Querying for Video Events by Semantic Signatures from Few Examples

Masoud Mazloom, Amirhossein Habibian and Cees G. M. Snoek
ISLA, Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
{m.mazloom, a.habibian, cgmsnoek}@uva.nl

ABSTRACT

We aim to query web video for complex events using only a handful of video query examples, where the standard approach learns a ranker from hundreds of examples. We consider a semantic signature representation, consisting of off-the-shelf concept detectors, to capture the variance in semantic appearance of events. Since it is unknown what similarity metric and query fusion to use in such an event retrieval setting, we perform three experiments on unconstrained web videos from the TRECVID event detection task. It reveals that: retrieval with semantic signatures using normalized correlation as similarity metric outperforms a low-level bag-of-words alternative, multiple queries are best combined using late fusion with an average operator, and event retrieval is preferred over event classification when less than eight positive video examples are available.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Video retrieval, multiple queries, fusion

1. INTRODUCTION

The goal of this paper is to retrieve complex events from web video using only a handful of video query examples, like the ones in Figure 1. So far, the common approach to event retrieval is to represent a video in terms of fused audiovisual features and to learn a ranker from hundreds of positives and negative labeled examples *e.g.*, [5, 9]. We differ from this supervised classification solution to event retrieval in three ways. 1) We focus on the more realistic scenario where only a few positive examples for an event are available. In contrast to [6, 7] deliberately ignoring the negative examples as these are not necessarily available at query time. 2) We prefer a high-level over a low-level video representation as events are often characterized by similarity in semantics rather than appearance. 3) We query directly by the example videos rather than a textual label.

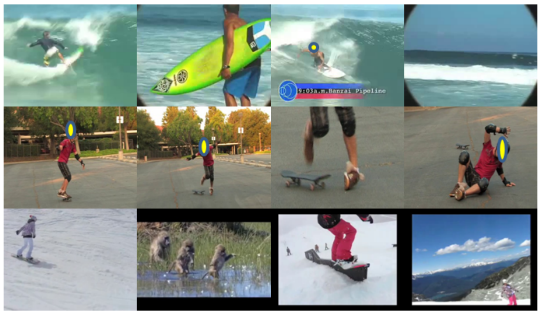

Figure 1: Each row shows a video query for the event *Attempting a board trick*. Due to the variance in both audiovisual and semantic appearance of events in video, a multiple query approach to retrieval, as we propose and evaluate, seems mandatory.

We draw inspiration from progress in query-by-image retrieval, where it is known to be advantageous to fuse multiple queries for retrieving concepts [10] or specific objects [1], but only for a low-level feature representation. In this paper we study the influence of combining multiple queries as well, but for the problem of retrieving complex events in video. In addition, we consider a *semantic signature* representation, consisting of off-the-shelf concept detectors [2, 12], which we deem more suited for capturing the variation in semantic appearance of events. We extend upon existing semantic image search approaches [11] by explicitly addressing fusion of multiple queries and pooling of the semantic representation for video. While approaches for video event classification using semantic representations have recently started to appear in the literature [3, 8], to the best of our knowledge this is the first work on query-by-video retrieval using semantic representations with multiple examples.

Since it is unknown what similarity metric and query fusion to use in such an event retrieval setting, as well as the influence of having only a handful of video query examples available, this paper conducts an experimental study to shed light on the matter.

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502160 .

2. SEMANTIC QUERY FUSION

2.1 Semantic Signatures

A video representation based on semantic signatures has the advantage of evaluating video similarity at a higher level of abstraction, and therefore potentially better semantic generalization than what is possible with a low level bag-of-words representation. To arrive at a semantic representation for video retrieval, we consider a lexicon consisting of n concept detectors that include *scenes, objects, people*, and *activities*. Thanks to efforts like TRECVID [12] and ImageNet [2] these detectors are commonly available these days, *e.g.*, [3, 8]. We compute concept detector scores per video frame, which are extracted once every s seconds. By concatenating and normalizing the detector outputs, each frame is represented by a concept score histogram of n elements. Finally the concept score histograms are aggregated into a video-level representation by average-pooling, which is known to be a stable choice for video classification [8]. We call the final histogram of concept detector scores a semantic signature.

