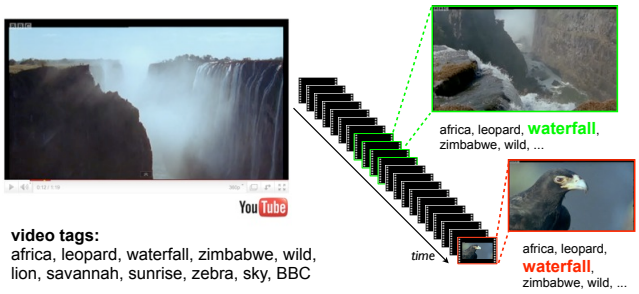

Figure 1: ***left)*** **Example of a YouTube video with its related tags;** ***right)*** **localization of tags in shots.**

Several efforts have been recently done in the area of content-based tag processing for social image retrieval [7]. The main focus of these works has been put on three aspects: *i)* tag ranking/relevance, *ii)* tag refinement and *iii)* tag-to-region assignment. Li *et al.* [4] addressed the problem of tag relevance estimation by accumulating votes from visually similar images. The basic idea is that if different users label similar images with the same tags, these tags truly represent the actual visual content. This method has been recently extended to a multi-feature relevance learning approach, obtained by combining global and local features to better represent visual image content [5]. In the tag ranking approach proposed by Liu *et al.* [6], tags are ranked according to their relevance to the content of images. First is adopted a Kernel Density Estimation approach to estimate the initial relevance score for each tag, then score values are refined by using a random walk process that explores the relationship of tags (represented using a graph model).

Most of the recent works on internet videos have addressed problems like near duplicate detection [11], training concept detectors [15] or topic detection [12]. Currently, only a few works have considered the problem of tag suggestion and

[1]For instance, YouTube reported in March 2011 that about 24 hours of video are uploaded every minute and more than 2 billion views per day have been registered.

MM'11, Nov 28–Dec 1 , 2011, Scottsdale, Arizona, USA.

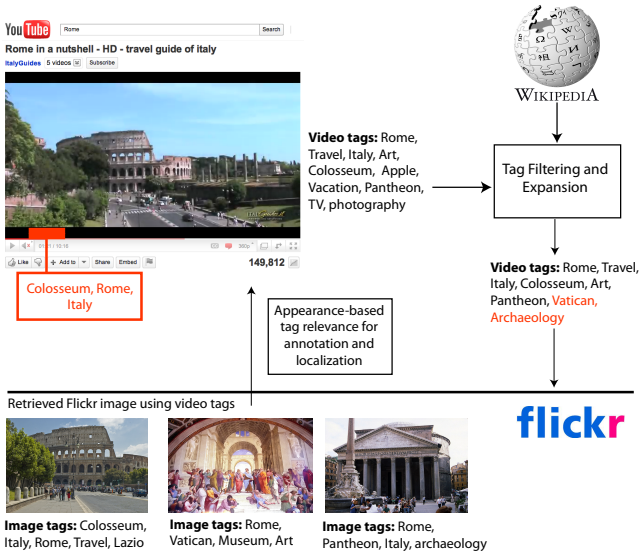

Figure 2: Overview of the system.

localization (see Fig. 1) in internet videos [1, 3, 10]. In [1] shots of YouTube videos are automatically annotated using Flickr images, with a tag relevance algorithm that, exploiting visual similarity of keyframes and images, can also add new tags that were not originally available in videos. Localization of video tags is addressed in [3]; a multiple instance learning approach that considers semantic relatedness of co-occurring tags is used to model shots and videos. In [10] video shots are annotated with 34 concept detectors, using their results to build a semantic representation for each shot. The same detectors are applied to Flickr images and semantic similarity with video keyframes is used to suggest tags selected from those of the images.

