Zero-Shot Video Event Detection with High-Order Semantic Concept Discovery and Matching

Yang Jin, Wenhao Jiang, Yi Yang and Yadong Mu

***Abstract*—Multimedia event detection aims to precisely retrieve videos that contain complex semantic events from a large pool. This work addresses this task under a zero-shot setting, where only brief event-specific textural information (such as event names, a few descriptive sentences, etc.) is known yet none positive video example is provided. Mainstream approaches to tackling this task are middle-level semantic concept-based, where meticulously-crafted concept banks (*e.g.*, LSCOM) are adopted. We argue that these concept banks are still inadequate facing video semantic complexity. Existing semantic concepts are essentially first-order, mainly designed for atomic objects, scenes or human actions, etc. This work advocates the utilization of high-order concepts (such as subject-predicate-object triplets or adjective-object). The main contributions are two-fold. First, we harvest a comprehensive albeit compact high-order concept library through distilling information from three large public datasets (MS-COCO, Visual Genome, and Kinetics-600), mainly related to visual relations and human-object interactions. Secondly, zero-shot events are often only briefly and partially described via textual input. The resultant semantic ambiguity makes the pursuit of the most indicative high-order concepts challenging. We thus design a novel query-expanding scheme that enriches ambiguous event-specific keywords by searching over either large common knowledge bases (*e.g.*, WikiHow) or top-ranked webpages retrieved from modern search engines. This way sets up a more faithful connection between zero-shot events and high-order concepts. To our best knowledge, this is the first work that strives for concept-based video search beyond first-order concepts. Extensive experiments have been conducted on several large video benchmarks (TRECVID 2013, TRECVID 2014, and ActivityNet-1.3). The evaluations clearly demonstrate the superiority of our constructed high-order concept library and its complementariness to existing concepts.**

***Index Terms*—Multimedia event detection, zero-shot learning, high-order concept**

Fig. 1. Illustration of the deficiency of first-order information in zero-shot event detection. We list two complex events: "Attempting a bike trick" and "Horse riding competition". The first presented video is a true example of the query event. The second one is a highly-ranked false example retrieved by leveraging first-order concepts. The left column represents the relevant first-order concepts for the query event. See the main text for more explanation.

I. Introduction

With the increasing ubiquity of video-capturing devices and social media, an enormous number of user-generated videos have been uploaded to the Internet. These videos often capture daily-life events with varying semantic complexities. To intelligently understand and search events presented in a video, the task of Multimedia Event Detection (MED) [1] has been proposed and attracted tremendous attention from researchers. Given a specific event, the goal of MED is to retrieve most semantically-related videos from a large-scale multimedia corpus. Compared with traditional semantic concept detection task, MED is more challenging with major complications from the compositional essence of a multimedia event. A typical high-level multimedia event is composed of a large number of atomic objects, scenes, human-to-object or human-to-human interactions, etc. For instance, the event "marriage proposal" can be evident by identifying a few indicative concepts, such as ring (object), restaurant (scene), and hugging (action).

In recent years, a variety of approaches [2], [3], [4], [5], [6] have been proposed to address the MED task. A majority of existing works have assumed the availability of sufficient annotated positive video examples for all interested events. For example, in a typical setting of the TRECVID MED competition [1], [7], a toolkit with 100+ positive videos per event and corresponding event-level textual description is provided for the model-training purpose. The sufficiency of training data ensures the good generalization ability of learned models. However, conducting video annotation is tedious and time-consuming, which hinders the wide coverage of annotated video events. In real-world scenarios, users of a MED system are often allowed to search an arbitrary event. Considering the tremendous number of possible event categories, it is infeasible to train a separate detector for each event in advance. In

Yang Jin is with Center for Big Data Research, Peking University, Beijing, China. Yadong Mu is with Wangxuan Institute of Computer Technology, Peking University, Beijing, China. Email: jiny@stu.pku.edu.cn, myd@pku.edu.cn.

Yi Yang is with the Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Ultimo, Sydney, Australia. Email: Yi.Yang@uts.edu.au.

Wenhao Jiang is with Tencent AI Lab, Shenzhen, China. Email: wenhao-jiang@tencent.com.

All correspondence is to Yadong Mu (myd@pku.edu.cn).

fact, a large body of user-generated videos on social media are typically unlabeled or come with weak noisy accompanying text. Therefore, detecting events without leveraging any labeled training data, called zero-shot multimedia event detection [8], [9], [10], has been strongly motivated and serves as a promising technique in video analysis.

In current literature, the dominating models for zero-shot multimedia event detection are semantic concept-based. Since no annotated video examples are available, the major challenge of zero-shot multimedia event detection is bridging the (typically succinct) event description and diverse video content. In practice, existing mainstream methods represent a video with a few middle-level attributes (concepts). Importantly, these semantic concepts are often pre-trained using external data sources and expected to generalize well to many application domains. The querying events are also mapped to the concept library, possibly with varying relatedness scores for different concepts. This way formulates the task of zero-shot multimedia event detection as concept mapping and matching and thereby enables detecting video events with zero-effort of data labeling. For a variety of events, this method proves to be highly feasible since the pre-trained concepts are usually meticulously chosen for comprehensively describing some complex events in a collective manner.

There are two main deficiencies in all existing works that clearly motivate our work:

First, most of them detect events based on first-order concepts like objects and scenes. High-order information such as the visual relationships between objects has been rarely explored in current works. We argue that first-order concepts lack enough semantic information and predictive power for detecting a complicated event. Some examples are shown in Figure 1. Even all the objects and scene concepts (people, bicycle, and outdoor) relevant to the querying event ("Attempting a bike trick") are precisely detected, the retrieved videos are irrelevant. Obviously, the system cannot distinguish relevant videos without leveraging discriminative high-order information (person-jump-bicycle).

The second deficiency is how to find semantically relevant concepts, which are crucial for concept-based methods. Irrelevant concepts will bring poor detection performance. Existing methods tackle concept selection by matching event name or pre-defined event description with all concepts. However, both event name and pre-defined description have inadequacy for concept selection. Event names usually cannot comprehensively represent the semantics of an event, resulting in the omission of some related concepts. Although the event description provides some vital clues for event detection, it needs to be defined by users in advance, which is very inconvenient and cumbersome in real-world scenarios.

To solve the first problem mentioned above, this work constructs a comprehensive concept library of high-order concepts. This idea is inspired by EventNet proposed in [11]. In specific, there are two different types of concepts in our concept library: visual relationship and complex human actions. Especially, for the visual relationship concepts, we resort to two large public datasets: MS-COCO [12] and Visual Genome [13]. Both of them involve detailed descriptions of image contents. By parsing this textual information with the help of NLP tools [14], we obtain a bulk of relationship triplets. After that, we propose a clustering-based approach to mine semantic concepts. As for human action concepts, we adopt the large human action dataset: Kinetics [15], which are widely used in action recognition. Finally, we train a detector for each concept.

Additionally, we propose a practical zero-shot detection framework to address the concept selection problem. The whole framework can be decomposed into three independent modules and only take event name as input. The first component harnesses large common knowledge bases to expand the query event, and thereby we can obtain a more comprehensive semantic expression of an event. The second component matches the expansion results with our high-order concept library by calculating the semantic similarities between them and picks out related concepts for the query event. Matching based on the expansion results helps to discover concepts that cannot be obtained from the event name. The third component leverages the selected concepts to retrieve the most relevant videos from the video corpus. Since no pre-defined information is required, the entire detection framework can be applied to any unseen event category.

Briefly, our contributions can be summarized as below:

1) As the first work of its kind, we explore high-order concepts in the zero-shot video event detection task. An extensive high-order concept library of the visual relationship and human action is constructed and proven effective in our experiments.

2) We propose an effective framework to detect complex multimedia events. The framework expands the event query by searching large common knowledge bases and can be applied to any unseen event. Competitive performance on TRECVID [7], [16] and ActivityNet-1.3 [17] benchmarks prove the superiority of our approach.

The remainder of this paper is organized as follows: Section II reviews the related works on MED task and analyzes the deficiencies of these methods. Section III provides the details of constructing high-order concept library. Section IV presents our concept-based framework for zero-shot event detection. The experimental analysis and performance are presented in Section V. Finally, the paper is concluded in Section VI.