2.2 Similarity Metrics

There are many well known methods for measuring the similarity distance between two histograms, see Table 1. However, it is not clear what metric is the most suited for the purpose of querying video by semantic signatures. In a semantic signature the presence of a concept is much more descriptive than absence of the concept. Hence, we expect that a similarity metric which only relies on the difference between elements while ignoring their values, such as the chi-square and histogram intersection, perform worse for retrieval. To illustrate, consider the cases where two videos contain a certain concept with the same probability. For the first case suppose this probability is 1 for both videos, and for the second case they are 0. From the chi-square point of view for both cases the similarity is identical. However, we know that for the second case the video content is different. A different semantic signature that results in the same similarity is problematic for fusing multiple queries. Since normalized correlation explicitly considers the value of the histogram elements, we expect this metric to be more suited. The same holds for metrics inspired by information theoretic divergence such as Kullback-Leibler and Jensen-Shannon.

2.3 Query Fusion

Since a single video query cannot cover all possible semantic variations of an event (see Figure 1), we consider a retrieval scenario where a limited set of video example queries is available. In such a multi query scenario, it is important to select the appropriate query fusion. We consider early and late query fusion. In early query fusion we simply combine the semantic signatures of each query into a single signature, which we call *signature pooling*. To arrive at a single signature, we consider both the average and max operator. In early query fusion, we query videos in the test set based on a single semantic signature, making it a very efficient query method. In contrast, for late query fusion, we search the test set based on each query individually making it slightly more demanding than early query fusion. For late query fusion we combine the retrieval results afterwards by *score pooling*. Again, we consider both the average and max operator. We anticipate that the average operator is preferred for both early and late query fusion, as it decreases the influence of noisy or irrelevant concept scores. Continuing the example from Figure 1, the average operator decreases the effect of the concepts *monkey* and *animal* which appear in the query at the bottom.

Table 1: Similarity metrics that we consider when querying for video events by a semantic signature containing n concept detector scores, where q and d are two semantic signatures for the query and a retrieved video, respectively.

Similarity Metric	Formula
Normalized Correlation	$\frac{\sum_{i=1}^{n} q_i * d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} * \sqrt{\sum_{i=1}^{n} d_i^2}}$
Kullback-Leibler	$\sum_{i=1}^{n} q_i \log(\frac{q_i}{d_i})$
Histogram Intersection	$\frac{\sum_{i=1}^{n} min(q_i, d_i)}{min(\sum_{i=1}^{n} q_i, \sum_{i=1}^{n} d_i)}$
Chi-Square	$\frac{1}{2} \sum_{i=1}^{n} \frac{(q_i - d_i)^2}{(q_i + d_i)}$
Jensen-Shannon	$\frac{1}{2} \sum_{i=1}^{n} q_i \log(\frac{2q_i}{q_i + d_i}) + \frac{1}{2} \sum_{i=1}^{n} d_i \log(\frac{2d_i}{d_i + q_i})$

In order to establish the effectiveness of using query fusion with semantic signatures for video event retrieval, we perform three experiments on a large corpus of challenging real-world web video.

3. EXPERIMENTAL SETUP

3.1 Data Sets

Video data For the event retrieval experiments, we rely on the web video training corpus from the TRECVID 2012 Multimedia Event Detection task [12]. It comes with ground truth annotations at video level for 25 real-world events, including life events, instructional events, sport events, etc.. Following the protocol for event classification outlined in [3], we split the corpus into two partitions: consisting of 1,736 and 4,434 videos respectively. In this paper we use the first partition as the *query set* that consist of several queries for each of 25 event classes (and an additional set of 7,104 negatives which we use for comparison against event classification approaches). We report all results on the second partition, the independent *test set*.

Concept detectors For ease of comparison we adopt the publicly available concept detector scores for this dataset provided by [3]. It consists of a lexicon of 1,346 concept detectors. The 1,346 concept detectors are trained using the training data for 346 concepts from the TRECVID 2012 Semantic Indexing task [12] and for 1,000 objects from the ImageNet Large-Scale Visual Recognition Challenge 2011 [2]. The detectors are trained using a linear SVM atop a standard bag-of-words of densely sampled color SIFT with Fisher vector coding and spatial pyramids, for implementation details see [3]. The concept scores are computed for 1 frame every 2 seconds. We convert them into semantic signatures using the procedure outline in Section 2.1. We experimented with the length of the signature and observed only a small retrieval performance difference when increasing length from 500 to 1,346 concepts. For all the reported experiments we use the full, length as it provided the best retrieval result overall.

Table 2: Experiment 1: Single Query Baseline. The results are averaged over 25 events and repeated 500 times. Best MAP result in bold. We observe only a minimal variance for all events, therefore not shown.

