Searching Informative Concept Banks for Video Event Detection

Masoud Mazloom, Efstratios Gavves, Koen E.A. van de Sande and Cees G.M. Snoek
ISLA, Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH, Amsterdam, The Netherlands
{m.mazloom, egavves, ksande, cgmsnoek}@uva.nl

ABSTRACT

An emerging trend in video event detection is to learn an event from a bank of concept detector scores. Different from existing work, which simply relies on a bank containing all available detectors, we propose in this paper an algorithm that learns from examples what concepts in a bank are most informative per event. We model finding this bank of informative concepts out of a large set of concept detectors as a rare event search. Our proposed approximate solution finds the optimal concept bank using a cross-entropy optimization. We study the behavior of video event detection based on a bank of informative concepts by performing three experiments on more than 1,000 hours of arbitrary internet video from the TRECVID multimedia event detection task. Starting from a concept bank of 1,346 detectors we show that *1.)* some concept banks are more informative than others for specific events, *2.)* event detection using an automatically obtained informative concept bank is more robust than using all available concepts, *3.)* even for small amounts of training examples an informative concept bank outperforms a full bank and a bag-of-word event representation, and *4.)* we show qualitatively that the informative concept banks make sense for the events of interest, without being programmed to do so. We conclude that for concept banks it pays to be informative.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Measurement

Keywords

Event recognition, concept detection, cross-entropy optimization

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.

1. INTRODUCTION

Automated understanding of events in unconstrained video has been a challenging problem in the multimedia community for decades [16]. This comes without surprise as providing access to events has great potential for many innovative applications [4,33]. Traditional event detectors represent an event by a carefully constructed explicit model [9,13]. In [9], for example, Haering *et al.* propose a three-layer inference process to model events in wildlife video. In each layer event-specific knowledge is incorporated ranging from object-level motion, to domain-specific knowledge of wildlife hunting behavior. While effective for detecting hunting events, such a knowledge-intensive approach is unlikely to generalize to other problem domains. Hence, event representations based on explicit models are well suited for constrained domains like wildlife and railroad monitoring, but they are unable, nor intended, to generalize to a broad class of events in unconstrained video like the ones in Figure 1.

Recently, other solutions have started to emerge. We group related works based on the type of representation used: bag-of-words and bank-of-concepts.

1.1 Event as bag-of-words

Inspired by the success of bag-of-word representations for object and scene recognition [14,31], there are several papers in the literature that exploit this low-level representation for event detection. In [15] the team of Columbia University, showed that state-of-art event detection performance is feasible by combining bag-of-words derived from SIFT descriptors, with bag-of-words derived from both MFCC audio features and space-time interest points. Their idea of combining multi-modal bag-of-words was further extended by Natarajan *et al.* [23] and Tamrakar *et al.* [29], who adhere to a more is better approach to event detection by exhaustively combining various visual descriptors, quantization methods, and word pooling strategies. In [12] Inoue *et al.* stress the importance of Principal Component Analysis to reduce the dimensionality of the growing amount of visual descriptors. Their event detection results are benchmarked with state-of-the-art results also, but it requires less computation than [23, 29]. In benchmarks like TRECVID's multimedia event detection task [30] the bag-of-words representation has proven it's merit with respect to robustness and generalization, but from the sheer number of highly correlated descriptors and vector quantized words, it is not easy to derive how these detectors arrive at their event classification. Moreover, events are often characterized by similarity in semantics rather than appearance. Our goal is to find an

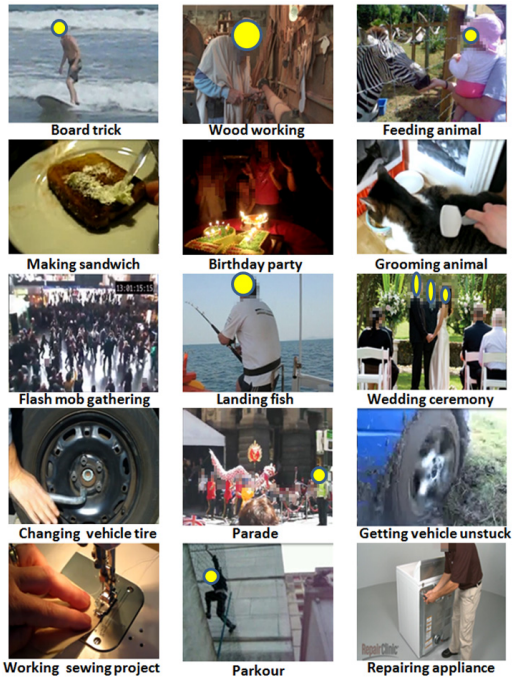

Figure 1: Eexample of fifteen arbitrary events in internet video content.

informative representation able to recognize, and ultimately describe, events in arbitrary video content. We argue that to reach that long-term goal a more semantic representation is urged for.

