On-the-Fly Video Event Search by Semantic Signatures

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

ABSTRACT

In this technical demonstration, we showcase an event search engine that facilities instant access to an archive of web video. Different from many search engines which rely on high dimensional low-level visual features to represent videos, we rely on our proposed *semantic signature*. We extract semantic signature as the detection scores obtained by applying a vocabulary of 1,346 concept detectors on videos. The semantic signatures are compact, semantic and effective, as we will demonstrate for on-the-fly event retrieval using only a few positive examples. In addition, we will show how the signatures provide a crude interpretation on why a certain video has been retrieved.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Video event detection, semantic video representation

1. INTRODUCTION

The goal of this demonstration is to showcase the value of a video representation based on *semantic signatures* for on-the-fly event retrieval. Semantic signature are extracted as the detection scores obtained by applying a vocabulary of 1,346 concept detectors on videos. Different from existing video event detection systems relying on high dimensional low-level features, like the ones active in TRECVID [15], our video representation has several benefits. First, semantic signatures are substantially more compact compared to high dimensional low-level features, allowing us to train event models in real-time. Second, semantic signatures transfer prior knowledge from the annotations used for the concept detectors, which is useful when few positive video event examples are available. Third, the semantic signatures are much easier to interpret by human users, see Figure 1. While others have also considered the value of semantic representations for video analysis, *e.g.,* [3–5, 7, 8, 10–12, 18], we are the first to demonstrate their benefit in a video search engine for on-the-fly event training and classification.

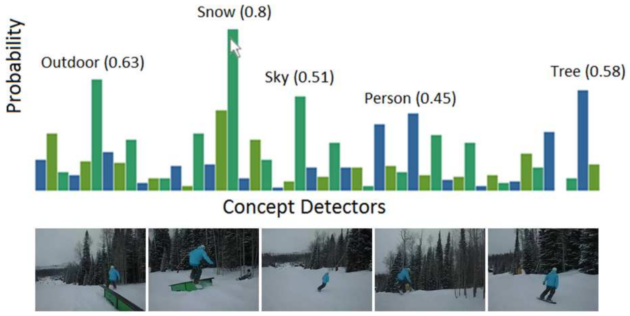


Figure 1: Semantic signature for a web video showing the event *Attempting a board trick.* Since the semantic signature is both expressive and compact, it can be exploited for on-the-fly event detection. What is more, it provides a crude form of recounting.

2. SEMANTIC SIGNATURE

To extract the semantic signature, we need a set of pre-trained concept detectors. For this purpose, we use a pool of 1,346 concept detectors trained on two publicly available datasets: the TRECVID 2012 Semantic Indexing task [1,15] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [2]. The former has annotations for 346 semantic concepts from web videos, and the latter has annotations for 1,000 semantic concepts on photos. We follow the well-known bag-of-words implementation of the 1,346 concept detectors. We use densely sampled SIFT, OpponentSIFT and C-SIFT descriptors [17] with Fisher vector coding [13]. The codebook used has a size of 256 words. As a spatial pyramid we use the full image and three horizontal bars [6]. The feature vectors representing the training images form the input for a fast linear Support Vector Machine [14].

After training the concept detectors, a semantic signature is computed by applying the concept detectors on video. We compute all concept detector scores per video frame, which are extracted once every 2 seconds. By concatenating and normalizing the detector outputs, each frame is represented

ICMR'14, April 1–4, 2014, Glasgow, United Kingdom.

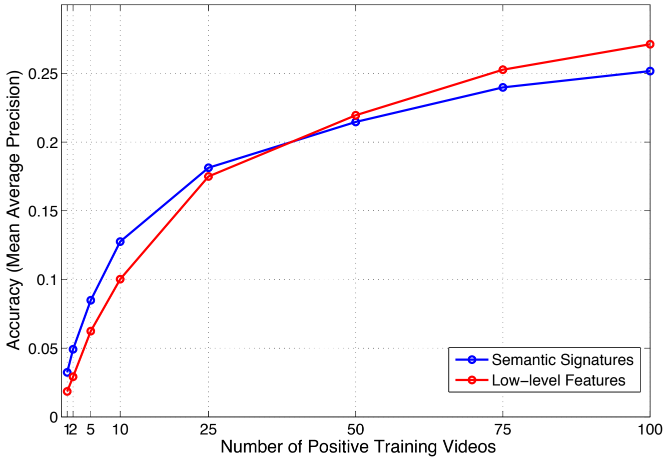


Figure 2: Experiment 2: Comparing event detection accuracies of semantic signatures and low-level features under varying training conditions. Semantic signatures outperform low-level features when there are less than 25 positive examples available for training the event classifier.

by a concept score histogram of 1,346 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 [11]. We call the final histogram of concept detector scores a semantic signature. When training an event model on top of the semantic signatures we consider a linear SVM classifier. Since we use a compact semantic signature for representing the videos, the process of training a model and retrieving the videos on unseen video collections can be done in real-time.

