

Dynamic Concept Composition for Zero-Example Event Detection

Xiaojun Chang¹, Yi Yang¹, Guodong Long¹, Chengqi Zhang¹ and Alexander G. Hauptmann²

¹Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney.

²Language Technologies Institute, Carnegie Mellon University.

{cxj273, yee.i.yang}@gmail.com, {guodong.long, chengqi.zhang}@uts.edu.au, alex@cs.cmu.edu

Abstract

In this paper, we focus on automatically detecting events in unconstrained videos without the use of any visual training exemplars. In principle, zero-shot learning makes it possible to train an event detection model based on the assumption that events (*e.g.* *birthday party*) can be described by multiple mid-level semantic concepts (*e.g.* “blowing candle”, “birthday cake”). Towards this goal, we first pre-train a bundle of concept classifiers using data from other sources. Then we evaluate the semantic correlation of each concept w.r.t. the event of interest and pick up the relevant concept classifiers, which are applied on all test videos to get multiple prediction score vectors. While most existing systems combine the predictions of the concept classifiers with fixed weights, we propose to learn the optimal weights of the concept classifiers for each testing video by exploring a set of online available videos with free-form text descriptions of their content. To validate the effectiveness of the proposed approach, we have conducted extensive experiments on the latest TRECVID MEDTest 2014, MEDTest 2013 and CCV dataset. The experimental results confirm the superiority of the proposed approach.

Introduction

In multimedia event detection (MED), a large number of *unseen* videos is presented and the learning algorithm must rank them according to their likelihood of containing an event of interest, such as *rock climbing* or *attempting a bike trick*. Compared to traditional recognition of visual concepts (*e.g.* actions, scenes, objects, *etc.*), event detection is more challenging for the following reasons. First, an event is a higher level abstraction of video sequences than a concept and consists of multiple concepts. For example, an event, say *birthday part*, can be described by multiple concepts (*e.g.* “birthday cake”, “blowing candle”, *etc.*) Second, an event spreads over the entire duration of long videos while a concept can be detected in a shorter video sequence or even in a single frame. As the first important step towards automatic categorization, recognition, search, indexing and retrieval, MED has attracted

more and more research attention in the computer vision and multimedia communities (Chen et al. 2014; Chang et al. 2015b; Cheng et al. 2014; Lai et al. 2014; Li et al. 2013; Chang et al. 2015c; Ma et al. 2012; Yan et al. 2015a; 2015b).

Current state-of-the-art systems for event detection first seek a compact representation of the video using feature extraction and encoding with a pre-trained codebook (Lowe 2004; Bay, Tuytelaars, and Gool 2006; Wang and Schmid 2013). With labeled training data, sophisticated statistical classifiers, such as support vector machines (SVM), are then applied on top to yield predictions. With sufficient labeled training examples, these systems have achieved remarkable performance in the past (Lai et al. 2014; Sun and Nevatia 2014; Li et al. 2013; Cheng et al. 2014). However, MED faces the severe data-scarcity challenge: only very few, perhaps even none, positive training samples are available for some events, and the performance degrades dramatically once the number of training samples falls short. Reflecting this challenge, the National Institute of Standards and Technology (NIST) hosts an annual competition on a variety of retrieval tasks, of which the Zero-Exemplar Multimedia Event Detection (0Ex MED in short) in TRECVID 2013 (TRECVID 2013) and 2014 (TRECVID 2014) has received considerable attention. Promising progress (Dalton, Allan, and Mirajkar 2013; Habibian, van de Sande, and Snoek 2013; Habibian, Mensink, and Snoek 2014; Chang et al. 2015a) has been made in this direction, but further improvement is still anticipated.

In this paper, we aim to detect complex event without any labeled training data for the event of interest. Following previous work on zero-shot learning (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009), we regard an event as compositions of multiple mid-level semantic concepts. These semantic concept classifiers are shared among events and can be trained using other resources. We then learn a skip-gram model (Mikolov et al. 2013) to assess the semantic correlation of the event description and the pre-trained vocabulary of concepts, based on which we automatically select the most relevant concepts to each event of interest. This step is carried out without any visual training data at all. Such concept bundle view of event also aligns with the cognitive science literature, where humans are found to conceive objects as bundles of attributes (Roach and Lloyd

1978). The concept prediction scores on the testing videos are combined to obtain a final ranking of the presence of the event of interest. However, most existing zero-shot event detection systems aggregate the prediction scores of the concept classifiers with fixed weights. Obviously, this assumes all the predictions of a concept classifier share the same weight and fails to consider the differences of the classifier’s prediction capability on individual testing videos. A concept classifier, in fact, does have different prediction capability on different testing videos, where some videos are correctly predicted while others are not. Therefore, instead of using a fixed weight for each concept classifier, a promising alternative is to estimate the specific weight for each testing video to alleviate the individual prediction errors from the imperfect concept classifiers and achieve robust detection result.