		Similarity Metric				
Video representation	*Random*	*Normalized Correlation*	*Kullback-Leibler*	*Histogram Intersection*	*Chi-Square*	*Jensen-Shannon*
Semantic signature	0.011	**0.059**	0.054	0.033	0.030	0.053
Bag-of-words	0.011	0.046	0.040	0.032	0.024	0.038

Table 3: Experiment 2: Early vs Late Query Fusion. For the fusion we sample eight queries per iteration. The results are averaged over 25 events and repeated 500 times. Best MAP result in bold. We observe only a minimal variance for all events, therefore not shown.

			Early Query Fusion		Late Query Fusion	
Event category	**Available Queries**	**Single Query**	*AVG*	*MAX*	*AVG*	*MAX*
Attempting a board trick	98	0.062	0.123	0.058	**0.135**	0.117
Feeding an animal	95	0.023	0.056	0.052	**0.069**	0.030
Landing a fish	71	0.086	0.120	0.102	**0.128**	0.061
Wedding ceremony	69	0.078	0.232	0.089	**0.294**	0.037
Working on a woodworking project	79	0.026	**0.092**	0.086	0.087	0.037
Birthday party	121	0.055	0.151	**0.184**	0.159	0.054
Changing a vehicle tire	75	0.045	0.106	0.078	**0.118**	0.027
Flash mob gathering	115	0.175	0.254	0.224	**0.305**	0.115
Getting a vehicle unstuck	85	0.083	0.147	0.101	**0.159**	0.055
Grooming an animal	91	0.068	0.195	0.071	**0.214**	0.016
Making a sandwich	83	0.047	**0.141**	0.087	0.138	0.020
Parade	105	0.120	**0.221**	0.202	**0.221**	0.055
Parkour	75	0.082	**0.231**	0.121	0.219	0.037
Repairing an appliance	85	0.082	0.213	0.146	**0.221**	0.028
Working on a sewing project	86	0.046	0.094	0.052	**0.101**	0.042
Attempting a bike trick	43	0.057	0.108	0.101	**0.110**	0.046
Cleaning an appliance	43	0.024	0.069	0.070	**0.072**	0.051
Dog show	43	0.059	0.116	0.067	**0.130**	0.083
Giving directions to a location	43	0.022	0.102	0.010	**0.114**	0.006
Marriage proposal	43	0.012	0.019	0.009	**0.025**	0.005
Renovating a home	43	0.054	0.200	0.130	**0.207**	0.023
Rock climbing	43	0.042	0.071	**0.103**	0.065	0.019
Town hall meeting	43	0.036	0.060	**0.113**	0.055	0.076
Winning a race without a vehicle	43	0.065	0.083	0.074	0.085	**0.121**
Working on a metal crafts project	43	0.021	0.029	0.022	**0.030**	0.015
Mean average precision		0.059	0.129	0.095	**0.138**	0.047

3.2 Experiments

*Experiment 1: **Single query baseline*** In this experiment we compare semantic signatures with a standard bag-of-words using densely sampled SIFT descriptors with VLAD difference coding [4] using a 1024 words codebook. We consider all five similarity metrics from Table 1. We use all available queries once and report the average of their retrieval accuracy.

*Experiment 2: **Early vs Late Query Fusion*** To assess the effect of using multiple queries instead of using only one query, we compare the early and late query fusion schemes described in section 2.3. We perform the experiment using semantic signatures with the best similarity metric from experiment 1. We follow [1, 10, 11] and use eight queries per event, which we select randomly from all available queries. We compute the retrieval accuracy using early and late fusion. We repeat this process 500 times.

*Experiment 3: **Event Retrieval vs Event Classification*** In this experiment we compare event retrieval with event classification when only a limited number of positive examples are available. We vary the number of video queries (or positive examples) from 1 to 20 by randomly sampling from our pool of positive query examples. For event retrieval we use the best performing fusion scheme from experiment 2. For event classification we employ a linear Support Vector Machine on top of each semantic signature, similar to [3]. We consider the video query examples per event jointly as positive examples and the other videos from the query set as negative examples. We also report results using an exemplar-SVM [7] trained on each individual positive query and all available negative examples, and then fused over all queries using average pooling. We measure event retrieval and event classification performance on the test set. We repeat this process 50 times for both event retrieval and event classification with different number of positive examples, and a fixed number of negatives for the event classification scenario. Note that we do not rely on negative examples for the event retrieval scenario.