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), their visual similarity and the Wikipedia folksonomy, to suggest new tags that can be associated at the shot level to a particular keyframe. Fig. 2 shows an overview of the system. The rest of the paper is organized as follows: tag filtering and suggestion is described in Sect. 2; visual analysis and tag relevance are presented in Sect. 3; experimental results are presented in Sect. 4. Conclusions are finally drawn in Sect. 5.

2. SOCIAL AND SEMANTIC TAG FILTERING AND EXPANSION

The first step of our approach is the expansion of the tags associated with the video to be annotated. This is required because, as noted in [16], YouTube videos are annotated with an average of five tags, a number that would not allow to produce a thorough annotation of all the shots. Tag expansion is also needed to ease the alignment of different folksonomies in YouTube and Flickr, to select the images that will be used to associate the tags to keyframes.

Filtering. We filter the video tags that are candidate for expansion, to reduce the risk of semantic drift. Given a video V let $U = \{u_1, \cdots u_n\}$ be the user-defined tags, after discarding stopwords, dates and numbers. We determine a relevance score based on the following method. The tags u_t that appear in the video title get the maximum score ($score(u_t) = 1$), a behavior similar to that of web search engines, while the scores of other tags are determined using the related videos, provided by YouTube.

The basic intuition is that the score of a tag in the video can be inferred from tags of the related videos: the more frequently a tag occurs in the related videos, the more relevant it might be. In particular, consider the m related videos of V and their tags. Let n_u be the number of occurrences of tag u in the related videos, we compute its relevance as $r_u = n_u/m$. Tags with low values are discarded ($r < 0.15$), while tags with high relevance ($r > 0.85$) take scores equal to 1 and are called "strong" tags. For all the other "weak" tags we consider their co-occurrence and semantic relation with the "strong" tags. Co-occurrence between two tags is the number of videos where both tags are used. This value is not very meaningful, as it does not consider the frequency of the individual tags. Therefore we normalize the co-occurrence using the asymmetric normalization method, i.e. with the frequency of one of the tags as in: $o(u_1, u_2) = n_{(u_1,u_2)}/n_{u_1}$, where $n_{(u_1,u_2)}$ is the number of times that the tag u_1 co-occurs with tag u_2. This normalization has been found to improve the diversity of the tags [13]. In particular we compute the co-occurrence between each "weak" tag and the "strongest" tag (i.e. the tag with maximum r). Then we evaluate the *semantic relatedness* with "strong" tags, considering the hyperlinks between the corresponding Wikipedia articles using the method in [9]; the maximum value, s_u, is considered. Finally, the scores of the "weak" tags is computed as the weighted sum of their relevance in related videos, their co-occurrence with the "strongest" tag $\bar{u}$ and their semantic relationship with the "strong" tags as:

$$score(u) = w_1 \cdot r_u + w_2 \cdot o(u, \bar{u}) + w_3 \cdot s_u$$

Tags with a score less than a threshold ($\tau_{filtering}$) are discarded, while the others are used in the next step.

Expansion. Tag expansion is done considering two aspects: *i)* social information, using tags of the related videos, and *ii)* semantic information and folksonomies, using Wikipedia. For the first aspect we consider the occurrences of the tags in the related videos: those with a high number of occurrences, that are not in the list of filtered tags of the analyzed video, are inserted. For the second step we use Wikipedia articles to expand semantically the tags. First we choose the search terms to select Wikipedia resources. Search terms are defined by a single tag or by a combination of two tags (initial experiments have shown that larger combinations are ineffective). The combination of tags is useful for the disambiguation of concepts (e.g. consider the combination of "golden" and "gate"). Two tags are combined if their co-occurrence in related videos, $o(u_1, u_2)$, is high; experimentally we found that an effective threshold is 0.9. Search terms are used to select relevant Wikipedia articles, using Wikipedia Miner toolkit [8]. For each Wikipedia resource we consider the list of *anchors*, i.e. text used within links to Wikipedia articles, as candidate tags. The anchors that are more frequently used are added to the tag list.