The problem of learning specific weights of all the semantic concept classifiers for each testing video is challenging in the following aspects: Firstly, it is unclear how to determine the specific weights for the unlabeled testing video since no label information can be used. Secondly, to get a robust detection result, we need to maximally ensure positive videos have higher scores than negative videos in the final ranking list. Note that the goal of event detection is to rank the positive videos above negative ones. To this end, we propose to learn the optimal weights for each testing video by exploring a set of online available videos with free-form text descriptions of their content. Meanwhile, we directly enforce that positive testing videos have the highest aggregated scores in the final result.

The main building blocks of the proposed approach for zero-example event detection can be described as follows. We first rank the semantic concepts for each event of interest using the skip-gram model, based on which the relevant concept classifiers are selected. Then following (Liu et al. 2013; Lai et al. 2015), we define the aggregation process as an information propagation procedure which propagates the weights learned on individual on-line available videos to the individual unlabeled testing videos, which enforces visually similar videos have similar aggregated scores and offers the capability to infer weights for the testing videos. To step further, we use the L_∞ norm infinite push constraint to minimize the number of positive videos ranked below the highest-scored negative videos, which ensures most positive videos have higher aggregated scores than negative videos. In this way, we learn the optimal weights for each testing video and push positive videos to rank above negative videos as possible.

Contributions: To summarize, we make the following contributions in this work:

1. We propose a novel approach for zero example event detection to learn the optimal weights of related concept classifiers for each testing video by exploring a set of on-line available videos with free-form text descriptions of their content.
2. Infinity push SVM has been incorporated to ensure most positive videos have the highest aggregated scores in the final prediction results.
3. We conduct extensive experiments on three real video

datasets (namely MEDTest 2014 dataset, MEDTest 2013 dataset and CCV_{sub}), and achieve state-of-the-art performances.

Related Works

Complex event detection on unconstrained web videos has attracted wide attention in the field of multimedia and computer vision. Significant progress has been made in the past (Lai et al. 2014; Li et al. 2013; Sun and Nevatia 2014). A decent video event detection system usually consists of a good feature extraction module and a highly effective classification module (such as large margin support machines and kernel methods). Various low-level features (static, audio, etc.) already achieve good performances under the bag-of-words representation. Further improvements are obtained by aggregating complementary features in the video level, such as coding (Boureau et al. 2010; Perronnin, Sánchez, and Mensink 2010) and pooling (Cao et al. 2012). It is observed that with enough labeled training data, superb performance can be obtained. However, when the number of positive training videos falls short, the detection performance drops dramatically. In this work, we focus on the more challenging zero-exemplar setting where *no* labeled training videos for the event of interest are provided.

Our work is inspired by the general zero-shot learning framework (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009; Mensink, Gavves, and Snoek 2014), which arises from practical considerations such as the tremendous cost of acquiring labeled data and the constant need of dealing with dynamic and evolving real-world object categories. On the event detection side, recent works have begun to explore intermediate semantic concepts (Chang et al. 2015a), and achieved limited success on the zero-exemplar setting (Dalton, Allan, and Mirajkar 2013; Habibiian, van de Sande, and Snoek 2013; Habibiian, Mensink, and Snoek 2014) also considered selecting more informative concepts. However, none of these works consider discovering the optimal weights of different concept classifiers for each *individual* testing video.

The Proposed Approach

In this paper, we focus on the challenging zero-exemplar event detection problem. In a nutshell, we are given a sequence of unseen testing videos and also the event description, but without any labeled training data for the event of interest. The goal is to rank the testing videos so that positive videos (those contain the event of interest) are ranked above negatives. With this goal in mind, we first associate a query event with related semantic concepts that are pre-trained using other sources. Then we aggregate the individual concept prediction scores using the proposed dynamic composition approach.

Semantic Query Generation

Our work is built upon the observation that each event can be described by multiple *semantic concepts*. For example, the *marriage proposal* event can be attributed to several concepts, such as “ring” (object), “kissing” (action), “kneeling

down” (action) and “cheering” (acoustic). Since semantic concepts are shared among different events and each concept classifier can be trained independently using data from other sources, zero-example event detection can be achieved by combining the relevant concept prediction scores. Different from the pioneer work (Lampert, Nickisch, and Harmeling 2009), which largely relies on human knowledge to decompose classes (events) into attributes (concepts), our goal is to automatically evaluate the semantic similarity between the event of interest and the concepts, based on which we select the relevant concepts for each event.

