Conceptlets: Selective Semantics for Classifying Video Events

Masoud Mazloom, Efstratios Gavves, and Cees G. M. Snoek, *Senior Member, IEEE*

***Abstract*—An emerging trend in video event classification 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, which we call the *conceptlet*. We model finding the conceptlet out of a large set of concept detectors as an importance sampling problem. Our proposed approximate algorithm finds the optimal conceptlet using a cross-entropy optimization. We study the behavior of video event classification based on conceptlets by performing four experiments on challenging internet video from the 2010 and 2012 TRECVID multimedia event detection tasks and Columbia's consumer video dataset. Starting from a concept bank of more than thousand precomputed detectors, our experiments establish there are (sets of) individual concept detectors that are more discriminative and appear to be more descriptive for a particular event than others, event classification using an automatically obtained conceptlet is more robust than using all available concepts, and conceptlets obtained with our cross-entropy algorithm are better than conceptlets from state-of-the-art feature selection algorithms. What is more, the conceptlets make sense for the events of interest, without being programmed to do so.**

***Index Terms*—Concept detection, cross-entropy optimization, event recognition.**

Fig. 1. Example videos for the events *Assemblinga shelter Board trick*, and *Birthday*. Despite the challenging diversity in visual appearance, each event maintains specific semantics in a consistent fashion. This paper studies whether a selective and descriptive event representation based on concept detectors can be learned from video examples.

I. Introduction

AUTOMATED understanding of events in unconstrained video has been a challenging problem in the multimedia community for decades [1]. This comes without surprise as providing access to events has great potential for many innovative applications [2]–[4]. Traditional classifiers represent an event by a carefully constructed explicit model [5], [6]. In [6], 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 classifying 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 Fig. 1.

Recently, other event classification solutions have started to emerge. Inspired by the success of bag-of-word representations for object and scene recognition [7], [8], several papers in the literature exploit this low-level representation for event classification [9]–[16]. In [9] Jiang *et al.* show that robust event classification accuracy 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 is further extended by Natarajan *et al.* [10] and Tamrakar *et al.* [11], who adhere to a more is better approach to event classification by exhaustively combining various visual descriptors, quantization methods, and word pooling strategies. In [13], [14] the robustness and efficiency of various low-level features for event classification are compared. In challenging benchmarks like TRECVID's multimedia event detection task [17] and Columbia University's Consumer Video dataset [18] the bag-of-words representation has proven it's merit with respect to robustness and generalization, but from the sheer number

Manuscript received January 24, 2014; revised June 13, 2014; accepted September 15, 2014. Date of publication September 23, 2014; date of current version November 13, 2014. This work was supported in part 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 under Contract D11PC20067. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chong-Wah Ngo.

The authors are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands (e-mail: m.mazloom@uva.nl; efstratios.gavves@gmail.com; cgmsnoek@uva.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2359771

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. In this paper we attempt to find a video representation able to recognize, and ultimately describe, events in arbitrary content. We argue that to reach that long-term goal a more semantic representation than bag-of-words is urged for.

Inspired by the success of semantic concept detectors such as 'Car', 'Animal', and 'Indoor' for image retrieval [19], object recognition [20], [21], action recognition [22], and video retrieval [23], [24] several papers in the event classification literature exploit a bank of concept detector scores as the video representation [25]–[32]. Ebadollahi *et al.*, for the first time, explored the use of semantic concepts for classifying events [25]. For creating their bank-of-concepts, they employed the 39 detectors from the Large Scale Concept Ontology [33]. 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 [33], [34], several others have recently also explored the utility of bank-of-concept representations for event classification [26], [27], [31], [32]. In [26] Merler *et al.* argue to use all available concept detectors for representing an event. Based on a video representation containing 280 concept detector scores, and a support vector machine for learning, the authors show that competitive event classification results can be obtained on the challenging internet video clips from the TRECVID 2010 Multimedia event detection collection. In [32] Habibian *et al.* arrive at a similar conclusion as [26] using a concept bank consisting of 1,346 concepts for event classification on a partition of the TRECVID 2012 Multimedia event detection collection. We note that in all these works [25], [26], [32] the resulting event detector operates on all concepts simultaneously, making it hard to pinpoint what concepts are most informative for each event under consideration.

Rather than using as many concepts as one can obtain, Liu *et al.* [31] show that by characterizing events using only a small set of carefully selected concepts, competitive results are feasible as well. It means that we do not necessarily need a large set of concept detectors to represent events. Rather than exploiting prior knowledge to manually specify a concept-subset for each event, we aim to learn the most informative concepts for an event from examples. We are inspired by the concept bank approach to event representation [25]–[29], [31], [32], so we start with a set of concept detectors as well. However, instead of using all available concepts, we attempt to learn from examples for a given event what concepts are most informative to include in its concept bank, which we call the *conceptlet*. Before detailing the contributions of our work, we first discuss related work on concept selection that we consider most relevant to this paper.

II. Related Work

In our survey of related work, we consider concept selection in the context of the multimedia retrieval literature and the feature selection literature.

A. Concept Selection by Multimedia Retrieval

Concept selection has been studied extensively in the video retrieval literature [35]–[41]. These selections automatically translate an input query into a weighted list of concepts which are then used for the retrieval. In [35] Natsev *et al.* consider text-based, visual-based and result-based selections. Using these three algorithms they find three rankings of concepts and use them for selection. In [36] Snoek *et al.* use text and visual analysis to select the single best concept for a query. Concepts are ranked according to their similarity to the query using the vector space model [42]. In [41] Wei *et al.* propose a semantic space to measure concept similarity and facilitate the selection of concept detectors based on the cosine similarity between the concepts and the query in the semantic space. Compared to [36], their approach combines detector scores from multiple selected concepts. Li *et al.* in [40] are inspired by tf-idf, which weights the importance of a detector according to its appearance frequency. In [19] Rasiwasia *et al.* rank concepts based on the scores the detectors obtained on the visual query images. In [38] Rudinac *et al.* make a ranking of concepts based on the frequency, variance, and kurtosis of concepts in the video queries. Using these three criteria, they select concepts. We observe that, in general, concept selection in multimedia retrieval, ranks a bank of concepts using text and video analysis and selects the single best or multiple concepts from the top of the obtained ranking.

All these existing selections evaluate the concept detectors individually and optimize a ranking of concepts per query. However, none of them considers the co-occurrence between the selected concepts. One can reasonably expect that for the events *feeding an animal* and *grooming an animal*, the concept 'cat' is important, but to differentiate the two events 'cat' has to co-occur with either 'food' or 'bathtub'. Rather than evaluating concepts individually, we aim in this paper to evaluate subsets of selected concepts simultaneously. We strive to select a near optimal concept-subset for each event category. We propose an algorithm that learns from examples what concepts in a bank are most informative per event.

B. Concept Selection by Feature Selection

A second perspective on concept selection considers it as feature selection, as common in the machine learning literature. Feature selection reduces data dimensionality by removing the irrelevant and redundant features using either unsupervised or supervised approaches. An example of unsupervised feature selection in the context of event classification by concept detectors is the work by Gkalelis *et al.* [27] who propose Mixture Subclass Discriminant Analysis to reduce a bank-of-concepts consisting of 231 detector scores to a subspace best describing an event. Since the algorithm alters the original concept representation it can no longer describe the semantics of the events of interest. Different from their work, we focus here on the problem of supervised feature selection, where the class labels, in our case event labels, are known beforehand.

Supervised feature selections are commonly classified into three categories, depending on their integration into the classifier [43], [44]: filters, embedders and wrappers.

Filters [45], [46], evaluate each feature separately with a measure such as mutual information or the correlation coefficient between features and class label. Hence, filters ignore dependencies between a set of features, which may lead to decreased classification performance when compared to other feature selections. Moreover, filters are usually computationally efficient and they produce a feature set which is not optimized for a specific type of classifier. Finally, filters provide a feature ranking rather than an explicit best feature subset, which demands a cut off point that needs to be set during cross validation. A strong filter is the Minimum Redundancy Maximum Relevancy proposed by Peng *et al.* [46], which uses mutual information and correlation between features for selection. When applied for selecting concepts for representing events, this method selects concepts that are mutually far away from each other while they still have high correlation to the event of interest. The feature selection computes a score for each concept based on the ratio of the relevancy of the concept to the redundancy of the concepts in the concept bank. Then it provides a concept ranking and removes the low scoring concepts. However, there may exist some concepts which are ranked low when considered individually but are still useful when considered in relationship with other concepts. In fact, Habibian *et al.* [32] presented an analysis that showed effective video event classification can be achieved, even when individual concept detector accuracies are modest, if sufficiently many concepts are combined. Hence, instead of selecting concepts by relying purely on detector scores, as Minimum Redundancy Maximum Relevancy does, we prefer to be less sensitive to the performance of the concept detectors. If the presence of a concept, either an accurate or inaccurate one, improves the accuracy of an event classifier, we strive to maintain it.

Embedders [47], [48], consider feature selection within the classifier construction. Compared to filters, the embedders can better account for correlations between features. State-of-the-art embedders are L1 norm SVM methods such as L1-Regularized Logistic Regression proposed by Ng [48]. During constructing of a linear classifier this embedder penalizes the regression coefficients and pushes many of them to zero. The features which have non-zero regression coefficients are selected as the informative features. The algorithm is most effective when there are many more redundant features than training examples. Furthermore, by definition of sparsity, one should expect and target for minimizing the number of non-zero elements in the solution vector. This condition is equivalent to employing the L0 norm for regularization. The L0 norm is accompanied by non smooth derivatives, which cannot be minimized in a gradient descent based setting. As an approximation, in [48] the L0 norm is replaced with the L1 norm. However, the L1 norm does not necessarily return the most optimal sparse solution. In this paper we attempt to solve the L0 problem directly and obtain a truly sparse, optimal solution. Not by using filters or embedders, but by a wrapper.

