Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

DECK: Discovering Event Composition Knowledge from Web Images for Zero-Shot Event Detection and Recounting in Videos

Chuang Gan,[†*] **Chen Sun,**[‡*] **Ram Nevatia** [§]
[†] IIIS, Tsinghua University
[‡] Google Research
[§] University of Southern California

Abstract

We address the problem of zero-shot event recognition in consumer videos. An event usually consists of multiple human-human and human-object interactions over a relative long period of time. A common approach proceeds by representing videos with banks of object and action concepts, but requires additional user inputs to specify the desired concepts per event. In this paper, we provide a fully automatic algorithm to select representative and reliable concepts for event queries. This is achieved by discovering event composition knowledge (DECK) from web images. To evaluate our proposed method, we use the standard zero-shot event detection protocol (ZeroMED), but also introduce a novel zero-shot event recounting (ZeroMER) problem to select supporting evidence of the events. Our ZeroMER formulation aims to select video snippets that are relevant and diverse. Evaluation on the challenging TRECVID MED dataset show that our proposed method achieves promising results on both tasks.

1 Introduction

We address the tasks of multimedia event detection (MED) and recounting (MER) from unconstrained user-generated videos of the kind we may find in social media sites. The goal of MED is to generate a single high-level event label for a given video. A high-level event label represents a broad category of activities, *e.g.* a *birthday party*. Each such event is typically composed of a number of lower-level actions depicting human-human or human-object interactions but these actions and their temporal ordering may vary widely for the same event in different videos. The high intra-class variations of such events make the task of MED very challenging. MER adds a further layer of complexity by aiming to provide event-specific supporting evidence in the form of short video snippets accompanied with text descriptions. It is useful for users to quickly focus on relevant video snippets and decide if the video contents meet their needs.

Existing MED and MER systems usually require video examples for training. However, it is cumbersome to collect training videos for a user to start classifying and recounting videos from a database; users would prefer to find videos by just giving an event name. Due to this observation, we approach MED and MER tasks under the zero-shot setting: we assume no positive video examples are available during training; rather, only the event query itself can be used. We name the zero-shot MED task as ZeroMED, and zero-shot MER task as ZeroMER. Although previous work exists for ZeroMED, to the best of our knowledge none have studied the important problem of ZeroMER.

Zero-shot event detection and recounting are both challenging tasks. A common approach used for zero-shot object and action recognition (Lampert, Nickisch, and Harmeling 2009; Mensink, Gavves, and Snoek 2014) is to represent the categories by banks of human-interpretable semantic concepts (*e.g. attributes*, *actions*, *objects*), where each concept is associated with a pre-trained classifier. It then detects an unseen category by specifying a subset of related concepts manually or following hand-designed rules, and combining their confidence scores. For event detection, (Wu et al. 2014) follows this approach and requires the users to provide a list of related concepts for every event query, thus making it difficult to explore the large-scale video collections. Moreover, users are usually unfamiliar with the internal operations of the system, and unable to determine which concepts will work well with the system. To avoid manual specification of concepts, an alternative approach aims at constructing classifiers of unseen categories directly from classifiers of observed categories (Frome et al. 2013; Norouzi et al. 2014). This is achieved by computing *semantic* similarities of category labels based on their continuous word representations (Mikolov et al. 2013). However, this approach implicitly assumes that unseen categories have *semantically* similar counterparts in the observed categories, which may not always hold.

In light of the above challenges, we adopt the concept-based zero-shot learning scheme, but design a fully automatic algorithm to select relevant concepts for an event. We name the information used to select relevant concepts as *event composition knowledge*, and propose to discover such knowledge from *web images*. Given an event query, we pass it to an Web image search engine (*e.g.* Google) and collect the highest ranking returned images. We then apply our bank of image-based concept classifiers to these images, the concepts with the highest average responses are chosen as event-relevant. For example, the web images retrieved with keyword *birthday party* have high responses of *candle*, *balloon*

[1]The first two authors contributed equally to this work.

and *dining table* classifiers on average. In contrast to other webly-supervised approaches (Singh et al. 2015), the collected web images are used only to select relevant concept classifiers from an existing pool, they are not used to train concept classifiers. We have shown empirically that tens of web images per event query is sufficient for our application.