3. EXPERIMENTAL EVALUATION

3.1 Event Dataset

We perform our experiments on the challenging TRECVID Multimedia Event Detection 2013 corpus, as the largest publicly available video corpus in the literature for event detection [15, 16]. We rely on three partitions of videos defined by NIST in our experiments: MED test, Event Kit, and Background including around 27K, 2K, and 5K videos, respectively. The videos come with ground truth annotation at video level for 20 event categories, such as *Marriage proposal*, *Attempting bike trick* and *Making sandwich*. We use the Event Kit and Background partitions as the train set and we use the MED test partition as the test set.

3.2 Experiments

We investigate the advantages of semantic signatures over low-level features by performing three experiments. In our experiments, the low-level features are extracted by densely sampling SIFT, OpponentSIFT and C-SIFT descriptors from video frames sampled every 2 seconds. The extracted descriptors are encoded by a Fisher vector, with the codebook of 256 words, and averaged over video frames to obtain the video-level representations. For both the semantic signature and the low-level features we train event detectors by a fast linear Support Vector Machine.

Table 1: Experiment 1: Comparing time and memory efficiency of semantic signatures vs low-level features. Semantic signatures are considerably more time and memory efficient.

	Time (milliseconds)		Memory (megabytes)	
	Train	Test	Train	Test
Semantic Signature	364	59	107	277
Low-level	22,086	3,363	6,538	16,891

1. **Efficiency:** We compare the time and memory efficiency of semantic signatures vs low-level features. Our comparisons are focused on efficiency of training and testing event classifiers and do not include feature extraction computations. Obviously, extracting semantic signatures requires an extra step to apply vocabulary concept detectors on video low-level features. However this step can be performed efficiently by using available pre-trained concept detectors. In this experiment, we measure the time required for training and testing each of the 20 event classifiers. Moreover, we report the memory required to train and test each event classifier, to assess the memory efficiency. We repeat this process 50 times and report the averaged results on a computer with a CPU Intel Xeon E5-2690@2.90GHZ and 256GB of memory.

2. **Accuracy:** We evaluate the event detection accuracies of low-level features and semantic signatures in various training conditions: We start from using one positive example during training, then gradually increasing the number of positives. At each iteration, the positive examples are randomly selected from the available positive examples per event. To compensate for the random effect we repeat the experiments 100 times and report the average performances. The event detection accuracies are reported in terms of mean average precision.

3. **Interpretability:** We showcase two examples of the capabilities of semantic signatures in explaining video content: *video summarization*, where each video is summarized by identifying its dominant concepts, and *video translation*, where a set of textual descriptions are generated for each video as we proposed in [4].