Wrappers [49], [50], search through the feature space and evaluate each feature subset by a classifier. To search the space of all feature subsets, a search algorithm is wrapped around the classification model. However, as the space of feature subsets grows exponentially with the number of features, heuristic methods are often used to conduct the search for an optimal subset. The advantages of wrappers are the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. Wrappers usually provide the proper feature set for that particular type of model. A common drawback of wrappers is that they have a higher risk of overfitting than other selections and are computationally more intensive. We demonstrate that the increased computation pays off for more accurate event classification. In our previous work [51], we propose a wrapper to find an informative concept-subset per event. For each keyframe in a video the most representative concepts are selected and eventually aggregated to video level. While effective, the algorithm cannot determine the optimal bank size and its iterative procedure on frame-level is computationally demanding. In this paper we address these two drawbacks. Inspired by the *wrappers* and our previous work [51], we attempt to find what concepts in a bank are most informative per event.

C. Contribution

We make three contributions in this paper. First, we model selecting the conceptlet out of a large set of concept detectors as an importance sampling simulation. Second, we propose an approximate solution that finds the near optimal conceptlet using a cross-entropy optimization. Third, we show qualitatively that the found conceptlets make 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 concepts for video event classification, other than our initial version of this work [51]. Note especially the algorithmic difference with concept selection by multimedia retrieval [19], [35]–[38], [41]. In the multimedia retrieval scenario the selected detector score is exploited directly for search. In our approach, the conceptlet is optimized for *learning* to classify an event. We study the behavior of conceptlets by performing several experiments on more than 1,000 hours of arbitrary internet video from the TRECVID Multimedia Event Detection tasks 2010 [17], 2012 [17], and Columbia's Consumer Video dataset [18]. But before we report our experimental validation, we first introduce our algorithm which learns from video examples the conceptlet for video event classification.

III. Conceptlets

Our goal is to arrive at an event representation containing informative concept detectors only, which we call a *conceptlet*. However, we first need to 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 to create a discriminative event model, and thus informative. We start from a large bank of concept detectors for representing events. Given a set of video exemplars of an event category, the aim is to find a (smaller) conceptlet that accurately describes this event. In this Section we describe our algorithm for selecting the conceptlet for each event category.

A. Preliminary

We first introduce some basic notation. Suppose we have a concept bank consisting of m concepts, $C = \{c_1, \ldots, c_m\}$ and C^n represents a concept-subset with length n, where $(n \ll m)$. Given a set of exemplar videos, a conceptlet C^* for an event is sampled from the space D^n of all possible concept-subsets with size n, that is $C^n \in D^n$, to best describe the video exemplars of the event. For a subset C^n, a concept is selected according to the probability density function $p(\alpha_i; .)$. Here, α_i denotes the binary variable that corresponds to whether concept c_i was selected or not, which is $\alpha_i = \{0, 1\}$. We denote the parameter that controls the probability of $\alpha_i = 1$, with θ_i, that is $p(\alpha_i; \theta_i)$.

The number of different concept-subsets with length n out of m concepts, *i.e.,* $|D^n|$, is equal to $\binom{m}{n}$, which grows combinatorially with increasing m and decreasing n. Thus, the probability of finding an informative conceptlet becomes very small. This inspires us to model the problem of finding the rare conceptlet, *i.e.,* C^*, in the space D^n, as an importance sampling problem [52]. We use the cross-entropy [53], proven to yield robust results in a variety of estimation and optimization problems [51], [54], [55] without depending too much on parameters and their initialization. As the cross-entropy requires only a small number of parameters, chances of overfitting are minimized during the process of finding the informative concepts during training. Moreover, convergence is relatively fast and a near-optimum solution is guaranteed [53].

Suppose that the sample of a random subset C^n is drawn from space D^n using the probability density function p. Since, every concept will either be sampled, or not, we assume function p to follow a one-trial binomial distribution:

$$p(\alpha_i, \theta_i) = \theta_i^{\alpha_i}(1 - \theta_i)^{1-\alpha_i}. \tag{1}$$

Moreover, assume that there is a neighborhood $\epsilon \subset D^n$ containing the concept-subsets with size n, accurately describing the video exemplars. Let l be the probability of sampling a concept-subset C^n from the ϵ neighborhood. Each of the of concept-subsets C^n has a limited capacity of accurately representing the video exemplars. Let $f(C^n)$ be the score function which measures the capacity that the concept-subsets C^n accurately represents the video exemplars. Suppose s is the lowest score of all concept-subsets in the neighborhood ϵ, according to the score function f, *i.e.,* $s = \min f(C^n), C^n \in \epsilon$. Then, the concept-subset with probability l will be the informative conceptlet C^n for which

$$l = P_\theta(f(C^n) \geq s) \tag{2}$$

We approximate this probability with the expectation:

$$l = E_\theta I(f(C^n) \geq s), \tag{3}$$

where $I(f(C^n) \geq s)$, is an indicator function, referring to the set of concept-subset C^n for which the condition $f(C^n) \geq s$ holds. The straightforward way for estimating l is to use conventional sampling methods, such as crude Monte Carlo. Since the space of all possible concept-subsets is huge, estimating the probability l of a concept-subset C^n in ϵ using the density function p is impractical.

B. Cross-Entropy Formulation

An alternative way is based on importance sampling simulation. To illustrate, suppose a different probability density function h exists, which draws samples from neighborhood ϵ with high probability. Using h has the advantage of drawing more concept-subsets C^n from ϵ. Indeed, h is used as an importance sampling density function to estimate the expectation of l, denoted $\hat{l}$, using a likelihood ratio estimator. More precisely, for N concept-subsets C^n samples, $\hat{l}$ is equal to:

$$\hat{l} = \frac{1}{N}\sum_{r=1}^{N} I(f(C_r^n) \geq s)\frac{p(C_r^n; \theta)}{h(C_r^n)}, \tag{4}$$

where C_r^n denotes the r^{th} concept-subset with size n. The expectation $\hat{l}$ is then optimally estimated when the right side of Eq. (4) is equal to l, which means the value of expression inside sigma has to be equal to l, $I(f(C_r^n) \geq s) \cdot \frac{p(C_r^n;\hat{\theta})}{h(C_r^n)} = l$. For this reason the value of density function $h(C_r^n)$ has to be equal to:

$$h(C_r^n) = \frac{I(f(C_r^n) \geq s)p(C_r^n; \theta)}{l}. \tag{5}$$

Since Eq. (5) depends on the unknown quantity l, an analytical solution is impossible. Instead, the solution is to be found in an iterative approximation. Let us assume that there exists an optimal conceptlet C^*, controlled by the parameter vector θ^*.

Using C^*, the maximum score with respect to a specific video event classification accuracy is given by s^*. We denote this theoretical conceptlet state as $\langle C^*, \theta^*, s^* \rangle$ and all other estimated concept-subsets C^n as $\langle \hat{C}^n, \hat{\theta}, \hat{s} \rangle$. The goal is to find $\langle \hat{C}^*, \hat{\theta}^*, \hat{s}^* \rangle$ which best approximates $\langle C^*, \theta^*, s^* \rangle$, *i.e.,* the theoretical optimal conceptlet. In order to reach the goal state, $\langle C^*, \theta^*, s^* \rangle$, we generate multiple $\langle \hat{C}^n, \hat{\theta}, \hat{s} \rangle$ at each iteration. At each iteration, the concept-subsets C^n that perform best are used to update the search parameters θ. The iterations gradually converge to neighborhood ϵ with high probability. To guarantee convergence towards the goal state, the distance between p and h should be decreased after each iteration. This is achieved by adapting the importance sampling density function h via updating the parameters θ of the iteration's best performing subsets. A particularly convenient measure of distance between two densities h and p is the *Kullback-Leiblerdistance*, which is also termed the *cross-entropy* between h and p.

The cross-entropy is defined as:

$$D_{CE}(h, p) = \int h(x) \ln \frac{h(x)}{p(x)} \mathrm{d}x. \tag{6}$$

Given that the sampling distributions p and h of concept-subsets follow a one-trial binomial distribution, the cross-entropy between the density function h and density function p is reduced for:

$$\hat{\theta}_i^t = \frac{1}{\rho N}\sum_{r=1}^{N} I(f(C_r^n) \geq \hat{s}_t)C_{r,i}^n, \tag{7}$$

where $\hat{\theta}_i^t$ denotes the probability of concept i in iteration t and $C_{r,i}^n = \{0, 1\}$ denotes the existence of concept i in the rth concept-subset C^n. The parameter $\hat{\theta}_i^t$ directly shows the impact of concept i in video event classification at iteration t. Larger $\hat{\theta}_i^t$

makes the presence of concept i in the optimal solution more likely. The parameter ρN, $\rho \in (0, 1)$ defines the percentage of best performing concepts C^n scoring higher than s taken into account during each iteration.

C. Algorithm

The sample and search strategy for concept-subsets at iteration t has three steps:

(1) **Sampling of concept-subsets** C^n. Based on the current parameter values θ^{t-1}, sample N concept-subsets C^n using $p(.;\theta^{t-1})$ that is:

$$\{C_1^{n(t)}, \ldots, C_N^{n(t)}\} \sim p(.;\theta^{t-1}). \quad (8)$$

(2) **Adaptive updating of score** s^t. At iteration t, evaluate each of sample $C_j^{n(t)}$ using score function $f(.)$ and find the ρN samples $C_j^{n(t)}$ that scored best on the $f(.)$. After having sampled N concept-subsets and sorted them in descending order by performance: $s_{(1)} \geq \ldots \geq s_{(N)}$, the smallest score value is used as the next iterations' reference score s^t, namely $\hat{s}^t = s_{\lfloor \rho N \rfloor}$. All samples $C_j^{n(t)}$ taken into account should perform at least as good as $\hat{s}^t$.