1.2 Event as bank-of-concepts

Inspired by the success of semantic concepts for improving video retrieval [10, 28], several papers in the literature exploit a bank of semantic concepts as the representation for learning event detectors. Ebadollahi *et al.*, for the first time, explored the use of semantic concepts for learning events [7]. For creating their bank-of-concepts, they employed the 39 detectors from the Large Scale Concept Ontology [22]. Each frame in their broadcast news video collection is then represented as a vector describing the likelihood of the 39 concept detectors. To arrive at an event classification score they employ a Hidden Markov Model. Due to the availability of large lexicons of concept annotations [5, 22], several others have recently also explored the utility of bank-of-concept representations [1, 8, 17, 21, 25]. In [21] Merler *et al.* argue to use all available concept detectors for event representation. Based on a keyframe representation containing 280 concept detector scores, and a support vector machine for learning, the authors show that competitive event detection results can be obtained. In [8] Gkalelis *et al.* propose to reduce, with the help of Mixture Subclass Discriminant Analysis, a bank-of-concepts consisting of 231 detector scores to a subspace best describing an event. Because for both [21] and [8] the resulting event detector operates on all concepts simultaneously, the precise explanation of an event cannot be provided. We are inspired by the concept bank approach to event representation [1, 7, 8, 21]. Our goal is to arrive at a more precise concept bank, while improving event detection accuracy. To that end we investigate whether we can learn for a given event what concepts are most informative to include in its concept bank.

1.3 Contribution

We make three contributions in this paper. First, we model finding the bank of informative concepts out of a large set of concept detectors as a rare event search. Second, we propose an approximate solution that finds the near optimal concept bank using a cross-entropy optimization [18]. Third, we show qualitatively that the found concept bank makes sense for the events of interest, without being programmed to do so.

To the best of our knowledge no method currently exists in the literature able to determine the most informative concept bank for learning to detect an event. Note especially the algorithmic difference with concept selection for video retrieval [11,32]. In the retrieval scenario the selected detector score is exploited directly for search. In our approach, the bank of automatically found informative detectors is optimized for *learning* to recognize an event. We study the behavior of our informative concept banks by performing several experiments on more than 1,000 hours of arbitrary internet video from the TRECVID 2011 multimedia event detection task. But before we report our experimental validation, we first introduce our algorithm which learns from video examples the informative concept bank for video event detection.

2. INFORMATIVE CONCEPT BANKS

Our goal is to arrive at an event representation containing informative concept detectors only. However, we need to first define what is informative. For example, one can reasonably expect that for the event "feeding an animal", concepts such as "food", "animal" or "person" should be more important, and thus informative.

We start from a large bank of concept detectors for representing events. Given a set of exemplar keyframes of an event category, the aim is to find a smaller bank of informative concepts that accurately describe this event. Suppose that the cardinality of the bank of all available concepts is S. Then the number of concept subsets within this set is 2^S. When S increases, the process of finding the best subset, i.e., the informative concept bank, will be very hard. In fact the problem of finding the best concepts from a large lexicon is an NP-complete problem for which approximation methods are the only viable solution.

We consider the problem of finding the best subset of informative concepts as a rare event in the concepts space. Hence, searching for the bank of informative concepts becomes a rare event search that is properly modeled by a rare event simulation. For solving rare event search problems, in general, the cross-entropy optimization [24] is a well known and frequently used solution [18]. As the cross entropy requires only a small number of parameters, chances of overfitting are minimized. Moreover, convergence is relatively fast and a near-optimum solution is guaranteed. We first describe briefly the theory behind cross-entropy, and then present our learning algorithm based on cross-entropy optimization for finding the informative concept banks for event detection.

