Encoding Concept Prototypes for Video Event Detection and Summarization

Masoud Mazloom[†], Amirhossein Habibian[†], Dong Liu[⋆], Cees G.M. Snoek[†‡], Shih-Fu Chang[⋆]
[†]University of Amsterdam [‡]Qualcomm Research Netherlands [⋆]Columbia University

ABSTRACT

This paper proposes a new semantic video representation for few and zero example event detection and unsupervised video event summarization. Different from existing works, which obtain a semantic representation by training concepts over images or entire video clips, we propose an algorithm that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations, and use them for representing an event video. We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames from both positive and negative training videos of a target concept. We study the behavior of our video event representation based on concept prototypes by performing three experiments on challenging web videos from the TRECVID 2013 multimedia event detection task and the MED-summaries dataset. Our experiments establish that i) Event detection accuracy increases when mapping each video into concept prototype space. ii) Zero-example event detection increases by analyzing each frame of a video individually in concept prototype space, rather than considering the holistic videos. iii) Unsupervised video event summarization using concept prototypes is more accurate than using video-level concept detectors.

Categories and Subject Descriptors

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

Keywords

concept prototype, video event detection and summarization

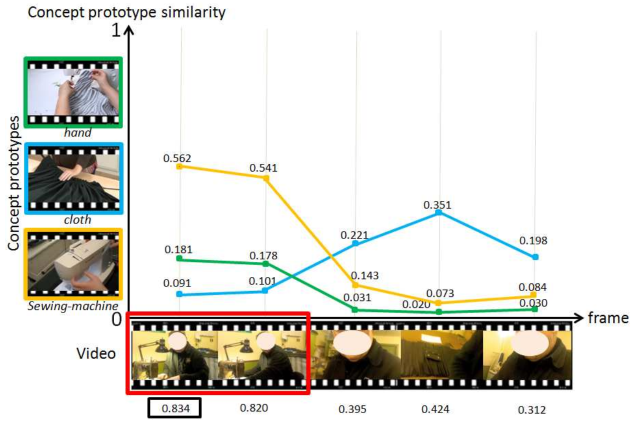


Figure 1: We propose an algorithm that learns from web data a set of relevant frames to the concept, which we call concept prototypes (y-axis). We encode videos as concept prototypes (x-axis) and show that for unseen videos it is able to detect events from few or zero examples. In addition, we demonstrate how the new representation can be exploited for unsupervised video event summarization (red border).

1. INTRODUCTION

The goal of this paper is to detect an event from very few examples and to summarize a video containing the event in its most descriptive frames. The common tactic for few- and zero-example event detection is to rely on a video representation consisting of concept detector scores indicating presence of concepts like *cake, rock* or *tire* in the video [4, 5, 7, 10, 11, 17, 19, 22, 25]. Such a representation is known to outperform the traditional low-level representation when examples are scarce [3, 9, 18, 23]. In [4, 8, 12, 24], it is further shown that semantic representations are capable of event detection without examples, from just a textual description of an event (and the corresponding event-to-concept mapping). We observe that concept detectors are included in the representation whenever they are available and are learned from either image- or video-level annotations. In this paper we propose a new video representation for few- and zero-example event detection, which also uses concept detectors, but we place special emphasis on what concepts to include and from what video-frames they are learned.

In [25] Merler *et al.* used a mixture of 280 relevant and irrelevant concepts to the events which are trained from thousands of labeled web images. In [10] Habibian *et al.* used a collection of concepts consisting of 1,346 relevant and irrelevant concepts to the event trained from ImageNet [1] and TRECVID [27]. In [22] Mazloom *et al.* proposed a learn-

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749402.

ing algorithm to find conceptlets, a set of concept detectors deemed relevant to an event, from the same 1,346 concepts used in [10]. In [4] Chen *et al.* use a collection of 2,000 concepts trained on Flickr images, deemed relevant to an event from analyzing the textual description. We note that in all these works [4,10,22,25] concept detectors are trained from a collection of images, which have the appealing ability that they can be applied on each frame of a video. However the source of training data used in these methods is images from Flickr and ImageNet, applying them on frames extracted from web video results in an unreliable video representation due to the domain mismatch between images and videos. To counter the domain mismatch, recently several works propose to use web video for training concept detectors [2,8,9,19,31], which is not only much more efficient, but also suffers less from the domain mismatch problem. However, the drawback of these works is that they operate on a video-level, although concepts and events may be visible in a part of the video only. Hence, they include many irrelevant frames in their pooled representation leading to a sub-optimal representation.

In [14–16] methods are proposed that find relevant video parts by formulating the problem of video event detection as multiple instance learning. These works show the efficiency of training an event model only on relevant parts of video. Inspired by [14–16], in this paper we attempt to find an informative representation per concept by considering only the frames from web-video that are relevant to the concept, without using frame-level annotations. We call the informative subset of frames per concept the concept prototypes and use them to represent video for few- and zero-example event detection.