(3) **Adaptive updating of parameter vector** θ^t. Given the ρN good performing samples $C_j^{n(t)}$ found in step 2, the updated parameter set θ^t is estimated as a function of the parameter vectors of these samples using Eq. (7). Informative concept-subsets are best captured by the concepts represented by high value of θ^t.

In the first step, we sample the concept-subsets C^n based on the parameters from iteration $t-1$. The second step aims at keeping at each iteration the top performing concept-subsets C^n, sampled in the first step. Finally, the parameter vector θ^t is updated according to Eq. (7) in the third step, in a way that the distance $D_{CE}(h, p)$ is reduced. Updating parameters θ using Eq. (7) is equivalent to finding the frequency of a concept in the top performing concept-subsets C^n at iteration $t-1$. It means that the probability of those concepts that together improve the event classification accuracy are increased after each iteration. Repeating these three steps for each iteration leads the search towards the conceptlet $\langle C^*, \theta^*, s^* \rangle$ in neighborhood ϵ. The selection process is illustrated in Fig. 2.

In the above analysis, θ plays an important role. Since θ controls the binomial distribution p, the initial values of the parameter vector $\theta^{t=0}$ regulates the size of the conceptlet. Typically, these initial values are set to the same value for all concepts. Lower initial values lead to smaller concept-subsets. Moreover, due to the randomness of sampling, the exact size of the final conceptlet is not known *a priori*. As a result, other than uniformly setting θ to a constant for all concepts, different values can be assigned to favor a certain subset of the concepts. In general, different sources of prior knowledge will lead to different concept-subsets. Thus the initialization of θ will influence both the size and the informativeness of the resulting conceptlet.

For the purpose of event classification, the score function $f(.)$ typically needs labeled training data to quantify the accuracy of various concept-subsets C^n. To do so, we split the training data into a training and validation set. An event classifier is then learned from the conceptlet in the training set and validated on the validation set. We use average precision to reflect the accuracy on the validation set. To find the optimum conceptlet per event, we also change the initialization value of θ by considering different values of n, the size of the concept-subsets. Our supervised selection algorithm for obtaining the conceptlets is summarized in Table I.

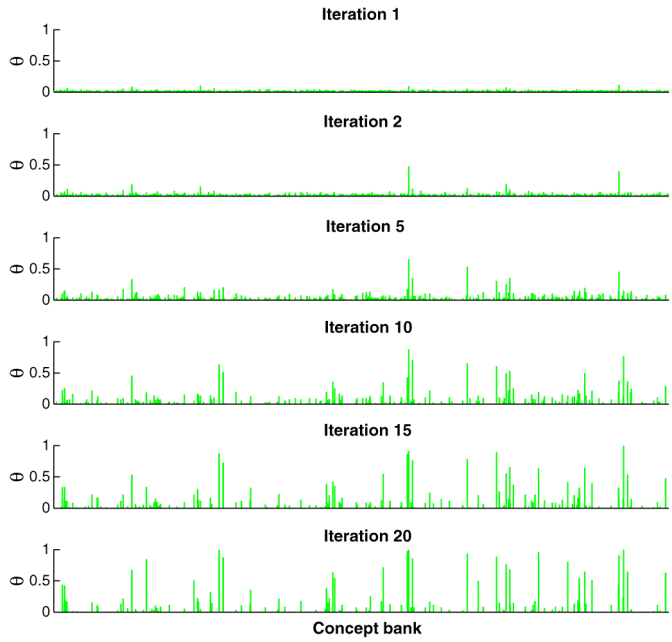

Fig. 2. Concept selection for an event using the values of the parameter vector θ at several iterations. At the beginning θ has a uniform low value, indicating that all concepts have the same low probability to be selected. After a few iterations, some concepts emerge as more probable for selection than others. After 20 iterations spikes are clearly visible, implying that the corresponding concepts are considered in the conceptlet.

IV. Experimental Setup

A. Data Sets

We investigate the effectiveness of conceptlets for video event classification by performing four experiments on three large datasets of challenging real-world web video for event classification: the TRECVID 2010 Multimedia Event Detection dataset [17], the partition of the TRECVID 2012 Multimedia Event Detection dataset [17] used in [32], and the Columbia Consumer Video dataset [18].

TRECVID MED 2010 [17] contains 3,465 internet video clips with over 115 hours of user generated videos content. The dataset contains ground truth for three event categories: *Assembling a shelter, Batting a run* and *Making a cake*. We train and evaluate our event classifiers on the train and test set that consist of 1,723 and 1,742 video examples respectively.

MediaMill MED 2012 [32] is a partition of the TRECVID 2012 Multimedia Event Detection dataset [17] defined by Habibian *et al.* [32]. It consist of 1,500 hours of unconstrained videos provided in MPEG-4 format taken from the web with challenges such as high camera motions, different view points, large intra class variation and poor quality with varying resolution. The dataset comes with ground-truth annotations at video

TABLE I
THE PROPOSED ALGORITHM WHICH MODELS FINDING A CONCEPTLET FOR VIDEO EVENT CLASSIFICATION AS A CROSS-ENTROPY OPTIMIZATION

INPUT: Number of iterations (T), samples (N), size of concept bank (m), percentage of best performing concepts (ρN), index of events ($event$), labeled event examples, k different sizes of concept-subsets (n)=$\{n_1, n_2, ..., n_k\}$ for finding the optimum conceptlet per event
OUTPUT: Conceptlet per event

1. for each $event$
2. $Max = -1$
3. for each $i = 1, \ldots, k$
4. Initialize $\Theta^{(0)}$: $\Theta^{(0)} = n(i)/m$
5. for $t = 1, \ldots, T$
6. **Sampling of concept-subset**: Generate N samples $\{C_1^{n(i)(t)}, ..., C_N^{n(i)(t)}\}$ by using current parameter $\Theta^{(t-1)}$.
7. **Adaptive updating of score** s^t: Find the ρN samples that perform best given the score function *f(.)* and the labeled examples. Sort the samples in descending order by performance: $s_1 \geq \ldots \geq s_{\lfloor \rho N \rfloor}$ and update $s^t = s_{\lfloor \rho N \rfloor}$.
8. **Adaptive updating of parameter vector** $\Theta^{(t)}$: Based on the best concept samples from step 7, update parameter set $\Theta^{(t)}$ by using Eq. 7
9. end
10. $C^* \leftarrow \Theta^{(T)}$
11. if $f(C^*) \geq Max$
12. $Max = f(C^*)$
13. Conceptlet = C^*
14. end
15. end
16. Return Conceptlet
17. end

level for 25 real-world complex events, such as *Attempting a board trick, Flash mob gathering, Town hall meeting*, etc. Following the setup of [32], we extract two partitions consisting of 8,840 and 4,434 videos from the annotated part of the development set. In this paper we use the first partition as the train set, on which we train our event classifiers, and we report all results on the second partition.

Columbia CV [18] consists of 9,317 user-generated YouTube videos with over 210 hours of content. The dataset is annotated with 20 semantic categories, where 15 of them are events, such as *Basketball, Ice skating, Birthday* etc. As we focus exclusively on events, the five object and scene categories in this dataset are excluded from our experiment. We use the split suggested by the authors which consist of 4,625 train videos and 4,637 test videos.

Table II summarizes the statistics of the training and test sets per event for the three video datasets. For a visual impression of characteristic event examples we refer to Fig. 1 showing two examples for the events *Assembling a shelter, Board trick*, and *Birthday* in the TRECVID 2010 MED, the MediaMill 2012 MED, and the Columbia CV dataset.

B. Implementation Details

Concept Bank In the TRECVID 2010 MED dataset we represent each video by a histogram of the output of 280 concept detectors, defined and provided by Merler *et al.* [26]. From the videos one frame is extracted every two seconds and represented as a histogram of 280 concept detectors scores. Then, the histograms are aggregated using average-pooling to arrive at a representation per video. For representing the videos in the MediaMill 2012 MED and Columbia CV datasets we use a concept bank that consists 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 [56] and for 1,000 objects from the ImageNet Large Scale Visual Recognition Challenge 2011 [57]. Although some of the detector names overlap, we prefer to keep all 1,346 as their training data is different. The detectors are trained using a linear SVM atop a standard bag-of-words of densely sampled color SIFT [8] with Fisher vector coding [58] and spatial pyramids [59]. The 1,000 concepts from ImageNet are trained one versus all. The negative examples for each concept from the TRECVID 2012 Semantic Indexing task are the positive examples from other concepts and several examples without label. We compute concept detector scores per video frame, which are extracted once every two seconds. By concatenating and normalizing the detector outputs, each frame is represented 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 [26].

Event classification As we focus on obtaining an informative representation for video event classification, we are for the moment less interested in the accuracy optimizations that may be obtained from various kernel settings [60]–[62]. Hence, we train for each event a one-versus-all linear support vector machine [63] and an approximated histogram intersection kernel map [64]. We find the optimal parameter settings using 5-fold cross-validation.

Cross entropy parameters After initial testing on small partitions of the data, we set the parameters of our algorithm to find the conceptlets for each event as follows: number of iterations $T = 20$, number of concept samples in each iteration $N = 1,000$, and a percentage of best performing concept samples $\rho = 0.1$, leaving 100 best performing concept samples per iteration for updating the sampling parameters. For finding the best conceptlet size per event, we consider various sizes of n, *i.e.,* the size of concept-subsets, during 5-fold cross-validation within the training data only.