2.1 Cross-Entropy

We want to maximize the contribution of the individual

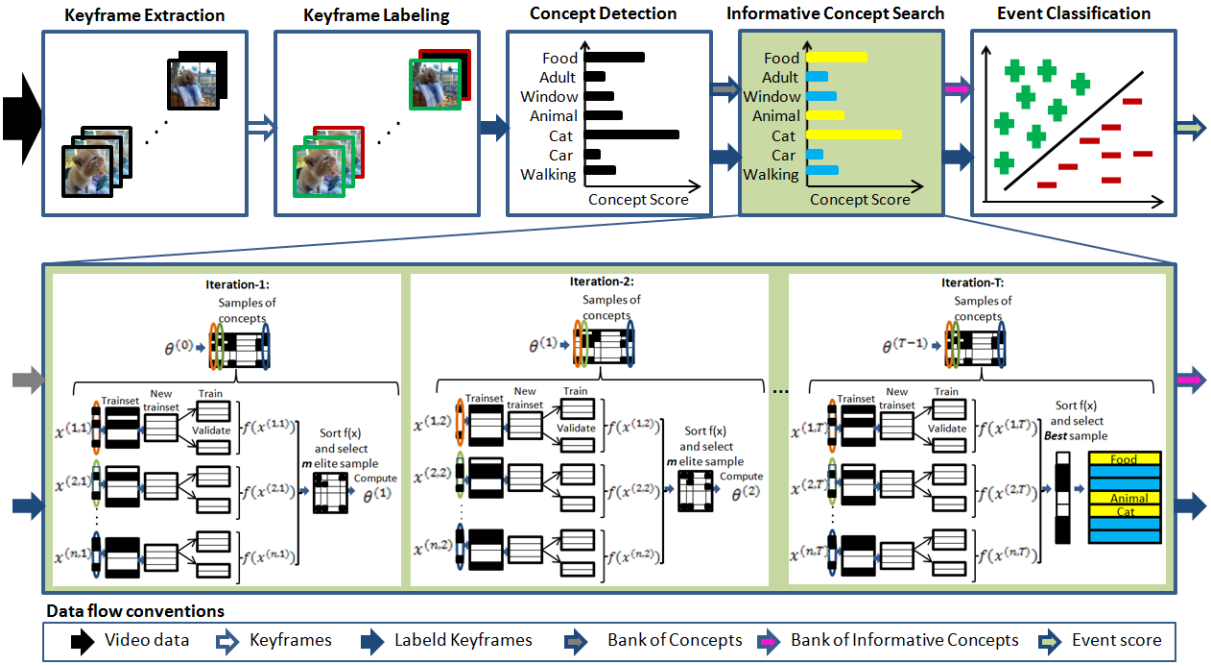

Figure 2: Flow chart for video event detection using an informative concept bank. We extract keyframes, label them based on event presence, and classify individual keyframes with a large number of concept detectors. We determine the informative concepts using the algorithm in Table 1. It selects random banks of concepts and determines their informativeness with the aid of an objective function (e.g., average precision) and cross-validation. After each iteration we update importance sampling parameter Θ^q. For the event *feeding an animal* the selection algorithm finds the concepts 'animal', 'cat', and 'food' to be the most informative. We represent each keyframe of the videos by this informative concept bank and use a classifier to infer a final event score.

concept detectors to the final classification of a video to one or the other event categories. We adopt the standard for the event detection benchmarks, mean average precision metric to measure the accuracy of our representation. Hence our objective function, $f(:)$, is the average precision.

We want to maximize the objective function $f(x)$ with respect to x, where x is a subset of the concept detectors that we have in our object bank. We thus want to identify the configuration x^* of concept detectors, which will return the maximum score $f(x^*)$.

The cross-entropy optimization [24] models the random variable x with the distribution $p(x;\Theta)$, where Θ is a variable that represents the distribution parameters. Since $x = \{x_i\}, i = 1, ..., S$ stands for the all S concept detectors, Θ_i is the variable that controls the participation of each of these concepts x_i to the final score $f(x)$. The cross entropy optimization estimates the optimal solution x^* for maximizing $f(x)$ as the expected value of $p(x;\Theta)$. Obviously, Θ plays an important role for x^*. We determine the optimal value for Θ by solving the following problem:

$$\Theta^* = \arg\max_{\Theta} \int_x I(f(x) \geqslant \alpha) p(x;\Theta) dx, \qquad (1)$$

where in this equation I is an indicator function, and α is a threshold that determines the minimum accuracy that a possible solution x should exhibit. Eq. 1 cannot be solved analytically, hence we need to derive an iterative approximation to Θ via the following three steps:

(1) Use $p(x;\Theta)$ to randomly generate n samples, that is:

$$x^1, ..., x^n \sim p(x;\Theta) \qquad (2)$$

(2) Evaluate each of x^j using $f(x)$. Then, sort the n samples descending order and select the top m samples $\{\hat{x^1}, ..., \hat{x^m}\}$, the *elite samples*.

(3) Finally, use the m elite samples to re-estimate Θ as the maximum likelihood estimators for maximizing $f(x)$.

The parameter Θ is updated in step 3 using the information from the m elite samples. Based on eq. 1 the solution vector Θ^q at iteration q minimizes the cross entropy distance between our current best model $p(x;\Theta^q)$ and the optimal $p(x;\Theta^*)$ one. From eq. 1, we observe that having more accurate elite samples generates solutions closer to the optimal Θ^*. Through the progression of iterations subsequent elite samples will exhibit higher and higher accuracy, leading to a better and better estimation of Θ_i^q. Repeating this update rule iteratively leads x to convergence towards x^* [24]. The stopping criterion for the algorithm may either be the accuracy standstill over the last iterations or reaching a maximum number of iterations.

2.2 Searching the informative concept bank

Now we present our algorithm for searching the informative concept bank for each event category. Since each concept is either selected, or not, we model function $p(x;\Theta)$ as

Table 1: The proposed algorithm which models finding an informative concept bank for video event detection as a cross-entropy optimization.