In [6], Ding *et al.* propose a method to generate a text which summarizes the important information in a video of an event. The algorithm selects relevant concepts to the event using lots of video examples. We also aim for video event summarization, but we use concept prototypes by selecting the relevant frames of a video to the event without relying on any video examples. In [28], Potapov *et al.* proposed video event summarization by learning from lots of positive and negative video examples. Each event model is applied on various segments of a video, the segments with the highest scores are returned as summary. Since they train a model over the video-level representation of videos in the training set, including all irrelevant parts, applying their model on small video segments are unlikely to reflect an accurate score. Since our prototypes are a frame-level representation of concepts deemed relevant to the event, we have the ability of applying them on each video frame and computing a reliable score per frame without the need for any event model. We simply consider the highest scoring frames as the video summary, see Figure 1. To the best of our knowledge no method currently exists in the literature for summarizing a video event using only a textual event definition.

We make the following contributions:

- We model selecting the concept prototypes out of a large set of frames of positive videos per concept, as seeking the frames closest to the densest region in the feature space of positive and negative videos.
- We introduce concept prototypes as a new semantic video representation for few example event detection.
- We show the effectiveness of the video representation by concept prototypes in few- and zero-example event detection as well as unsupervised event summarization.

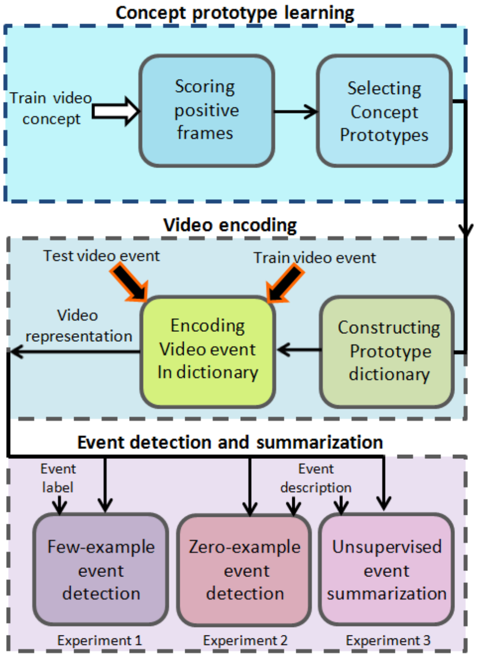

Concept prototype learning
Train video concept
Scoring positive frames
Selecting Concept Prototypes
Video encoding
Test video event
Train video event
Video representation
Encoding Video event In dictionary
Constructing Prototype dictionary
Event detection and summarization
Event label
Event description
Few-example event detection
Zero-example event detection
Unsupervised event summarization
Experiment 1
Experiment 2
Experiment 3

Figure 2: The flow chart for our proposal which consist of three modules. At first, we propose an algorithm which learns from a video dataset and corresponding captions a set of prototypes per concept. Secondly, we construct a prototype dictionary by collecting concept prototypes of all concepts and encode the video event in the dictionary. At the end we show the ability of our video representation for few- and zero-example event detection and unsupervised event summarization.

We organize the remainder of this paper as follows. We present our encoding using concept prototypes in Section 2. We introduce the experimental setup on the challenging TRECVID 2013 Multimedia Event Detection [26] and MED-summaries datasets [28] in Section 3. Results are presented in Section 4. We conclude in Section 5.

2. ENCODING CONCEPT PROTOTYPES

In this paper we aim to arrive at a semantic space with the ability for video event detection and summarization. We propose concept prototypes and leverage it for a new video representation. In this work, we define concept prototypes as a set of diversified video frames with variant visual appearance depicting the same concept in videos. For example, one can reasonably expect that for the concept *car*, prototypes are those frames related to various types and models of cars such as frames containing *city cars, sport car, truck,*. Because of their diversity, prototypes are beneficial over using a set of frames which describe only a specific type of car, e.g. *sport car*, or a set of frames which includes irrelevant frames to the concept *car*. In other words, concept prototypes for an arbitrary concept are those frames which appear in the most informative parts of the positive videos and not in the negative videos of the concept.

Since the concept prototypes are frame-level detectors, we have the ability of applying them in each frame of a video, which has several benefits. Firstly, by representing each frame of a video by concept prototypes, we can analyze each frame of the video individually. It gives us the opportunity of proposing a powerful video representation by aggregating all frames of a video in concept prototype space. Secondly, it allows us to do zero-example video event detection by considering the score of the frames in a video to each concept prototypes, and therefore help localize the key information in the video for zero shot event detection. It also offers us to to do unsupervised video event summarization by selecting the relevant frames of video to the event.

Our framework contains three parts, schematically illustrated in Figure 2. At first, we learn from a video dataset harvested from the web a set of concept prototypes per concept. Secondly, we construct a prototype dictionary by collecting concept prototypes of all concepts and encode each video from a train and test video event set in the prototype dictionary. At the end we show the ability of the concept prototype encoding for event video detection and summarization. Each part is detailed next.

2.1 Concept prototype learning

Our algorithm for learning concept prototypes of an arbitrary concept consists of two steps. i) Scoring positive frames which aims to compute a score for each frame of a positive videos containing the concept. ii) Selecting a subset of positive frames and consider them as concept prototypes.