TABLE II
NUMBER OF POSITIVE VIDEOS IN THE TRECVID 2010 MED, MEDIAMILL 2012 MED, AND COLUMBIA CV DATASETS USED IN OUR EXPERIMENTS, SPLIT PER EVENT. THE NUMBER OF NEGATIVE VIDEOS FOR EACH EVENT ARE AROUND 1,600, 8,800, AND 4,500, RESPECTIVELY

TRECVID MED 2010			**MediaMill MED 2012**			**Columbia CV**		
Event	Train	Test	*Event*	Train	Test	*Event*	Train	Test
Assembling a shelter	50	48	*Board trick*	98	49	*Basketball*	182	181
Batting a run	52	50	*Feeding animal*	75	48	*Baseball*	150	151
Making a cake	58	48	*Landing fish*	71	36	*Soccer*	161	162
			Wedding ceremony	69	35	*Ice skating*	192	193
			Wood working	79	40	*Skiing*	197	196
			Birthday party	121	61	*Swimming*	199	202
			Changing vehicle tire	75	37	*Biking*	136	137
			Flash mob gathering	115	58	*Graduation*	143	145
			Getting vehicle unstuck	85	43	*Birthday*	158	160
			Grooming animal	91	46	*Wedding reception*	129	130
			Making sandwich	83	42	*Wedding ceremony*	111	110
			Parade	105	50	*Wedding dance*	174	176
			Parkour	75	38	*Music performance*	403	403
			Repairing appliance	85	43	*Non-Music performance*	345	346
			Working on sewing project	86	43	*Parade*	191	194
			Attempting bike trick	43	22			
			Cleaning an appliance	43	22			
			Dog show	43	22			
			Giving directions to location	43	22			
			Marriage proposal	43	22			
			Renovating home	43	22			
			Rock climbing	43	22			
			Town hall meeting	43	22			
			Winning race without vehicle	43	22			
			Working on metal crafts project	43	22			

Evaluation criteria For both the objective function $f(.)$ in our conceptlet algorithm, as well as the final event classification evaluation, we consider as criterion the average precision (AP), which is a well known and popular measure in the video retrieval literature [56]. We also report the average performance over all events as the mean average precision (MAP).

C. Experiments

In order to establish the effectiveness of conceptlets for video event classification, we perform four experiments.

Experiment 1: ***Influence of individual concepts*** To evaluate the maximum effect of individual concept detectors on event classification accuracy, we perform an oracle experiment by simply evaluating each individual concept detector as if it was an event classifier. We evaluate all individual concepts on all events. Then we sort the list of concepts by their classification accuracy for each of the events in the three datasets.

Experiment 2: ***Influence of concept bank size*** To assess the effect of a growing number of concepts in a bank on video event classification performance, we randomly sample a concept-subset from our concept bank. For TRECVID 2010 MED we randomly select concepts from the concept bank with 280 concepts defined by Merler *et al.* [26] with a step size of 10. For both MediaMill 2012 MED and Columbia CV dataset we randomly select concepts from our 1,346 concept bank with a step size of 100. Each video in our dataset is then represented in terms of the detector scores from the concepts in this random subset. To cancel out the accidental effects of randomness, we repeat this procedure 20 times for each subset size.

Experiment 3: ***Conceptletsversus all concepts*** In this experiment we compare our proposed conceptlets to a bank based on all available concept detectors [26], [32]. As the baseline, we represent each video in TRECVID 2010 MED as a 280D vector of detector scores [26] and each video in MediaMill 2012 MED and Columbia CV datasets as a 1,346D vector of detector scores [32] (see Section IV-B). For finding the conceptlet per event, we apply the cross-entropy optimization as described in Section III-C on the training set only. To find the best conceptlet size, we vary parameter n. For events in the TRECVID 2010 MED we consider n in the range [10, 20, …, 100]. In the MediaMill 2012 MED and Columbia CV datasets, we consider values of n in the range [10, 20, …, 100, 200, 300,400,500]. We train an event detector on the found conceptlet and report its performance on the (unseen) test set.

Experiment 4: ***Conceptletsversus other selections*** In this experiment we compare conceptlets obtained with our cross-entropy algorithm to conceptlets obtained from state-of-the-art feature selection algorithms: Minimum Redundancy Maximum Relevancy [46] and L1-Regularized Logistic Regression [48]. To select the concepts per event by Minimum Redundancy Maximum Relevancy, at first we rank all concepts. Then we conduct a 5-fold cross validation on the training set with a varying number of selected concepts ranging from 10 to 1,000 with a step size of 10. In L1-Regularized Logistic Regression, since the regularization parameter controls the sparsity of

TABLE III
EXPERIMENT 1. INFLUENCE OF INDIVIDUAL CONCEPTS ON VIDEO EVENT CLASSIFICATION ACCURACY. WE LIST THE FIVE BEST CONCEPTS FOR THREE EVENTS PER DATASET, TOGETHER WITH THE NUMBER OF POSITIVE TRAINING EXAMPLES USED TO TRAIN THE CONCEPT DETECTORS. NOTE THE SEMANTIC CORRESPONDENCE BETWEEN GOOD PERFORMING CONCEPTS AND EVENTS. CONCEPTS IN ITALICS ARE ALSO AUTOMATICALLY SELECTED BY THE CONCEPTLET ALGORITHM IN EXPERIMENT 3

Assembling a shelter			**Batting a run**			**Making a cake**		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
Snow scene	0.158	1,138	*Baseball cricket*	0.326	1,000	*Cake*	0.141	230
Outdoors	0.121	1,000	*Hockey*	0.214	1,998	*Food*	0.126	794
Mountain scene	0.105	3,972	*Diamond*	0.202	1,000	*Table desk*	0.097	1,000
Forest	0.103	12	*Running*	0.184	896	Building	0.053	2,354
Water scene	0.082	2,746	Suit	0.176	1,000	Room	0.052	2,90
Board trick			**Wedding ceremony**			**Flash mob gathering**		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
Skating	0.194	1,300	*Church*	0.396	1,300	*Crowd*	0.280	2,341
Road	0.171	1,096	*Altar*	0.324	1,300	*3 or more people*	0.214	2,099
Snow	0.162	1,013	*Gown*	0.306	1300	*People marching*	0.205	624
Snowplow	0.123	540	*Groom*	0.288	1,280	Street battle	0.202	1,300
Ski	0.119	1,096	*Suit*	0.251	1,300	Meeting	0.186	340
Basketball			**Swimming**			**Parade**		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
Basketball	0.488	1,300	*Swimming*	0.698	1,300	*People marching*	0.318	624
Throw ball	0.485	811	*Swimming pool*	0.621	1,300	*Urban scenes*	0.155	1,403
Throwing	0.432	1,300	*Underwater*	0.432	1,300	*Police van*	0.150	1,300
Indoor sport venue	0.355	1,300	Stingray	0.227	1,300	*3 or more people*	0.138	2,099
Gym	0.337	153	*Waterscape/Waterfront*	0.211	604	*Streets*	0.135	1,300

concepts, we conduct a 5-fold cross validations on the training set by varying this parameter from 1 to 100 with step 5 to select the concepts per event. For both feature selections we train an event classifier with a linear SVM on the selected concepts and report its performance on the (unseen) test set.

V. RESULTS

A. Influence of Individual Concepts

We show the results of experiment 1 in Table III. We observe that the best detectors per event also make sense, most of the time. When we consider the event *Wedding ceremony*, for example, the best possible concepts are 'Church', 'Altar', 'Gown', 'Groom' and 'Suit'. For the event *Makinga cake*, concepts like 'Cake', 'Food', 'Table desk', 'Building' and 'Room' are the oracle choice. However, for the event *Battinga run* we find an irrelevant concept in the top of the concept ranking: 'Hockey'. We explain this by the fact that 'Hockey' shares many low-level visual characteristics with *Baseball e.g.,* both sports are played on a green field. It is also interesting to note that some of the relevant concept detectors obtain the good event classification accuracy by having only a few positive training examples, consider for example 'Forest' for the event *Assembling a shelter*, which has only 12 positive examples. This result shows that there are individual concepts that are more discriminative and descriptive than others for representing events in internet video.

B. Influence of Concept Bank Size

We plot the results of experiment 2 on the three datasets in Fig. 3. As expected the event classification accuracy increases when more and more concept detectors are part of the bank.

For the TRECVID 2010 MED dataset (Fig. 3(a)), the increase in event classification accuracy is close to linear up to approximately 40 (random) concept detectors, afterwards it saturates to the end value of 0.361 MAP when using all 280 available concept detectors. Interestingly, the plot reveals that there exist an outlier concept-subset, containing only 70 concepts, which performs better than using all 280 concepts (compare the MAP of 0.389 with the maximum MAP of 0.361 when using all concepts). This result shows that some concept-subsets are more informative than others for video event classification. The results on the other two datasets confirm this conclusion. For the MediaMill 2012 MED dataset (Fig. 3(c)), there is an outlier concept-subset, containing only 800 concepts, which performs better than using all 1,346 concepts (compare the MAP of 0.312 with the maximum MAP of 0.292 when using all concepts). Also for the Columbia CV dataset (Fig. 3(e)), we find that there is an outlier concept-subset, containing only 600 concepts, which performs better than using all 1,346 concepts (compare the MAP of 0.531 with the maximum MAP of 0.507 when using all concepts). These results indicate much is to be expected from *a priori* search for the conceptlet of an event.