INPUT: Number of iterations (T), samples (n), elite samples (m), index of events ($event$)
OUTPUT: Informative concept bank per event (x^*)

1. for each $event$
2. Initialize $\Theta^{(0)}$
3. for $q = 1, \ldots, T$
4. **Concepts sampling**: Generate n samples $\{x^{(1,q)}, ..., x^{(n,q)}\}$ by using current parameter $\Theta^{(q-1)}$.
5. **Samples selection**: Find the m samples that perform best given the objective function $f(x)$.
6. **Update parameter vector** $\Theta^{(q)}$: Based on the best concept samples from step 5, update parameter set $\Theta^{(q)}$ by using Eq. 4
7. $x^* \leftarrow \Theta^{(T)}$

an one-trial binomial distribution, that is

$$x_i = Binomial(1, \Theta_i), for \quad i = 1, ..., S. \tag{3}$$

Each concept x_i follows a distribution $p(x_i; \Theta_i)$, and $\Theta = \{\Theta_i\}$ $i = 1, ..., S$. Given Θ we generate at the q-th iteration n samples $x^{(1,q)}, ..., x^{(n,q)}$ for all concepts $i = 1, ..., S$. Each of these samples $x^{(j,q)}$ in reality is a binary vector, with $x_i^{(j,q)} = 1$ when a concept i is part of the solution for this concept and 0 otherwise. The parameters Θ_i^q of our binomial distributions directly measure the impact of concept i in the process of event detection for each event. Larger Θ_i^q makes the presence of concept i in the optimal solution more likely. In the end, the majority of concepts should not participate in finding an event category, so that their binomial parameter Θ_i^q is equal to 0.

For the purpose of event detection, the objective function typically needs labeled training data to quantify the accuracy of various banks. To do so, we separate the training data into a training and validation set. An event classifier is learned from the selected concepts in the training set and validated on the validation set. We use average precision to reflect the accuracy on the validation set. After each iteration we update Θ_i^q by maximum likelihood estimation on the m elite samples. For a Binomial distribution, this accounts to averaging over the elite samples:

$$\Theta_i^{(q)} = \frac{1}{m} \sum_{j=1}^{m} x_i^{(j,q)}. \tag{4}$$

We visualize the overall flow chart for searching informative concept banks for video event detection in Figure 2. Our learning algorithm for obtaining the bank of informative concepts is summarized in Table 1.

3. EXPERIMENTAL SETUP

We investigate the effectiveness of informative concept banks for video event detection by performing a series of experiments on a large corpus of challenging real-world web video.

Table 2: Our experiments are evaluated on the TRECVID 2011 Multimedia event detection corpus. The training set is based on the provided event kits only. Number of video and extracted keyframes per event detailed.

	Training set				Test set			
	Positive		Negative		Positive		Negative	
Name of event	Video	Frame	Video	Frame	Video	Frame	Video	Frame
Board trick	161	1,592	555	11,673	114	2,334	4,177	36,758
Feeding animal	162	1,332	554	12,220	114	401	4,177	38,691
Landing fish	122	996	594	12,261	85	1,291	4,206	37,801
Wedding ceremony	128	1,595	588	11,147	89	2,766	4,202	36,326
Wood working	143	1,304	573	12,191	100	945	4,191	38,147
Birthday party	173	1,216	1,175	24,628	172	2,032	31,863	251,699
Changing a vehicle tire	111	1,124	1,237	24,716	113	1,244	31,922	252,487
Flash mob gartering	172	1,893	1,176	23,942	135	1,933	31,900	251,798
Getting a vehicle unstuck	132	1,269	1,216	24,581	83	504	31,952	253,227
Grooming an animal	138	1,411	1,210	24,424	81	521	31,954	253,210
Making a sandwich	126	1,504	1,222	24,350	137	1,885	31,898	251,846
Parade	138	1,491	1,210	24,340	187	1,556	31,848	252,175
Parkour	112	1,873	1,236	23,978	102	2,943	31,933	250,788
Repairing an appliance	123	1,518	1,225	24,296	88	1,366	31,947	252,365
Working on sewing project	120	1,171	1,228	24,687	82	1,046	31,953	252,685

3.1 Data set

TRECVID Multimedia Event Detection For our experiments we adopt the large-scale publicly available video data set from TRECVID's 2011 multimedia event detection corpus [30]. This corpus contains a collection of 38,387 internet video clips, totaling 1,229 hours. The MPEG-4 formatted video data consist of user-generated content posted to various Internet video hosting sites. TRECVID divided the data set into three collections, an event kit containing a textual description of the events together with labeled training video. A development collection[1] containing test video for the events *Board trick, Feeding animal, Landing fish, Wedding ceremony,* and *Wood working*, and an opaque collection containing test video for the events *Birthday party, Changing vehicle tire, Flash mob gathering, Getting a vehicle unstuck, Grooming animal, Making sandwich, Parade, Parkour, Repairing appliance,* and *Working on sewing project.* Since the groundtruth annotations are defined on different partitions of the data, we group the fifteen events into two groups. The first five events defined on the development collection are in *group 1* and the ten remaining events are in *group 2*. For a visual impression of characteristic event examples we refer to Figure 1.

