Few-Example Video Event Retrieval using Tag Propagation

Masoud Mazloom
ISLA, Faculty of Science,
University of Amsterdam
Science Park 904, 1098 XH
Amsterdam, The Netherlands
m.mazloom@uva.nl

Xirong Li
Key Lab of Data Engineering
and Knowledge Engineering
Renmin University of China
100872 China
xirong@ruc.edu.cn

Cees G. M. Snoek
ISLA, Faculty of Science,
University of Amsterdam
Science Park 904, 1098 XH
Amsterdam, The Netherlands
cgmsnoek@uva.nl

ABSTRACT

An emerging topic in multimedia retrieval is to detect a complex event in video using only a handful of video examples. Different from existing work which learns a ranker from positive video examples and hundreds of negative examples, we aim to query web video for events using zero or only a few visual examples. To that end, we propose in this paper a tag-based video retrieval system which propagates tags from a tagged video source to an unlabeled video collection without the need of any training examples. Our algorithm is based on weighted frequency neighbor voting using concept vector similarity. Once tags are propagated to unlabeled video we can rely on off-the-shelf language models to rank these videos by the tag similarity. We study the behavior of our tag-based video event retrieval system by performing three experiments on web videos from the TRECVID multimedia event detection corpus, with zero, one and multiple query examples that beats a recent alternative.

Categories and Subject Descriptors

1.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Video event retrieval, tag propagation

1. INTRODUCTION

Finding web videos showing a specific event is an emerging topic in multimedia retrieval. Different from video concept detection that focuses on searching for single concepts, finding a video event such as 'feeding an animal' requires the co-occurrence of multiple concepts including 'animal', 'person', and an action of 'feeding'. Existing works on video event retrieval mostly follow a supervised learning approach [2, 7, 8], with the prerequisite that hundreds of training examples w.r.t. the query event are available at hand. In practice, however, only very few examples might be provided. The limited availability of training examples restricts the applicability of the supervised approach for video event retrieval.

For searching unlabeled videos with few query examples, an effective representation of both the unlabeled videos and the query is crucial. Low-level features such as bag of visual words are considered in [4, 8], while concept vectors generated by concept detection are used in [2,7]. A recent comparison [6] shows that the concept vectors outperform the bag of visual words feature for query-by-video retrieval. Nevertheless, the list of concepts has to be predefined. Hence, they could be suboptimal for describing novel videos and events.

In contrast to the fixed concepts, social tags contributed by many users serve as an up-to-date description of many videos. In particular, we observe that for videos showing an event of interest, social tags are often more versatile than the concepts. Thus, we hypothesize that representing videos by tags is better than the concept-based representation for video event retrieval. As we aim for searching in unlabeled data, the question arises as how to propagate tags from socially tagged videos to the unlabeled videos under consideration?

Propagating social tags between images has been actively studied. A representative algorithm is neighbor voting [5]. Given an image, the algorithm first retrieves the nearest neighbors in terms of low-level visual similarity. Tags are sorted in descending order in terms of their occurrence frequency in the neighbors, and the top ranked tags are propagated to the given image. Propagating tags between videos of YouTube categories has also been studied, e.g., in [1, 11]. To the best of our knowledge, studying tag propagation for video event retrieval on arbitrary video of complex event categories has not been done yet.

In this paper we propose the use of automatically generated tags for video event retrieval in unlabeled data. We improve the neighbor voting algorithm for selecting relevant tags from many socially-tagged videos. All these efforts lead to a video event retrieval system that beats recent alternatives for few-example search scenarios [6].

2. OUR PROPOSAL

Given a set of unlabeled videos, we aim to build a system that can automatically find videos showing a specific event. We use x to denote an unlabeled video and $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of n videos. Let q be a query with respect to an

ICMR'14, April 01–04 2014, Glasgow, United Kingdom.

event of interest. At the heart of our system is a conditional probability function $p(x|q)$ which measures the possibility of the video x containing the event. Accordingly, the system retrieves video events by sorting videos in $\mathcal{X}$ in descending order according to $p(x|q)$.