When we zoom in on individual events the connection between concept-subsets and event definitions can be studied. We inspect the box plot also for all the individual events of the

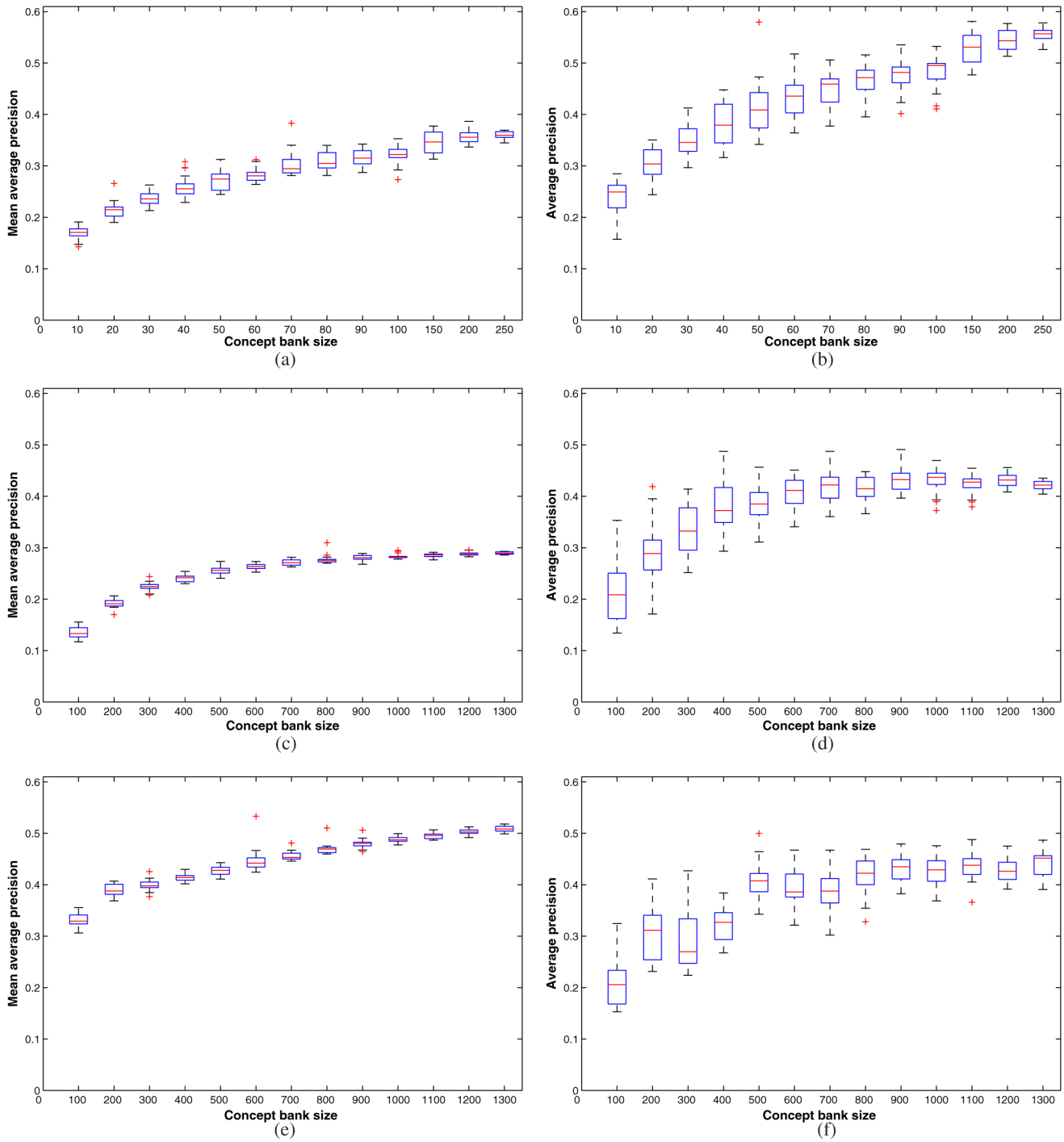

Fig. 3. Experiment 2. Influence of concept bank size in (a) TRECVID 2010 MED, (c) MediaMill 2012 MED, and (e) Columbia CV: Event classification accuracy increases with the number of concepts in the bank, but the variance suggests that some concept-subsets are more informative than others. (d), (b) and (f) Influence of concept bank size for the particular events: *Landing a fish*, *Batting a run*, and *Wedding ceremony*. For these events a small subset outperforms the bank using all available concepts. Indicating that much is to be expected from a priori search for the most informative conceptlet for an event.

three datasets (data not shown). The plots reveal several positive outliers using just a small number of concepts in the subset. Figs. 3(b), (d), (f) detail the box plot for the specific events *Battinga run, Landing a fish*, and *Wedding ceremony*. For event *Battinga run* (Fig. 3(b)) we perceive an outlier subset with an AP of 0.590 containing only 50 randomly selected concepts (compare to the maximum of 0.553 when using all 280 concepts). For event *Landinga fish* (Fig. 3(d)) the box plot reveals that there exist a subset, containing only 400 concepts, which performs better than using all 1,346 concepts (compare the top of the whisker at 400 concepts, with an MAP of 0.489 with the maximum MAP of 0.433 when using all concepts). Also for the event *Wedding ceremony* (Fig. 3(f)) we observe an outlier subset with an AP of 0.500 containing only 500 randomly selected concepts (compare to the maximum of 0.473 when using all 1,346 concepts). The results of experiment 2 on three

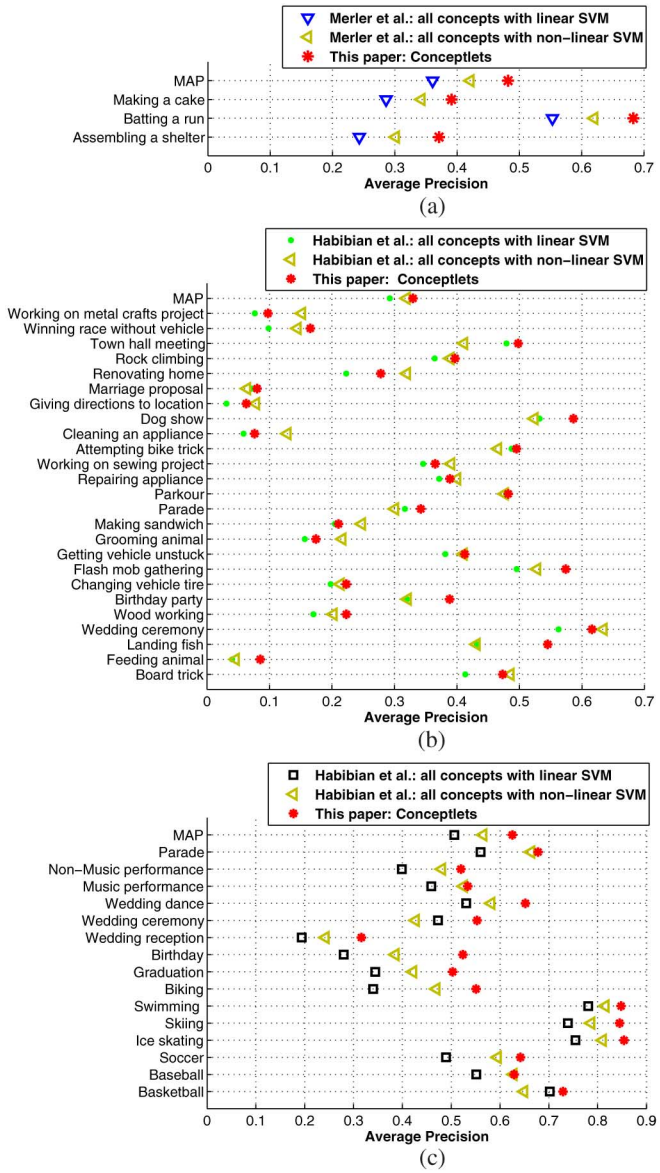

Fig. 4. Experiment 3. Conceptlets versus all concepts. A conceptlet outperforms a bank containing all available concept detectors for the large majority of event categories when using either a linear or a non-linear SVM for video event classification. (a) *TRECVID 2010 MED*. (b) *MediaMill 2012 MED*. (c) *Columbia CV*.

datasets with two different concept banks show that, in general, the event classification accuracy increases with the number of semantic concepts in the bank. However, it also shows that some concept-subsets are more informative than others for specific events, and this may result in improved event classification accuracy.

C. Conceptlets Versus All Concepts

We plot the result of experiment 3 in Fig. 4. For the large majority of event categories, conceptlets with selective semantics are better than using all available concepts.

On the TRECVID 2010 MED dataset (Fig. 4(a)), we achieve a 0.483 MAP in event classification by conceptlets, where the result is 0.361/0.421 MAP when using all 280 concepts [26]. Conceptlets obtain a relative improvement of 34.0% over the linear SVM and 14.7% over the non-linear SVM with only 83 concepts per event on average. We observe a considerable improvement for all three events using only a fraction of the available concept detectors (90, 70, and 90). Fig. 5(a), (b) shows the conceptlets for the events *Batting a run* and *Making a cake*. Our algorithm selects concepts such as 'Baseball', 'Cricket', 'Field', 'Running', and 'Sport' that make sense for the event *Batting a run*, without being programmed to do so. However, the conceptlet also contains some irrelevant semantic concepts, such as 'Hockey' and 'Soccer', which share several visual characteristics to the event (see Fig. 6). Similar conclusions hold for the event *Makinga cake*.