Evaluation criteria The retrieval performance is measured in terms of the well known 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).

4. RESULTS

*Experiment 1: **Single query baseline*** We present the results of experiment 1 in Table 2. While the overall MAP using a single query is modest, but always much better than random, the video representation using semantic signatures outperforms bag-of-words for all metrics. As expected, the semantic signatures can generalize better than the bag-of-words. Table 2 also confirms that the normalized correlation metric and two information theory based metrics, are better than the histogram intersection and chi-square metrics for computing the similarity between events, as predicted in

Section 2.2. For the remaining experiments we consider the semantic signatures with normalized correlation.

Experiment 2: ***Early vs Late Query Fusion*** Table 3 presents the results of experiment 2. The results show considerable improvement in event retrieval performance when we use multiple queries instead of a single query (0.138 vs 0.059). For both fusion methods using the average operator is better than the maximum. We explain this by the fact that the average operator reduces the effect of noisy and irrelevant concepts that may occur accidentally in one or two event queries, as noted in Section 2.3. By contrast, the max operator is sensitive to irrelevant and noisy concepts scores with a high value. This results in a big drop in retrieval performance (from 0.138 to 0.047), even lower than the single query baseline. However, for some events such as *Rock climbing, Town hall meeting* and *Birthday party* early query fusion with the max operator results in the best retrieval performance. We attribute this to the presence of reliable relevant concepts in one of the queries. Indicating, that much is to be expected from adaptive semantic signatures that determine concept relevance at query time. The results of experiment 2 show that using multiple queries improves the performance of event video retrieval in comparison to a single query, especially for late query fusion. Moreover, we observe that in multiple query video retrieval the average operator is preferred for the fusion.

Experiment 3: ***Event Retrieval vs Event Classification*** We plot the result of experiment 3 in Figure 2. As expected the accuracy of both event retrieval and classification increases when more and more queries or positive event example are available. However, when we use only a limited number of queries, *i.e.*, from 1 to 8, we observe that the accuracy of event retrieval is higher than event classification. After increasing the number of queries, *i.e.*, from 8 to 20, we see the difference in accuracy increasing in favor of event classification. Using the available positive examples jointly is a better choice than bagging them individually with an exemplar-SVM. The exemplar-SVM needs up to 16 positive examples (and a bunch of negative examples) before it outperforms event retrieval. We conclude that when more than eight positive examples are available per event, as well as a set of negative examples, it pays off to learn a classifier. However, for more realistic event video retrieval scenarios where only a handful of positive examples are available, it is advantageous to rely on late query fusion retrieval with an average operator.

5. CONCLUSIONS

This paper studies the behavior of multiple video queries for event retrieval, by performing three experiments on an unconstrained web video collection. The result of experiment 1 provides an indication that querying for event video using semantic signatures generalizes better than a low-level bag-of-words alternative. In addition, we find that normalized correlation is a suitable similarity metric when considering retrieval using semantic signatures. The results of experiment 2 shows that video retrieval using multiple event queries in combination with late query fusion and an average operator outperforms event retrieval using only a single query. With only a handful of examples the results increase from 0.059 to 0.138. A considerable improvement. Finally, experiment 3 reveals that it is advantageous to rely on event retrieval rather than event classification, when the number of available video examples is less than eight.

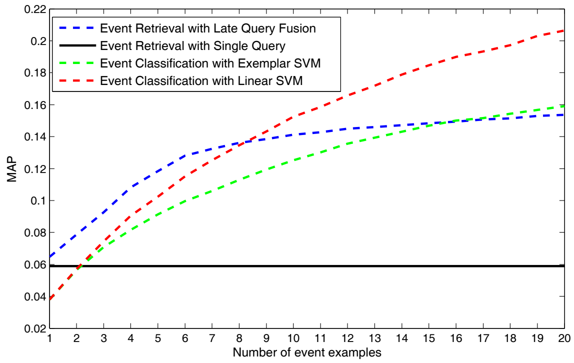

Figure 2: Experiment 3: Event Retrieval vs Event Classification. When only a handful of semantic signature queries are available per event, classification is outperformed by retrieval.

Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

6. REFERENCES

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.

[2] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li. ImageNet large scale visual recognition challenge 2011. `http://www.image-net.org/challenges/LSVRC/2011`.

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

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

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

[6] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM MM*, 2012.

[7] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

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

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

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

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

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