Training set In our experiments we adopt the event kit as our training set, which corresponds to 2,061 video clips with an approximate duration of 92 hours. We report our results for events in group 1 on the development collection which contains 4,291 video clips corresponding to 146 hours. For events in group 2 we report our result on the test collection that contains 32,035 video clips with an approximate duration of 991 hours. We shot segment the video and designate the middle frame as as keyframe. To assure sufficient training data, especially from single-shot video, we require at least 10 frames per video from the event kit. As an arbitrary internet video may contain several non-relevant frames like black frames, over-exposed frames and extreme close-ups, we manually verify all the extracted keyframes from the positively labeled videos in the event kit. We label a keyframe

[1]To be precise, we use part 1 of the development collection and ignore part 2 which contains background video clips only.

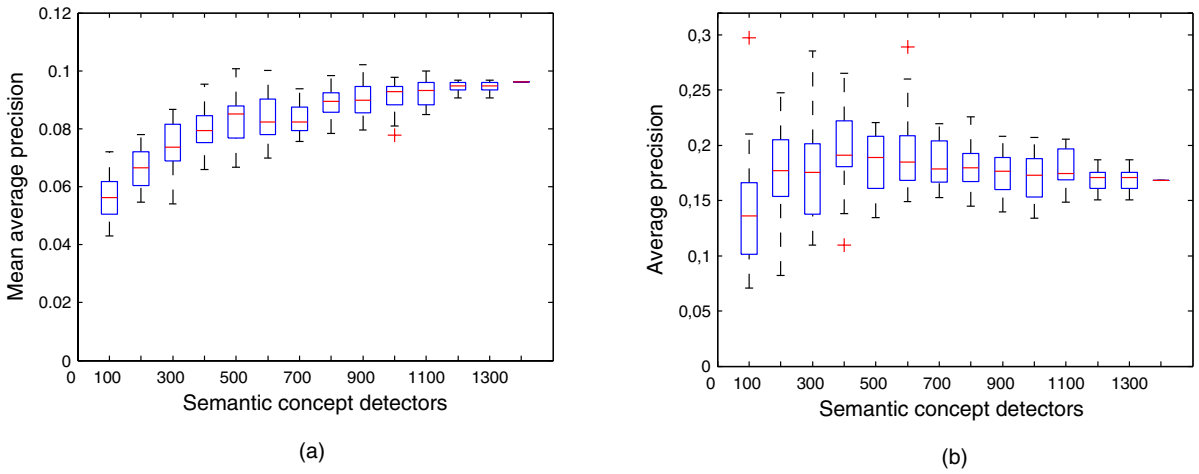

Figure 3: Experiment 1. (*a*) ***Influence of concept bank size*: Event detection accuracy increases with the number of concepts in the bank, but the variance suggests that some concept banks are more informative than others.** (*b*) ***Influence of concept bank size for "Landing fish"*: For the event *Landing fish* a small bank of 100 (random) concepts clearly outperforms the bank using all 1,346 concepts. Indicating that much is to be expected from a priori search for the most informative concept bank for an event.**

as positive if the context of the event is observable, if not we label it as negative. All the keyframes of negatively labeled videos are simply considered as additional negatives also.

Test set Similar to the training data we shot segment the videos in the development and opaque collection. To reduce computation we extract a fixed number of six frames per shot. Table 2 summarizes the number of labeled videos and keyframes available for each event in both our training and test sets.

3.2 Implementation details

Concept Bank We classify each keyframe in our data set with a bank of 1,346 concept detectors. The detectors are trained using annotations for 346 concepts from the TRECVID 2011 Semantic Indexing Task [2] and 1,000 concepts from the ImageNet Large Scale Visual Recognition Challenge 2011 [5]. We implement them using a bag-of-words with SIFT [19], OpponentSIFT and RGB-SIFT descriptors extracted at Harris-Laplace keypoints and dense sampled points, at every 6 pixels for two scales, using the Color Descriptor software from [31]. The codebook size is 4,096 and we employ a 1x3 spatial pyramid subdivision. As classifier we employ a Support Vector Machine with a fast approximate histogram intersection kernel [20].