On the MediaMill 2012 MED dataset (Fig. 4(b)), our conceptlets reach to a 0.329 MAP in event classification, where using all 1,346 concepts results in 0.292/0.317 [32]. A relative improvement of 13.0% for the linear SVM and 3.7% for the non-linear SVM using about 245 concepts on average per event. Conceptlets obtain a considerable improvement for events such as *Landing fish, Dog show* and *Flash mob gathering* using only 300, 200, and 40 of the concept detectors available. When relevant concepts are unavailable in the concept bank we started with, the results will not improve much, as can be seen for the events *Attempting bike trick, Marriage proposal*, and *Making sandwich*, but often better than using all. Fig. 5(c), (d) shows the conceptlet for *Landing a fish* and *Flash mob gathering*. The conceptlet for the event *Landing a fish* consist of general concepts such as 'Adult male human', 'Hand', '3-or-more-people', 'Sea-Mammal' and event-specific concepts such as 'Hook' and 'Reel'. The conceptlet for *Flash mob gathering* shows several concepts that seem semantically relevant as well, such as 'Walking-running', 'Crowd', 'Daytime-outdoor'. However, we also observe some concepts whose semantic connection is less apparent, such as 'Water-bottle' and 'Ground-combat'. Note that the concepts are selected automatically from provided event examples only.

On the Columbia CV dataset (Fig. 4(c)), we observe that conceptlets obtain 0.625 MAP, where using all 1,346 concepts results in 0.507/0.565 MAP. Conceptlets are always better and obtain a relative improvement of 23.2%/10.6% with only 93 concepts per event on average. Conceptlets obtain a considerable relative improvement for events such as *Soccer, Biking* and *Graduation* using only 50, 50, and 200 of the available concept detectors. Interestingly, for the event *Birthday* the improvement compared to the linear SVM is as much as 87.2% (0.524 MAP against 0.280 MAP) using only 30 concepts. Fig. 5(e), (f) highlights the conceptlets for *Biking* and *Birthday*. We observe that most of the selected concepts for event *Biking* in Fig. 5(e), such as 'Bicycling', 'Bicycle', 'Daytime outdoor', 'Road', 'Legs', are semantically relevant to this event. In Fig. 5(f), we show the conceptlet for *Birthday*. Beside semantically relevant concepts such as 'Candle' we observe several semantically irrelevant concepts such as 'Abacus'. When we inspect the ImageNet images used for training of this concept detector in Fig. 6, we observe that the color beads of the abacus are visually similar to typical *Birthday* objects such as candles and balloons. When the quality of concept detectors further improves, we expect better selection in the conceptlets.

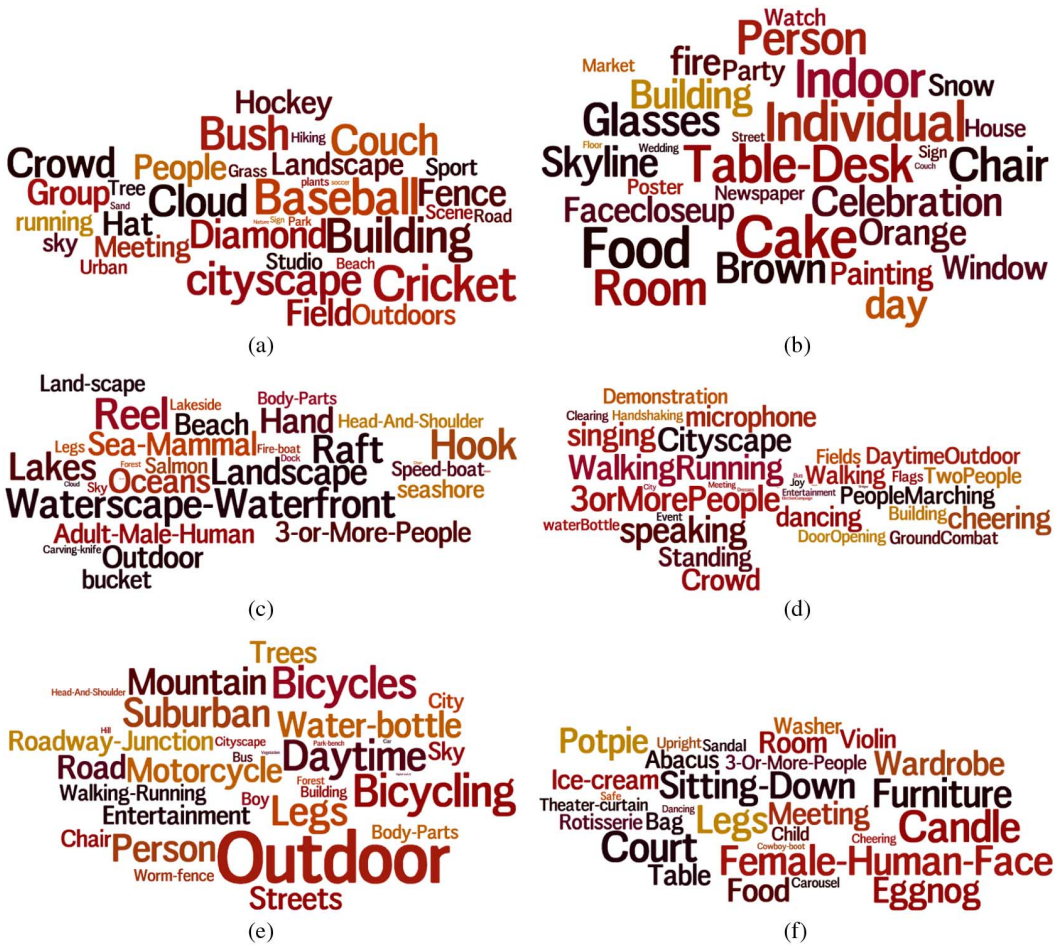

Fig. 5. Conceptlets for various events as automatically selected from event video examples by our algorithm. Font size correlates with automatically estimated informativeness. Note that the algorithm finds concepts that make sense, most of the time, without being programmed to do so. (a) *Batting a run*. (b) *Making a cake*. (c) *Landing a fish*. (d) *Flash mob gathering*. (e) *Biking*. (f) *Birthday*.

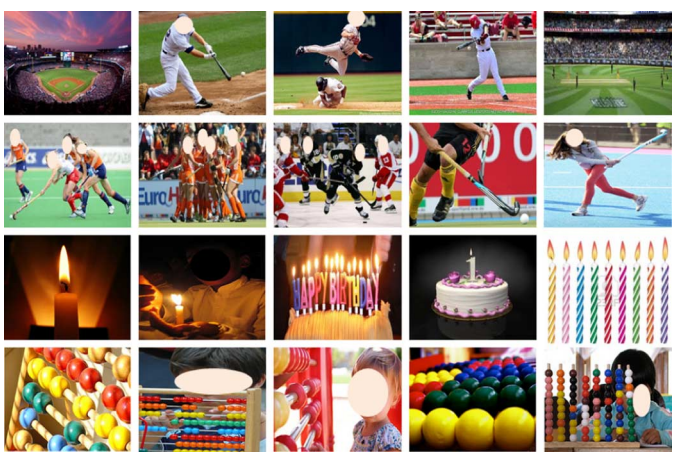

Fig. 6. Example images used for training the concept detectors: 'Baseball', 'Hockey', 'Candle' and 'Abacus' (top to bottom). The visual similarity between 'Baseball' and 'Hockey' causes that our algorithm mistakenly selects 'Hockey' into the conceptlet for the event *Battingin run*. Likewise, 'Abacus' results in a high probability in the videos containing *Birthday* events since the color beads of the abacus are visually similar to typical *Birthday* objects such as candles and balloons.

To explore the correlation between the selected concepts and the source of these concepts when using both the TRECVID and ImageNet annotations, we plot the fraction of selected concepts by their training source for all 25 events of MediaMill MED 2012 and the 15 events of Columbia CV in Fig. 7. As can be observed, conceptlets automatically select the most informative concepts independent of their training source.

Since conceptlets need event video training examples we also investigated how many event videos are sufficient for conceptlet discovery. On the TRECVID MED 2010 dataset we train event classifiers using 10, 20, 30, and 40 positive examples, and repeat the process of selecting examples 10 times. We compare event classifiers using all concepts with a linear SVM [26] versus our conceptlets with a linear SVM. The result in Fig. 8 shows that training size affects the performance, more examples are beneficial for both the baseline and conceplets, but conceptlets are always better, even when examples are scarce.

For future reference we also evaluate our conceptlets on the TRECVID 2013 MED test and Kindred sets [17]. While low-level feature representations are known to be more accurate on these benchmarks [17], we focus in our comparison on semantic representations and consider the state-of-the-art approach by Habibian *et al.* [32]. We observe similar behavior as before. Event classification using all concepts with a linear SVM (0.270, 0.289) is outperformed by all concepts with a non-linear SVM (0.296, 0.309), but conceptlets are best (0.309, 0.325).

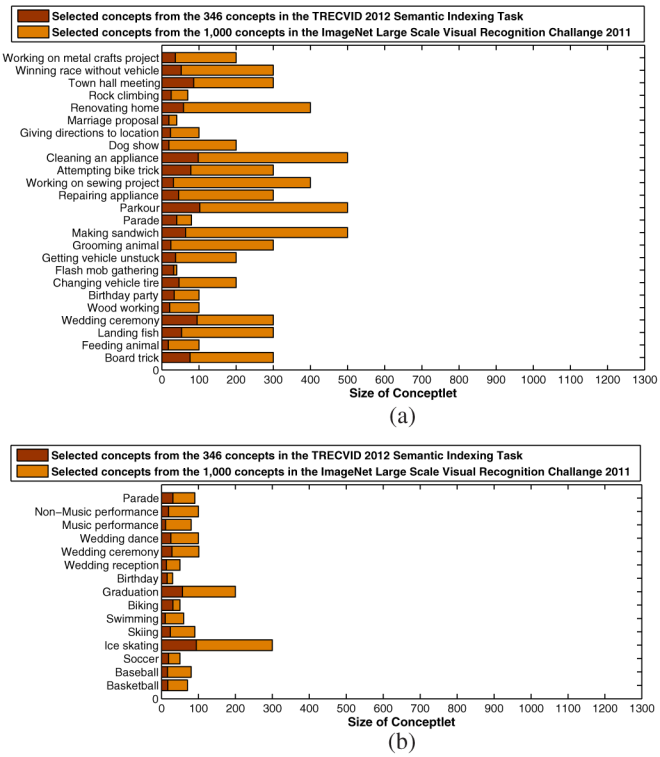

Fig. 7. Correlation between the selected concepts and their training source per event in (a) MediaMill 2012 MED and (b) Columbia CV. Conceptlets automatically select the most informative concepts independent of their training source. (a) *MediaMill 2012 MED*. (b) *Columbia CV*.