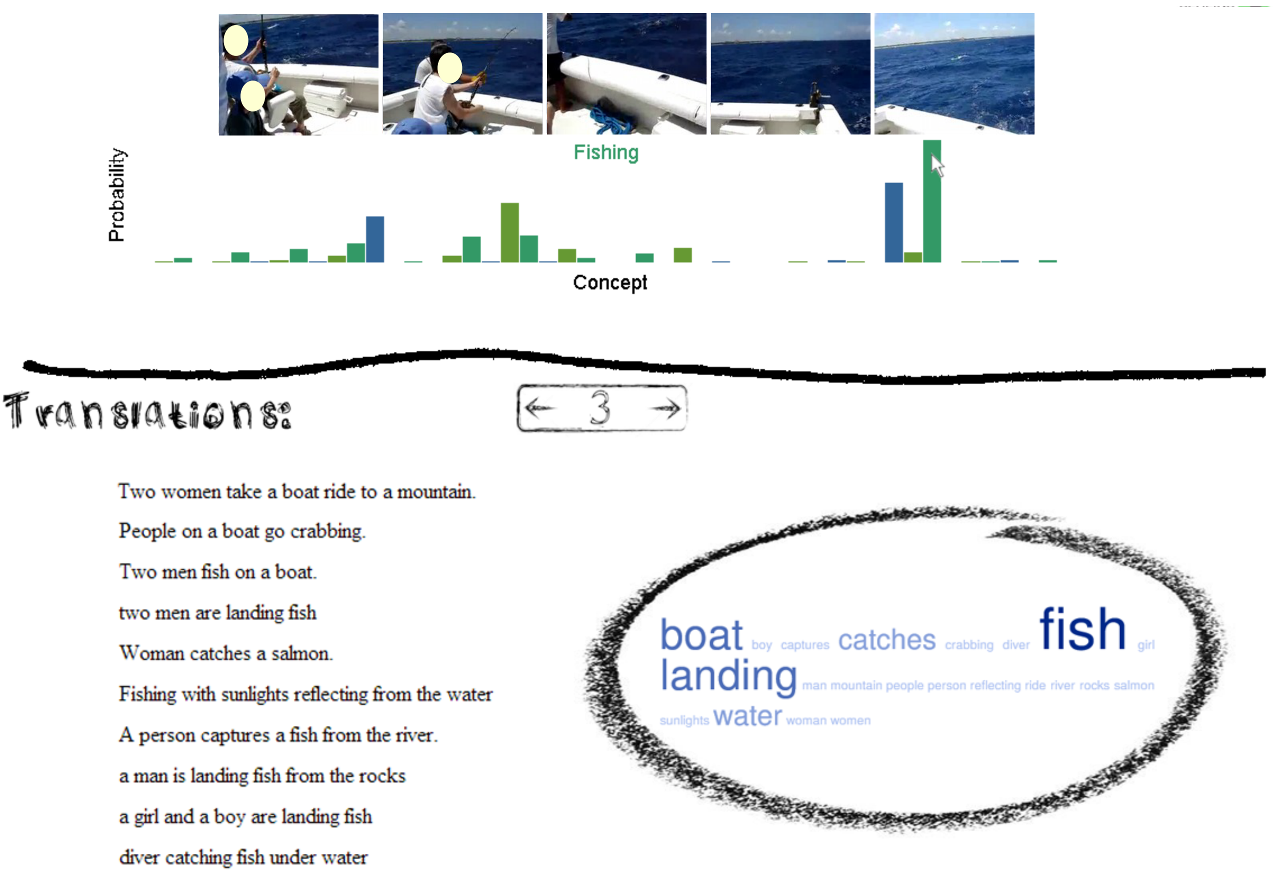
4. RESULTS

4.1 Efficiency

The results of this experiment are shown in Table 1. Training and testing of event classifiers with semantic signatures instead of low-level features requires 60 times less memory. Moreover, training and testing event classifier from semantic signature is respectively 61 and 57 times faster than low-level features. The results demonstrate the effectiveness of semantic signature for real-time search of video events without the need of large memory resources.

4.2 Accuracy

As Figure 2 shows, semantic signatures outperform low-level features when there are less than 25 positive examples available for training the classifiers. *i.e.* by using 10 positive examples, semantic signatures lead to an accuracy of 0.127,


Fishing
Probability
Concept
Translations:
← 3 →
boat boy captures catches crabbing diver fish girl
landing man mountain people person reflecting ride river rocks salmon
sunlights water woman women

Two women take a boat ride to a mountain.

People on a boat go crabbing.

Two men fish on a boat.

two men are landing fish

Woman catches a salmon.

Fishing with sunlights reflecting from the water

A person captures a fish from the river.

a man is landing fish from the rocks

a girl and a boy are landing fish

diver catching fish under water

Figure 3: Experiment 3: The interface of our event video search engine for explaining video content. The upper part provides a summary of the video by plotting the histogram of its dominant concepts. This histogram can be used to explain why the video has been retrieved. The lower part of the figure, demonstrates video translation in terms of textual descriptions. Moreover, the tag-cloud indicates the most frequent terms extracted from the translations [4].

where low-level features obtain 0.100, in terms of mean average precision. We explain it by the fact that semantic signatures transfer knowledge from the training examples used for their concept detectors. The transferred knowledge is especially effective when the number of positive examples is low. It demonstrates the suitability of semantic signatures for real-world event search scenarios, where only a few positive examples are provided by the user.

4.3 Interpretability

Video Summarization: As shown in Figure 3, a compact summary of video is obtained by selecting the most dominant concepts within semantic signatures for the video. This video summarization can be used to explain why a video has been retrieved for a specific event.