Event detection As we focus on obtaining an informative concept bank for video event detection in this paper, we are for the moment less interested in the accuracy optimizations that may be obtained from various kernel settings [3, 6, 34]. Hence, we train for each event a one-versus-all linear support vector machine [26] and fix the value of its regularization parameter C to 100. We train and test the linear support vector machine on keyframe level. To arrive at a decision at video level, we employ max pooling over the classification scores per keyframe.

Cross entropy parameters After initial testing on small partitions of the data, we set the parameters of our cross-entropy learning algorithm to find the informative concept banks for each event as follows: number of iterations 20, number of concept samples in each iteration 1,000, and number of elite samples in each iteration 200. Inside the objective function we use 5-fold cross-validation.

Evaluation criteria For both objective function $f(x)$ in our learning algorithm, as well as the final event detection performance we consider as evaluation criterion the average precision (AP), which is a well known and popular measure in the video retrieval literature [27]. We also report the average performance over all fifteen events as the mean average precision (MAP).

3.3 Experiments

In order to establish the effectiveness of informative concept banks for video event detection, we perform three experiments.

Experiment 1: ***Influence of concept bank size*** To assess the effect of a growing number of concepts in a bank on video event detection performance, we randomly sample a bank

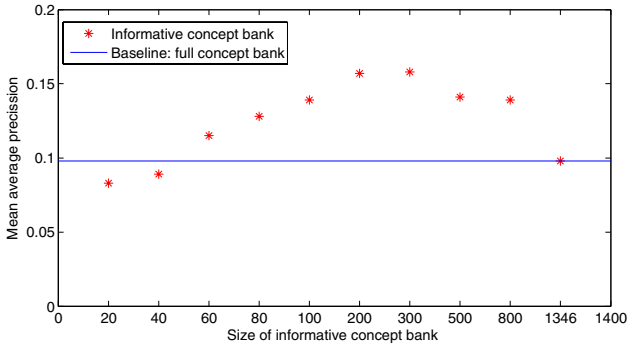

Figure 4: Experiment 2. The result of using different size of informative concept bank. The result shows that there is an informative concept bank composed of 300 concept detectors that reach 0.158 MAP in video event detection.

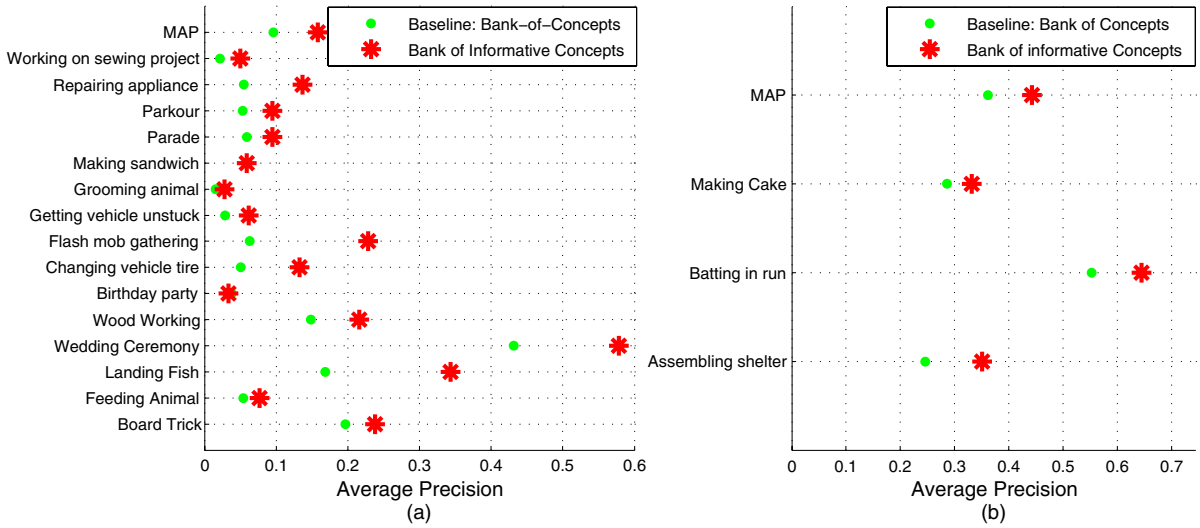

Figure 5: Experiment 2. (a) **An informative concept bank always outperforms a bank containing all available concept detectors for video event detection. On average the relative improvement is 65%.** (b) **Repeating experiment 2 on the dataset provided by [21] confirms the conclusion of (a).**

of concepts from our 1,346 concept-lexicon with a step size of 100. Each keyframe in our dataset is then represented in terms of the detector scores from the concepts in this random bank. We repeat this procedure 20 times for each bank size.

Experiment 2: ***All concepts versus informative concepts*** In this experiment we compare a bank based on all available concept detectors to a bank containing informative concepts. As the baseline, we represent each keyframe in our data set as a 1,346D vector of detector scores (see section 3.2). For finding the informative concept bank per event, we apply the cross-entropy optimization as described in section 2.2 on the training set only. We train an event detector on the most informative concept bank and report its performance on the (unseen) test set. Since we can fix the sample size inside our algorithm, we evaluate the following bank sizes: 20, 40, 60, 80, 100, 200, 300, 500, 800 to find the most appropriate setting for our dataset based on the MAP.