II. Related Work

Several lines of research are highly related to our work:

Zero-shot learning: Zero-shot learning (ZSL) aims to recognize or detect unseen samples during testing. The idea is to learn from seen samples and then transfer the knowledge to unseen samples with the use of semantic information. There are two types of semantic information widely used in ZSL, including common attributes and word embeddings of seen and unseen classes. Common attributes are descriptions of samples, including size, shape, color, etc. Attributed methods such as [18], [19] pre-learned a set of attribute classifiers from seen samples and recognized unseen samples based on their attribute representations. Since it is time-consuming to train each attribute classifier independently, Akata *et al.* [20]

proposed an attribute label embedding approach that takes all attributes as a whole to tackle this problem. Although attribute-based ZSL methods have gained promising results, common attributes are usually defined by human experts. To reduce manual annotation of common attributes, there are some ZSL methods such as [21], [22] based on unsupervised word embeddings. For example, Bucher *et al.*[23] proposed an approach that generates visual features from word embedding to tackle the zero-shot semantic segmentation task.

Zero-shot Multimedia Event Detection: Video event detection (or multimedia event detection) [24], [25], [26], [27] aims to retrieve videos based on semantic similarity to the given event description. Event detection systems usually first extract and quantify features to get the video feature representation, then training classifiers with labeled data [28]. Various features such as SIFT [29], trajectory feature [30] are widely used in event detection methods. With sufficient training data, event detection methods [31], [32], [33], [34] can achieve excellent performance. However, when some events only have few or no positive training examples, the detection performance tends to degrade dramatically. On account of labeled multimedia content is scarce, Ma *et al.* [35] proposed a knowledge adaptation approach that only uses few positive examples. In order to further reduce the dependence on labeled samples, some works [9], [10], [36] focus on zero-shot setting where no labeled training example is provided. Most zero-shot event detection methods are based on the idea that events can be detected with the help of individual concept responses. Ye *et al.* [11] generated concept-based representations of videos based on a large concept library. Chang *et al.* [37] evaluated the semantic correlation of concepts then fuse the individual concept scores with the help of a rank aggregation framework. Li *et al.* [10] introduced a novel integration algorithm to effectively exploit the event-concept relevance by assigning adaptive weights to different concepts. To further enhance the representative capacity of semantic concepts, Zhang *et al.* [36] proposed a well-designed ranking aggregation algorithm. However, all these related works suffer from the semantic insufficiency and fuzziness of first-order concepts, which severely deteriorate the detection performance. By contrast, we try to address this deficiency by exploiting the great potential of high-order concepts and propose a novel concept matching framework.

Concept learning: Visual concept detection is a vital task in the computer vision field. [38] investigated a higher-order pooling strategy that aggregates over co-occurrences of visual objects. [39] tried to utilize external knowledge to expand the concepts detected by the visual classifier. In recent years, some recent development [40], [41] tended to use scene-graph to represent the high-order concept information in Images. Furthermore, there has been much work [9], [42] that explored the concepts detection in the zero-shot scenario. Generally, Complicated multimedia events can be composed of several middle-interpreted semantic concepts. Consequently, concept learning methods have been widely concerned by researchers in the multimedia field. A lot of researches [43], [44] investigated the semantic concepts for detecting events in video data. Inspired by the achievements of previous studies, our work discovers high-order concepts based on three large public datasets, which precisely reflect the semantic information of multimedia events.

III. Construct concept library

As mentioned above, in most cases, first-order concepts (e.g., objects and scenes) lack enough representative capacity. Therefore, we plan to leverage high-order information. To this end, a comprehensive large concept library is constructed, which contains two types of concepts: visual relationship and human action. Specifically, for the action concept, we directly adopt Kinetics [45], which is a large human action video dataset. For the relationship concept, we resort to two large image datasets: MS-COCO [12] and Visual Genome [13], which contain detailed descriptions for interactions and relationships between objects in an image. By fully mining this visual information, a bulk of relationship concepts are generated. Next, we will describe the construction procedures in detail.

A. Discovering Relationship Triplet

Visual relationships between objects offer a comprehensive visual content understanding beyond objects. We accomplish the construction of a high-level concept library by mining existing visual-text data corpus rather than building it from scratch. The chosen datasets include Visual Genome and MS-COCO. The former has already provided manually-annotated relationship triplets for each image. However, in the MS-COCO dataset, only five short textual captions are available for each image. It is thus desired to devise a scheme for distilling representative triplets from the raw captions. To this end, we utilize Stanford CoreNLP tools [14] to perform syntactic parsing on each image description. For each image caption, a dependency parse tree is generated that reflects the grammatical relationships among different sentence components. Phrases with ⟨*subject-predicate-object*⟩ syntax in the dependency tree are identified and extracted as potential relationship triplets. Overall, each image is found to be associated with an average of roughly 10 relationships in the MS-COCO dataset and 18 relationships in the Visual Genome dataset.

B. Mining Semantic Concept

Through the previous step, massive triplet-style relationships can be extracted from image captions. However, the collection of raw relationships unavoidably suffer from redundancy and noise, mainly caused by the variety of descriptions in natural language. In particular, different visual relationships, such as ⟨*bicycle, park on, road*⟩ and ⟨*bike, sit on, street*⟩, may actually convey very similar semantic meaning. In order to filter out semantically near-duplicate concepts and obtain a more concise concept library, we here propose an effective albeit simple clustering-based approach, described as below:

Step I: relationship triplet encoding. A commonly-adopted practice for evaluating the affinity among relationship triplets is embedding them into some well-designed semantic space and calculating the distance of semantic vectors, which is typically linear and additive. To achieve this goal, we extract

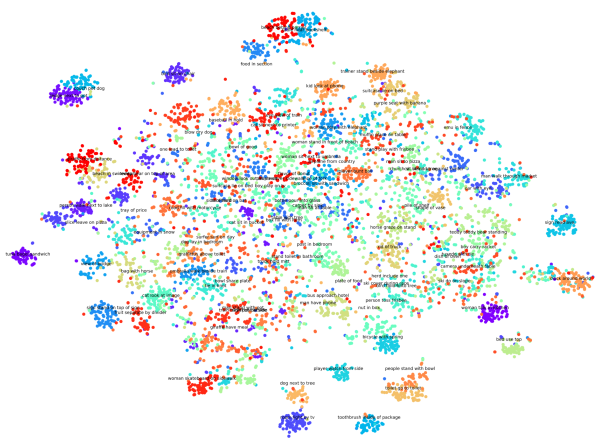


Fig. 2. Visualization of clustering results on the MS-COCO dataset. Due to the limited space, only 100 clusters (represented by different colors) are presented. We also show a specific triplet example for a cluster. Better viewing if enlarging the image.

a language feature f_{li} and a visual feature f_{vi} as the semantic embedding for each triplet r_i.

To capture the f_{li}, we directly borrow BERT [46] as the workhorse. Other alternative contextualized representations beyond BERT may operate similarly, yet we omit more empirical comparisons. Practically, we treat each relationship triplet r_i as a single sentence and feed it into a Google-released version of the BERT model. As a pre-processing step, the input sentence is tokenized by WordPiece tokenizer into an ordered token set $\langle x_1, ..., x_n \rangle$, where x_k is a one-hot encoding of the k-th token. The BERT model outputs a sequence of hidden representations $\langle z_1, ..., z_n \rangle$. Importantly, z_1 (corresponding to [CLS] in implementation) is a vector for conveying all-sentence context in a compressed manner. Therefore, we directly treat z_1 as the language feature $f_{li} \in \mathbb{R}^{768}$ for r_i.

The BERT model is essentially trained for text-processing tasks. The resultant inter-phrase affinity is not ensured to be precisely aligned with the true visual co-occurrence or visual similarity. To compensate for the bias brought by the visual-textual semantic gap, we propose to further extract a visual feature f_{vi} for a triplet r_i. To this end, ResNet-18 [47] pre-trained on ImageNet [48] is harnessed to generate a fixed-length vector v_j for each image, where v_j is from the last global average pooling layer. Then, the visual feature $f_{vi} \in \mathbb{R}^{256}$ is calculated by:

$$f_{vi} = \frac{1}{|I_{r_i}|} \sum_{j \in I_{r_i}} v_j, \tag{1}$$

where I_{r_i} is the image set corresponding to the relationship triplet r_i. After that, we generate the final triplet representation f_i by concatenating language and visual feature:

$$f_i = L_2Norm([f_{li}; f_{vi}]), \tag{2}$$

where $L_2Norm(\cdot)$ denotes L_2 normalization and $f_i \in \mathbb{R}^{1024}$.