3. APPEARANCE-BASED TAG RELEVANCE

Videos are segmented in shots using a fast algorithm that

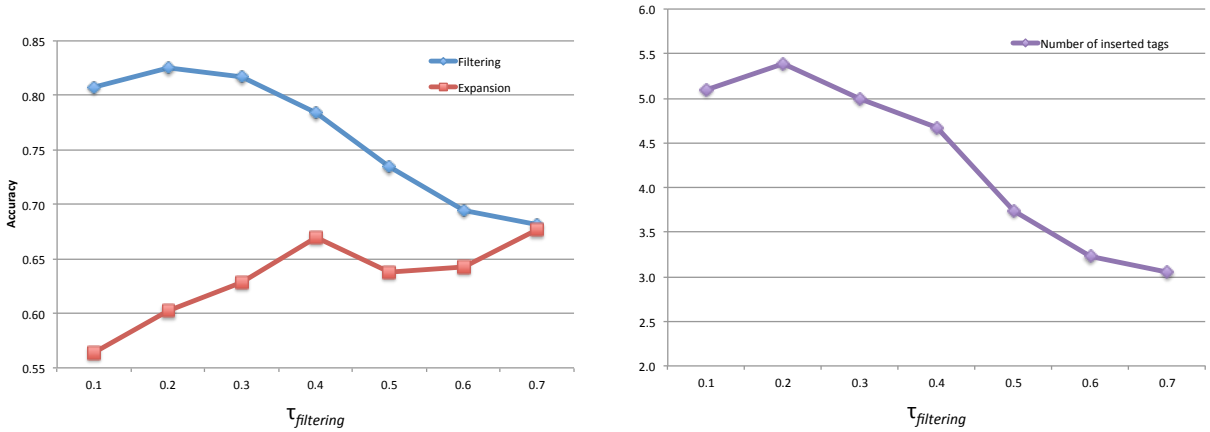

Figure 3: ***left)*** **Mean accuracy of the filtering and expansion steps at different score thresholds;** ***right)*** **Mean number of tags added correctly to a video.**

analyses the frame luminance. For each shot we extract the middle frame, creating a set of keyframes $K = \{K_1, \ldots, K_l\}$. The list of filtered and expanded tags obtained in the previous step is used to start the annotation process, selecting from Flickr a set of images $I = \{I_1, \ldots, I_m\}$ that have been annotated at least with one of them; each I_d has the tags $\{u_1^i, \ldots u_p^i\}$, and the union of these tags is the vocabulary used for video annotation. Due to the fact that user-generated tags are noisy, it is necessary to determine the tag relevance in order to obtain a correct annotation. For this reason we have followed, in principle, the approach for evaluating tag relevance in similar images of [4], adapting it to deal with the problem of locating tags in videos. In our approach keyframes do not have any tag associated to them, since the tags associated to the whole video may not refer to the specific content of a shot; applying the approach of [4] to the shots and tags of a video would simply result in their reordering [1]. The relevance of a tag u_j^i is computed by counting the presence of u_j^i in the visual neighbors of the keyframe K_l, minus its prior frequency. This is based on the consideration that tags that occur frequently in the visual neighborhood of K_l are important for the keyframe being analyzed. In our approach we compute relevance as the weighted sum of the presence of a tag in the neighbors, where weights are inversely proportional to the visual distance between keyframe K_l and the I images. Since we are treating keyframes as unlabeled images, we estimate tag relevance for each candidate tag w.r.t. the keyframe and select those whose relevance score is above a threshold $\tau_{relevance}$. We found that this approach is more consistent with the distance-based weighting of tag relevance, rather than selecting a fixed number of the highest ranking tags independently from their distance from the keyframe, as in [4]. Moreover, it has to be considered that we are working with images from very diverse sources and tagging practices, so that we want to be able to avoid adding tags that are not relevant enough, or add all the relevant tags in the neighborhood. All the tags added to the shots in this step are used to annotate the video also at the global level.