Experiment 3: ***Influence of event training examples*** To investigate the stability of informative concept banks for video event detection under limited number of training examples, we compare it with a bank containing all available concepts and an appearance-based bag-of-words using densely sampled SIFT descriptors, which are vector quantized into a 4K codebook. In all cases we employ a linear Support Vector Machine for event classification. We vary the number of positive training examples from 1 to 900 keyframes. The positive event training examples are randomly sampled from our pool of positively labeled keyframes, the negative examples are fixed per event (see Table 2). For each (random) set of positive examples we measure event detection performance on the test set and repeat this process 20 times.

4. RESULTS

4.1 Influence of concept bank size

We plot the results of experiment 1 in Figure 3(a). As expected the event detection accuracy increases when more and more concept detectors are part of the bank. Up to approximately 500 (random) concept detectors the increase in event detection accuracy is close to linear, afterwards it saturates to the end value of 0.096 MAP when using all 1,346 available concept detectors. Interestingly, the box plot reveals that there exist a bank, containing only 500 concepts, which performs better than using all concepts (compare the top of the whisker at 500 concepts, with an MAP of 0.102 with the maximum MAP of 0.096 when using all concepts). This result shows that some banks of concepts are more informative than others for video event detection.

When we zoom in on individual events the connection between concept banks and event definitions can be studied. We inspect the box plot of Figure 3(a) also for the 15 individual events (data not shown). The plots reveal several positive outliers using just a small number of concepts in the bank. Noticeable examples are obtained for the events *Landing fish, Wedding ceremony, Flash mob gathering*, and *Parkour*. Figure 3(b) details the box plot for *Landing fish*. For this event we observe an outlier bank with an AP of 0.292 containing only 100 randomly selected concepts (compare to the maximum of 0.170 when using all concepts). The results of experiment 1 show that, in general, the event detection accuracy increases with the number of semantic concepts in the bank. However, it also shows that some banks of concepts are more informative than others for specific events, and this may result in improved event detection accuracy.

4.2 All concepts versus informative concepts

We plot the result of experiment 2 in Figure 4 and Figure 5. The result in Figure 4 shows that by using the concept bank with size less than 40 concepts, the performance of event detection is below the baseline of 0.098. When we increase the size of concept bank from 60 to 300 concepts, the event detection accuracy also increases from 0.118 to 0.158. However, more is not better, as further increasing the size from 300 to 800 results in a decrease in MAP again from

Figure 6: Experiment 2. Informative concept banks for the events (*a*) *Flash mob gathering* and (*b*) *Batting in run*. Font size correlates with automatically estimated informativeness. Note that the algorithm found concepts that make sense without being programmed to do so.

0.158 to 0.133. The result in Figure 4 show that by selecting an informative concept bank with size 300 we can reach to the 0.158 MAP. We plot the result of using the informative concept bank of size 300 in the Figure 5(a). We observe that on average, the bank of informative concepts relatively improve the normal bank-of-concepts method 65% (0.158 vs 0.096 MAP). We can see that in all event categories, our representation based on the informative concept bank is better than using a representation using all concepts of bank. When we focus on the result of Figure 5(a) we find a considerable improvement for events such as *Landing fish, Wedding ceremony* and *Flash mob gathering*, where the improvements are 88%, 59%, and 175% respectively. Recall that we reach this result by using an informative concept bank containing only 23% of the concept detectors available. When relevant concepts are unavailable in the concept bank, the results will not improve, as can be seen for the event *Making sandwich*. Figure 6(a) highlights the informative concept bank for the event *Flash mob gathering*.

For sake of comparison with the state-of-the-art we also repeat experiment 2 for TRECVID's 2010 multimedia event detection corpus. This data set consists of three events: *Assembling a shelter, Batting in run, and Making cake*. Here we adopt the 280 concept bank provided by Merler *et al.* [21]. Again, we employ our cross-entropy algorithm for finding the most informative concept bank per event. The results

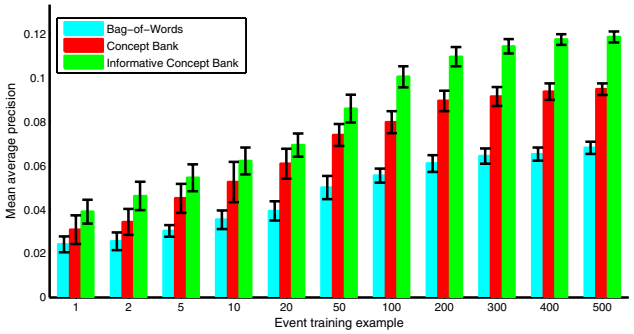

Figure 7: Experiment 3. An informative concept bank outperforms a full concept bank and bag-of-words, even for small amounts of training examples.

in Figure 5(b) confirm the results of experiment 2. Again the informative concept bank outperforms the baseline for all three events (0.443 vs 0.360 MAP) and uses only 36% of the available concepts (100 vs 280). Figure 6(b) shows the automatically selected concepts for the event *Batting in run*.