Events come with textual side informatin, *e.g.*, an event name or a short description. For example, the event *dog show* in the TRECVID MEDTest 2014 (TRECVID 2014) is defined as “a competitive exhibition of dogs”. With the availability of a pre-trained vocabulary of concept classifiers, we can evaluate the semantic correlation between the query event and each individual concepts. Specifically, we learn a skip-gram model (Mikolov et al. 2013) using the English Wikipedia dump¹. The skip-gram model infers a D -dimensional vector space representation by fitting the joint probability of the co-occurrence of surrounding contexts on large unstructured text data, and places semantically similar words near each other in the embedding vector space. Thus it is able to capture a large number of precise syntactic and semantic word relationships. For short phrases consisting of multiple words (*e.g.*, event descriptions), we simply average its word-vector representations. After properly normalizing the respective word-vectors, we compute the cosine distance of the event description and all individual concepts, resulting in a correlation vector $\mathbf{w} \in [0, 1]^m$, where w_k measures a priori relevance of the k -th concept and the event of interest. Based on the relevance vector, we select the most informative concept classifiers for each event.

Weak Label Generation

According to the NIST standard, we utilize the TRECVID MED *research* dataset to explore the optimal weights for the testing videos. All the videos in the *research* set come with a sentence of description, summarizing the contents contained in the videos. Note that all the videos in the *research* set has no event-level label information. We further collect a set of videos with free-form text descriptions of their content, which are widely available online in websites such as YouTube and NetFlix.

As the descriptions of the videos in both *research* set and online website are very noisy, we apply standard natural language processing (NLP) techniques to clean up the annotations, including removal of the common stop words and stemming to normalize word inflections.

Similar to the steps of SQG, we measure the semantic correlation between the cleaned sentence and each concept description, and use it as a weak label for each concept. In the next section, we will learn the optimal aggregation weights for the individual testing video by exploiting the supervision information with weak label, which accounts for the differences in the concept classifiers’ prediction abilities on the

individual testing video, and hence achieve robust aggregation results.

Dynamic Composition

Up until now, we have l videos with weak label and u testing videos. We propose to learn an aggregation function $f_i(\mathbf{s}_i) = \mathbf{w}_i^T \mathbf{s}_i$ for each testing video ($i = 1, \dots, l + u$), where $\mathbf{w} = [w_1^1, \dots, w_1^m]^T$ is a non-negative aggregation weight vector with w_i^j being the aggregation weight of s_i^j . Clearly, it is straightforward for us to learn the optimal aggregation weights for the videos in the collected set based on the weak label information. However, it is challenging to derive the optimal aggregation weights for the *unlabeled* videos since no label information is available.

To achieve this goal, we build our model based on the local smoothness property in graph-based semi-supervised learning, which assumes visually similar videos have comparable labels within a local region of the sample space (Liu et al. 2013; Lai et al. 2015). To exploit this property, we build a nearest neighbor graph based on the low level features (In this paper, the improved dense trajectories feature (Wang and Schmid 2013) is used). Note that other graph construction methods (Li et al. 2015) can be used in our approach. We put an edge between the video x_i and the video x_j if x_i is among k nearest neighbors of x_j or x_j is among k nearest neighbors of x_i . The weight of the edge G_{ij} in the graph matrix \mathbf{G} is calculated by

$$\mathbf{G}_{ij} = \begin{cases} \exp(-\frac{d(x_i, x_j)}{\sigma}), & \text{if } i \in \mathcal{N}_k(j) \text{ or } j \in \mathcal{N}_k(i), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathcal{N}_k(i)$ denotes the index set of the k nearest neighbors of video x_i (in this paper, we empirically set k to 5), in which $d(x_i, x_j)$ denotes the distance calculated based on the low-level feature. σ is the radius parameter of the Gaussian function. In our experiment, following (Liu et al. 2013; Lai et al. 2015) we set it as the mean value of all pairwise average distances among the videos.