To compute visual similarity between keyframes K and Flickr images I we use a 370-dimensional feature vectors that includes local and global features. This feature vector is composed by a 50 dimensional color correlogram computed in the HSV color space, a 80 dimensional vector for the MPEG-7 Edge Histogram Descriptor and a 240 dimension vector for the TOP-SIFT descriptor. This latter descriptor is a variation of TOP-SURF [14], a compact image descriptor that combines interest points with visual words, designed for fast content-based image retrieval.

The Flickr images are clustered using k-means, to use the cluster centers as indexes for a fast approximate nearest neighbor search. For each keyframe of the video the nearest cluster center based on the visual similarity is retrieved. Images belonging to this cluster are considered as neighbors.

4. EXPERIMENTS

We evaluate our approach on a dataset created to represent the diversity of content on YouTube. As in [1] the dataset is composed by four YouTube videos for each of the 15 categories (*Auto & Vehicles, Comedy, Education, Entertainment, Film and Animation, Gaming, Howto & Style, Music, News and Politics, Nonprofits & Activism, Pets & Animals, Science & Technology, Sports, Travel & Events*). The total duration of videos is three hours and eight minutes and the number of detected shots is 4196. The number of tags per video varies from 8 to 22.

Experiment 1: Tag Filtering/Expansion. In the first experiment we analyzed the performance of our system to filter and expand the initial tags related to a video. Evaluation results is reported in terms of accuracy, computed as the ratio between the number of correct operations (addition/deletion of tags) and the number of tags. Fig. 3-*top* presents the mean accuracy of filtering and expansion steps at different threshold scores ($\tau_{filtering}$); Fig. 3-*bottom* shows the mean number of the tags correctly added to the video. Note that accuracy of expansion increases for high values of the threshold. This is due to the fact that with a high threshold the filtering step maintains only very relevant tags, thus reducing the semantic drift that may happen during the expansion process. However, in this case only few new tags are added. To select a good threshold it also necessary to consider the mean number of tags added correctly. Based on these data a good value for $\tau_{filtering}$ is 0.40, that allows to have a mean accuracy of 78.4% (filtering) and of 67% (expansion) adding an average of 4.6 tags to the original set.

YouTube category	$\tau_{relevance}$=1		$\tau_{relevance}$=3		$\tau_{relevance}$=5		$\tau_{relevance}$=7		$\tau_{relevance}$=11	
	Acc.	Tags	Acc.	Tags	Acc.	Tags	Acc.	Tags	Acc.	Tags
Auto & Vehicles	0.41	10.99	0.65	4.09	0.78	2.13	0.86	1.36	0.93	0.66
Comedy	0.58	5.49	0.85	2.68	0.95	1.68	0.92	0.89	0.77	0.16
Education	0.49	3.97	0.62	1.83	0.76	0.84	0.72	0.39	0.69	0.11
Entertainment	0.60	4.46	0.84	2.98	0.99	1.94	1	0.89	1	0.03
Film & Animation	0.54	2.16	0.93	1.28	0.99	0.59	1	0.19	1	0.01
Gaming	0.47	3.85	0.85	2.13	0.93	0.97	0.99	0.60	1	0.2
Howto & Style	0.39	3.91	0.61	2.02	0.69	1.04	0.71	0.45	0,71	0.31
Music	0.39	2.48	0.69	0.48	1	0.10	1	0.012	1	0.06
News & Politics	0.62	5.32	0.87	2.40	0.97	1.04	1	0.46	1	0.04
No-profit & Activism	0.61	2.62	0.93	1	0.98	0.42	1	0.17	1	0.04
People & Blogs	0.40	5.70	0.67	2.74	0.79	1.22	0.82	0.58	0.50	0.15
Pets & Animals	0.56	4.83	0.75	2.28	0.86	1.04	0.85	0.55	0.94	0.23
Science & Technology	0.44	4.80	0.64	1.67	0.81	0.84	0.89	0.44	0.87	0.16
Sport	0.41	4.49	0.74	2.63	0.82	1.39	0.92	0.62	0.94	0.14
Travel & Events	0.61	12.57	0.79	7.34	0.87	4.21	0.91	2.45	0.98	1.18
Average	0.50	5.18	0.76	2.50	0.88	1.30	0.91	0.67	0.90	0.23

Table 1: Results for tag localization and suggestion for each YouTube category, in terms of accuracy and average number of correctly added tags, as $\tau_{relevance}$ varies.