We propose to realize $p(x|q)$ in a tag space which covers a broader range of visual semantics than a restricted set of pre-trained concept detectors can cover. Let $\mathcal{T} = \{t_1, \ldots, t_m\}$ be a vocabulary of m tags, where each tag corresponds to a dimension in the tag space. Next, we describe how to map both the videos and the query into the tag space, followed by a language model approach for computing $p(x|q)$.

2.1 Tag Propagation

Given an unlabeled video x, we map it into the tag space by propagating relevant tags from a set of labeled videos. We refer to this set as a *source set*, consisting of l videos. Each video in the source set is labeled with some tags from $\mathcal{T}$ by Internet users.

For tag propagation between videos, we adapt the neighbor voting algorithm [5]. Where the traditional neighbor voting relies on low-level visual features to find visual neighbors [1, 11], we prefer concept vectors as these were recently shown to be more reliable for similarity matching of videos [6]. Moreover, we assign different weights to the neighbors in terms of their similarity to the test video. Let $\{s_1, \ldots, s_k\}$ be the k nearest neighbors retrieved from the source set. For each tag $t \in \mathcal{T}$, we compute its relevance score w.r.t. the video x as:

$$rel(t;x) = \frac{\sum_{i=1}^{k} w(x,s_i) \cdot \alpha(t,s_i)}{k} - \frac{\sum_{i=1}^{l} w(x,s_i) \cdot \alpha(t,s_i)}{l}, \quad (1)$$

where $\alpha(t, s_i)$ returns 1 if the video s_i is labeled with the tag, and 0 otherwise. The weighting function $w(x, s_i)$ is the similarity between the video of x and s_i. As the last term in Eq. 1 shows, we introduce a weighted prior to suppress frequently used tags. Using $rel(t; x)$, we sort all the tags in descending order, and select the top h ranked tags as the propagated tags for the video x. Figure 1 demonstrates some of the tag propagation results.

For a query q expressed in terms of a video, we apply the same tag propagation technique on the query video.

2.2 Video Retrieval by Propagated Tags

Once we have propagated tags to all videos on which we want to perform retrieval, we are ready to compute $p(x|q)$ in light of the similarity between the tags of the query q and the tags of videos in test set $\mathcal{X}$. Since both the query q and the videos in $\mathcal{X}$ are now mapped into the common tag space, a number of well established text retrieval techniques can be leveraged with ease. In this work, we choose the popular Jelinek-Mercer language model approach [13].

Given that $p(x)$ follows a uniform distribution, ranking videos by $p(x|q)$ amounts to ranking videos by $p(q|x)$. To compute $p(q|x)$, we use a unigram model, that is,

$$p(q|x) = \prod_{t \in q} p(t|x)^{c(t,q)}, \quad (2)$$

where $c(t, q)$ returns the occurrence frequency of a tag t in the query q. Following the Jelinek-Mercer approach [13], we compute $p(t|x)$ as

$$p(t|x) = \lambda \cdot \frac{c_k(t,x)}{c(x)} + (1-\lambda) \cdot p(t|\mathcal{X}), \quad (3)$$

where the parameter λ controls the influence of the prior $p(t|\mathcal{X})$, which is $\frac{c(t,\mathcal{X})}{c(\mathcal{X})}$. In this formula $c(z)$ is the number of tags in z. If t is one of the tags propagated to x, $c_k(t, x)$ returns the frequency of t in k-nearest neighbor of x, and 0 otherwise.

3. EXPERIMENTAL SETUP

3.1 Data Set

For the event retrieval experiments we rely on the challenging web video corpus from TRECVID 2013 [12]. It comes with ground truth annotations at video level for 20 real-world events, including life events, instructional events, sport events, etc. The TRECVID corpus comes with many partitions optimized for event classification. For the purpose of our experiments we consider as our *source set* the Research partition (10K), the positive event videos (2K) and the (negative) background videos (8K). We report all our results on the MED-test partition which contains 27K videos. Besides the video files, TRECVID also provides for many videos a sentence level description of the content. We would like to stress that in our experiments we only rely on sentence-level descriptions belonging to videos in the source set, we do not consider any sentences that are provided for the test set.