Our dynamic composition model learns the optimal aggregation weight vectors $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{l+u}]$ for both the collected videos and the testing videos by solving the following problem:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \Omega(\mathbf{W}) + \lambda \ell(\{f_i\}_{i=1}^l; \mathfrak{P}, \mathfrak{N}), \\ \text{s.t.} \quad & \mathbf{w}_i \geq 0, i = 1, \dots, l + u, \end{aligned} \quad (2)$$

where λ is a trade-off parameter among the two competing terms, \mathfrak{P} denotes positive examples and \mathfrak{N} denotes negative examples. The normalized weight matrix is defined as $\mathbf{E} = \mathbf{U}^{-\frac{1}{2}} \mathbf{G} \mathbf{U}^{-\frac{1}{2}}$, where \mathbf{U} is diagonal matrix with its element defined as $U_{ii} = \sum_{j=1}^{l+u} \mathbf{G}_{ij}$. Noting that visually similar videos are presumed to have the same label information and thus the same composition weights, we propose a regularization term to propagate the aggregation weights (Belkin and Niyogi 2001):

$$\Omega(\mathbf{W}) = \sum_{i,j=1}^{l+u} (\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{E}_{ij}, \quad (3)$$

¹<http://dumps.wikimedia.org/enwiki/>

By simple algebra formulation, the regularization term Equation (3) can be rewritten as:

$$\Omega(\mathbf{W}) = (\pi(\mathbf{W}))^T \mathbf{L}(\pi(\mathbf{W})), \quad (4)$$

where $\mathbf{L} = \mathbf{I} - \mathbf{E}$ is the graph laplacian. $\pi(\mathbf{W})$ is a vector calculated by $\pi(\mathbf{W}) = ((\mathbf{W}^T \mathbf{S}) \circ \mathbf{I})\mathbf{1}$, where \circ is the Hadamard matrix product. To make visually similar videos have comparable aggregation scores and similar aggregation weights, we propose to minimize (4), which enforces a smooth aggregation score propagation over the constructed graph structure.

To step further, we incorporate an infinite push loss function (Rakotomamonjy 2012) to achieve robust aggregation result. The goal of infinite push loss function is to minimize the number of positive videos which are ranked below the highest scored negative videos. The infinite push loss function has shown promising performance for event detection problem in (Chang et al. 2015c). In fact, the number of positive videos ranked below the highest scored negative videos equals to the maximum number of positive videos ranked below any negative videos. Hence, we define it as follows:

$$\ell(\{f_i\}_{i=1}^l; \mathfrak{P}, \mathfrak{N}) = \max_{j \in \mathfrak{N}} \left(\frac{1}{p} \sum_{i \in \mathfrak{P}} I_{f_i(\mathbf{s}_i^+) < f_j(\mathbf{s}_j^-)} \right), \quad (5)$$

where I is the indicator function whose value is 1 if $f_i(\mathbf{s}_i^+) < f_j(\mathbf{s}_j^-)$ and 0 otherwise. The maximum operator over j equals to calculating the l_∞ -norm of a vector consisting of n entries, each of which corresponds to one value based on j in the parentheses of Equation (5). By minimizing this penalty, positive videos tend to score higher than any negative videos. This essentially ensures positive videos have higher combined scores than the negatives, leading to more accurate combined results.

For computational tractability we upper bound the discrete 0-1 loss $I(\delta < 0)$ by the *convex* hinge loss $(1 - \delta)_+$, where as usual $(\delta)_+ := \max(\delta, 0)$ is the positive part. Since we usually pay more attention, if not exclusively, to the top of the rank list, we focus on minimizing the maximum ranking error among all negative exemplars $j \in \mathfrak{N}$:

$$\ell(\{f_i\}_{i=1}^l; \mathfrak{P}, \mathfrak{N}) = \max_{j \in \mathfrak{N}} \left(\frac{1}{p} \sum_{i \in \mathfrak{P}} (1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+ \right), \quad (6)$$

Finally, the objective function can be written as:

$$\begin{aligned} & \min_{\mathbf{W}} (\pi(\mathbf{W}))^T \mathbf{L}(\pi(\mathbf{W})) \\ & + \lambda \max_{j \in \mathfrak{N}} \left(\frac{1}{p} \sum_{i \in \mathfrak{P}} (1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+ \right), \quad (7) \\ & \text{s.t. } \mathbf{w}_i \geq 0, i = 1, \dots, l + u. \end{aligned}$$

The above objective function is convex, and thus can achieve the global optimum. Thanks to the proximal map (Yu 2013), we employ the faster ADMM proposed in (Chang et al. 2015c) for efficient solution.

Experiments

In this section, we conduct extensive experiments to validate the proposed **Dynamic Concept Composition** for zero-exemplar event detection task, abbreviated as **DCC**.