Experiment 2: Localizing Tag. In the second experiment we analyze the performance of the system in adding and locating of relevant tags to shots. The performance is measured in terms of accuracy: i.e. ratio between the number of tags correctly suggested and the total number of suggested tags. For each tag, resulting from the filtering and expansion process, the system downloads the first 15 Flickr images ranked according the "relevance" criterion provided by the Flickr API. Table 1 reports, for different relevance threshold scores, the accuracy and the mean number of correctly suggested tags for shot. The overall performance of the system is promising. We can observe that the mean accuracy on the entire dataset increases until score equals to seven and slightly decreases for higher scores, remaining close to 0.9; while the mean number of suggested tags correctly decreases significantly for high scores (e.g. when requiring a threshold above 5). From the experimental results we can also note that some categories are more tractable than the others. In the "Auto & Vehicle" and "Travel & Events" categories, the extracted Flickr images are very relevant and similar to the shots analysed. This can be seen from the number of suggested tags which is quite large. In "Film & Animation" we saw that it is difficult to retrieve Flickr images similar to trailer scenes of feature films. "Howto & Style" collects very diverse content that is hard to be correctly annotated.

5. CONCLUSIONS

In this work we have presented a method for semantic video annotation based on social knowledge embedded in YouTube, Flickr and Wikipedia. The tags provided by the users that share videos on internet 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 and with evaluation of visual similarity among videos. We plan also to further improve the use of semantic relations between tags, adding other structured sources of social knowledge, like DBPedia, to improve tag expansion.

Acknowledgments. This work is partially supported by the EU euTV Project (Contract FP7-226248).

6. REFERENCES

[1] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra. Tag suggestion and localization in user-generated videos based on social knowledge. In *Proc. of ACM WSM*, 2010.

[2] 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.

[3] G. Li, M. Wang, Y.-T. Zheng, and T.-S. Chua. ShotTagger: Tag location for internet videos. In *Proc. of ACM ICMR*, 2011.

[4] 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.

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

[6] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. of WWW*, 2009.

[7] D. Liu, X.-S. Hua, and H.-J. Zhang. Content-based tag processing for internet social images. *Multimedia Tools and Applications*, 51(1):723–738, 2011.

[8] D. Milne and I. Witten. An open-source toolkit for mining Wikipedia. In *Proc. of NZCSRSC*, 2009.

[9] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of AAAI*, 2008.

[10] H.-S. Min, J. Choi, W. De Neve, Y. M. Ro, and K. N. Plataniotis. Semantic annotation of personal video content using an image folksonomy. In *Proc. of IEEE ICIP*, 2009.

[11] S. Paisitkriangkrai, T. Mei, J. Zhang, and X.-S. Hua. Scalable clip-based near-duplicate video detection with ordinal measure. In *Proc. of ACM CIVR*, 2010.

[12] J. Shao, W. Yin, S. Ma, and Y. Zhuang. Topic discovery of web video using star-structured k-partite graph. In *Proc. of ACM Multimedia*, 2010.

[13] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of WWW*, 2008.

[14] B. Thomee, E. M. Bakker, and M. S. Lew. TOP-SURF: a visual words toolkit. In *Proc. of ACM Multimedia*, 2010.

[15] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel. Learning automatic concept detectors from online video. *Computer Vision and Image Understanding*, 114(4):429–438, 2010.

[16] W. Zhao, X. Wu, and C. Ngo. On the annotation of web videos by efficient near-duplicate search. *IEEE Transactions on Multimedia*, 12(5):448 – 461, 2010.