3.2 Implementation

For finding the k-nearest neighbors in the tag propagation process, we follow the two implementations provided by [6]. The first feature is a concept vector containing 1,346 concept detector scores for one frame every 2 seconds and aggregated into a video-level representation by average-pooling [7]. The second feature is a standard bag-of-words using densely sampled SIFT descriptors with VLAD difference coding [3] using a 1024 words codebook. As suggested in [6], we use normalized correlation as the similarity metric. We empirically set the number of neighbors k to 50 and the number of preserved tags t to 20. We set λ in Eq.(3) to 0.5 as suggested by [13].

3.3 Experiments

Experiment 1: ***Event retrieval using zero examples.*** In this experiment we focus on tag-based event retrieval without video examples. We use the TRECVID provided text description per event as a query. After extracting the individual terms and removing stop words, we represent each query as a tag vector. Then we retrieve the videos in the test set using the similarity between the query tags and the propagated tags for the videos in the test set using Eq. (3). For the similarity used for tag propagation, we compare both the concept vector and the bag of visual words features.

Experiment 2: ***Event retrieval using one example.*** In this experiment we assume that we have only one example video per event. We compare the proposed tag-based retrieval system with a recent concept-based retrieval system [6]. For concept-based retrieval, we rank videos in the test set using the concept vector of the query and the concept vector of the test videos. For tag-based retrieval, we retrieve the videos in the test set using the similarity be-

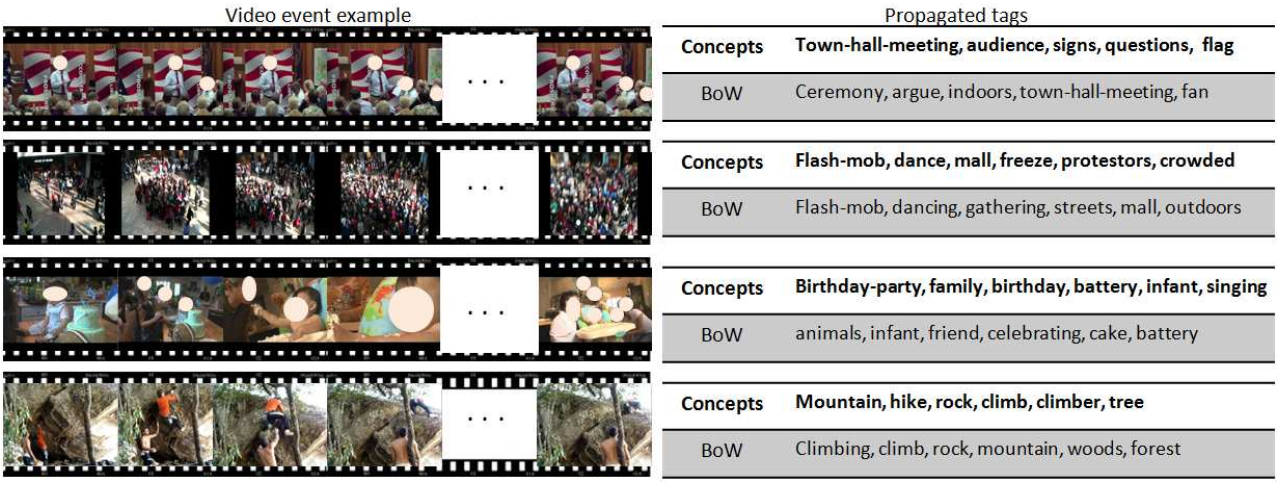

Figure 1: Example videos for the events 'town hall meeting', 'flash mob gathering', 'birthday party', and 'rock climbing', from top to bottom, with the automatically propagated tags using concept vector features and bag of visual word features, respectively. Concept vector features result in more meaningful tags.

tween the propagated tags of the query and the propagated tags of the test videos.