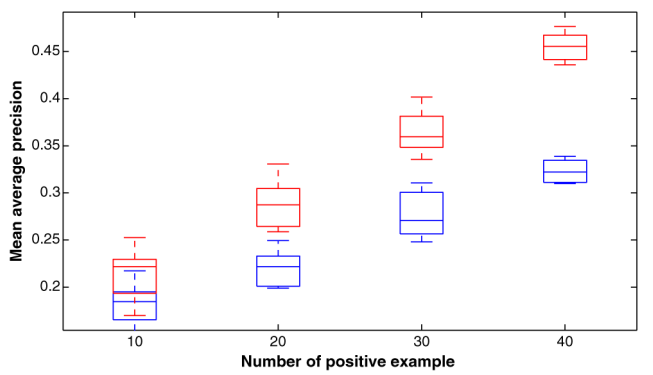

Fig. 8. Effect of number of positive event examples on classification with conceptlets (red) or an all-concept baseline (blue). Conceptlets are always better, even when examples are scarce.

The results of experiment 3 affirm that event classification using conceptlets outperforms a bank using all concepts and always contains significantly less semantic concepts, which often appear descriptive.

D. Conceptlets Versus Other Selections

We plot the result of experiment 4 in Fig. 9. On all three datasets conceptlets outperform selections based on Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression.

On the TRECVID 2010 MED dataset (Fig. 9(b)), conceptlets score 0.483 MAP where Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression score 0.406

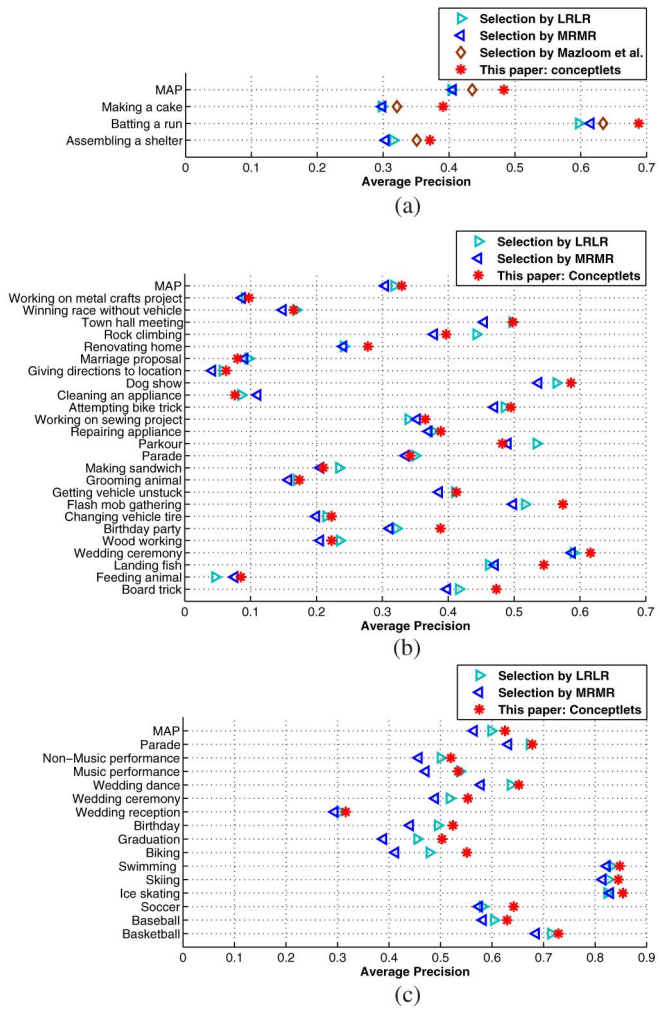

Fig. 9. Experiment 4. Conceptlets versus other selections. For video event classification, conceptlets selected with our proposed cross-entropy algorithm outperform alternative concept selection algorithms. (a) *TRECVID 2010 MED*. (b) *MediaMill 2012 MED*. (c) *Columbia CV*.

and 0.403. Where our algorithm selects 83, the others select 120 and 165. On this dataset we also compare with the initial version this work [51], with a MAP of 0.443. We improve our previous work by 10.0%, which we attribute to the fact that we optimize the conceptlet size, where we previously relied on fixed sizes only.

On the MediaMill 2012 MED dataset Minimum Redundancy Maximum Relevancy scores 0.304 MAP and L1-Regularized Logistic Regression scores 0.318 MAP, where conceptlets obtain 0.329 MAP (Fig. 9(b)). Again we reach better numbers with fewer selected concepts. Where conceptlets need on average 245 concepts to reach this results, the others need many more (458 vs 374). For the event *Flash mob gathering*, a conceptlet with 40 concepts obtains 0.574 AP where Minimum Redundancy Maximum Relevancy reaches 0.499 AP with 600 concepts and L1-Regularized Logistic Regression scores 0.516 AP with 298 concepts. The results on the Columbia CV dataset in Fig. 9(c) show similar behavior. Conceptlets obtain 0.625 MAP using on average 93 concepts per event where the others perform worse with more concepts selected (0.565 MAP/452 concepts vs 0.598 MAP/440 concepts).

TABLE IV
PROPERTIES OF CONCEPTLETS AS A FUNCTION OF BANK SIZES CONSIDERED DURING CROSS-ENTROPY OPTIMIZATION. WE CONSIDER SIZES IN THE RANGE 10 TO 90, WITH A STEP SIZE OF 20. ALL RESULTS AVERAGED OVER THREE EVENTS FROM THE TRECVID 2010 MED DATASET USING THE CONCEPT BANK OF MERLER *ET AL.* [26]. IN ADDITION TO L1-REGULARIZED LOGISTIC REGRESSION (LRLR) AND MINIMUM REDUNDANCY MAXIMUM RELEVANCY (MRMR) AND THE INITIAL VERSION OF THIS WORK [51] WE ALSO REPORT FOUR OTHER RECENT FEATURE SELECTIONS. CONCEPTLETS ARE COMPUTATIONALLY MORE DEMANDING THAN THE OTHER SELECTIONS, BUT ALWAYS RESULT IN THE BEST EVENT CLASSIFICATION ACCURACY

	Conceptlet					LRLR [48]	MRMR [46]	CIFE [65]	SFFS [66]	RELIEF [67]	DISR [68]	[51]
	[10]	[10,30]	[10,50]	[10,70]	[10,90]	–	–	–	–	–	–	–
Size	10	30	50	70	83	165	120	120	70	50	100	100
Time(m)	5	17	32	52	76	10	6	18	38	29	10	19
MAP	0.306	0.359	0.403	0.447	0.482	0.403	0.406	0.392	0.396	0.395	0.395	0.443

When we inspect the selected concepts for the event *Landing fish* for Minimum Redundancy Maximum Relevancy (data not shown) we find several redundant concepts such as 'Adult-Female-Human', 'Female-Human-Face-Closeup', 'Female-Person', 'Single-Person-Female', and 'Two-People' which have a negative effect on the event classification result. Moreover, the redundancy makes the description less precise. Recall that Minimum Redundancy Maximum Relevancy ranks the concepts first based on the ratio of the relevance of the concept to the redundancy of the concepts in the set. Making this selection sensitive to the accuracy of the concept detectors. Moreover, some low ranked concept detectors may have poor average precision in isolation, but when combined with other concepts lead to a better event classification. Our algorithm is less sensitive to the performance of concept detectors. If the presence of a concept, either an accurate or inaccurate one, improves the accuracy of the event classifier it will try to maintain it in the conceptlet. Selection by L1-Regularized Logistic Regression is more competitive, but the selection always contains more concepts than our conceptlets. In addition to being less accurate than conceptlets, selection by L1-Regularized Logistic Regression is less descriptive.

In Table IV we report the run time and mean average precision of conceptlets as a function of the number of bank sizes considered. We consider sizes in the range 10 to 100, with a step size of 10. In addition to L1-Regularized Logistic Regression, Minimum Redundancy Maximum Relevancy and the initial version of this work [51] we also report four other recent feature selections. As expected, the best result is obtained when we consider all sizes in the range, and the worst result is obtained when we only consider banks of 10 concepts. When we consider all ranges, without any preselection, conceptlets need 90 minutes per event on average. Using a non-linear instead of a linear SVM with conceptlets increases the MAP from 0.482 to 0.513, but needs 425 minutes per event on average. This is longer than the other selections, but can be sped up by parallelization when execution time is an issue. Conceptlets always result in better mean average precision.

We conclude that conceptlets, at the expense of a longer run time, are more effective than Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression by considering more discriminant and less redundant concepts.

VI. CONCLUSION

We study event classification based on banks of concept detectors. Different from existing work, which simply includes in the bank all available detectors, we propose an algorithm that learns to find from examples the most informative concept-subset per event, which we call the conceptlet. We formulate finding the conceptlet as an importance sampling problem which can be solved with a near-optimal cross-entropy optimization. We study the behavior of conceptlets by performing four experiments on three unconstrained web video collection from the 2010 and 2012 TRECVID Multimedia Event Detection task and the Columbia Consumer Video datasets using a total of 1,346 pre-trained concept detectors.