Step II: semantic vector clustering. Based on the relationship triplets and corresponding feature representations f_i, we construct a fully-connected affinity graph $G = \{R, E\}$, where R denotes vertices, *i.e.*, relation triplets, and E denotes edges. Let $e_{i,j}$ denote the edge between r_i and r_j, its weight $w_{i,j}$ is calculated by:

$$w_{ij} = \exp\left(-\frac{\|f_i - f_j\|_2^2}{2\sigma^2}\right), \tag{3}$$

where the parameter σ is empirically estimated from the averaged pairwise distances.

To obtain a compact representation of high-order concepts, we conduct a clustering procedure to split all concepts into C groups. Spectral clustering [49] is adopted in our practice since it also investigates the problem from a graph aspect and admits a moderate time complexity (quadratic with respect to the graph node number and linear to C). Importantly, choosing an optimal value for target cluster number C is non-trivial. To address this issue, we follow the intuition that a library with ~ 1000 concepts highly likely strikes a good balance between the usefulness of each individual concept and the richness of data annotation. We further opt for Silhouette Coefficient[1], which can quantify the consistency among different clusters and thus serve as an index for adaptively determining the near-optimal cluster number C^*. Specially, we determine C^* according to:

$$C^* = \arg\max_{C} SC(C), \tag{4}$$

where $SC(\cdot)$ is the Silhouette Coefficient for a specific number of clusters C. Figure 2 presents a visualization of cluster results on the MS-COCO dataset. It can be seen that relationship triplets with similar meanings are grouped to the same cluster with high probability. The action concepts are already well-defined on Kinetics [45] dataset, therefore we omit the clustering step.

C. Training Concept Detectors

Based on the results of semantic vector clustering, all the relationship triplets are mapped into different clusters. For each cluster, we select the data point closest to cluster centroid and adopt its corresponding triplet r_i as a relationship concept. In general, our high-order concept library has 1,299 relationship concepts and 600 action concepts in total.

We train a detector $f_{c_i}(\cdot)$ for each semantic concept c_i. To be specific, for the relationship concept, the ResNet-101 architecture [47] pre-trained on the ImageNet is adopted as a visual feature extractor. We replace the last layer with a fully-connected layer with a sigmoid function and adopt the binary cross-entropy loss to train the relationship concept detector. For the action concept, we utilize the I3D [50] architecture pre-trained on Kinetics to capture actions in the video. These concept detectors will yield probability distributions that reflect the appearance confidence of a given concept. Finally, our high-order concept library can be formulated as $\mathcal{L}$:

$$\mathcal{L} = \{(c_i, f_{c_i}(\cdot))\}_{i=1,2,...,J} \tag{5}$$

where c_i is the textual name of i-th concept, $f_{c_i}(\cdot)$ is the corresponding concept detector, J is the size of our high-order concept library.

[1] https://en.wikipedia.org/wiki/Silhouette_(clustering)

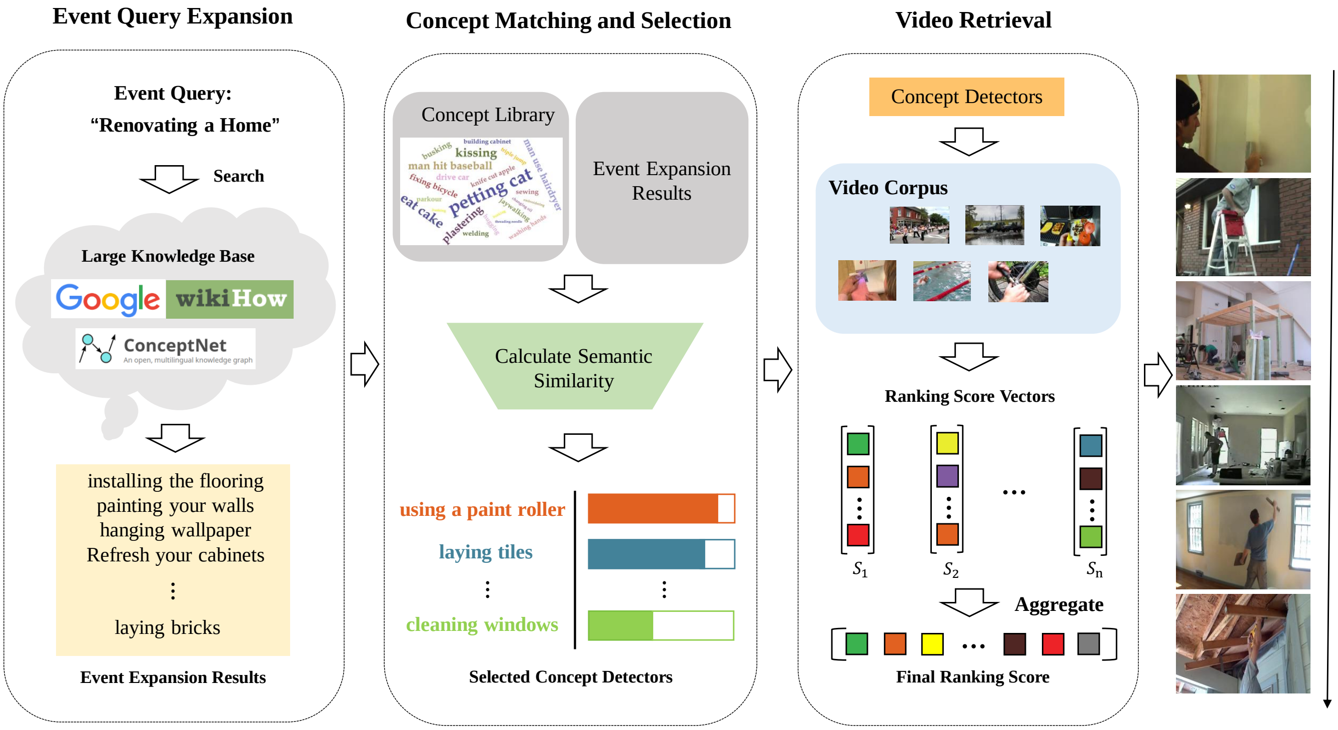


Fig. 3. Illustration of our proposed method for zero-shot video event detection. The overall framework only takes event name as input and contains three key components: Event Query Expansion, Concept Matching and Selection, Video Retrieval. See the main text for more explanation.

IV. Zero-Shot Event Detection

The architecture of our approach for zero-shot video event detection is illustrated in Figure 3. Given a multimedia event e, the overall framework only takes the event name e_ξ as input and retrieves the most related videos from the video corpus. In a nutshell, the complete procedure is split into three major components: *event query expansion*, *concept matching and selection*, and *video retrieval*. The framework firstly expands the event name through searching large external knowledge bases to enrich the textual description of an event query. After that, the results of query expansion are used to match our concept library. The most relevant concepts will be chosen. The last component scores the videos with selected concepts and produces a ranking list.

A. Event Query Expansion

The event name is often brief and ambiguous, lacking sufficient information to describe the event content. Take the event "cleaning an appliance" as an example, the definition of an appliance includes microwave, dishwasher, refrigerator, stove, etc. The cleaning operation usually contains "use a towel", "wash hands", etc. However, all the above-mentioned information is not included in the event name. In order to enrich the event representation and ameliorate the concept mismatch problem in existing works, we expand the origin event name e_ξ to several semantically related terms through external common knowledge bases. The whole expanding procedure is illustrated in Figure 4. This involves the following sub-steps:

Construct event-related corpus. To obtain abundant descriptive information related to event e, we resort to the Internet knowledge bases: WikiHow[2] and Google Search. WikiHow is an online wiki-style community containing extensive how-to articles in regard to daily life. The event name e_ξ is searched on the WikiHow website to get the corresponding articles. Specifically, when the event name is a sub-string of the title of one returned article, it is kept for future use. In addition, we also query event name through the Google search engine and select the top-10 most relevant articles. All these selected articles are crawled from the website, and we can obtain an article set A_e. These articles constitute a corpus that contains useful information such as actions and relationships related to the event.

Generate the query items. Generally, each event name is a short phrase that consists of several words. We extract several query terms from the initial e_ξ. Especially, the Part Of Speech (POS) analysis is firstly conducted on the lemmatized event name. Then, items with nouns, verb+nouns, or adjective+nouns grammatical form are extracted based on the POS tags, and we keep these as the query items for subsequent expansion, which is denoted as Q_e. For example, for the event "Changing a vehicle tire", the generated queries are [*vehicle*, *tire*, *vehicle tire*, *change vehicle tire*].