Scoring positive frames: Suppose D is a dataset of web videos which are annotated for d concepts in a collection $C = \{c_1, c_2, ..., c_d\}$. Let consider $X_c \subset D$ is a set of n video/label pairs for arbitrary concept $c \in C$. We denote positive videos as V_i^+ which consist of n_i^+ frames v_{ij}^+, $j = 1, ..., n_i^+$. l^+ shown the number of positive videos in X_c, $X_c = \{\{V_i^+\}_{i=1}^{l^+}, \{V_i^-\}_{i=1}^{l^-}\}$. Similarly, V_i^-, n_i^-, v_{ij}^-, and l^- represent negative videos, the number of frames in the video, the j^{th} frames in the video, and number of negative video in X_c respectively. We suppose all frames v_{ij} belong to the feature space $\mathbb{R}^m$. We collect all frames of positive videos V_i^+ of concept c in P_c , $P_c = \{v_{ij}\}(i = 1, \ldots, l^+, j = 1, \ldots, n_i^+)$.

To compute a score per frame $v \in P_c$ we need a function to precisely return a value for v indicating its confidence of being applied as a prototype of the concept. It has to describe how relevant the frame v is w.r.t. concept c in the positive videos, V_i^+, and while departing away from the negative videos, V_i^-. We use a diverse density function which is well known and frequently used in machine learning [20,21]. The Diverse Density function (DD) at frame $v \in P_c$ returns a real value which expresses the probability that frame v agrees with the densest regions within the frames of both positive and negative videos. It reflects the coherence with the visual appearance of the concept and is defined as:

$$DD(v) = Pr(v|V_1^+, ..., V_l^+, V_1^-, ..., V_l^-) \tag{1}$$

where $Pr()$ represents the probability function. By applying Bayes'rule to Eq. 1 and assuming that all videos are conditionally independent given v, we write Eq. 1 as:

$$DD(v) = \prod_{i=1}^{l^+} Pr(v|V_i^+) \prod_{i=1}^{l^-} Pr(v|V_i^-) \tag{2}$$

To estimate $Pr(v|V_i^+)$ and $\mathrm{Pr}(v|V_i^-)$ we use the proposed way in [20]:

$$Pr(v|V_i^+) = \max_j \quad exp\left(-\|v_{ij}^+ - v\|\right) \tag{3}$$

$$Pr(v|V_i^-) = 1 - \max_j \quad exp\left(-\|v_{ij}^+ - v\|\right) \tag{4}$$

Table 1: Pseudocode of learning concept prototypes for concept category c.

```
INPUT: Positive videos for concept $c$ ($V_i^+$),
       Negative videos for concept $c$ ($V_i^-$),
       Threshold ($T$), and coefficient $\alpha$
OUTPUT: Concept prototypes for concept $c$ ($CP_c$)
1. Set $P_c$ be the set of all frames of videos in $V_i^+$
2. $H = []$ an array for keeping $DD$ values
3. $M = length(P_c)$
4. for $q = 1, \ldots, M$
5.     $H(q) =$ compute $DD(P_c(q))$ using Eq. 2
6. end
7. $i = 1$
8. Repeat
9.     Select concept prototype, a frame of $P_c$ corespond to
       maximum value of DD
       $CP_c(i) = P_c(find(H = max(H)))$
10.    Remove from H all elements $a$ satisfying
       $\|CP_c(i) - P_c(a)\| < \alpha\|CP_c(i)\|$ OR $H(a) < T$
11.    i = i + 1
12. While ($P_c \neg = []$)
13. Output: $CP_c$, concept prototypes for concept $c$
```

Eq.(3) uses the maximum similarity between all frames in a positive video and the given frame as the probability of choosing v as a prototype of video V_i^+. On the contrary, Eq.(4) measures how dissimilar a negative video should depart from the candidate prototype frame. After considering Eq.(3) and Eq.(4) in Eq.(2) the DD function computes a real value between 0 and 1 for frame v. A larger value of DD at frame v indicates a higher probability that the frame v fits better with the frames from all positive videos than with those from all negative videos, and therefore have higher chance to be chosen as a concept prototype.

Selecting concept prototypes: After computing a score per each frame of P_c, we select the concept prototypes for concept c by considering two constraints: i) They need to be distinctive from each other so that the prototype collection has no redundancy. ii) They need to have large DD values. Applying the first constraint, we remove some of the frames that are similar to each other. In our algorithm we control the repetition of frames by considering coefficient α. It is a parameter which defines the degree of closeness of frames to concept prototypes. Its value is between 0 and 1. Closer to 1 causes our algorithm generate small number of prototypes per concept. The second constraint limits concept prototypes to those that are most informative in terms of co-occurrence in different positive videos. In our algorithm, we consider it by picking frames with DD value greater than a threshold T.

Our learning algorithm for obtaining concept prototypes for concept category c is summarized in Table 1. The inputs are a set of positive and negative videos for concept c, threshold T and coefficient α which are parameters to define

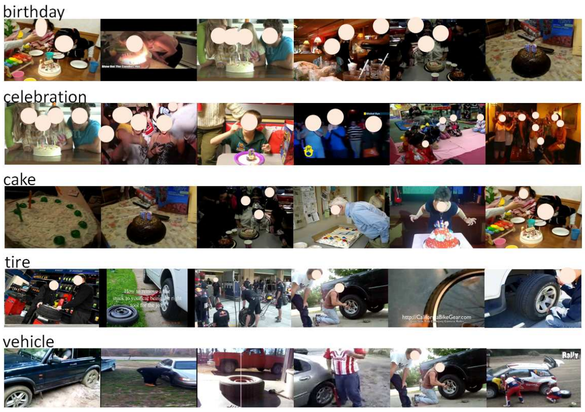


Figure 3: Learned prototypes for five arbitrary concepts. Prototypes ordered from left (best) to right.

