Video Event Recognition Using Concept Attributes

Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, Harpreet Sawhney
SRI International Sarnoff
Princeton, NJ, USA 08540
jingen.liu@sri.com, hui.cheng@sri.com

Abstract

We propose to use action, scene and object concepts as semantic attributes for classification of video events in InTheWild content, such as YouTube videos. We model events using a variety of complementary semantic attribute features developed in a semantic concept space. Our contribution is to systematically demonstrate the advantages of this concept-based event representation (CBER) in applications of video event classification and understanding. Specifically, CBER has better generalization capability, which enables to recognize events with a few training examples. In addition, CBER makes it possible to recognize a novel event without training examples (i.e., zero-shot learning). We further show our proposed enhanced event model can further improve the zero-shot learning. Furthermore, CBER provides a straightforward way for event recounting/understanding. We use the TRECVID Multimedia Event Detection (MED11) open source event definitions and datasets as our test bed and show results on over 1400 hours of videos.

1. Introduction

Recognizing atomic human actions from videos "in the wild" has received considerable attention in the past few years [17, 15, 18]. However, atomic actions, such as "walking", "kissing", "placing an object" are too primitive to be used for search of internet videos. In Internet searches users usually look for events such as "wedding ceremony","woodworking" or "birthday party", but rarely retrieve videos of a simple action such as "person walking" or "person bending". In this work, we characterize an *event* as a juxtaposition of various actions, scenes and objects, which is more descriptive and meaningful. Our goal is to recognize complex *events* from large-scale open source videos.

In order to accurately recognize an event, an effective event representation is required. Unlike an action, the visual contents of a video event are usually very diverse. For example, a "wedding ceremony" consists of various concepts including actions such as "hugging" and "kissing", scenes such as "church" and "garden", and objects such as "cake" and "ring", as shown in Fig. 1. In such a scenario, the low-level feature based event representation (LLFeat, e.g., the bag-of-visual-word model) may have difficulty in handling the intra-class variability, especially when *the number of example event videos is small*. In addition, the numeric LLFeat is not suitable for high-level event analysis and understanding, such as event recounting.

Figure 1. Examples of Events (i.e., Event 02 "feeding an animal", Event 04 "wedding ceremony", Event 05 "woodworking", and Event 06 "birthday party"). Each row corresponds to one video selected from the event. The examples indicate that a diverse set of action, scene, and object concepts constitute these events.

We propose to represent events in a semantic space consisting of concepts related to actions, scenes and objects. Based upon this space, we are able to model an event with various concept features. Basically, this Concept-Based Event Representation (CBER) divides the event recognition problem into two parts: concept detection followed by event recognition. Concept detection is performed using low-level visual features, and it is not restricted to a specific event. In other words, the universal concept detectors are shared across different events. Meanwhile, event recognition is able to focus on high-level inference using semantic concept features.

In order to capture various aspects of the concept distribution over an event, we develop five complementary semantic concept features for event representation. We

demonstrate that CBER has stronger generalization capability versus to direct event recognition using low-level features. This property is particularly useful for learning events with just a few training examples. In fact, the generalization capability is due to the information sharing among events in terms of concepts. Thanks to CBER models, we further demonstrate how to recount and summarize an event for understanding beyond the recognition.

The notion of representing an event in a semantic concept space is inspired by recent work in object recognition [5, 14, 24, 23, 29], action recognition [16, 31], and image retrieval [25] with attributes. Indeed, we characterize an event by treating the action, scene and object concepts as event attributes. These semantic attributes usually embody the information of *who*, *what*, *where* and *how*, which is usually discriminative for an event. Therefore, CBER can not only be used for recognizing a familiar event with training examples, but it can also be used to describe and recognize a novel event *without* training examples.

To recognize a novel event, one needs to manually define/localize it in the semantic concept space according to its semantic description which indicates the presence or absence of concepts in the event. This manual process is commonly employed in most existing work to recognize novel objects [14, 24] and actions [16]. In general, however, the human-provided description is subjective and incomplete, thus the human-defined event models may not be accurate. For instance, "dancing" and "hugging" can be left out when one defines an event model vector for the "wedding ceremony", as they are not as common as concepts such as "kissing", "church", "bride", and "groom" for a wedding. To enhance the event model, we propose to use semantic similarity between concepts to augment the model with other concepts similar to the one provided by a user. Our idea is inspired by [7] on sharing semantic labels for image classification. Semantic similarity between concepts can be estimated from the statistical distribution of concepts on videos.