First-order query expansion. For each query q_e in Q_e, it's firstly expanded with low-level information (*e.g.*, objects and scenes relevant to the event) by searching in two large knowledge bases: WordNet [51] and ConceptNet5 [52]. When expanding from WordNet, only the hyponyms and synonyms

[2] http://www.wikihow.com/Main-Page

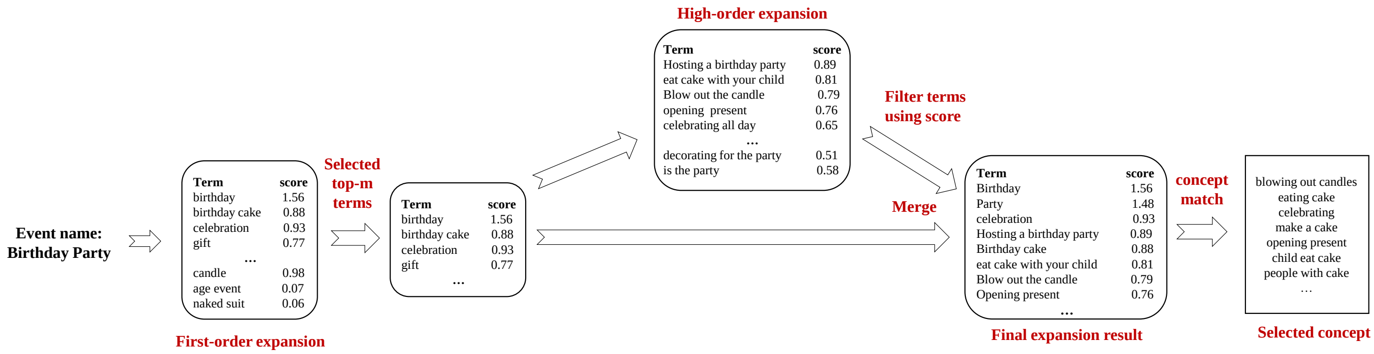


Fig. 4. Illustration of the query expansion procedure of a specific event "birthday party".

of q_e are taken into consideration. With respect to ConceptNet5, all concepts with the relation RelatedTo, CapableOf, AtLocation, UsedFor to the query q_e are selected. Finally, we harvest a wide range of semantically similar items, which is denoted as T_e. The expanded T_e will inevitably include redundant information. To distill contents that have a strong relationship with the event, we assign a correlation score for each element t_e in T_e based on frequency and relatedness. To be specific, the correlation score of t_e is calculated as follows:

$$Corre(t_e) = (1 + tf(t_e)) \cdot R(t_e, q_e) \tag{6}$$

where $tf(t_e)$ is the term frequency of expanded term t_e in article set A_e. $R(t_e, q_e)$ is the relatedness of t_e and its origin query q_e, which is calculated through the ConceptNet REST API[3]. Based on the $Corre(t_e)$, we collect the top-m terms to form the first-order expansions, denoted as $\mathcal{E}_{low,e}$. As an example, the first-order expansions for the event "Marriage Proposal" are "bended knee", "ring", "romantic", etc.

High-order query expansion. Intuitively, high-order phrases contain richer semantic information and thus may ameliorate the concept matching process. Therefore, the query event is further expanded with high-order information by extracting verb-noun phrases from the article set A_e based on the first-order expansion results. For example, based on "ring" in $\mathcal{E}_{low,e}$, we extract phrases such as "put the ring on her finger", "slip the ring box", etc. The detailed expanding procedure is presented in Algorithm 1. The high-order expansion $\mathcal{E}_{high,e}$ is merged with $\mathcal{E}_{low,e}$ to constitute the final semantic representation of event e, denoted as $\mathcal{E}_e$. Each term ζ_j in $\mathcal{E}_e$ has a correlation score $Corre(\zeta_j)$.

B. Concept Matching and Selection

In concept-based video event detection, it is a crucial step to map the user-generated event query to an internal, concept-based representation. To this end, we first evaluate the semantic similarity between expansion results $\mathcal{E}_e$ and concepts in constructed concept library $\mathcal{L}$. For each concept c_i and expansion item ζ_j, the cosine similarity between them is computed by:

$$sim(c_i, \zeta_j) = \frac{\theta(c_i)^T \theta(\zeta_j)}{\|\theta(c_i)\| \|\theta(\zeta_j)\|}, \tag{7}$$

[3]https://github.com/commonsense/conceptnet5/wiki/API

Algorithm 1 High-order Query Expansion

Input: the event name e_ξ; the first-order expansion $\mathcal{E}_{low,e}$; the article set A_e;
Output: the high-order expansion $\mathcal{E}_{high,e}$;

```
E_high,e ← ∅;
Parse all articles in A_e and get a phrase set P;
for ζ in E_low,e do
    for p in P do
        if ζ is the sub-string of p then
            calculate sim(e_ξ, p) using Eq.(7) as Corre(p);
            if Corre(p) > ρ then
                add p to set E_high,e;
return E_high,e
```

where $\theta(\cdot)$ is the embedding function. We adopt the same BERT architecture as $\theta(\cdot)$ as in Section III. Then, the final semantic similarity score s_i between the i-th concept and event e is defined as:

$$s_i = \sum_{j=1}^{|\mathcal{E}_e|} Corre(\zeta_j) \cdot sim(c_i, \zeta_j). \tag{8}$$

The s_i indicates the extent to which the concept c_i is relevant to the query event e. Mainstream approaches choose the top-k concepts by ranking s_i for an event. However, the optimal number of concepts is difficult to determine with respect to different events. As an example, for the event "Tuning musical instrument", the threshold k should be large due to the variety of musical instruments. While for the event "Attempting a bike trick", there will not be so many related concepts. Our method adopts a more robust strategy. The concepts are firstly ordered based s_i in descending order. Then, only concepts whose s_i are among top $\alpha\%$ of all similarity scores are remained. As a result, our method will select the different number of concepts for different query events. The value of α will be explored in subsequent experiments.

C. Video Retrieval

After concept matching and selection for the query event, our framework retrieves the most related videos from the video corpus based on pre-trained concept detectors. The whole video retrieval procedure can be formulated as follows:

$$S_{c_i,v} = f_{c_i}(X), \tag{9}$$

$$S_{e,v} = G(S_{c_1,v}, S_{c_2,v}, ..., S_{c_k,v}), \tag{10}$$

where $f_{c_i}(\cdot)$ is the pre-trained concept detector, X is the visual feature of video. k is the total number of selected concepts, $S_{c_i,v}$ is the concept detection score. $S_{e,v}$ is the final event detection score. $G(\cdot)$ is the aggregation function that combines all the concept detection results.

Recall that we build two kinds of concept detectors in Section III. For the relationship concept, we feed the keyframes of a video to the detector to obtain the detection score on frame level. The keyframes are sampled at the rate of one frame per 2s. For action concept, we decompose a video into segments of length L with $L/5$ overlapping (*e.g.*, L = 250). Then we apply the detector to each segment and obtain the segment-level detection score. A pooling operation is executed on the frame- and segment-level results to get the whole video detection score $S_{c_i,v}$. To make detection score for different concepts on the same scale, we adopt a min-max normalization:

$$S_{c_i,v} = \frac{S_{c_i,v} - \min_v \ S_{c_i,v}}{\max_v \ S_{c_i,v} - \min_v \ S_{c_i,v}}. \tag{11}$$

It is worth noting that both semantic relatedness and discrimination of a concept are crucial for event detection. The semantic relatedness is denoted as s_i, which is calculated in the previous procedure. The discrimination indicates the power of a concept for discerning specific events. Take the event "parking a vehicle" as an example, in our experiment, the concept "driving car" has a high semantic relatedness to this event, but it achieves high scores on a majority of videos. Therefore, it lacks enough discriminative ability for detecting this specific event. We should assign a smaller weight to it. To balance the semantic relatedness and discrimination, we design a weight w_i for each score $S_{c_i,v}$. The w_i has a similar form to TF-IDF [56] weight:

$$w_i = s_i \cdot \log \frac{N}{1 + |\{S_{c_i,v} \geq \delta\}|}, \tag{12}$$

where N is the total number of videos in video corpus, $|\{S_{c_i,v} \geq \delta\}|$ is the amount of the video whose $S_{c_i,v} \geq$ threshold δ. The w_i jointly considers semantic relatedness and discrimination and suppresses the contribution of concept that appears too frequently in all videos. We calculate the final detection score $S_{e,v}$ by $G(\cdot)$:

$$G(S_{c_1,v}, S_{c_2,v}, ..., S_{c_k,v}) = \sum_{i=0}^{k} w_i S_{c_i,v}. \tag{13}$$

By sorting all videos in a video corpus based on $S_{e,v}$, our framework finally returns the event-relevant videos as a response to the user query. Importantly, the concept detection score $S_{c_i,v}$ for any c_i from the concept library can be calculated in an off-line manner. Before a new event query comes, all videos in the database have their concept detection scores computed and properly scaled according to Eq. 11. The runtime computations thus mainly stem from the most relevant concept selection and aggression as in Eq. 13.