the number of concept prototypes for concept c. The pseudo code learns a set of concept prototypes for concept c which is represented by CP_c. In the pseudo code in Table 1, lines 1-6 compute the value of the DD function for all frames of positive videos in P_c. Line 7-12 describe an iterative process to select a collection of distinct concept prototypes. In each iteration, a frame of P_c with the maximum value of DD function is selected as a concept prototype (line 9). The frames of P_c that are close to the selected concept prototype or that have DD values lower than a threshold are removed from P_c (line 10). A new iteration starts when there is no any frame in P_c. In Figure 3 we visualize several concept prototypes.

2.2 Video encoding

In this section we first describe the process of constructing a prototype dictionary using concept prototypes, then we explain how to encode video events of train and test set into the prototype dictionary.

Constructing prototype dictionary: To make a prototype dictionary we first learn m concept prototypes for each of d concept $c_i \in C$ using the method which we explained in section 2.1. We consider $CP_{c_i} = \{cp_{i1}, cp_{i2}, ..., cp_{im}\}$ as concept prototypes for concept c_i. Then we construct a prototype dictionary PD by collecting all concept prototypes of concepts in $C, PD = \{CP_{c_1}, CP_{c_2}, ..., CP_{c_d}\}$ which element cp_{jk} is k^{th} concept prototype of j^{th} concept.

Encoding video in prototype dictionary: We consider each element cp_{jk} as a feature in new feature space $\mathbb{F}_{PD}$. We define a function $g_{jk}(:)$ to compute the value of the feature cp_{jk} on video event V, which consist of n frames $V = \{v_1, v_2, ..., v_n\}$ as:

$$g_{jk}(V) = Pr(cp_{jk}|V) = \max_{i=1,...,n} similarity(cp_{jk}, v_i). \quad (5)$$

We can interpret $g_{jk}(V)$ as a measure of similarity between the concept prototype cp_{jk} and the video event V. It is determined by the value of similarity between concept prototype cp_{jk} and the closest frame in the video V.

As a result, each video event V can be encoded in $\mathbb{F}_{PD}$ using all concept prototypes in PD. It can be defined as a point ,$\varphi(V)$, in feature space $\mathbb{F}_{PD}$ which define as:

$$\varphi(V) = [g_{11}(V), g_{12}(V), ..., g_{jk}(V), ..., g_{dm}(V)]^T \quad (6)$$

where each element of $\varphi(V)$ is defined by one concept prototype and one frame from the video event V, the frame that is *closest* to the concept prototype. Also it can be viewed as a measure of the degree that a concept prototype is visible in the video event V.

2.3 Event detection and summarization

After representing video events in concept prototypes space, we use it for the problem of few- and zero-example event detection, and unsupervised video event summarization.

Few-example event detection: Suppose $TR = \{V_1^+, ..., V_{n^+}^+, V_1^-, ..., V_{n^-}^-\}$ is a train set consist of n^+ positive videos and n^- negative videos, and $TE = \{W_1, ..., W_n\}$ is a test sets for an event category e. By mapping each video of train and test set to concept prototype space using Eq. 6 we define TR and TE as two matrix representation of all videos in $\mathbb{F}_{PD}$:

$$TR = [\varphi(V_1^+), ..., \varphi(V_{n^+}^+), \varphi(V_1^-), ..., \varphi(V_{n^-}^-)]^T \quad (7)$$

$$TE = [\varphi(W_1), ..., \varphi(W_n)]^T \quad (8)$$

Each row of TR and TE represents a video, k^{th} column of TR and TE represent k^{th} feature in $\mathbb{F}_{PD}$. We train a model on TR and report the result of event detection accuracy on TE.

Zero-example event detection: In this scenario there is only a textual query explaining the event. Suppose T is an explanation for event e which consist of n words, $T = \{w_1, w_2, ..., w_n\}$, $C = \{c_1, c_2, ..., c_d\}$ is a set of concepts, and PD is a prototype dictionary which consists of $m \cdot d$ elements (m concept prototypes for each of d concept in C). Let's consider $W_i = \{w_{i1}, w_{i2}, ..., w_{il}\}$, which consist of l frames, be a video of event e in test set TE. We use following steps for computing a score per each video W_i in TE. i) At first we define a binary representation for event e as a query, $\mathbf{Q} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_d}] \in \mathbb{F}_{PD}$, by comparing the concepts in C with the words in T. The value of a_i computes as:

$$a_i = \begin{cases} \overrightarrow{1} \in \mathbb{N}^m & \text{if } c_i \in T \\ \overrightarrow{0} \in \mathbb{N}^m & \text{if } c_i \notin T \end{cases}$$

ii) We map each frame of W_i to prototype dictionary PD by computing similarity between each frame of W_i and each of concept prototypes in PD. As a result, each frame of W_i is represented as a vector of concept prototype score, $W_i = \{x_{i1}, x_{i2}, ..., x_{il}\}, x_{ij} \in \mathbb{F}_{PD}$. iii) We compute a score per each frame of W_i by multiplying query Q with $x_{ij}, j = 1, ..., l$ and define vector $Y = [Q \cdot x_{i1}, Q \cdot x_{i2}, ..., Q \cdot x_{il}]$. The value of j^{th} element of Y describe how much the content of the j^{th} frame of W_i matches to event e. At the end iv) We score video W_i with maximum value of vector Y, $S(W_i) = max(Y)$.