1.1. Related Work

The definitions of *event* and *action* are ambiguous in computer vision literatures. As both event and action videos generally contain plenty of object motions, the terminologies of *event* and *action* are alternatively used [3, 13]. In this work, however, an *event* depicts a complex visual happening consisting of a number of actions, scenes and objects. For instance, our event definitions include life events such as "weddings" and "birthday parties", as well as how-to events such as "wood-working' [1]. In contrast, in past work on action recognition [21, 17, 9], actions largely consist of atomic actions such as "tennis swing", "jumping", etc. detection for surveillance by fixed cameras, usually appeal to object tracking and action analysis [8, 12, 2].

There are some recent works on recognizing events from still images [10, 20]. For example, Imran *et al.* [10] proposes to use PageRank to recognizing events from photo collections. Luo *et al.* [20] also combines GPS information with photos for event recognition. Since these works concentrate on images, they use scene information to distinguish events to a large extent. However, the action concept is a critical component in event classification. Therefore, both action and scene are important concepts for our CBER.

Our event recognition task is similar to [22, 4, 11, 27], but our goals are significantly different. [27] focus on event detection with low-level features. In [22, 4], the authors aim at comparing the performance of their Vector Models with that of various low-level feature based models for event detection with a relatively large number of training examples. As a comparison, we demonstrate that the CBER has better generalization capability (e.g., recognizing events with a few examples), and enables recognizing a novel event, as well as better event understanding (e.g., event recounting). To the best of our knowledge, all these advantages of CBER have not been discussed for video event recognition. In addition, we proposed several complementary semantic features rather than the Vector Model in [22, 4].

We treat concepts as the attributes of events in our CBER, which is related to the usage of attributes in object recognition [14, 24, 5, 23, 28], action recognition [16, 31], image retrieval [25], and event recognition in still images [26]. We explore more informative event representations derived from the semantic concept space, which capture not only the distribution of concepts, but also the co-occurrence relationship between concepts. Our goal is similar to [31], in which they learn an action base using sparse coding from still images. Moreover, unlike most existing works which usually directly employ manually defined models by a user to recognize novel categories, we further utilize the semantic similarity between concepts to enhance the human-defined event model. Our method differs from [23], which uses ranking technique to assign relative attributes to images. Our approach is data-driven, and no further information (e.g., attributes ranking) is required.

1.2. Contributions

Unlike the previous work in CBER, our work is the first to *systematically* demonstrate the advantages of representing events using concepts as event attributes in some event recognition and understanding applications. More specifically, we propose various semantic concept features and demonstrate three interesting applications well-supported by CBER: (1) We demonstrate that CBER improves the generalization capability of event models. (2) We devise and demonstrate an approach to improve human-defined event models for recognizing a novel event by enhancing the model components using semantic similarities between concepts. (3) Beyond recognition, we further present a method to recount the detected event for a video, on the

basis of which we can assess the strength of contribution of each concept towards the classification of the event.

We evaluate our approach on the TRECVID Multimedia Event Detection (MED11 [1]) dataset. This source is the first of its kind to make public large scale videos from open sources. We wish to emphasize that experimentation with such a large scale video dataset implies that our experimental results validate conclusions that could be broadly applicable to YouTube like videos.

2. Event Recognition in Concept Space

2.1. Learning Concept Detectors

We define our concept collection $\mathcal{C} = \{C_1, C_2, ..., C_K\}$, where K=101. It includes 81 action concepts, as well as 17 scene and object concepts such as "kitchen", "lake/pond", "wheel-closeup", and so on (The full list is attached in the supplemental material) . In addition, it contains three common object concepts "face", "car" and "person". For each concept, we acquire training examples (video segments for actions and keyframes for scene and objects) from our developmental dataset.

We employ well-established techniques for action, scene and object detection for building our concept detectors. In particular, static features (i.e., SIFT [19]), dynamic features (i.e., STIP [15] and Dense Trajectory Based features [30]), and the bag-of-word representations [30, 14] defined over codebooks of these features are used to represent action, scene and object concepts. Binary SVM classifiers with Histogram Intersection kernel are used for concept classification. While the concept detectors of "face", "car" and "person" are adopted from some publicity available detectors such as [6] used in our work.

In the rest of this paper, we represent a concept detector as φ_i for concept C_i. The inputs $\mathbf{x}$ for action concept detectors are short video segments, and keyframes for the other concept detectors. The output is the detection confidence $\varphi_i(\mathbf{x})$. Sec. 2.3 describes how to apply a concept detector to long length videos (e.g., an event video), where we localize the concepts temporally. The performance of detectors on unconstrained videos is varied. In the next section, we design more robust concept features from these detectors for event classification.