V. Experiments

In this section, we conduct extensive experiments on three large video benchmarks. The comparisons with state-of-the-art methods demonstrate the superiority of our constructed high-order concept library and proposed detection framework. Additionally, all source code and deep models for our proposed high-order concept detection have been released for non-commercial free use by the multimedia community. More details can be found at https://github.com/Rain-coder1/video_zsl.

A. Experimental Setup

Since we focus on the zero-shot scenario, the overall detection procedures are conducted without using any positive examples. Next, we will introduce the setup of our experiments.

Dataset. We adopt three large video benchmarks in our experiments. (1) TRECVID 2013 Multimedia Event Detection (MED2013) [7]: It's a large publicly available user-generated video dataset for event detection released by NIST. The whole dataset contains 20 pre-defined complex events. We adopt its official test split, named MED13Test, which includes around 25,000 unconstrained videos. (2) TRECVID 2014 Multimedia Event Detection (MED2014) [16]: Similar to the MED2013 settings, MED14Test contains around 24,000 videos for 20 event categories (10 events overlapping with MED2013 benchmark). (3) ActivityNet-1.3 [17]: To make the experimental results more comprehensive, we include the more recent ActivityNet-1.3 dataset, which contains 19,994 unconstrained videos that cover 200 different complex human activities in daily life. We treat each activity label as an event query and detect it separately from the whole dataset.

Concept Detectors. All the relationship and human action concepts in our constructed high-order concept library $\mathcal{L}$ are adopted in our experiments. In addition, we still leverage two types of low-level concepts (ImageNet and Places [57]) to explore the influence of incorporating concepts at different semantic levels.

Evaluation Metrics. According to the official metric of NIST, each event is detected separately. The whole framework returns a ranked video list as the final detection result. The average precision (AP) is adopted as the evaluation metric to measure the detection performance of each event in the test dataset. Eventually, the mean Average Precision (mAP) of all event classes is calculated to evaluate the overall performance.

B. Performance Comparison

Comparison Methods. To demonstrate the advantage of proposed method, we compare the event detection results on the aforementioned benchmarks with existing state-of-the-art works. For MEDTest 2013 and 2014 benchmarks, the following methods is considered: Prim [8], Sel [58], Bi [59], EventNet [11], Fu [8], PCF [37], DCC [9], TagBook [53], CP [54], EACI [10], I-w2v [60],VSF [55] and GVC [36]. All these baselines are concept-based zero-shot event detection methods. They utilize the first-order semantic concepts such

TABLE I
COMPARISON RESULTS OF DIFFERENT METHODS ON MEDTEST 2013 DATASET. A LARGER MAP INDICATES BETTER PERFORMANCE.

	MEDTest 2013							
Event Name	Prim[8]	EventNet[11]	PCF[37]	TagBook[53]	CP[54]	GVC[36]	VSF[55]	Ours
Birthday party	7.6	9.5	16.3	15.5	15.4	-	**24.6**	23.2
Changing a vehicle tire	1.8	32.3	3.5	33.7	32.0	-	43.9	**58.1**
Flash mob gathering	37.3	0.5	**43.4**	17.4	27.1	-	14.5	6.5
Getting a vehicle unstuck	5.5	1.3	9.6	31.2	**40.6**	-	40.2	8.4
Grooming an animal	0.9	2.1	1.5	20.1	9.5	-	18.7	**29.5**
Making a sandwich	7.9	5.4	9.6	9.9	16.4	-	19.4	**29.4**
Parade	22.4	27.8	35.9	18.5	24.0	-	17.6	**38.5**
Parkour	2.2	18.6	4.5	21.5	11.2	-	26.1	**70.1**
Repairing an appliance	2.5	4.7	5.8	21.1	21.3	-	**39.8**	18.3
Working on a sewing project	1.5	1.1	1.2	9.8	8.9	-	30.8	**50.2**
Attempting a bike trick	2.2	1.1	3.3	6.6	6.1	-	8.8	**18.0**
Cleaning an appliance	0.8	3.4	1.5	2.3	2.6	-	**8.2**	6.8
Dog show	0.1	**46.1**	0.8	20.0	1.1	-	4.0	22.9
Giving directions to a location	2.5	0.1	**4.1**	0.5	0.8	-	0.6	1.6
Marriage proposal	0.2	0.7	0.7	0.3	0.5	-	0.3	**3.8**
Renovating a home	2.3	0.6	4.5	1.8	2.6	-	5.2	**5.4**
Rock climbing	14.7	7.5	**21.3**	2.6	3.6	-	1.6	6.6
Town hall meeting	1.5	**16.7**	3.4	14.8	3.5	-	1.9	15.7
Winning a race without a vehicle	13.6	0.1	19.8	9.9	10.1	-	9.4	**20.6**
Working on a metal crafts project	0.6	0.4	1.2	0.2	1.4	-	1.6	**27.1**
mAP (%)	6.4	8.9	9.6	12.9	11.9	15.3	15.9	**23.1**

as ImageNet (objects) and Places (scenes). Some of these methods consider action concepts like Sports-1M [61] and UCF-101 [62]. However, these datasets only contain sport-related concepts, and are limited by smaller scale, lacking enough discerning power for detecting complex multimedia events. As for ActivityNet-1.3, since there is no related work to study zero-shot event detection task on this benchmark, we construct three baselines based on mainstream approaches (Bi-Concept [59], I-w2v [60] and EACI [10]) for comparison.

Quantitative Analysis. We present the full experimental results on the MED13Test benchmark in Table I and also comparison results on MED14Test in Table II. For fair comparison, all results in Table I and Table II are cited from the original papers. From the shown results, we can find that the proposed method is consistently superior to all the state-of-the-art baselines. Especially for the MED13Test benchmark, our method outperforms the best baseline (VSF) by a large margin, with the mAP increased from 15.9% to 23.1%. Moreover, due to the full use of high-order information, our approach achieves optimal results on most of the event categories. Performances on the rest events are still comparable. The mAP of event "Felling a tree" (Table II) significantly improves compared to other baselines (7.3% v.s. 2.1% achieved by Bi). The reason is that by query expansion, our method discovers some vital information for detecting this event, such as "using a chainsaw". However, the compared baselines detect this event mainly depending on the concept "tree". Obviously, this concept frequently appears in videos and lacks enough discriminative power. Another reason for the better performance of our method is that the selected concepts are highly related to the query event. For example, for "Attempting a bike trick" (Table I), we discover concepts like "jumping bicycle", "people riding a bike" and "falling off bike", which are crucial elements of this event.

The performance on event "Tailgating" (Table II) is not so satisfactory since the query expansion module introduces irrelevant information. This is mainly caused by the ambiguity of its event name. The expansion results for "Tailgating" are like "car collision" or "traffic accident", while the event in MED14Test indicates "tailgate party". Moreover, we can also observe that, for the event "Giving directions to a location" (Table II), all the approaches achieve poor performance. This is due to the fact that none of them take advantage of audio features, which are more discerning than visual features for detecting this event.

We also conduct experiments on ActivityNet-1.3 benchmark, Table III shows the comparison results. Due to the limited space, we only present 20 complex events. From the results, we can observe that our approach achieves significant improvement on this benchmark compared with baselines that leverage conventional low-order concept libraries. Among all the listed baselines, Bi [59] has the worst performance. This is reasonable since it only leverages atom concepts like objects and scenes. In contrast, I-w2v [60] and EACI [10] borrow more action information that contributes to detecting sports-related events (see the last three rows in Table III). Besides, the mAP of our approach on the ActivityNet-1.3 benchmark is much higher than that on TRECVID 2013 and 2014. This is because events on the TRECVID dataset are more complex and have higher semantics compared to ActivityNet-1.3. Moreover, our concept library has some concepts that exactly match events on ActivityNet-1.3.

By summarizing all experimental results, we can conclude that our approach is competitive for the task of zero-shot event detection. It is worth mentioning that unlike numerous existing works that request a pre-defined textual description of the event query, our approach only takes a brief event name as input but performs the best.