Some of the concept classifiers cannot be directly applied on web images (*e.g.* action classifiers trained with motion features). For those concepts, we first retrieve the relevant image-based concepts using the above framework, then compute the semantic similarities between the names of the image-based concepts and all other concepts, and keep those with highest similarities. The semantic similarity is computed by the cosine similarity of word2vec embeddings (*e.g.* for birthday party, *blow candle* action classifier is selected as its name is semantically similar to *candle*). We name this framework to discover event composition knowledge from web images and select relevant concepts as DECK.

Once the relevant concepts have been selected for each event, ZeroMED proceeds by computing weighted sum of relevant concept detection scores on video-level. For ZeroMER, a naïve approach is to directly extend ZeroMED to video snippets. However, users might have different preferences even for the same event queries. It is important to present a diverse video segments covering different aspects of the query events. Take *renovating a home* event as an example, some users are interested in laying the floor while others are more interested in tiling the roof. By providing diverse results, users can quickly locate the clips they are interested in. For this purpose, we treat the selection of each segment within a video as a binary variable, and aim to maximize the confidence scores of the desired concepts as well as the diversity of selected segments. We relax this integer programming problem into linear programming, and show that the approximated version offers good performance.

Most previous work for MER relies on subjective evaluation performed by humans (Sun et al. 2014; Sun and Nevatia 2014; Gan et al. 2015b). We designed a quantitative metric to evaluate MER automatically, and annotated 200 videos over 20 event categories from the TRECVID MED'14 dataset for evaluation. We also studied the ZeroMED performance on the full MEDTest 13 and 14 dataset with around 25,000 videos respectively. Our proposed method performs competitively on both tasks. In summary, our work makes three contributions:

- We propose the DECK framework to select representative and reliable concepts automatically from web images.
- We introduce the zero-shot event recounting problem, and propose a framework to generate event recounting results that are relevant, diverse and compact.
- We propose a set of quantitative metrics for recounting evaluation and also provide a set of annotations.

2 Related Work

Zero-shot learning. The seminal work by Lampert et al. (Lampert, Nickisch, and Harmeling 2009) demonstrates the effectiveness of zero-shot classification using attributes. In video domain, such attributes are usually named as concepts (Sadanand and Corso 2012), and used for applications like zero-shot event detection (Jiang et al. 2015; Wu et al. 2014; Gan et al. 2015a), and event recounting with training samples (Liu et al. 2013). To our best knowledge, there is no previous work on zero-shot multimedia event recounting. To find relevant attributes for unseen categories, linguistic knowledge databases (Rohrbach et al. 2010), web search hit counts (Rohrbach, Stark, and Schiele 2011) and semantic embeddings (Jain et al. 2015) can be used. The use of web images for concept selection is yet to be explored.

Concept discovery from web images. Images from web search engines have been used to discover and train concept detectors (Divvala, Farhadi, and Guestrin 2014; Chen, Shrivastava, and Gupta 2013). For event recognition, Ye et al. (Ye et al. 2015) have applied a concept ontology to collect web videos and images, and train CNN classifiers from the collected data. Chen et al. (Chen et al. 2014) define concept list by using event descriptions provided by users, and collect Flickr images to train concept detectors. Such detectors can be enhanced by pseudo relevance feedback from test videos (Singh et al. 2015). Rather than training concept detectors directly from web images, DECK treats them as a source of event composition knowledge.

Multimedia event recounting. Most existing approaches on MER (Sun et al. 2014; Liu et al. 2013) apply concept detectors or low-level visual features to localize key evidence. They rely on training videos with event labels to train video-level classifiers, which are then used to rank the video segments. Such approaches assume implicitly that video-level classifiers can be used to distinguish segments, which may not always hold. To delve into segments, Lai et al. (Lai et al. 2014a; 2014b) formulated MIL problems. Sun et al. (Sun and Nevatia 2014) proposed a latent SVM framework to learn segment-level classifiers using video-level labels. To use video descriptions, Habibian et al. (Habibian, Mensink, and Snoek 2014b) proposed the VideoStory pipeline to learn embeddings from words and video features, it can be used to generate video descriptions. Recently, Potapov et al. (Potapov et al. 2014) studied the problem of event-specific video summarization. Its focus is more on temporal segmentation of videos, and does not distinguish different types of evidence during evaluation. All these approaches require videos with annotations for training.