2.2. Concept Space Definition

We define a concept space $\mathbb{C}^K$ as an K-dimensional semantic space, in which each dimension encodes the value of a semantic property. This space is spanned by K concepts $\mathcal{C} = \{C_1, C_2, ..., C_K\}$. In order to embed a video $\mathbf{x}$ into the K-dimensional space, we define a set of functions $\Phi = \{\phi_1, ..., \phi_K\}$, where ϕ_i assigns a value $c_i \in [0, 1]$ to a video indicating the confidence of the i^{th} concept presence in it. The definition of ϕ_i depends on the application. *Note that ϕ_i is not necessary the concept detector φ_i.* If the concept detector φ_i take the whole video as one *single* input, then we can treat ϕ_i and φ_i same. However, if the detector is applied to a video by means of sliding window (i.e., split a video into W input windows, and thus produces W outputs), then we need to define ϕ_i (i.e., max function in Sec. 2.3) to convert W outputs of φ_i into one ***single*** confidence value c_i. As a result, the function set $\Phi(\mathbf{x})$ embeds a video $\mathbf{x}$ in the K-dimensional semantic space as a vector $(c_1, ..., c_K)$. Semantically similar videos form a cluster in the space. Thus we can perform event recognition by training a classifier in this space. In fact the event classification is decomposed into phases: (1) Embedding a given video in the concept space; and (2) Classifying the event with features derived from the embedding, as discussed in Sec. 2.3.

On the other hand, we also can define a new event in the space by manually assign a confidence value c_i (e.g., 1 or 0) to i^{th} concept based on our knowledge to this event. In other words, it is possible to define a novel event without looking at video examples. This fact enables to recognize a novel event without training examples, as discussed in Sec. 2.4.

2.3. Event Modeling over Concept Space

Since the goal of this work is to assess the efficacy of the CBER, we focus on understanding how the presence or absence of episodic concepts during the course of an event influences recognition of the event. We assume detectors for a suite of concepts that we define for the events at hand are available to us. Any of the well-known past works can be used to create detectors for this concept suite.

We apply a sliding-window (i.e., an *XYT* cube) based detection scheme for action concepts, while scene and object concepts are detected on sampled frames of the video. Fig. 2 depicts our overall approach for description and recognition of a video event in terms of a set of concepts. Since the concept detection is noisy for videos "in the wild", our method uses the atomic concept detectors as filters that are applied to a given *XYT* segment of a video to capture the similarity of content to the given concept. So as a first step towards representing a video with concepts, each concept detector is applied to each *XYT* window in a video to obtain an $K \times W$ matrix $\mathbf{C}$ of scores, where $C_{ij} \propto p(c_i|w_j)$. Each C_{ij} is the detection confidence of concept i applied to window j. $\mathbf{C}$ represents the complete embedding of the video in the space of concepts. W is determined by the video length and the sliding window size which is set to the average segment length of all training video segments in our work. Other approaches such as shot detection can be used to determine the detection windows too. So far, the fixed size sliding window works for our case.

There may be many ways in which occurrence of concepts determines the presence or absence of a video event. We exploit a number of increasingly complex features derived from $\mathbf{C}$ to model and classify events. These features span the spectrum from counting the occurrence of concepts to statistics of concept confidences to co-occurrence and co-

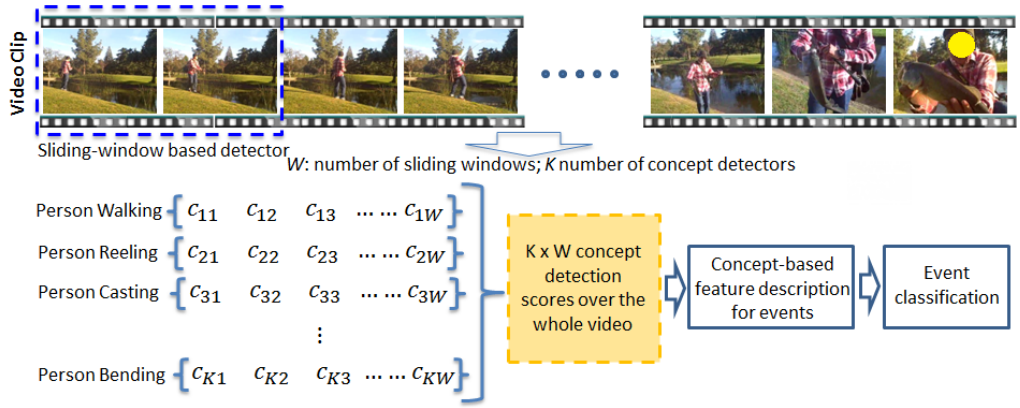

Figure 2. Event Descriptions using Concepts. K concepts detectors are applied within each moving window over the whole video clip to generate and $K * W$ matrix of concept detection scores. These are transformed to generate feature descriptors as event representations. Classifiers are trained with the feature descriptors to detect events.

occurrence strengths of the confidences. In particular, we explore the following five feature representations:

Max Concept Detection Score(Max): This method selects the maximum detection score C_i^{max} over all sliding windows as the detection confidence of detector φ_i. As a result, a video is mapped to a K-dimensional vector $\mathbf{C}^{max} = (C_1^{max}, C_2^{max}, ..., C_K^{max})$. Since the maximum detection score provides information on the presence of a concept, this feature is useful for some applications such as novel event recognition as discussed in Sec. 2.4.