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS ON MEDTEST 2014 DATASET. A LARGER MAP INDICATES BETTER PERFORMANCE.

Event Name	MEDTest 2014 Sel[58]	Bi[59]	Fu[8]	PCF[37]	DCC[9]	EACI[10]	I-w2v[60]	GVC[36]	Ours
Attempting a bike trick	3.0	2.6	4.0	4.6	6.4	12.8	6.2	7.9	**18.0**
Cleaning an appliance	1.0	0.8	1.5	1.5	2.9	5.5	6.2	3.6	**8.0**
Dog show	36.9	35.2	40.9	41.8	44.3	65.5	**76.6**	46.8	23.7
Giving directions to a location	3.8	3.0	4.9	4.9	6.1	4.8	0.6	**7.7**	1.7
Marriage proposal	0.8	0.6	1.4	1.0	1.3	1.0	1.0	1.9	**4.2**
Renovating a home	1.6	1.3	3.0	2.7	4.2	5.0	0.2	5.7	**6.3**
Rock climbing	13.6	12.5	16.3	16.5	19.6	20.4	**30.9**	21.5	12.6
Town hall meeting	0.7	1.1	2.0	2.3	4.0	10.7	14.8	5.3	**16.1**
Winning a race without a vehicle	10.7	12.2	14.9	14.8	**17.7**	21.3	2.1	16.8	15.0
Working on a metal crafts project	0.6	0.5	1.0	0.5	0.5	1.3	0.5	0.7	**27.3**
Beekeeping	53.2	45.9	69.7	72.6	77.5	10.0	62.0	79.1	**79.6**
Wedding shower	5.9	4.4	8.5	8.7	**11.4**	6.7	0.5	10.8	2.5
Non-motorized vehicle repair	20.2	18.5	22.2	23.3	26.6	38.6	0.6	28.3	**50.2**
Fixing musical instrument	0.5	0.4	0.8	0.8	0.9	4.7	**14.7**	1.1	9.9
Horse riding competition	13.3	11.1	16.7	18.7	21.8	31.0	11.9	23.1	**39.7**
Felling a tree	2.6	2.1	3.4	3.8	5.5	3.5	4.2	7.2	**7.3**
Parking a vehicle	4.5	3.8	6.9	6.8	8.5	15.3	**22.0**	9.6	2.9
Playing fetch	0.7	0.6	1.2	2.3	2.9	**4.9**	0.1	2.6	3.2
Tailgating	0.6	0.4	0.8	0.9	2.3	11.4	**23.2**	1.9	0.4
Tuning musical instrument	1.0	0.7	1.6	1.8	3.1	5.3	5.2	3.7	**8.5**
mAP (%)	9.6	7.9	11.1	11.4	13.4	14.0	14.2	14.7	**16.8**

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON ACTIVITYNET-1.3 DATASET. A LARGER MAP INDICATES BETTER PERFORMANCE.

Event Name	ActivityNet-1.3 Bi [59]	I-w2v [60]	EACI [10]	Ours
Changing car wheel	19.6	0.5	18.2	**72.8**
Putting on makeup	0.5	4.7	0.8	**43.6**
Making an omelette	6.3	0.3	5.3	**25.4**
BMX	14.1	92.9	**94.7**	75.9
Painting furniture	**10.4**	7.3	8.1	5.4
Assembling bicycle	17.4	18.8	20.2	**67.1**
Mixing drinks	14.6	0.7	0.7	**17.1**
Fixing the roof	8.1	3.5	12.7	**22.7**
Cutting the grass	0.6	8.5	6.4	**37.4**
Trimming branches	0.5	1.8	0.4	**21.3**
Getting a haircut	8.5	0.6	11.9	**70.4**
Cleaning sink	5.7	0.3	1.8	**22.9**
Doing motocross	4.8	**63.6**	56.5	61.0
Hanging wallpaper	0.5	0.8	1.1	**21.0**
Clipping cat claws	**64.5**	0.3	59.8	43.3
Disc dog	9.5	82.4	**91.2**	53.7
Making a cake	7.3	20.6	5.4	**44.6**
Layup drill in basketball	0.4	**33.7**	29.7	25.2
Snow tubing	0.7	**15.3**	13.5	13.4
Doing kickboxing	0.5	38.5	**55.9**	19.7
mAP (%)	8.0	22.6	27.9	**57.6**

Qualitative Analysis. To intuitively present the performance of the proposed method, we visualize the qualitative results of some event examples in Figure 5, including the top-ranked videos and the most related concepts. To save space, we only present ten events and their top-5 ranked videos. It can be seen that the videos retrieved by our method are accurate and visually related to the event query. Besides, the discovered concepts for these events are very reliable and discriminative. For instance, for the event "Grooming an animal", we discover the concepts "bathing dog", "cutting nails", "woman holds a dog", etc., which are crucial clues.

C. Ablation Studies

To isolate the contribution of different parts in our method, we conduct some ablation studies to verify the effectiveness of two key components: *Event Query Expansion*, *Concept Matching and Selection*.

TABLE IV
ABLATION EXPERIMENT RESULTS OF COMBINING DIFFERENT PARTS OF EVENT EXPANSION MODULE ON THREE BENCHMARKS.

Methods	mAP (%) MED13Test	MED14Test	ActivityNet-1.3
Event Name	8.7	6.5	44.6
Only First-order	18.8	13.9	55.3
Only High-order	22.0	16.3	48.5
All	**23.1**	**16.8**	**57.6**

Event Query Expansion. The semantic ambiguity of event name makes matching high-order concepts challenging. Therefore, the query expansion module is leveraged for enriching the textual description of an event query, which is helpful for discovering the most related concepts, especially for the event whose name contains scarce information. To explore the influence of different parts (first-order and high order expansion) in the query expansion module on the overall performance, we ablate it from the entire framework.

The comparison results are listed in Table IV. It can be seen that the detection performance on all three benchmarks has dropped significantly compared with directly using event name (e.g.,from 23.1% to 8.7% on MED13Test). Moreover, both first-order and high-order expansion significantly surpass event name and alleviate the concept mismatch problem, while the latter contributes more to boost detection performance. The reason is that high-order expansion brings some phrases with richer semantic information, which is conducive to matching more relevant high-order concepts. The combination of these