The results of experiment 2, and the same experiment on the dataset provided by [21], show that event detection using an automatically found bank of informative concepts outperforms a bank using all concepts, and always contains significantly less semantic concepts.

4.3 Influence of event training examples

We plot the results of experiment 3 in Figure 7. As expected the event detection accuracy increases when more and more positive event example are used for training the classifier. Independent of the number of training examples used, the accuracy of the informative concept bank outperforms both the concept bank using all available detectors and the bag-of-words. Moreover, the difference in accuracy between the three methods is increasing when the number of event training examples grows. For example, when we use only 1 positive event training example the difference between informative concept banks is small with 0.01 compared to a full concept bank and 0.015 compared to bag-of-words. When using 500 event keyframe examples the differences increases to 0.028 compared to bank of concepts and 0.058 compared to bag-of-words.

The result of experiment 3 shows an increasing video event detection accuracy when increasing the number of positive training examples. More surprising, concept banks outperform bag-of-words for small amounts of training examples. Moreover, we observe that independent of the number of positive training example used, the accuracy of the informative concept bank tends to be better than both the full concept bank and bag-of-words. We conclude that, compared to competing approaches, an informative concept bank is most robust under a limited number of training examples.

5. CONCLUSION

We study event detection based on banks of concept detectors. Different from existing work, which simply includes in the bank all available detectors, we propose a cross-entropy inspired algorithm that learns to find from examples the bank of most informative concepts. We study the behavior

of informative concept banks by performing three experiments on the unconstrained web video collection from the TRECVID 2011 multimedia event detection task using a total of 1,346 concept detectors.

The result of experiment 1 gives an indication that large banks of concept detectors are important for covering a variety of complex events, as they may appear in unconstrained video. In general, the event detection accuracy increases with the number of concept detectors in the bank. However, it also shows that some concept banks are more informative than others for specific events, and this may result in improved event detection accuracy. The results of experiment 2, and the same experiment on the dataset provided by Merler *et al.* [21], show that event detection using an informative concept bank outperform banks using all concepts, and always contains significantly less detectors. Finally, experiment 3 reveals that our informative concept bank outperforms both a bank using all concepts and a bag-of-words for small amounts of training examples. What is more the concepts in the informative concept bank appear to have a semantic relation with the events they model. We conclude that for video event detection using concept banks it pays to be informative.

Acknowledgments This research is supported by the STW STORY project, the BeeldCanon 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] T. Althoff, H. O. Song, and T. Darrell. Detection bank: An object detection based video representation for multimedia event recognition. In *ACM Multimedia*, 2012.

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

[3] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Video event classification using string kernels. *MTAP*, 48(1), 2010.

[4] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *MTAP*, 51, 2011.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *TPAMI*, 34(9), 2012.

[7] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *ICME*, 2006.

[8] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *CBMI*, 2011.

[9] N. Haering, R. Qian, and I. Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *TCSVT*, 2000.

[10] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 2008.

[11] B. Huurnink, K. Hofmann, and M. de Rijke. Assessing concept selection for video retrieval. In *ACM MIR*, 2008.

[12] N. Inoue et al. TokyoTech+Canon at TRECVID 2011. In *NIST TRECVID Workshop*, 2011.

[13] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *TPAMI*, 22(8), 2000.

[14] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *TMM*, 12(1), 2010.

[15] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[16] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in videos. *TSMC*, 39(5), 2009.

[17] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[18] X. Li, E. Gavves, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Personalizing automated image annotation using cross-entropy. In *ACM Multimedia*, 2011.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[20] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[21] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia*, 14(1), 2012.

[22] M. R. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3), 2006.

[23] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[24] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, 2004.

[25] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[26] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Math. Program.*, 127(1), 2011.

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

[28] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *FnTIR*, 2(4), 2009.

[29] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. S. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

[30] TRECVID Multimedia Event Detection Evaluation Track, 2011. `http://www.nist.gov/itl/iad/mig/med.cfm`.

[31] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9), 2010.

[32] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang. Selection of concept detectors for video search by ontology-enriched semantic spaces. *TMM*, 10(6), 2008.

[33] L. Xie, H. Sundaram, and M. Campbell. Event mining in multimedia streams. *Proceedings of the IEEE, 96*, 2008.

[34] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *TPAMI*, 30(11), 2008.