Experiment Setup

Dataset: To evaluate the effectiveness of the proposed approach, we conduct extensive experiments on the following three large-scale event detection datasets:

- **TRECVID MEDTest 2014 dataset (TRECVID 2014):** This dataset has been introduced by the NIST for all participants in the TRECVID competition and research community to perform experiments on. There are in total 20 events, whose description can be found in (TRECVID 2014). We use the official test split released by the NIST, and strictly follow its standard procedure (TRECVID 2014). To be more specific, we detect each event *separately*, treating each of them as a binary classification/ranking problem.
- **TRECVID MEDTest 2013 dataset (TRECVID 2013):** The settings of MEDTest 2013 dataset is similar to MEDTest 2014, with 10 of their 20 events overlapping.
- **Columbia Consumer Video dataset (Jiang et al. 2011):** The official Columbia Consumer Video dataset contains 9,317 videos in 20 different categories, including scenes like “beach”, objects like “cat”, and events like “basketball” and “parade”. Since the goal of this work is to *search complex events*, we only use the 15 event categories.

According to the standard of the NIST, each event is detected separately and the performance of event detection is evaluated using the mean Average Precision (mAP).

Concept Detectors: 3,135 concept detectors are pre-trained using TRECVID SIN dataset (346 categories) (Over et al. 2014; Jiang et al. 2014), Google sports (478 categories) (Karpathy et al. 2014; Jiang et al. 2014), UCF101 dataset (101 categories) (Soomro, Zamir, and Shah 2012; Jiang et al. 2014), YFCC dataset (609 categories) (YFC ; Jiang et al. 2014) and DIY dataset (1601 categories) (Yu, Jiang, and Hauptmann 2014; Jiang et al. 2014). The improved dense trajectory features (including trajectory, HOG, HOF and MBH) are first extracted using the code of (Wang and Schmid 2013) and encode them with the Fisher vector representation (Perronnin, Sánchez, and Mensink 2010). Following (Wang and Schmid 2013), the dimension of each descriptor is first reduced by a factor of 2 and then use 256 components to generate the Fisher vectors. Then, on top of the extracted low-level features, the cascade SVM (Graf et al. 2004) is trained for each concept detector.

Competitors: We compare the proposed approach with the following alternatives: 1). Prim (Habibian, Mensink, and Snoek 2014): Primitive concepts, separately trained. 2). Sel (Mazloom et al. 2013): A subset of primitive concepts that are more informative for each event. 3). Bi (Rastegari et al. 2013): Bi-concepts discovered in (Rastegari et al. 2013). 4). OR (Habibian, Mensink, and Snoek 2014): Boolean OR combinations of Prim concepts. 5). Fu (Habibian, Mensink, and Snoek 2014): Boolean AND/OR combinations of Prim

Table 1: Experiment results for 0Ex event detection on MEDTest 2014, MEDTest 2013, and CCV_{sub}. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

MEDTest 2014								
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	DCC
E021	2.12	2.98	2.64	3.89	3.97	3.12	4.64	6.37
E022	0.75	0.97	0.83	1.36	1.49	1.15	1.48	2.85
E023	33.86	36.94	35.23	39.18	40.87	38.68	41.78	44.26
E024	2.64	3.75	3.02	4.66	4.92	4.11	4.87	6.12
E025	0.54	0.76	0.62	0.97	1.39	0.84	1.01	1.26
E026	0.96	1.59	1.32	2.41	2.96	1.96	2.65	4.23
E027	11.21	13.64	12.48	15.93	16.26	15.12	16.47	19.63
E028	0.79	0.67	1.06	1.57	1.95	1.72	2.25	4.04
E029	8.43	10.68	12.21	14.01	14.85	13.19	14.75	17.69
E030	0.35	0.63	0.48	0.91	0.96	0.36	0.48	0.52
E031	32.78	53.19	45.87	69.52	69.66	67.49	72.64	77.45
E032	3.12	5.88	4.37	8.12	8.45	7.54	8.65	11.38
E033	15.25	20.19	18.54	22.14	22.23	21.53	23.26	26.64
E034	0.28	0.47	0.41	0.71	0.75	0.53	0.76	0.94
E035	9.26	13.28	11.09	16.53	16.68	15.82	18.65	21.78
E036	1.87	2.63	2.14	3.15	3.39	2.88	3.76	5.47
E037	2.16	4.52	3.81	6.84	6.88	5.42	6.83	8.45
E038	0.66	0.74	0.58	0.99	1.16	0.85	1.12	2.89
E039	0.36	0.57	0.42	0.69	0.77	0.64	0.85	2.26
E040	0.65	0.98	0.72	1.57	1.57	1.24	1.76	3.12
mean	6.40	9.55	7.89	10.76	11.05	10.21	11.44	13.37
MEDTest 2013								
mean	7.07	7.94	6.92	9.45	9.88	8.43	9.96	12.64
CCV _{sub}								
mean	19.05	19.40	20.25	21.16	21.89	23.08	23.87	24.36

concepts, w/o concept refinement. 6). Bor: The Borda rank aggregation with equal weights on the discovered semantic concepts. 7). PCF (Chang et al. 2015a): The pair-comparison framework is incorporated for zero-shot event detection.