3 Zero-shot Event Recognition

Given a pool of semantic concept classifiers, the DECK framework first selects those concepts relevant to an event by using web images. It then breaks videos into shot segments, and represents each video segment using the detection scores from a bank of semantic concepts. Classifiers of the selected concepts are used to generate detection scores for videos (ZeroMED) and segments (ZeroMER). For ZeroMER, we introduce a diversity term and aim to generate recounting results that are both relevant and diverse. The overall approach is depicted in Figure 1.

3.1 Video Representation

We first segment long videos into short clips using off-the-shelf shot boundary detectors (Yu et al.), and choose the middle frame in each segment as the representative

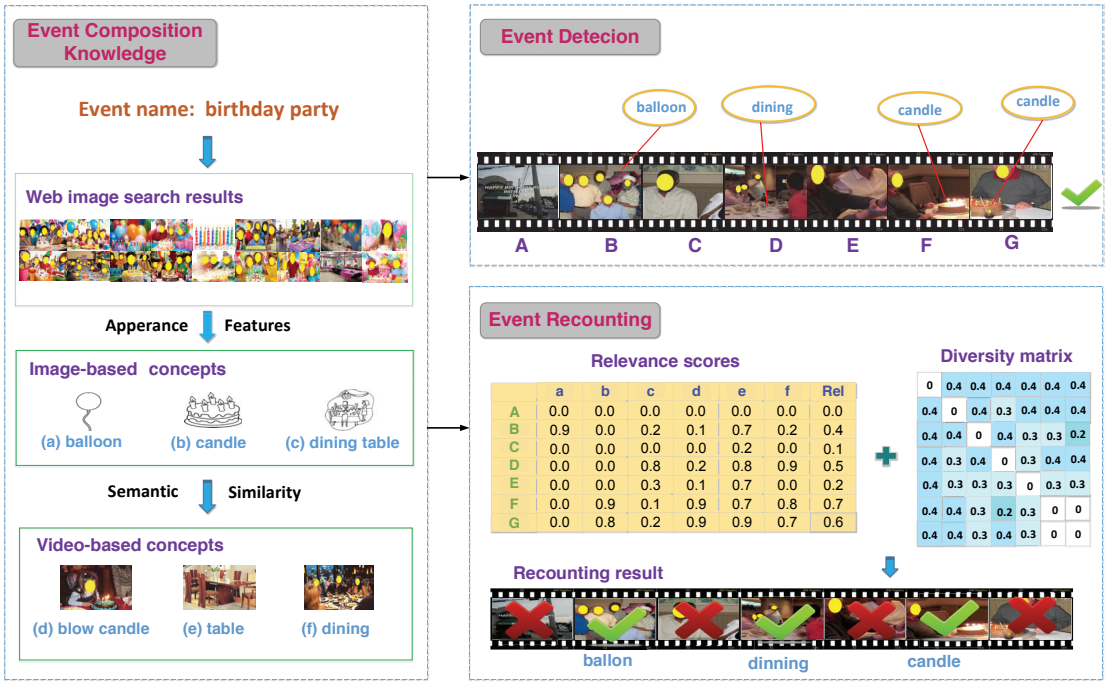

Figure 1: An illustration of our event recognition framework. Each video segment is represented by a bank of semantic concepts. The concept classifiers can be trained using image features (image-based) or motion features (video-based). Given an event query, we select the relevant image-based concepts by discovering event composition knowledge from web images. We then select the video-based concepts based on their semantic similarities with the selected image concepts. The classifier outputs of selected concepts are used to compute event relevance scores. They are pooled over whole videos to generate video classification scores for ZeroMED. For ZeroMER, consider both relevance and diversity of the selected segments.

key frame. To allow zero-shot detection and recounting, we represent video segments by a bank of pre-defined semantic concepts, which ranges from objects, scenes to actions. Assume we have pre-trained a concept detector $f_{C_i}(\cdot) \in \mathbb{R}$ for each semantic concept C_i, a video segment v_i is mapped into the concept space by $C(v_i) = [f_{C_1}(x_i), f_{C_2}(x_i), \ldots, f_{C_K}(x_i)]$, where K is the total number of concepts and x_i is the visual feature for v_i.