After computing a score S per each video $W_i, i = 1, ..., n$ of TE, we sort the videos with their scores and compute the accuracy of zero-example event detection for event e.

Unsupervised video event summarization: We demonstrate the feasibility of using our concept prototypes, for summarizing the content of a video event by selecting relevant frames to an event category. Suppose W_i is a positive video of event category e consist of l frame, $W_i = \{w_{i1}, w_{i2}, ..., w_{il}\}$. We follow the first three steps of zero-example event detection, to compute vector Y which the value of j^{th} element of Y show how much the content of the j^{th} frame of W_i is match to event e. We sort the vector Y

from the best frame, that completely matches the content of the event, to the worst frame. We select the best frames based on how many percent of frames of video W_i we want to report as summarization of W_i.

3. EXPERIMENTAL SETUP

3.1 Datasets

In our experiments we report on two large corpora of challenging real-world web video: the TRECVID 2013 Multimedia Event Detection dataset [26], and MED-summaries provided by Potapov *et al.* [28].

Event detection: TRECVID 2013 MED [26]. This corpus contains 56k user-generated web videos with a large variation in quality, length and content of 20 real-world events. This corpus consists of several partitions, we consider the Event Kit training, Background training, and MED test set, containing about 200, 5K, and 27K, videos. We follow the *10Ex evaluation procedure* outlined by the NIST TRECVID event detection task [27]. It means that for each event the training data consist of 10 positive videos from the Event Kit training data along with about 5K negative videos from the Background training data. We also consider the *0Ex evaluation procedure* as specified by NIST, where we don't have any positive video example available. We rely only on a provided brief textual definition of the event containing some evidences of concepts which can be expected to be visible in positive video examples of a particular event.

Event summary: MED-summaries [28]. This set consist of 160 videos selected from the TRECVID 2011 Multimedia Event Detection task: the validation set consist of 60 videos and the test set consist of 100 videos. Every segment of all these 160 videos is annotated with a category-specific importance value (0, 0.333, 0.666, and 1). All segments of all videos are annotated by different users to obtain the ground truth video summary for evaluation. High scores indicate the human annotators consider the segment is close to the event of interest. We report all results on the test set containing 10 videos for a total of 10 events. We adapt the provided annotation format from segment level to frame level by simply propagating the segment annotation to each of its frames.

3.2 Implementation details

Concept prototypes training data. To train our concept prototypes, we use a video dataset that is disjunct from both the TRECVID 2013 MED and MED summaries collections. We adapt the VideoStory46K from Habibian *et al.* [9] which contains a collection of 46k video from YouTube. Every video has a short title caption provided by the user who uploaded the video. From these captions we generate individual terms which serves as positive concept labels for the corresponding video. In total there are 19,159 unique terms in the captions. We filter the terms by: removing stop words, noisy terms, terms which are not visually detectable such as *God* (we used the visualness filter from [9]), terms with frequency less than 100, and keep those terms which match with the words in the definition text of the events. After this step we reach to a list of 479 different concept labels.

Features. For representing the frames in all of our datasets we use deep learning features. At first we extract the frames of videos uniformly every 2 seconds. Then, similar to [13], we use a pre-trained convolutional neural network for mapping every frame of videos to a 4,096-dimensional feature vector which is the output of the second fully connected layer of the convolutional neural network. The network is pre-trained on all the 15,0254 categories in the ImageNet dataset, for which at least 50 positive examples are available.

Concept prototypes learning. For learning concept prototypes for each of the 479 concepts we repeat the following steps. i) We construct a set of videos for each concept label by considering all its positive videos and supplement it with negatives up to ten times the number of positives by random sampling from other concept labels. ii) We apply our concept prototypes learning algorithm (Section 2.1). We set parameter T as the average of the maximum and minimum of the DD function in H (Table 1), and we empirically found 0.05 to be a good value for α. We map the videos from TRECVID 2013 MED and MED-summaries to concept prototypes space using cosine similarity.

Event detection. We train for each event a one-versus-all linear support vector machine [29] and fix the value of its regularization parameter C to 1.

3.3 Experiments

In order to establish the effectiveness of our concept prototypes for event detection and summarization we perform three experiments.

Experiment 1: ***Few-example event detection*** In this experiment we compare the event detection accuracy of our proposed concept prototype encoding versus five baselines. *Baseline 1:* DeepNet. Video representation using deep learning features. *Baseline 2:* 15k-Concepts. A video event representation consisting of 15k concept detectors trained on ImageNet (inspired by [10]). *Baseline 3:* Conceptlets [22]. Video representation consisting of the most informative ImageNet 15K concepts per event. *Baseline 4:* 479-Concepts. Video representation using 479 video-level concept detectors trained on VideoStory46k (inspired by [11]). *Baseline 5:* VideoStory [9]. Video event representation which strives to encode the caption of a video rather than a single word. All six methods are evaluated by their event detection performance on the TRECVID 2013 MED test set.