Video Translation: We use semantic signatures to generate textual descriptions of videos. Following our previous work [4], we first identify the dominant concepts per video, as proposed in [9]. Then, we concatenate captions of the dominant concepts and create a query. Finally, we search for the created query in a large pool of textual descriptions to find the sentences which best match the video content. Figure 3 shows some examples of the retrieved descriptions for a video. In this figure, the tag-cloud shows the most frequent terms in the retrieved descriptions.

5. DEMONSTRATION

We demonstrate a semantic video search engine that allows user to create a model for an arbitrary event on-the-fly. Our system shows different videos to an interacting user (Figure 4 (left)). In order to create a model for an event, such as *attempting a board trick*, *cleaning an appliance* or *birthday party*, the user selects a handful of positive video instances. The positive examples are supplemented with random negatives, to train an event model. Based on this model our video event search engine ranks the videos in an unseen test video collection by their event classification scores (Figure 4, right). Users may inspect the semantic signatures of the retrieved videos by clicking the appropriate button. Then, a summary of the video as well as its generated textual descriptions are returned to user to provide a better understanding about why the video has been retrieved, see Figure 3. Taken together, the search engine provides a means to interact with human to get instant access to complex events in video collections.

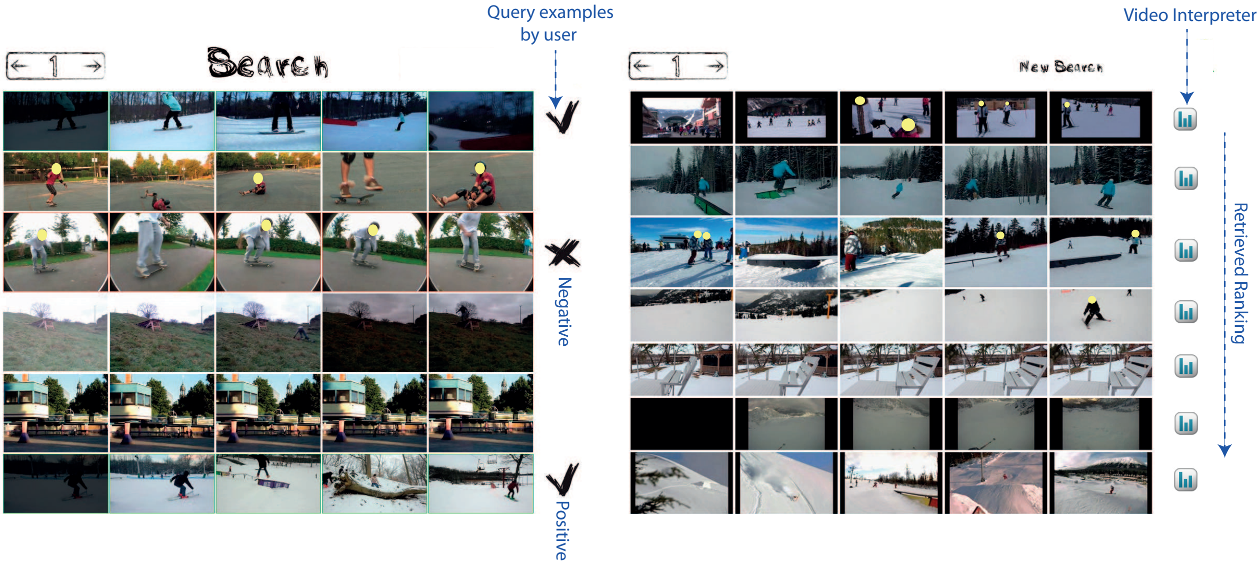


Figure 4: User interface of our event video search engine, based on semantic signatures, with a few user marked examples (left). The results of our on-the-fly event training model are displayed on unseen web video (right). Note that with only a handful of positive event examples, we can obtain a quite accurate retrieval result. The user may inspect the video summary and its textual translations by using the video interpreter, see Figure 3.

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] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *ECIR*, 2008.

[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] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013.

[4] A. Habibian and C. Snoek. Video2sentence and vice versa. In *ACM MM*, 2013.

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

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[7] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[8] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.

[9] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

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

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

[12] G. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. Koelma, K. van de Sande, A. Smeulders, and C. Snoek. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 25(1):17–32, 2014.

[13] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[14] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.

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

[16] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous audio visual internet collection. In *LREC*, 2012.

[17] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE T-PAMI*, 32(9):1582–1596, 2010.

[18] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Multimedia event recounting with concept based representation. In *ACM MM*, 2012.