Statistics of Concept Score(SCS): For some application, knowing the maximum detection score is not enough. We also need the distribution of the scores to model a specific event. Therefore, we further compute the following parameters of the detection scores $(c_{max}, c_{avg}, c_{std})_i = (max_j(c_{ij}), \frac{1}{W}\Sigma_j c_{ij}, \frac{1}{W}\Sigma_j (c_{ij} - c_{avg})^2)$

Bag of Concepts(BoC): Akin to the bag of words descriptors used for visual word like features, a bag of concepts feature measures the frequency of occurrence of each concept over the whole video clip. To compute this histogram feature, the SVM output of each concept detector is binarized to represent the presence or absence of each concept in each window.

Co-occurrence Matrix(CoMat): A histogram of pairwise co-occurrences is used to represent the pairwise presence of concepts independent of their temporal distance.

Max Outer Product(MOP): Since concepts represent semantic content in a video, the max value of each concept across the whole video represents the confidence in the presence of a concept in a video. The outer product of the vector of max values of each of the concepts represents both the strength of the presence of each concept (diagonal values) as well as the strength of co-occurrence of pairwise concepts (off-diagonal values): $MOP = \mathbf{C}^{max} \times (\mathbf{C}^{max})^T$.

2.4. Modeling and Recognizing Novel Events

Binary Event Model. Suppose we have n events with training videos, which are known events. Based on the videos from these events, we annotate K concepts (in terms of video segments or keyframes), and train K concept detectors. Thus, we build a K-dimensional concept space. Given this, our problem is how we recognize the other z events which don't have training videos, called novel events. This is very important issue for event retrieval, since web users can be potentially interested in tens of thousands of different events. Collecting a large training set for each event is not feasible.

Given the constructed K-dimensional concept space, we need to define (localize) each novel event in the space. There are no videos for the novel event, but let's assume we know the description of the event in terms of K-concepts. And then we can make a K-dimensional vector with each bin having a binary value indicating the presence or absence of a concept related to this event. This vector is the model of the novel event in the space. For example, in Fig. 3 there is a space consisting of 10 concepts including "kitchen", "person pointing", "person kissing", etc. By mapping "Yes" and "No" to number *1* and *0* respectively, we acquire a 10-dimensional vector $(1, 0, 1, 0, 0, 1, 1, 1, 0, 0) \in \mathbb{C}^{10}$, which is the position of "making a sandwich" in the semantic space. Now, as both the novel event and its videos (by concept detectors) are embedded in a common semantic space, we can tell if a video belongs to an event by computing their semantic similarity in the space. Thus, we can recognize novel events based on its description.

Suppose $\mathbf{C}^y = (c_1^y, c_2^y, ..., c_n^y)$ is the location, estimated from human knowledge, of *Event* y in the concept space, then given an event video $\mathbf{x}$, its event label $y*$ is estimated by $y* = argmax_{y \in Y} \mathcal{S}(\Phi(\mathbf{x}), \mathbf{C}^y)$, where $\mathcal{S}$ is a function measuring the semantic affinity between two points in the space. In our experiments, it is defined using a Gaussian kernel as follows,

$$\mathcal{S}(\mathbf{z}_1, \mathbf{z}_2) = Exp(- \parallel \mathbf{z}_1 - \mathbf{z}_2 \parallel /(2\sigma)). \quad (1)$$

The assumption on novel event recognition is that the novel event can be described by K concepts, which means the novel event shares concepts with other known events. The concept detectors are not necessary to be trained on annotated short video segments (or keyframes) extracted from the n known events. They can be trained on any dataset.

Enhancing Event Model. The manually defined binary

Figure 3. An example of characterizing an event in terms of concepts. Based on common knowledge on "making a sandwich", we can mark its relevant concepts ("Yes") from a set of pre-defined concepts.

event model suffers from two issues. One issue is that a human's knowledge to an event may be subjective and incomplete, thus the embedded position derived from this knowledge will be biased. The other issue is sometimes the binary coding is restrictive and unnatural [23]. Consider for example "wedding ceremony", it is hard to tell if "dancing" should occur in a wedding video. In order to handle these issues, we propose to improve the binary setting of concepts using the semantic similarity between any two concepts. The underlying assumption is that semantically similar concepts co-occur frequently.