The results of experiment 1 show that there are individual concepts that are more discriminative and descriptive than others for representing events in internet video. The results of experiment 2 give 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 classification accuracy increases with the number of concept detectors in the bank. However, we show that some concept-subsets are more informative than others for specific events, and this may result in improved event classification accuracy. The results of experiment 3 confirm that event classification using conceptlet outperform banks using all concepts, and always contains significantly less concept detectors. Finally, results of experiment 4 reveal that our conceptlets, at the expense of a longer run time, are more effective than selected concepts by state-of-the-art feature selection algorithms, by considering more discriminant and less redundant concepts. What is more, the conceptlets make sense for the events of interest, without being programmed to do so.

Further improvements of conceptlets with respect to the reduction of training examples, run time efficiency, classification accuracy and event descriptiveness can be envisioned, for example by the use of semantic query-by-video solutions [69]. We plan to elaborate on this direction in future work. For the moment we conclude that selective use of semantic concepts, by means of conceptlets, is beneficial for classifying video events and opens up the possibility to automatically describe and explain why a particular video was found.

ACKNOWLEDGMENT

The authors would like to thank D. Koelma and K. E. A. van de Sande for providing concept detectors.

REFERENCES

[1] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in videos," *IEEE Trans. Syst., Man, Cybern.*, vol. 39, no. 5, pp. 489–504, May 2009.

[2] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Event detection and recognition for semantic annotation of video," *Multimedia Tools Applic.*, vol. 51, no. 1, pp. 279–302, 2011.

[3] L. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia streams," *Proc. IEEE*, vol. 96, no. 4, pp. 623–647, Apr. 2008.

[4] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimedia Inf. Retrieval*, vol. 2, no. 2, pp. 73–101, 2013.

[5] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.

[6] N. Haering, R. Qian, and I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 857–868, Jun. 2000.

[7] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 42–53, Jan. 2010.

[8] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[9] 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 *Proc. NIST TRECVID Workshop*, 2010.

[10] 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 *Proc. CVPR*, 2012, pp. 1298–1305.

[11] 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 *Proc. CVPR*, 2012, pp. 3681–3688.

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

[13] Y.-G. Jiang, "Super: Towards real-time event recognition in internet videos," in *Proc. ICMR*, 2012.

[14] G. K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. E. A. van de Sande, A. W. M. Smeulders, and C. G. M. Snoek, "Evaluating multimedia features and fusion for example-based event detection," *Mach. Vis. Applic.*, vol. 25, no. 1, 2014.

[15] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. G. A. Perera, M. Pandey, and J. I. Corso, "Multimedia event detection with multimodal feature fusion and temporal concept localization," *Mach. Vis. Applic.*, vol. 25, no. 1, 2014.

[16] D. Oneata, M. Douze, J. Revaud, S. Jochen, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. Mcguiness, S. Chen, N. O'Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, F. Basura, and T. Tuytelaars, "AXES at TRECVid 2012: KIS, INS, and MED," in *Proc. NIST TRECVID Workshop*, 2012.

[17] NIST TRECVID Multimedia Event Detection (MED) Evaluation Track [Online]. Available: http://www.nist.gov/itl/iad/mig/med.cfm

[18] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ICMR*, 2011.

[19] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.

[20] 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 *Proc. NIPS*, 2010, pp. 1378–1386.

[21] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. ECCV*, 2010, pp. 776–789.

[22] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. CVPR*, 2012, pp. 1234–1241.

[23] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *FnTIR*, vol. 2, no. 4, pp. 215–322, 2009.

[24] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

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

[26] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Jan. 2012.

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

[28] Z. Ma, Y. Yang, Z. Xu, N. Sebe, S. Yan, and A. Hauptmann, "Complex event detection via multi-source video attributes," in *Proc. CVPR*, 2013, pp. 2627–2633.

[29] E. Younessian, T. Mitamura, and A. G. Hauptmann, "Multimodal knowledge-based analysis in multimedia event detection," in *Proc. ICMR*, 2012.

[30] Z. Ma, "From concepts to events: A progressive process for multimedia content analysis," Ph.D. dissertation, Inf. Commun. Technol. Sch., Univ. Trento, Trento, Italy, 2013.

[31] 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 *Proc. Workshop Applic. Comput. Vis.*, 2013, pp. 339–346.

[32] A. Habibian and C. G. M. Snoek, "Recommendations for recognizing video events by concept vocabularies," *CVIU*, vol. 124, 2014.

[33] 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 Trans. Multimedia*, vol. 13, no. 3, pp. 86–91, May 2006.

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

[35] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proc. ACM Multimedia*, 2007, pp. 991–1000.

[36] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.

[37] B. Huurnink, K. Hofmann, and M. de Rijke, "Assessing concept selection for video retrieval," in *Proc. MIR*, 2008.

[38] S. Rudinac, M. Larson, and A. Hanjalic, "Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval," in *Proc. Int. J. Multimedia Inf. Retrieval*, 2012, pp. 263–280.

[39] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Proc. CIVR*, 2006.

[40] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Proc. CIVR*, 2007.

[41] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1085–1096, Oct. 2008.

[42] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[43] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[44] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1182, 2003.

[45] M. H. Lloyd, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proc. 12th Int. Florida Artif. Intell. Res. Soc.*, 1999, pp. 235–239.

[46] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Dec. 2005.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, vol. 58, pp. 267–288, 1994.

[48] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proc. ICML*, 2004.

[49] I. Inza, P. Larranaga, and B. S. R. Etxeberria, "Feature subset selection by Bayesian network-based optimization," *Artif. Intell.*, vol. 123, no. 1, pp. 157–184, 2000.

[50] W. Siedlecki and J. Sklansky, "On automatic feature selection," in *Proc. Int. J. Pattern. Recognit.*, 1988.

[51] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. G. M. Snoek, "Searching informative concept banks for video event detection," in *Proc. ICMR*, 2013.

[52] J. A. Bucklew, *Introduction to Rare Event Simulation*. Berlin, Germany: Springer, 2004.

[53] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Berlin, Germany: Springer, 2004.
[54] S. Mannor, D. Peleg, and R. Rubinstein, "The cross entropy method for classification," in *Proc. ICML*, 2005.
[55] X. Li, E. Gavves, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Personalizing automated image annotation using cross-entropy," in *Proc. ACM Multimedia*, 2011, pp. 233–242.
[56] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. MIR*, 2006.
[57] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li, "ImageNet large scale visual recognition challenge 2011," [Online]. Available: http://www.image-net.org/challenges/LSVRC/2011
[58] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek, "Image classification with the fisher vector: Theory and practice," in *Proc. IJCV*, 2013, pp. 222–245.
[59] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.
[60] D. Xu and S.-F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, Nov. 2008.
[61] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.
[62] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, "Video event classification using string kernels," *Multimedia Toops Applic.*, vol. 48, no. 1, pp. 69–87, 2010.
[63] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Math. Program.*, vol. 127, no. 1, 2011.
[64] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.
[65] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *Proc. ECCV*, 2006, pp. 68–82.
[66] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Proc. AISTATS*, 1995, pp. 199–206.
[67] , H. Liu and H. Motoda, Eds., *Computational Methods of Feature Selection*. London, U.K.: Chapman & Hall, 2008.
[68] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topcs Signal Process.*, vol. 2, no. 3, pp. 261–274, Mar. 2008.
[69] M. Mazloom, A. Habibian, and C. G. M. Snoek, "Querying for video events by semantic signatures from few examples," *MM*, 2013.

Masoud Mazloom received the B.Sc. degree in computer engineering from Azad University, Tehran-South Campus in 2002, and the M.Sc. degree in computer science from Sharif University of Technology, Iran, in 2005. He is currently working toward the Ph.D. degree in computer science at the University of Amsterdam, The Netherlands.

He was a Lecturer with the Department of Computer Engineering in Shahid Chamran University, Ahvaz, Iran (2006–2010). He was a Visiting Scientist with DVMM lab in Columbia University (summer 2014). His research interest is video and image retrieval, analysis and understanding, covering computer vision, and pattern recognition.

Efstratios Gavves received the B.Sc./M.Sc. Diploma in electrical and computer engineering from the Aristotle University of Thessaloniki in 2007, and the Ph.D. degree from the University of Amsterdam, The Netherlands.

After working as a Research Assistant with the Informatics and Telematics Institute in Thessaloniki, he joined the Intelligent Systems Lab Amsterdam, University of Amsterdam, The Netherlands. Currently he is a Post-doctoral Researcher in the VISICS lab of the KU Leuven, Belgium. His research interests focus on computer vision and machine learning.

Cees G. M. Snoek received the M.Sc. degree in business information systems and Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2000 and 2005, respectively.

He is currently an Associate Professor in the Intelligent Systems Lab at the University of Amsterdam. He was previously at Carnegie Mellon University, USA (2003) and UC Berkeley (2010-2011). His research interests focus on video and image retrieval.

Dr. Snoek is the lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He has published over 150 refereed book chapters, journal and conference papers. He is co-initiator and co-organizer of the VideOlympics and general co-chair of ACM Multimedia 2016 in Amsterdam. He is a senior member of ACM and IEEE. Dr. Snoek is member of the editorial boards for IEEE MultiMedia and IEEE TRANSACTIONS ON MULTIMEDIA Cees is recipient of an NWO Veni award (2008), a Fulbright Junior Scholarship (2010), an NWO Vidi award (2012), and the Netherlands Prize for ICT Research (2012). Several of his Ph.D. students have won best paper awards, including the IEEE TRANSACTIONS ON MULTIMEDIA PRIZE PAPER AWARD (2012) and the SIGMM Award for Outstanding Ph.D. Thesis in Multimedia Computing, Communications and Applications (2013).