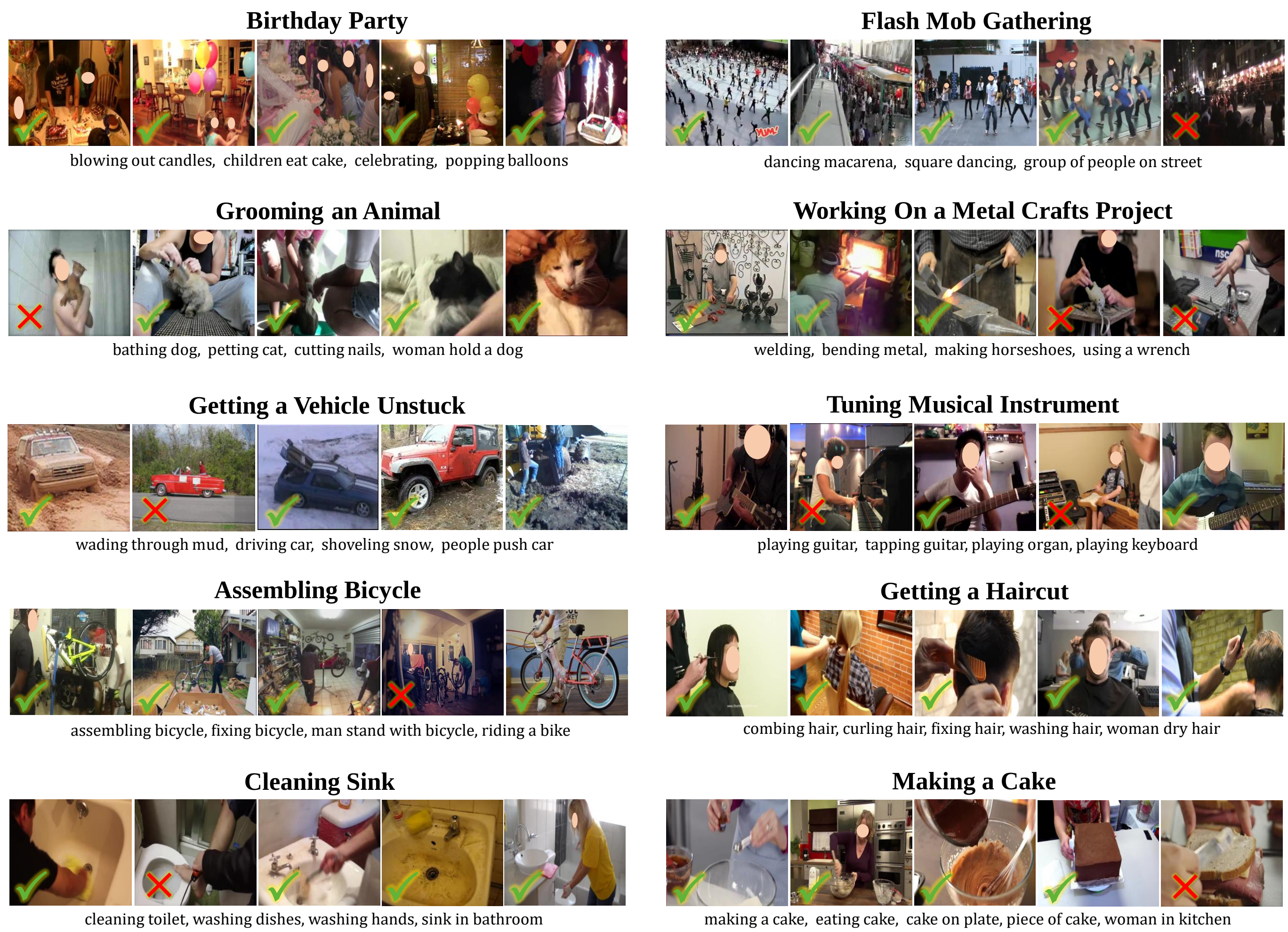


Fig. 5. Visualization of the top-5 ranked videos and the most relevant concepts for some event queries on larger video benchmarks. True/False labels are marked at the left bottom of each video frame.

TABLE V
TOP 5 SELECTED CONCEPTS FOR SOME EVENT QUERIES WHEN USING EVENT NAME AND EXPANSION.

Event Query	Use Event Name	Use Event expansion
Flash Mob Gathering	dodgeball hockey stop mushroom foraging popping balloons throwing water balloon	square dancing dancing macarena people on street mosh pit dancing singing
Renovating a home	man at home decorating christmas tree building sandcastle base jumping dyeing hair	plastering using a paint roller man at home laying tiles installing carpet
Felling a tree	trimming trees people near tree climbing tree climbing a rope sawing wood	trimming trees climbing tree throwing axe sawing wood using circular saw

two expansions (the last row) achieves the optimal performances, which demonstrates the effectiveness of the query expansion module.

We present the top 5 matched concepts of different settings in Table V. It shows that, with the help of expansion, concepts in the third column are highly relevant and reasonable with respect to query events. For instance, the concepts "using a paint roller", "laying tiles", are indeed related actions when renovating a home. However, we can only obtain noisy or irrelevant concepts by directly leveraging the event name.

Influence of Concept Type. To explore the impact of different concept types, we adopt several concept combinations: (1) Object+Scene: use object and scene concepts (ImageNet+Places). (2) Relation: use relationship concepts in constructed concept library. (3) Action: use action concepts in constructed concept library. (4) Relation+Action: use the whole high-order concept library. (5) All: use all concept types. The results are presented in Table VI.

Carefully comparing the results in Table VI, we can make the following observations: (1) As expected, first-order concepts achieved the worst detection performance. The reason is that atom objects and scenes are usually not the key clues to distinguish a complex event. For example, we cannot simply conclude the event "Felling a tree" just because a tree is detected in videos. (2) Comparing the results of relationship and action concepts, we conclude that action concepts make mainly contribution to detection performance. This is because

TABLE VI
COMPARISON RESULTS OF ZERO-SHOT EVENT DETECTION WITH DIFFERENT CONCEPT COMBINATIONS ON THREE BENCHMARKS.

Concept Category	mAP (%) MED13Test	MED14Test	ActivityNet-1.3
Object+Scene	7.2	9.8	5.7
Relation	11.1	10.6	9.0
Action	20.7	14.0	56.1
Relation+Action	21.4	15.1	57.3
All	**23.1**	**16.8**	**57.6**

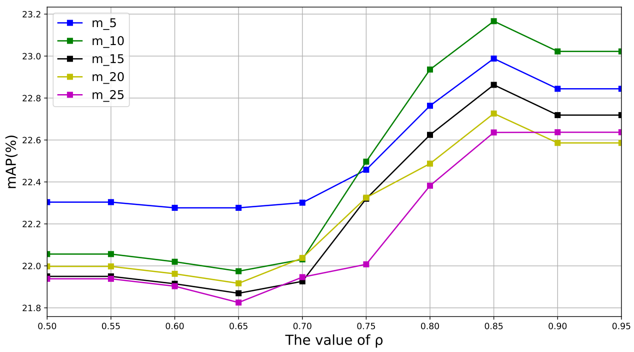


Fig. 6. The impact of different values of m and ρ for experimental performance on MED13Test benchmark.

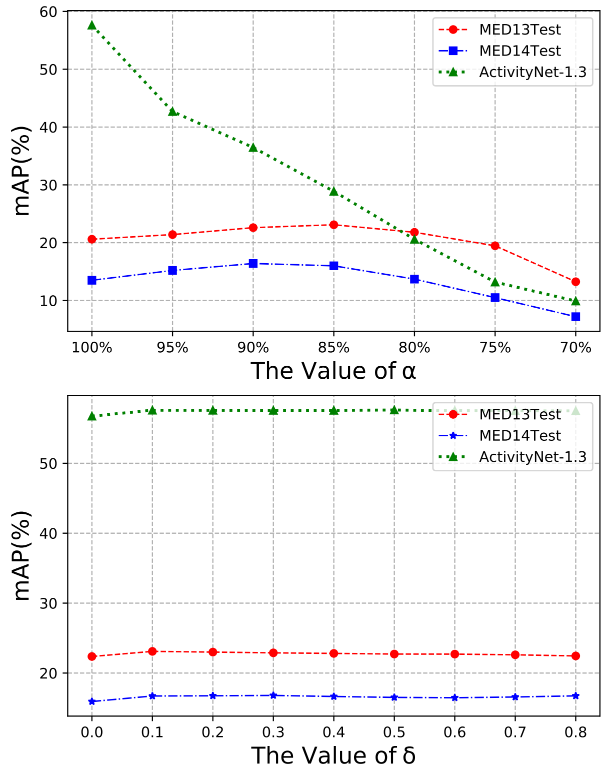


Fig. 7. The impact of different values of $\alpha\%$ and δ for experimental performance on three video benchmarks.

relationship concept detectors are trained based on images, lacking the ability to utilize temporal information in videos. In addition, our relationship concepts are obtained through clustering all relationship triplets. The noisy information in each cluster will inevitably degrade detection performance. (3) Nonetheless, our method obtains the best performance when combining two different types of concepts. Therefore, our proposed high-order concepts are indeed very comprehensive when collaboratively representing complex events. (4) Including object and scene concepts can slightly improve performance, which clearly demonstrates the complementariness of first-order concepts to our high-order concepts.

D. Parameter Sensitivity Studies

In this part, we perform a series of related experiments to explore the effect of different parameter settings of our proposed framework.

The impact of parameters in query expansion. There are two hyper parameters in our query expansion module: m and ρ. The parameter m is utilized for controlling the number of first-order expansion terms. And ρ is the threshold for filtering irrelevant noise in the high-order expansions. We conduct experiments on the MED13Test benchmark to explore the influence of different (m, ρ) values. From the results presented in Figure 6, we can see that the optimal values is $m = 10, \rho = 0.85$. A lower ρ and a higher m will bring irrelevant noise and deteriorate the performance.

The impact of $\alpha\%$. In the concept matching and selection module, only the concepts whose similarity score s_i are more than $\alpha\%$ of the highest one are remained. The quantitative results of different $\alpha\%$ on three benchmarks are shown at the top of Figure 7. It can be seen that our method achieves optimal results at different $\alpha\%$ with respect to per benchmark. It is worth noting that, for MED13Test and MED14Test, the performance first increases and then decreases as $\alpha\%$ decreases. This is because a higher value of $\alpha\%$ will produce fewer concepts, which cannot capture the full semantics of an event. When the $\alpha\%$ decreases, it will involve many irrelevant concepts and deteriorate the performance. However, for the ActivityNet-1.3 benchmark, the mAP keeps decreasing as $\alpha\%$ getting smaller. A possible reason is that the semantic level of the event in this benchmark is relatively low and does not require excessive concept representation.