The semantic affinity matrix S of concepts can be derived from some knowledge databases, such as WordNet and Wikipedia, or computed from video training examples of known events, say n-1 events. We estimate matrix S from the data $\mathbf{D} = (d_1, d_2, ..., d_M)^T$, an $M \times K$ matrix, where d_i is a K-dimensional vector representing a video in the semantic space. Note that $\mathbf{D}$ does't contain videos from the novel event, say the n-th event. On the other hand, we also treat each column $\mathbf{c}_j \in \mathbb{R}^M$ as a representation of the corresponding concept j. As a result, each entry $S(i,j)$ of the affinity matrix can be estimated using Eq. 1. Having the $K \times K$ symmetric affinity matrix S, we can replace the binary event model vector $\mathbf{y} = (y_1, y_2, ..., y_K) \in \mathbb{C}^K$ with $\mathbf{y}^T \times S$. Indeed, each concept location y_i is updated by $y_i = \sum_j y_j \times S(j,i)$. It means the concepts with $y_i = 1$ are copied between concepts and weighted according to their semantic affinity.

3. Semantic Concept Recounting

A video event is a complex activity occurring at a specific time. Such a video may contain a lot of irrelevant information. Thus, for each recognized event occurrence in a video, the goal of recounting is to describe the details of the occurrence. The recounting includes key observations regarding the scene, people, objects, and actions pertaining to the event occurrence. Such recounting provides user a semantic description that is useful to perform further analysis. As concept features that we use by definition contain semantic information, concept features are more appropriate for recounting purpose than low-level features.

As our event classification is based on SVMs, we present an approach to perform the recounting in the context of SVMs. Given the feature vector $\mathbf{x} \in R^n$ where n is the feature dimension, the decision function $h(\mathbf{x})$ is represented as $h(\mathbf{x}) = \sum_{l=1}^{m} \alpha_l K(\mathbf{x}, \mathbf{x}_l) + b$, where $\mathbf{x}_l$ is a support vector, K is the kernel function, α_l is the signed weight of $\mathbf{x}_l$ and b is the bias. If the kernel function has the form of $K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} f(x_i, z_i)$, where f can be any function and x_i, z_i are the values of the i-th dimension of $\mathbf{x}$ and $\mathbf{z}$. For example, intersection kernel satisfies such a form as $f_{INT} = \min(x,z)$. Linear kernel also follows this form. Now the decision function can be rewritten as follows,

$$h(\mathbf{x}) = \sum_{i=1}^{n} \sum_{l=1}^{m} \alpha_l f(x_i, z_i^l) + b.$$

Suppose $h_i(x) = \sum_{l=1}^{m} \alpha_l f(x_i, z_i^l)$, we can decompose the decision value as,

$$h(\mathbf{x}) = \sum_{i=1}^{n} h(x_i) + b, \tag{2}$$

where $h_i(x)$ encodes how much the i-th dimension/feature contributes to the final decision value. As each dimension has semantic information, we can retrieve the important evidences by sorting $h_i(x)$. We have shown our recounting approach in the context of SVMs. In fact, the approach can be applied to any *additive* classifiers as in Eq.2, which cover a wide spectrum of classification approaches.

4. Experiments and Discussion

4.1. Event Dataset and Experiment Setup

We evaluate CBER based event recognition on the TRECVID MED11 open source dataset [1], which includes over 45,000 YouTube-like videos with about 18-minutes length per video in average, i.e., over 1400 hours of video data approximately. This dataset contains 15 named event categories, such as "making a sandwich", "parkour", "Change a vehicle tire", and more as listed in Fig. 5, plus other unnamed negative events (UNE) other than the 15 events. All the videos are unconstrained videos.

We selected about $3,500$ videos as our development dataset (DEV), which includes about 2062 videos from Event 01-15 plus 1438 UNE videos. The number of videos for each event ranges from around 110 to 170. The rest of videos of MED11 dataset are used as our testing data, including 1751 videos from Event 01-15, ranging from 80 to 170 for each event. From the DEV data, we annotated about $4,000$ short video segments to develop 81 action concept detectors, and $5,000$ keyframes to develop the scene and object detectors, as discussed in Sec. 2.1. The main computation cost comes from low-level feature generation, which is common to all models. The results are reported in terms of Average Precision for the top ranked 1000 videos.

4.2. Experimental Results

We conducted the following experiments to demonstrate the advantages of CBER for event classification: (I) experiments showing the effectiveness of various semantic features; (II) experiments verifying CBER has better general-

	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15	mAP
SCS	0.467	0.123	0.535	0.533	0.327	0.213	0.338	0.707	0.232	0.149	0.240	0.434	0.577	0.300	0.249	***0.362***
CoMat	0.498	0.149	0.569	0.601	0.271	0.191	0.306	0.544	0.309	0.179	0.301	0.252	0.499	0.350	0.208	***0.348***
BOC	0.486	0.129	0.491	0.578	0.241	0.194	0.242	0.458	0.295	0.156	0.244	0.217	0.553	0.293	0.157	***0.316***
MOP	0.494	0.121	0.481	0.555	0.414	0.233	0.327	0.739	0.264	0.214	0.238	0.302	0.491	0.218	0.311	***0.360***
Combined	0.511	0.146	0.584	0.618	0.378	0.273	0.351	0.741	0.291	0.203	0.285	0.399	0.566	0.368	0.335	***0.403***
SparseBase	0.457	0.121	0.537	0.480	0.327	0.184	0.274	0.652	0.221	0.124	0.219	0.262	0.464	0.262	0.227	***0.321***