Zero-exemplar event detection

We report the full experimental results on the TRECVID MEDTest 2014 dataset in Table 1 and also a summary on the MEDTest 2013 dataset and CCV_{sub}. From the experimental results shown in Table 1, the proposed algorithm, **DCC**, performs better than the other approaches with a large margin (13.37% vs 11.44% achieved by PCF). The proposed approach gets significant improvement on some vents, such as *Dog Show* (E023), *Rock Climbing* (E027), *Beekeeping* (E031) and *Non-motorized Vehicle Repair* (E033). By analyzing the discovered concepts for these events, we find that their classifiers are very discriminative and reliable. For example, for the event *Rock Climbing*, we discovered the concepts “Sport climbing”, “Person climbing” and “Bouldering”, which are the most informative concepts for *Rock climbing* in the concept vocabulary. Figure 1 illustrates the top retrieved results on the *non-motorized vehicle repair* event. To save space, we only show the top 4 compared algorithms (OR, Fu, PCF and DCC). It is clear that videos retrieved by the proposed DCC are more accurate and visually coherent.

We make the following observations from the results shown in Table 1: 1). Sel significantly outperform Prim with mAP of 9.55% vs 6.40% on MEDTest 2014, which indi-

cates that selecting the most discriminative concepts generally improves detection performance than naively using all the concepts. 2). Comparing the results of Bi, OR, Fu and Bor, we verify that the selected informative concept classifiers are not equally important for event detection. It is beneficial to derivate the weights for each concept classifier. By treating the concept classifiers differentially, better performance is achieved. 3). Comparing the results of DCC with the other alternatives, we observe that learning optimal weights of concept classifiers for each testing video significantly improves performance of event detection. This confirms the importance of dynamic concept composition.

We made similar observations from the results on the TRECVID MEDTest 2013 dataset and CCV_{sub} dataset.

Extension to few-exemplar event detection

The proposed zero-example event detection framework can also be used for few-exemplar event detection: we aggregate the concept classifiers and the supervised classifier using the dynamic concept composition approach. In this section, experiments are conducted to demonstrate the benefit of this hybrid approach. Table 2 summarizes the mAP on both the MEDTest 2014 and 2013 datasets, while Figure 2 compares the performance event-wise.

According to the NIST standard, we consider the 10 Ex setting, where 10 positive videos are given for each event of interest. The improved dense trajectories feature (Wang and Schmid 2013) is extracted, on top of which an SVM classifier is trained. It is worthwhile to note that our DCC

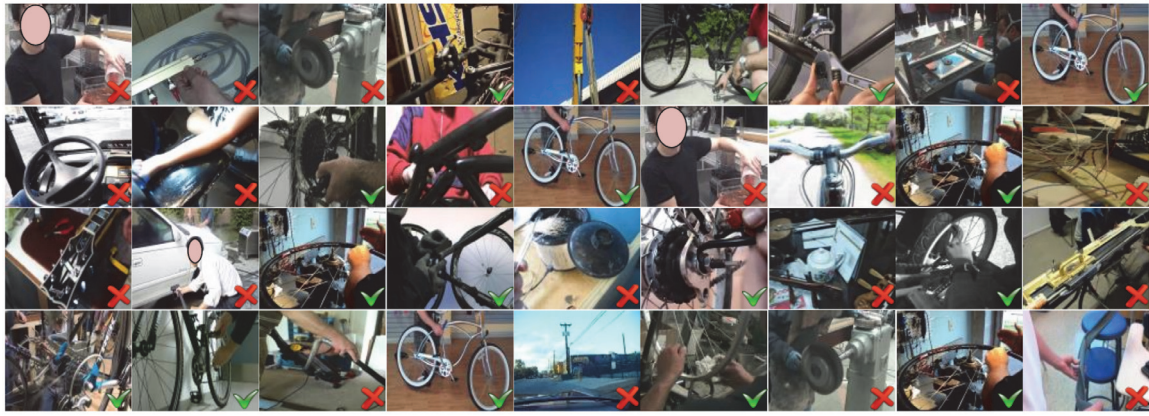


Figure 1: Top ranked videos for the event *non-motorized vehicle repair*. From top to below: OR, Fu, PCF and DCC. True/false labels (provided by NIST) are marked in the lower-right of each frame.