The impact of the weight in Eq.(12). Recall that w_i is the weight for aggregating concept detection scores in the video retrieval module. We explore the threshold δ in Eq.(12) and the results are shown at the bottom of Figure 7. It's worth noting that when $\delta = 0$, we don't use weight w_i and directly sum the concept detection scores, which deteriorates the mAP a little. The comparison results validate the effectiveness of w_i for balancing semantic relatedness and discrimination of different concepts. Moreover, when the value of δ increases, the overall performance basically does not change much, so we adopt the best set ($\delta = 0.6$) in all other experiments.

VI. Conclusion

In this paper, we highlight the high-order semantic concepts. By fully exploiting three large public datasets, we harvest a comprehensive albeit compact high-order concept library.

Besides, we propose a novel query-expanding scheme by searching several large common knowledge bases, which can map an event query to these high-order concepts. To our best knowledge, this paper is the first attempt in the multimedia community that explores high-order semantic concepts for the zero-shot event detection task. Our experiments report significant improvement on several standard benchmarks compared with conventional low-order concept libraries.

VII. Acknowledgement

This research is supported by National Key R&D Program of China (2018AAA0100702), National Natural Science Foundation of China (61772037), Beijing Natural Science Foundation (Z190001) and Tencent AI Lab Rhino-Bird Focused Research Program (JR202021).

References

[1] "Trecvid multimedia event detection evaluation track," http://www.nist.gov/itl/iad/mig/med.cfm, 2017.

[2] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney, "Video event recognition using concept attributes," in *IEEE Workshop on Applications of Computer Vision*, 2013, pp. 339–346.

[3] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, 2011.

[4] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE transactions on cybernetics*, vol. 47, no. 5, pp. 1180–1197, 2016.

[5] J. Geng, Z. Miao, and X.-P. Zhang, "Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection," *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 498–511, 2015.

[6] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super fast event recognition in internet videos," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1174–1186, 2015.

[7] P. Over, G. Awad, J. Fiscus, G. Sanders, and B. Shaw, "Trecvid 2013– an introduction to the goals, tasks, data, evaluation mechanisms, and metrics," in *TRECVID Workshop*, vol. 2, no. 7, 2013, pp. 1–15.

[8] A. Habibian, T. Mensink, and C. G. Snoek, "Composite concept discovery for zero-shot video event detection," in *Proceedings of International Conference on Multimedia Retrieval*, 2014, pp. 17–24.

[9] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann, "Dynamic concept composition for zero-example event detection," in *Proceedings of the Thirtieth Conference on Artificial Intelligence*, 2016, pp. 3464–3470.

[10] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, and H. Zhang, "Zero-shot event detection via event-adaptive concept relevance mining," *Pattern Recognition*, vol. 88, no. 1, pp. 595–603, 2019.

[11] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 471–480.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[13] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[16] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot, "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms, and metrics," in *2014 TREC Video Retrieval Evaluation,TRECVID*. NIST, 2014, pp. 1–52.

[17] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 951–958.

[19] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[20] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 819–826.

[21] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4166–4174.

[22] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943.

[23] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 466–477.

[24] Y. Cheng, Q. Fan, S. Pankanti, and A. Choudhary, "Temporal sequence modeling for video event detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2227–2234.

[25] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, and S.-F. Chang, "Modeling multimodal clues in a hybrid deep learning framework for video classification," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3137–3147, 2018.

[26] K. Kanagaraj and G. L. Priya, "A new 3d convolutional neural network (3d-cnn) framework for multimedia event detection," *Signal, Image and Video Processing*, pp. 1–9, 2020.

[27] A. Aslam and E. Curry, "Reducing response time for multimedia event processing using domain adaptation," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 261–265.

[28] H. Song, X. Wu, W. Yu, and Y. Jia, "Extracting key segments of videos for event detection by learning from web sources," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1088–1100, 2017.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3169–3176.

[31] Z. Ma, X. Chang, Z. Xu, N. Sebe, and A. G. Hauptmann, "Joint attributes and event analysis for multimedia event detection," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 7, pp. 2921–2930, 2017.

[32] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "Multi-task learning for acoustic event detection using event and frame position information," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 569–578, 2019.

[33] W.-Y. Lee, W. H. Hsu, and S. Satoh, "Learning from cross-domain media streams for event-of-interest discovery," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 142–154, 2017.

[34] H. Zhang and C.-W. Ngo, "A fine granularity object-level representation for event detection and recounting," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1450–1463, 2018.

[35] Z. Ma, Y. Yang, N. Sebe, and A. G. Hauptmann, "Knowledge adaptation with partially shared features for event detection using few exemplars." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1789–1802, 2014.

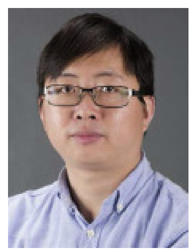
[36] Z. Li, X. Chang, L. Yao, S. Pan, G. Zongyuan, and H. Zhang, "Grounding visual concepts for zero-shot event detection and event captioning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 297–305.

[37] X. Chang, Y. Yang, A. Hauptmann, E. P. Xing, and Y.-L. Yu, "Semantic concept discovery for large-scale zero-shot event detection," in *Proceedings of the Twenty-fourth international joint conference on artificial intelligence*, 2015, pp. 2234–2240.

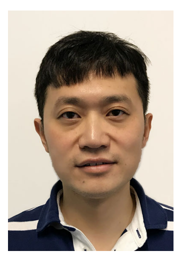
[38] P. Koniusz, F. Yan, P.-H. Gosselin, and K. Mikolajczyk, "Higher-order occurrence pooling for bags-of-words: Visual concept detection," *IEEE*

transactions on pattern analysis and machine intelligence, vol. 39, no. 2, pp. 313–326, 2016.
[39] B. Shi, L. Ji, P. Lu, Z. Niu, and N. Duan, "Knowledge aware semantic concept expansion for image-text matching." in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5182–5189.
[40] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.
[41] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.
[42] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *Asian Conference on Computer Vision*, 2018, pp. 547–563.
[43] P. Mettes, D. C. Koelma, and C. G. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 175–182.
[44] Z. Qin and C. R. Shelton, "Event detection in continuous video: An inference in point process approach," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5680–5691, 2017.
[45] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *CoRR*, vol. abs/1808.01340, 2018.
[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 4171–4186.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[49] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
[50] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
[51] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[52] R. Speer and C. Havasi, "Representing general relational knowledge in conceptnet 5." in *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2012, pp. 3679–3686.
[53] M. Mazloom, X. Li, and C. G. Snoek, "Tagbook: A semantic video representation without supervision for event detection," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1378–1388, 2016.
[54] M. Mazloom, A. Habibian, D. Liu, C. G. Snoek, and S.-F. Chang, "Encoding concept prototypes for video event detection and summarization," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 123–130.
[55] A. Habibian, T. Mensink, and C. G. Snoek, "Video2vec embeddings recognize events when examples are scarce," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 10, pp. 2089–2103, 2016.
[56] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1, 2003, pp. 29–48.
[57] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1452–1464, 2017.
[58] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek, "Searching informative concept banks for video event detection," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 2013, pp. 255–262.
[59] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi, "Multi-attribute queries: To merge or not to merge?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3310–3317.
[60] M. H. D. Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, and W. Kraaij, "Semantic reasoning in zero example video event retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 4, pp. 1–17, 2017.
[61] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
[62] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

Yang Jin is currently a master student in Center for Big Data Research, Peking University. He is developing various sophisticated techniques for large-scale video search.



Wenhao Jiang received his B.E. and M.E. in computer science from Shandong University and Harbin Institute of Technology Shenzhen Graduate School, China, in 2006 and 2009, respectively. He received the Ph.D. degree in the Department of Computing from the Hong Kong Polytechnic University in 2014 and has worked at Tencent AI Lab since 2016. His research interests include machine learning and computer vision.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University. He was a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, from 2011 to 2013. He is currently a Professor with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, School of Software, University of Technology Sydney. His research interests include multimedia, computer vision, and data science.

Yadong Mu is an Assistant Professor at Wangxuan Institute of Computer Technology, Peking University. He obtained both the B.S. and Ph.D. degrees from Peking University. Before joining Peking University, he had ever worked as research fellow at National University of Singapore, research scientist at Columbia University, researcher at Huawei Noah's Ark Lab in Hong Kong, and senior scientist at AT&T Labs. His research interest is in broad research topics in computer vision and machine learning.