Figure 4. The Average Precision of event recognition using various semantic features. The last column lists the mean Average Precision (mAP) over all events. The last row shows the performance of Sparse-Base approach, which is proposed in [31] for action recognition in still images.

E01	Attempting a Board Trick	**E02**	Feeding Animals	**E03**	Landing a Fish	**E04**	Wedding Ceremony	**E05**	Woodworking
E06	Birthday Party	**E07**	Changing a Tire	**E08**	Flash mob Gathering	**E09**	Getting Vehicle Unstuck	**E10**	Grooming an Animal
E11	Making a Sandwich	**E12**	Parade	**E13**	Parkour	**E14**	Repairing an Appliance	**E15**	Sewing Project

Figure 5. The fifteen events defined in MED11 dataset.

ization capability; (III) experiments on novel event recognition; (IV) experiments demonstrating event recounting on CBER for semantic event understanding.

I. Effectiveness of semantic features. As discussed in Sec. 2.3, the concept space enables the design of various semantic features, which capture a variety of complementary properties of concepts for an event. We train a binary SVM classifier for each event on DEV, and test it on the testing dataset. For BOC and CoMat, we use histogram intersection kernel, and RBF kernel for SCS, MOP and SparseBases [31]. The default parameters are used for SVM. Fig. 4 shows the Average Precision [31] for each event using different features. Although MOP and SCS obtain better performance in terms of mean AP, no specific feature wins for all events. This observation also means the features are complementary to each other. Combining all features, we achieve about 4% improvment in terms of mean AP. Moreover, each feature has a big variance in performance across events due to various diversities in visual contents of each event. For example, it is difficult to recognize "feeding an animal" and "grooming an animal", while easier to recognize "attempting a board trick" and "flash mob gathering". The direct comparison between semantic concept features and low-level features is exploited in next experiment **II**.

There is little work studying our problem on event classification. But as aforementioned, the goal of our semantic features is similar to the sparse bases learning approach, which is originally proposed in [31] to capture high order relationship between attributes and parts for action recognition in still images. So we apply this approach over our concept space too. A variety of bases are learnt from around 2,000 video segments, and the best performance is reported with 500 learned bases in Fig. 4 (i.e., the last row). Overall, MOP and SCS features work better than SparseBases. We conjecture the data is so noisy that it is hard to learn robust sparse bases, while our features deal with noise better. Note that we can not conduct the same experiments on the image dataset as [31] does, because our features are generated from videos.

II. Generalization capability. In this experiment, we evaluate event classification as a function of number of positive training examples for both concept-base (CBER) and low-level feature based (LLFeat) event representation. For a fair comparison, we extract DTF [30] and STIP [15] low-level dynamic features, which are also used to train our concept detectors, and represent an event as a bag of visual words using these features. The reported results are the fusion of that of DTF and STIP. Other classification setups are same to that of experiment I. Fig. 6 shows the performance comparison in terms of mean AP over all events (i.e., (a)), as well as the comparison in terms of AP for three other events (i.e., (b-d)). Obviously, CBER achieves much better performance than LLFeat when the number of positive training videos is small, which means semantic concept features are more generalized as compared to low-level features. This is because features with semantic meanings are more helpful for recognition. These observations are especially important for event retrieval in cases when large training samples are not available while concept detectors are available.

On the other hand, we notice the performance of CBER is worse than that of LLFeat when the number of positive training examples increases across some point. Ideally, CBER is able to accurately recognize events with a few representative examples due to its good generalization capability. In other words, increasing number of training examples does not necessarily gain better performance. In practice, however, due to the information loss caused by lower quality concept detectors, CBER also needs more training examples to gain more event information. But its performance grows slowly with increased training examples. In contrast, LLFeat directly acquires information from low-level features, so it performs better when a significant number of training examples are available. So in practice, improving concept detector can boost CBER's performance.