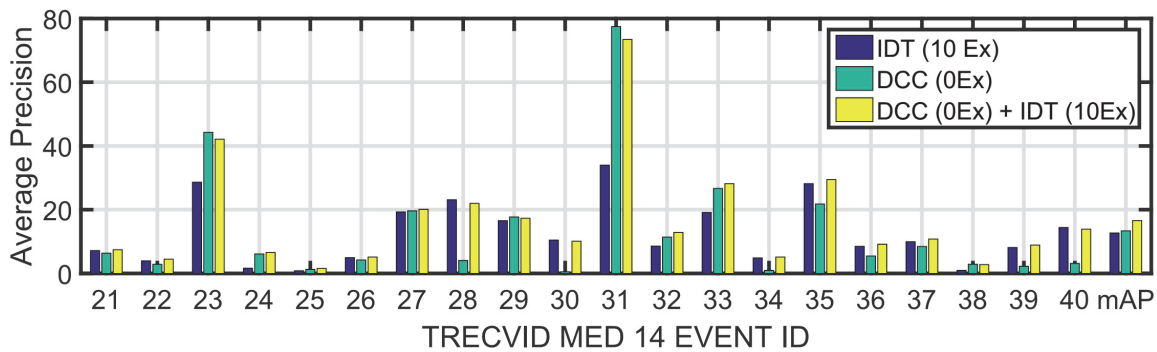


Figure 2: Performance comparison of IDT, DCC, and the hybrid of IDT and DCC.

Table 2: Few-exemplar results on MED14 and MED13.

DCC (0Ex)	13.37	12.64
IDT (10Ex)	13.92	18.08
DCC (0Ex) + IDT (10Ex)	16.37	19.24

which had no labeled training data can get comparable results with the supervised classifier (mAP 13.37% vs 13.92% on MEDTest 2014). This demonstrates that proper aggregation with optimal weights for each testing video can get promising results for event detection.

We compare DCC with IDT event-wise in Figure 2. From the results we can see that our DCC outperforms the supervised IDT on multiple events, namely E023, E024, E031 and E033. To be specific, on the event *Beekeeping* (E031), DCC significantly outperform the supervised IDT (77.45% vs 33.92%). This is not surprising, since DCC significantly benefited from the presence of informative and reliable concepts such as “apiary bee house” and “honeycomb” on the particular *Beekeeping* event.

Finally we combine DCC with IDT to get the final few-example event detection result. This improves the performance from 13.92% to 16.98% on the MEDTest 2014 dataset. As expected, the gain obtained from such simple

hybrid diminishes when combining with more sophisticated methods. Overall, the results clearly demonstrate the utility of our framework even in the few-exemplar setting.

Conclusions

To address the challenging task of zero-exemplar or few-exemplar event detection, we proposed to learn the optimal weights for each testing video by exploring the collected videos from other sources. Data-driven word embedding models were used to seek the relevance of the concepts to the event of interest. To further derive the optimal weights of the concept classifiers for each testing video, we have proposed a novel dynamic concept composition method by exploiting the textual information of the collected videos. Extensive experiments are conducted on three real video datasets. The experimental results confirm the efficiency of the proposed approach.

Acknowledgment

This paper was partially supported by the ARC DECRA project DE130101311, partially supported by the ARC discovery project DP150103008, partially supported by the US Department of Defense, U. S. Army Research Office (W911NF-13-1-0277) and by the National Science Foundation under Grant No. IIS-1251187. The U.S. Government is

authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

- Bay, H.; Tuytelaars, T.; and Gool, L. J. V. 2006. SURF: speeded up robust features. In *ECCV*.
- Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*.
- Boureau, Y.; Bach, F.; LeCun, Y.; and Ponce, J. 2010. Learning mid-level features for recognition. In *CVPR*.
- Cao, L.; Mu, Y.; Natsev, A.; Chang, S.; Hua, G.; and Smith, J. R. 2012. Scene aligned pooling for complex video recognition. In *ECCV*, 688–701.
- Chang, X.; Yang, Y.; Hauptmann, A. G.; Xing, E. P.; and Yu, Y. 2015a. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*.
- Chang, X.; Yang, Y.; Xing, E. P.; and Yu, Y.-L. 2015b. Complex event detection using semantic saliency and nearly-isotonic SVM. In *ICML*.
- Chang, X.; Yu, Y.; Yang, Y.; and Hauptmann, A. G. 2015c. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*.
- Chen, J.; Cui, Y.; Ye, G.; Liu, D.; and Chang, S. 2014. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*.
- Cheng, Y.; Fan, Q.; Pankanti, S.; and Choudhary, A. N. 2014. Temporal sequence modeling for video event detection. In *CVPR*.
- Dalton, J.; Allan, J.; and Mirajkar, P. 2013. Zero-shot video retrieval using content and concepts. In *CIKM*.
- Graf, H. P.; Cosatto, E.; Bottou, L.; Durdanovic, I.; and Vapnik, V. 2004. Parallel support vector machines: The cascade SVM. In *NIPS*.
- Habibian, A.; Mensink, T.; and Snoek, C. G. M. 2014. Composite concept discovery for zero-shot video event detection. In *ICMR*.
- Habibian, A.; van de Sande, K. E. A.; and Snoek, C. G. M. 2013. Recommendations for video event recognition using concept vocabularies. In *ICMR*.
- Jiang, Y.; Ye, G.; Chang, S.; Ellis, D. P. W.; and Loui, A. C. 2011. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *ICMR*.
- Jiang, L.; Meng, D.; Yu, S.; Lan, Z.; Shan, S.; and Hauptmann, A. G. 2014. Self-paced learning with diversity. In *NIPS*.
- Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- Lai, K.; Yu, F. X.; Chen, M.; and Chang, S. 2014. Video event detection by inferring temporal instance labels. In *CVPR*.
- Lai, K.; Liu, D.; Chang, S.; and Chen, M. 2015. Learning sample specific weights for late fusion. *IEEE Transactions on Image Processing* 24(9):2772–2783.
- Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.
- Li, W.; Yu, Q.; Divakaran, A.; and Vasconcelos, N. 2013. Dynamic pooling for complex event recognition. In *ICCV*.
- Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*.
- Liu, D.; Lai, K.; Ye, G.; Chen, M.; and Chang, S. 2013. Sample-specific late fusion for visual category recognition. In *CVPR*.
- Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.
- Ma, Z.; Yang, Y.; Cai, Y.; Sebe, N.; and Hauptmann, A. G. 2012. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM MM*.
- Mazloom, M.; Gavves, E.; van de Sande, K. E. A.; and Snoek, C. 2013. Searching informative concept banks for video event detection. In *ICMR*.
- Mensink, T.; Gavves, E.; and Snoek, C. G. M. 2014. COSTA: co-occurrence statistics for zero-shot classification. In *CVPR*.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
- Over, P.; Awad, G.; Michel, M.; Fiscus, J.; Sanders, G.; Kraaij, W.; Smeaton, A. F.; and Queenot, G. 2014. Trecvid 2014 an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*.
- Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NIPS*.
- Perronnin, F.; Sánchez, J.; and Mensink, T. 2010. Improving the Fisher kernel for large-scale image classification. In *ECCV*.
- Rakotomamonjy, A. 2012. Sparse support vector infinite push. In *ICML*.
- Rastegari, M.; Diba, A.; Parikh, D.; and Farhadi, A. 2013. Multi-attribute queries: To merge or not to merge? In *CVPR*.
- Roach, E., and Lloyd, B. B. 1978. *Cognition and categorization*.
- Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild.
- Sun, C., and Nevatia, R. 2014. DISCOVER: discovering important segments for classification of video events and recounting. In *CVPR*.
- TRECVID. 2013. Multimedia event detection. <http://www.nist.gov/itl/iad/mig/med13.cfm>.
- TRECVID. 2014. Multimedia event detection. <http://www.nist.gov/itl/iad/mig/med14.cfm>.
- Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *ICCV*.
- Yan, Y.; Yang, Y.; Meng, D.; Liu, G.; Tong, W.; Hauptmann, A. G.; and Sebe, N. 2015a. Event oriented dictionary learning for complex event detection. *IEEE Transactions on Image Processing* 24(6):1867–1878.
- Yan, Y.; Yang, Y.; Shen, H.; Meng, D.; Liu, G.; Hauptmann, A. G.; and Sebe, N. 2015b. Complex event detection via event oriented dictionary learning. In *AAAI*.
- The YFCC dataset. <http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>.
- Yu, S.; Jiang, L.; and Hauptmann, A. G. 2014. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*.
- Yu, Y. 2013. On decomposing the proximal map. In *NIPS*.