Experiment 3: ***Event retrieval using multiple examples.*** In this experiment we investigate the effect of using multiple video queries. Again we compare the tag-based retrieval system and the concept-based retrieval system. We follow [6,9,10] and use eight example videos per event, which are selected at random from all available queries. For concept-based retrieval, we adopt the score-pooling fusion [6]. In tag-based retrieval, we consider both early tag-fusion and late result-fusion. In early tag-fusion, we combine the tags of all the eight queries to create a single query. While in the late result-fusion, the retrieval results of the individual queries are combined using average pooling. To cancel out the accidental effects of randomness, we repeat the procedure of selecting eight queries 100 times and report the averaged performance.

Evaluation criteria Following the common practice in the literature, the retrieval performance is measured in terms of average precision (AP), which combines precision and recall into a single metric [12]. We also report the average retrieval performance over all events as the mean average precision (MAP).

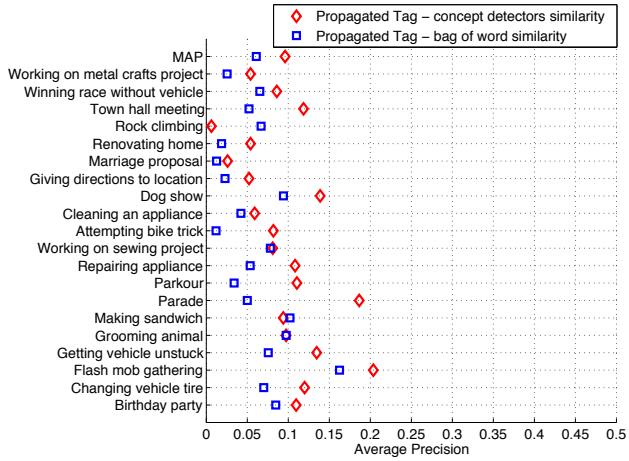

Figure 2: Experiment 1: Event retrieval using zero examples. Retrieval using tags propagated by concept vector similarity is better than retrieval using tags propagated by bag of visual words similarity.

4. RESULTS

Experiment 1: Event retrieval using zero examples. As shown in Figure 2, using tags propagated by the concept vector similarity, with an MAP of 0.096, is better than using tags propagated by the bag of visual word feature, with an MAP of 0.061. Looking into the individual events, for 18 out of the 20 events, tag propagation by the concept vector similarity surpasses its bag of visual words counterpart. Some qualitative results are given in Figure 1. So for the remaining two experiments we consider tag propagation by neighbor voting using concept vectors. The result also shows that with tag propagation, we can retrieve complex events from unlabeled videos using only textual queries.

Experiment 2: Event retrieval using one example. As shown in Figure 3, tag-based retrieval shows a considerable improvement over concept-based retrieval (0.197 vs 0.037). Since the event video examples are visually diverse, a single query alone cannot cover all possible semantic variations of an event. In contrast, as videos of the same event tend to share a set of similar tags, we mitigate the negative effect of visual diversity. Consider the event 'Dog show', for instance. Tag-based retrieval and concept-based retrieval scores an AP of 0.232 and 0.021, respectively. For concept-based retrieval, we observe that many of the retrieved videos contain animals, like the ones relevant to the event 'Grooming animal', which are visually similar to 'Dog show' videos. By contrast, tags such as 'competitive-exhibition', 'judges', 'group-of-dog', and 'Sheepdog-trial' are propagated to the 'Dog show' videos. Such tags provides a better discrimination between this event and the other events. From the results, we conclude that representing videos as a vector of

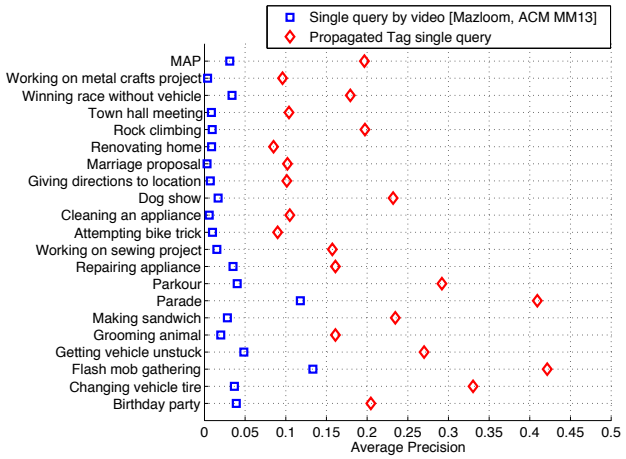

Figure 3: Experiment 2: Event retrieval using one example. The proposed tag-based retrieval system beats the concept-based retrieval system [6].