III. Recognizing novel events. In these experiments, we selected 81 action concepts to form the concept space. Given a novel event, we define an 81-dimensional model vector, in which each bin holds a binary value indicating the presence/absence of the corresponding concept, as one example shown in Fig. 3. (*The description for all events in terms of*

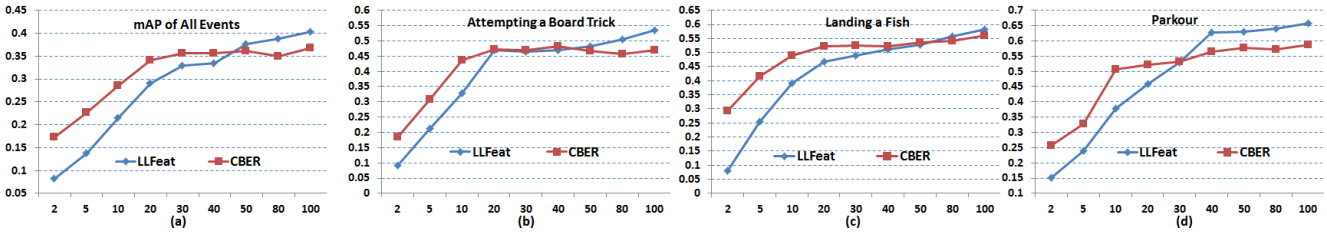

Figure 6. The performance of event recognition as a function of number of positive training examples. Red and blue curves depict the results of CBER and LLFeat (low-level feature) respectively. Sub-figure (a) shows the performance comparison in terms of mean Average Precision over all events. Sub-figures (b), (c) and (d) illustrate the results corresponding to three selected events.

concepts is included in the supplemental materials.) Note that in our experiments when Event A is selected as the novel event, any video segments from A in DEV will be excluded from the concept detector training, which means no concept detectors have seen information of A. In this way, Event A is novel to the system. In each experiment, we treated one event as the novel event, and returned a confidence value (estimated by Eq.1) for each of test videos. We repeat this experiment for all events.

Fig. 7 (a) shows the mean APs of three approaches. "Binary", as our baseline, represents the directly usage of human-defined binary event models for recognition. "SS-Based" is our approach of using semantic similarities between concepts to enhance binary event models. The semantic affinity matrix S of concepts is estimated from the DEV data **excluding videos from the novel event**. All of them are evaluated on the testing data. As a comparison, we also use the sparse bases to enhance binary models. As [31] does, we project both binary event models and testing data into the same sparse base before recognition. Overall, "SS-based" improves about 2.5% in terms of mAP than "Binary", which is significant considering the baseline is only 12.3%. "SparseBases" does not improve the baseline. It is interesting that CBER based zero-shot approaches obtained comparative or event better results, as compared to the event recognition of LLFeat with 5 positive and more than 1000 negative examples (i.e., "LLFeat-05" in Fig. 7 (a)). This is reasonable because classifier trained on LLFeat generalize event models from only five examples, while the model generalization capability of CBER is endowed by the pre-trained concept detectors, which serve as the bridge of information sharing across different events.

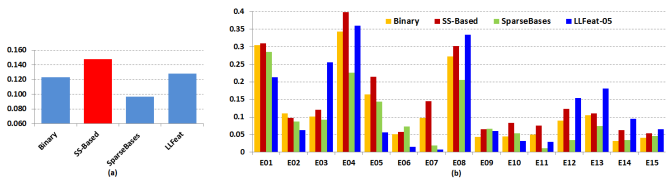

Figure 7. (a) The mean AP of recognizing novel events. As a comparison, last column shows the trained based approach using low-level features with 5 positive training examples. SparseBases represents the approach in [31]. (b) The performance comparison between various zero-shot learning approaches for each event. Note that "LLFeat-05" is training-based method using LLFeat with 5 positive training examples.

Although it seems CBER does not have training examples for the novel event, it actually gains more information from the known events. Fig. 7 (b) lists the detailed results for each of the events with different approaches. We can see some events, such as "Wedding ceremony" and "Flash mob gathering", which share more concept detectors with other events, achieve much better performance than the ones with less information sharing with other events.

IV. Event Recounting. We show two recounting examples, i.e. for Event 1 and Event 4, in Fig. 8. BOC concept feature is used to train the event classifier (a SVM classifier with intersection kernel). For a video that is correctly classified, we show the top recounted concepts in Fig. 8 (a) and (b). The concepts are shown together with the confidence contributed to the final event decision. The center frame of the sliding window with maximum detection confidence is shown as the exemplar. The red bounding box shows the detected object concept. It is worth noting that, although our concept detection contains a lot of noise in terms of both false alarm and miss detection, the top recounted concepts are all sort of relevant to the event.

To show Fig. 8 (a) and (b) are not by chance, we demonstrate the overall recounted concepts from both events. We collect the top recounted concepts from all positive videos and create a histogram shown in Fig. 8 (c) and (d) . Note that, the concepts that have significant hits are all relevant to the event to some extent.