Experiment 2: ***Zero-example event detection*** We compare the accuracy of our concept prototypes for zero-example event detection versus four baselines. *Baseline 1:* Flickr concepts. As a first baseline we report the result in [4] which used 2,000 relevant concepts per event as trained from flickr images. *Baseline 2:* 15k-Concepts. Same as above. *Baseline 3:* Composite concepts [8]. Video representation using 138 video-level concept detectors which are the result of combining several concepts using logical connectors. *Baseline 4:* 479-Concepts. Video representation using 479 video-level concept detectors trained on VideoStory46k. For all methods we build a query Q per event as a binary vector indicating presence or absence of the concept in the event description, as detailed in Section 2.3. We compute a score for a video in the test set by multiplying Q with a histogram containing the scores of concept detectors when applied on the video. From the resulting ranking, we compute the zero-example accuracy. We again evaluate the accuracy of all methods on the TRECVID 2013 MED test set.

Experiment 3: ***Unsupervised event summarization*** We compare the accuracy of our concept prototypes in video event summarization versus three baselines. *Baseline 1:*

Table 2: Experiment 1: Few-example event detection on TRECVID 2013 MED test set. Best AP result per event in bold.

Event	**DeepNet**	**15k-Concepts** [10]	**Conceptlet** [22]	**479-Concepts** [11]	**VideoStory** [9]	**This paper**
Birthday party	0.137	0.114	0.145	0.146	0.118	**0.188**
Changing a vehicle tire	0.391	0.388	0.401	0.444	0.103	**0.464**
Flash mob gathering	0.405	0.347	0.377	0.425	**0.535**	0.439
Getting a vehicle unstuck	0.334	0.323	0.390	0.377	0.319	**0.418**
Grooming an animal	0.084	0.108	0.110	0.126	0.151	**0.154**
Making a sandwich	0.031	0.074	0.075	0.100	0.074	**0.131**
Parade	0.171	0.109	0.145	0.246	**0.452**	0.303
Parkour	0.320	0.309	0.319	0.317	**0.721**	0.326
Repairing an appliance	0.169	0.127	0.133	**0.320**	0.184	0.244
Working on a sewing project	0.058	0.071	0.103	0.072	**0.151**	0.109
Attempting a bike trick	0.054	0.030	0.047	0.106	0.061	**0.144**
Cleaning an appliance	0.021	0.019	0.020	0.031	**0.078**	0.055
Dog show	0.232	0.134	0.165	**0.379**	0.354	0.313
Giving directions to a location	0.012	0.005	0.008	0.020	0.004	**0.022**
Marriage proposal	0.002	0.002	0.003	**0.004**	**0.004**	**0.004**
Renovating a home	0.019	0.024	0.032	**0.056**	0.051	0.033
Rock climbing	0.070	0.063	0.083	0.080	0.100	**0.110**
Town hall meeting	0.268	0.201	0.229	0.158	0.118	**0.290**
Winning a race without a vehicle	0.150	0.126	0.164	0.116	**0.217**	0.182
Working on a metal crafts project	0.054	0.068	0.078	0.116	0.118	**0.144**
MAP	0.150	0.135	0.156	0.181	0.196	**0.202**

Random. Video event summarization with relevant frames selected randomly. To cancel out the accidental effects of randomness, we repeat the random frame selection 10 times and report the mean. *Baseline 2:* 15k-Concepts. A video event summarization with relevant frames selected using 15k ImageNet concepts. *Baseline 3:* 479-Concepts. A video event summarization with relevant frames selected using concept detectors. *Oracle* Video event summarization using the ground truth user annotation value from [28], to create an upper bound for video event summarization. In *Baseline 2* and *3*, similar to our method in section 2.3, we make a vector Y with the value of each of its elements being the result of multiplying Q, a binary representation of event, and a histogram of applying concept detectors on each frame of video. Then we follow the steps which we explained in Section 2.3 for summarizing the video. We evaluated the accuracy of all methods in video event summarization on the MED-summaries dataset [28] by keeping 10, 20, 30, ..., 100 percent of the frames per video. For those videos annotated by more than one user, we report the mean accuracy. At the end we report the mean accuracy for all 10 videos of each of the 10 events.

Evaluation criteria For event detection performance we consider the average precision (AP) [30]. We also report the average performance over all events as the mean average precision (MAP). Similar to [28], for evaluating event summarization performance we propose a new metric, importance ratio. We defined information ratio (IR) as average of user annotation value of selected p frames of the video. The value of IR shows the quality of the video event summarization. We consider the average of the p highest user annotation value as a ground truth. The closeness of the value of IR to ground truth value shows the better summarization.

4. RESULTS

4.1 Few-example event detection

We show the result of experiment 1 in Table 2. Results demonstrate that the video event representation based on our concept prototypes performs better than all baselines. Our concept prototypes reach 0.202 MAP in event detection, where using DeepNet feature, 15k-Concepts, Conceptlet, 479-Concepts, and VideoStory results in 0.150, 0.135, 0.156, 0.181, and 0.196 respectively.