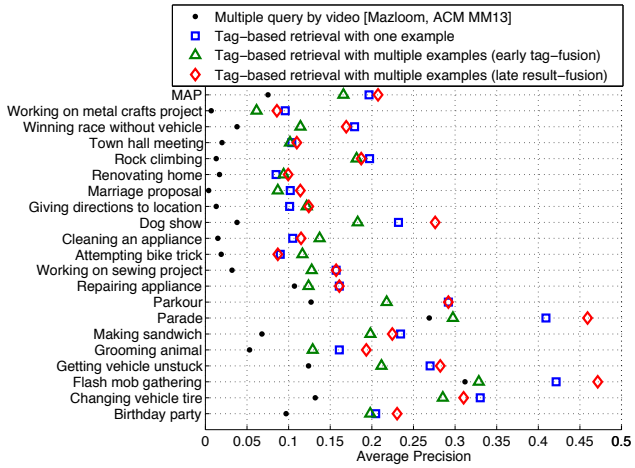

Figure 4: Experiment 3: Event retrieval using multiple examples. Tag-based retrieval with late result-fusion is the best choice for exploiting multiple examples.

propagated tags is superior to concept-based representation for one-example video event retrieval.

Experiment 3: Event retrieval using multiple examples. As shown in Figure 4, for event retrieval using multiple examples, the tag-based retrieval system is again the winner. Notice that when more video examples are provided, the accuracy of the concept-based retrieval system improves, with its MAP score increases from 0.037 to 0.079. The behavior of the tag-based system is different. Compared to the one-example scenario (as done in Experiment 2), the performance of the early tag-fusion strategy degenerates, with an MAP of 0.166. This is because tag propagation is a fully automatic process, which inevitably introduces noisy tags. When merging the tags of the individual queries to form a single tag vector, the impact of noisy tags increases. In contrast, the late result-fusion strategy reduces such impact, obtaining the highest MAP of 0.208. Hence, for searching with multiple query examples, tag-based retrieval with late result-fusion is a good choice.

5. CONCLUSIONS

For video event retrieval with few examples, we propose the use of tag-based video representation and retrieval. To that end, we develop an algorithm to propagate tags to unlabeled videos from many socially-tagged videos. The algorithm is founded on neighbor voting, and we improve it by constructing visual neighbors using concept vector features instead of low-level visual features. Moreover, we assign different weights to the neighbors in terms of their similarity to the test video. Tag propagation enables us to represent both the unlabeled videos and the query examples into a common tag space. Consequently, we employ an off-the-shelve language model for video event retrieval. Three experiments on the web video collection from the TRECVID event detection task support our findings as follows. First, for tag propagation, concept vector features are better than the bag of visual words feature. Using propagated tags, we can search for events in unlabeled videos using textual queries directly. Second, when example videos of a query event are provided, we justify that retrieving using the auto-propagated tags beats concept-based retrieval. Finally, given multiple example videos, we find that tag-based retrieval with late result-fusion is a good choice for video event retrieval.

Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, and the Chinese NSFC (No. 61303184), SRFDP (No. 20130004120006), and SRF for ROCS, SEM.

6. REFERENCES

[1] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Enriching and localizing semantic tags in internet videos. In *MM*, 2011.

[2] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.

[3] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 2012.

[4] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *MMM*, 2012.

[5] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 2009.

[6] M. Mazloom, A. Habibian, and C. G. M. Snoek. Querying for video events by semantic signatures from few examples. In *MM*, 2013.

[7] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *TMM*, 2012.

[8] P. Natarajan et al. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[9] A. Natsev, M. R. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In *MM*, 2005.

[10] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *TMM*, 2007.

[11] S. Siersdorfer, J. S. Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *SIGIR*, 2009.

[12] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR*, 2006.

[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TIS*, 2004.