5. Conclusion

In this paper we proposed a novel representation of events defined in terms of a semantic space of action, scene and object concepts. We demonstrated that our CBER representation requires fewer number of training examples versus low-level features for similar event classification performance. We also demonstrate that CBER based approach has good classification performance on new events that can be defined over the same concept space with zero training examples. Our approach was evaluated on the challenging TRECVID MED11 [1], which is completely unconstrained, and large scale dataset from open sources. Beyond event classification, we also presented a method to recount the most relevant semantic concepts for a video. The recountings indicate that the CBER approach accurately captures the semantic concepts related to videos

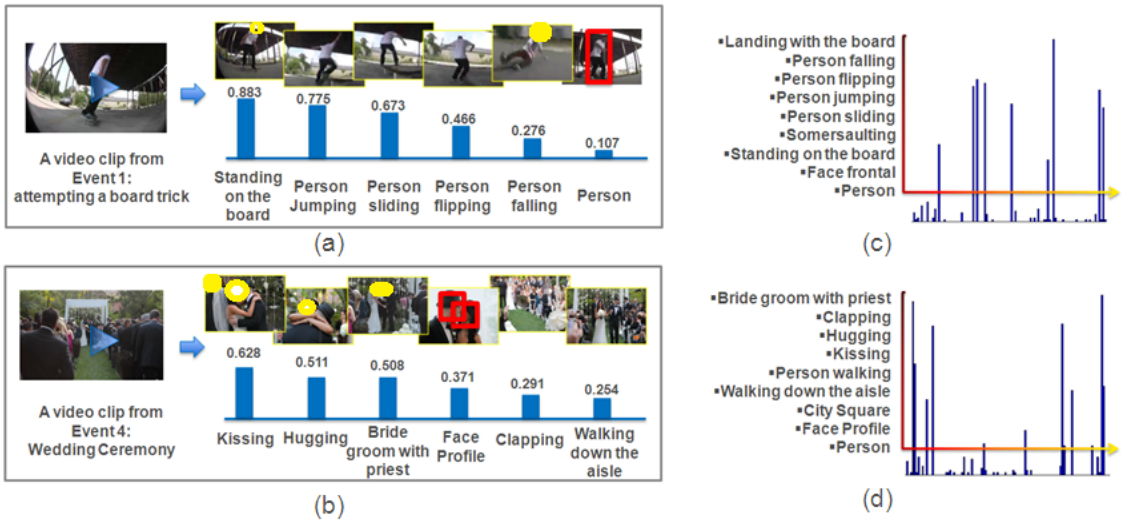

Figure 8. Event recounting examples. Given a video clip that is classified as a positive, the top six recounted concepts are shown in (a)(b). In (c)/(d), we show a list of the most recounted concepts among all positive videos of Event 01 or Event 04. The x axis shows the concept indices and y axis shows the number of positive videos in that of a concept ranks top. The list of concepts is in the same of order that the arrow line intersects the histogram bars.

"in the wild".

Acknowledgment This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not with-standing any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the of?cial policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

References

[1] http://www.nist.gov/itl/iad/mig/med11.cfm. 2, 3, 5, 7

[2] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Tran. on Multimedia*, 9(2):257–267, 2007. 2

[3] L. Duan, D. Xu, I. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010. 2

[4] S. Ebadollahi, L. Xie, S. Chang, and J. Smith. Visual event detection using multi-dimensional concept dynamics. In *ICME*, 2006. 2

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 3

[7] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, 2010. 2

[8] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *CVIU*, 96(2):129–162, 2004. 2

[9] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: combining multiple features for human action recognition. In *ECCV*, 2010. 2

[10] N. Imran, J. Liu, J. Luo, and M. Shah. Event recognition from photo collections via pagerank. In *ACM MM*, 2009. 2

[11] Y. Jiang, X. Zeng, and et al. Columbia-ucf trecvid 2010 multimedia event detection: combining multiple modalities, contextual concepts, and temporal matching. In *TRECVID*, 2010. 2

[12] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *IVC*, 14(8):609–615, 1996. 2

[13] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005. 2

[14] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2, 3

[15] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 3, 6

[16] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[17] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009. 1, 2

[18] J. Liu, Y. Yang, and M. Shah. Learning semantic features for action recognition via diffusion map. *CVIU*, 116(3), 2012. 1

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[20] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM MM*, 2008. 2

[21] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[22] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, (99), 2012. 2

[23] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 2, 5

[24] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? Semantic relatedness for knowledge transfer. In *CVPR*, 2010. 2

[25] B. Siddiquie, R. Feris, and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011. 2

[26] J. Stottinger, J. Uijlings, A. Pandey, N. Sebe, and F. Giunchiglia. (unseen) event recognition via semantic compositionality. In *CVPR*, 2012. 2

[27] A. Tamrakar and et al. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012. 2

[28] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. 2010. 2

[29] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *CVPR*, 2009. 2

[30] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 3, 6

[31] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2, 6, 7