We explain the good results of our concept prototypes from several observations. First, most of the concepts in 15k-concepts are irrelevant to the event categories. Moreover, as these detectors originate from ImageNet, they are trained on images and not on video leading to a domain-mismatch. Using Conceptlets [22] to find the most informative concepts, out of the 15k, per event improves the event detection accuracy, but it cannot compensate for the domain mismatch. Interestingly, learning the events directly from DeepNet features is advantageous over using 15k-concepts. However, when we learn the detectors from video annotations, as suggested in [11], result improve considerably (from 0.150 to 0.181). VideoStory further improves upon this result, but our concept prototypes are more accurate. There are two reasons. First, VideoStory trains detectors using all frames from both positive and negative training videos. By adding lots of irrelevant frames it leads to a noisy video concept representations. In contrast, our concept prototypes identify the relevant frames of video only, making the representation more clean. Secondly, since the concept detectors from VideoStory are trained over a video-level representation, they have to apply on a video-level representation for event detection as well. Since most of the events happen in a part of the video only, aggregating all frames is sub-optimal. In contrast, our concept prototypes have the ability to apply the concepts on individual frames, which decreases the effect of irrelevant frames.

The results of experiment 1 confirm that few-example event detection accuracy profits from using concepts that are relevant to the event and which are learned from a video, rather than the image dataset. Moreover, it is beneficial to represent video using concept prototypes as the influence of irrelevant frames is reduced.

4.2 Zero-example event detection

The results of this experiment are shown in Table 3. It indicates our concept prototypes outperform all four baselines in zero-example event detection. We reach to 0.119 MAP where the alternatives reach to 0.024, 0.032, 0.063, and 0.100 MAP. Since our concept prototypes are frame-

Table 3: Experiment 2: Zero-example event detection on TRECVID 2013 MED test set. Best AP result per event in bold.

Event	Flickr concepts [4]	15k-Concepts [10]	Composite Concepts [8]	479-Concepts [11]	This paper
Birthday party	0.032	0.022	0.076	0.096	**0.154**
Changing a vehicle tire	0.038	0.099	0.018	0.238	**0.320**
Flash mob gathering	0.058	0.104	**0.319**	0.266	0.271
Getting a vehicle unstuck	0.010	0.107	0.055	0.356	**0.406**
Grooming an animal	0.011	0.019	0.009	0.079	**0.095**
Making a sandwich	0.028	0.021	0.079	0.082	**0.164**
Parade	0.130	0.094	0.223	0.221	**0.240**
Parkour	0.027	0.020	0.021	0.077	**0.112**
Repairing an appliance	0.048	0.078	0.025	**0.221**	0.213
Working on a sewing project	0.012	0.016	0.014	0.087	**0.089**
Attempting a bike trick	0.003	0.017	0.021	0.056	**0.061**
Cleaning an appliance	0.013	0.006	0.006	0.022	**0.026**
Dog show	0.003	0.003	0.001	**0.011**	**0.011**
Giving directions to a location	0.002	0.004	**0.025**	0.004	0.008
Marriage proposal	0.009	0.004	0.002	0.003	**0.005**
Renovating a home	0.002	0.017	0.023	**0.029**	0.026
Rock climbing	0.014	0.003	**0.146**	0.015	0.036
Town hall meeting	0.020	0.008	0.014	0.023	**0.035**
Winning a race without a vehicle	0.028	0.012	**0.110**	0.095	0.101
Working on a metal crafts project	0.002	0.002	0.006	0.010	**0.014**
MAP	0.024	0.032	0.063	0.100	**0.119**

based we have the ability to apply them on each frame. As a result we describe how well the frame is related to an event. Irrelevant frames are simply ignored.

Flickr concepts are relevant to the event but they are trained on images of Flickr leading to a domain mismatch. The same holds for the 15k-Concepts trained on ImageNet. Moreover, most of its concepts are irrelevant to the event of interest. Both the composite concepts and 479-Concepts are trained on video-level labels. It means they suffer less from the domain mismatch leading to improved zero-example event detection results. However, both of these methods suffer from the fact that they treat all frames equally, which adds lots of irrelevant frames to the representation. Our concept prototypes contain concepts that are relevant to the event, are trained on video, and consider only the most relevant frames.

4.3 Unsupervised event summarization

We show the result of experiment 3 on MED-summaries [28] in Figure 4. The result demonstrates the effectiveness of our concept prototypes for unsupervised video event summarization against random frame selection, 15k-Concepts, and 479-Concepts for all summarization settings. When we request a summary containing 30% of the video we reach to 0.600 accuracy in IR, where random frame selection,15k-Concepts, and 479-Concepts reach to 0.221, 0.274, and 0.384 accuracy in IR. The accuracy of the oracle upper bond is 0.923 IR. Since our concept prototypes are frame-based, we can apply them on each frame of a video and reach to an accurate video representation. Where the others are video-level by nature which makes them less suited for frame-level classification, resulting in a less reliable summary.

We visualize the result of video event summarization for two videos of the events *birthday party* and *Changing a vehicle tire* using our concept prototypes and the 479-concept detectors, as the best baseline, in Figure 5. In both examples we see the effectiveness of using our concept prototypes against concept detectors in video event summarization.

5. CONCLUSIONS

In this paper we propose *concept prototypes* a new semantic video representation for few and zero example event detection and unsupervised video summarization. Different from existing works, which obtain a semantic representation by training concepts over images or an entire video, we propose an algorithm that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations.

We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames containing both positive and negative training videos. The prototypes represent a class of frames that are more likely to appear in positive videos than in negative videos. Since the concept prototypes are a frame-level representation of concepts, we have the ability of mapping each frame of a video in concept prototype space which has several benefits. First, few-example event detection accuracy increases when using relevant concept prototypes. Second, the accuracy of zero-example event detection increases by analyzing each frame of a video individually in concept prototype space. Finally, unsupervised video event summarization using concept prototypes is more accurate than recent alternatives.

We conclude that selective use of relevant frames of a video to the concept, by means of concept prototypes, is beneficial for detecting and summarizing video events.

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

6. REFERENCES

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

[2] S. Bhattacharya, M. Kalayeh, R. Sukthankar, and M. Shah.

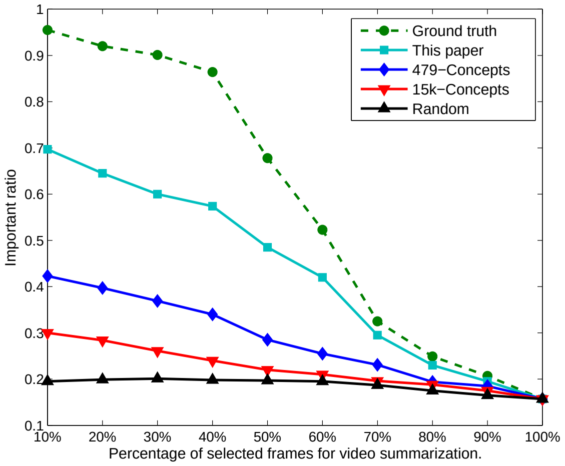
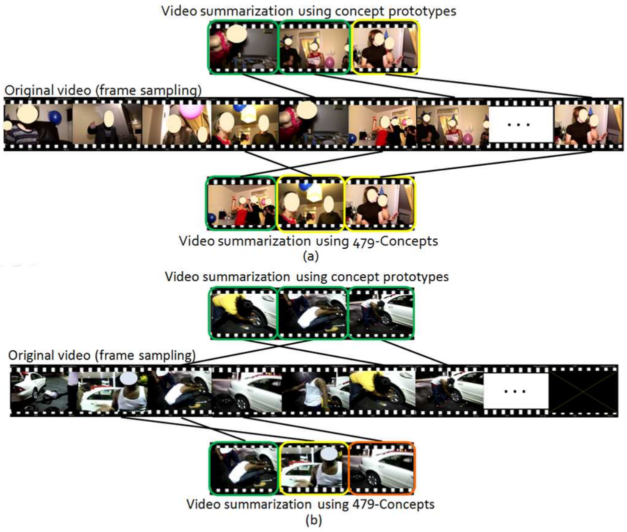


Figure 4: Experiment 3. Video event summarization on MED-summaries dataset. The result shows the effectiveness of our concept prototypes.



Video summarization using concept prototypes
Original video (frame sampling)
Video summarization using 479-Concepts
(a)
Video summarization using concept prototypes
Original video (frame sampling)
Video summarization using 479-Concepts
(b)

Figure 5: Experiment 3. Two examples of video event summarization using concept prototypes and 479-Concepts. We denote the ground truth annotation value of each frame with four colors. Green indicates 1, yellow 0.666, orange 0.333, and red 0 (higher is better). The color coding shows the effectiveness of concept prototypes.

Recognition of complex events exploiting temporal dynamics between underlying concepts. In *CVPR*, 2014.

[3] E. F. Can and R. Manmatha. Modeling concept dependencies for event detection. In *ICMR*, 2014.

[4] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014.

[5] Y. Cui, D. Liu, J. Chen, and S.-F. Chang. Building a large concept bank for representing events in video. *arXiv preprint arXiv:1403.7591*, 2014.

[6] D. Ding et al. Beyond audio and video retrieval: Towards multimedia summarization. In *ICMR*, 2012.

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

[8] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

[9] A. Habibian, T. Mensink, and C. G. M. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *MM*, 2014.

[10] A. Habibian and C. G. M. Snoek. Recommendations for recognizing video events by concept vocabularies. *CVIU*, 124:110–122, 2014.

[11] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.

[12] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[14] K.-T. Lai, D. Liu, M.-S. Chen, and S.-F. Chang. Recognizing complex events in videos by learning key static-dynamic evidences. In *ECCV*, 2014.

[15] K.-T. Lai, F. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.

[16] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013.

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

[18] Z. Ma, Y. Yang, Z. Xu, N. Sebe, and A. G. Hauptmann. We are not equally negative: fine-grained labeling for multimedia event detection. In *MM*, 2013.

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

[20] O. Maron. *Learning from Ambiguity*. PhD thesis, 1998.

[21] O. Maron and T. Lozano-Pérez. A framework for multiple instance learning. In *NIPS*, 1998.

[22] M. Mazloom, E. Gavves, and C. G. M. Snoek. Conceptlets: Selective semantics for classifying video events. *TMM*, 16(8):2214–2228, 2014.

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

[24] M. Mazloom, X. Li, and C. G. M. Snoek. Few-example video event retrieval using tag propagation. In *ICMR*, 2014.

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

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

[27] P. Over et al. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2013.

[28] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.

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

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

[31] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, 2014.