We include both image-based and video-based concepts in the bank. An *image-based concept detector* is trained from static images using appearance features. We feed the middle key frame of each video segment to it and use the output to represent the whole segment. An *video-based concept detector* is trained from and applied directly on video segments. Video-based concept detectors allow us to utilize motion information, which is important to recognize actions.

Training concept detectors: we use 1000 image-based concepts which are related to objects and over 900 video-based concepts which are related to actions and activities.

For image-based concepts, we obtain 1000 object concept detectors using a 19-layers very deep CNN architecture proposed by Simonyan et al. (Simonyan and Zisserman 2015) trained on the ImageNet ILSVRC-2014 dataset (Deng et al. 2009). After the CNN model is trained, we take the key frame of each test video segment as an input, make a forward-pass of the CNN and use the softmax outputs as the concept detection scores for the segment.

For video-based concepts, we use three publicly available datasets: UCF101 (Soomro, Zamir, and Shah 2012), TRECVID Semantic Indexing (SIN) and Google Sports1M (Karpathy et al. 2014). They contain 101, 346 and 487 categories respectively. To train concept classifiers, we extract the improved dense trajectory (Wang and Schmid 2013) features from videos, and aggregate the local features into video-level feature vectors by Fisher vectors (Sun and Nevatia 2013). The features are used to train linear SVM classifiers (Chang and Lin 2011) by fixing bias to 10 and soft margin cost to 1. We then apply them to test video following the same feature extraction step but on video shots.

Selecting image-based concepts: for each event, we query Google image search engine with the event names. We download the top ranked images with type *photo*, which helps remove non-realistic images (*e.g.* cartoon). We apply each image-based concept detector $f_{C_i}(\cdot)$ to the retrieved image set $\mathcal{I}$. To suppress noise from individual images, we compute the event matching score $h(C_i) = \frac{1}{|\mathcal{I}|}\sum_{I\in\mathcal{I}} f_{C_i}(I)$. We select T image-based concepts with top h values.

Expanding to video-based concepts: To go beyond image-based concepts and select video-based concepts, we measure the semantic similarity between an image-based concept name w_I selected in the previous stage and a

video-based concept name w_V. For this purpose, we use a data-driven similarity measurement based on the skip-gram model (Mikolov et al. 2013). The resulting word vectors have properties that semantically similar words are close to each other. We use the complete dump of English Wikipedia to train it. We set the embedding dimension to 300, and use default options for other parameters.

To measure the semantic similarity between w_I and w_V, we compute the cosine similarity between the normalized average of their corresponding word embeddings, as $sim(w_I, w_V) = \eta(w_I)^T \eta(w_V)$, where $\eta(w) = \sum_{i \in w} e_i / \left\| \sum_{i \in w} e_i \right\|_2$, and e_i is the word embedding for word i. We compute $sim(w_I, w_V)$ between the selected image-based concepts I and all video-based concepts V. The top T' video-based concepts with the highest similarities are selected as relevant video-based concepts.

By applying the image-based concept detectors directly to web images and taking the average of detector responses, concepts with less reliable detectors are filtered out implicitly. This selection process is also less sensitive to the naming of image-based concepts. Unlike previous work (Chen et al. 2014) which crawled web images to train concept detectors, we only use web images as a source to discover mid-level knowledge which decomposes events into concepts. As a result, only a small number ($\sim$ 90) of web images is needed for each event query. The whole process is very fast.

Implementation: We queried Google image search engine and downloaded the top 90 images for each query. To avoid the artificial images, only *photo* type images were kept. To comply with image query format, we replace all occurrences of *without*, *non-* and *not* with the minus sign. As recommended by (Jiang et al. 2014), the number of selected image-based concepts T and video-based concept T' are both fixed as 3 in all experiments. The embedding dimension of word2vec we used is 300.

3.2 Event Detection

After relevant concepts are selected, zero-shot event detection is straightforward. Given a event query E, the confidence score S_k^E of each testing video k is computed by summing *video-level* detection scores from selected concepts, computed as $S_k^E = \sum_{c \in \mathcal{C}_E} v_c(k)$, where $\mathcal{C}_E$ is the set of selected concepts for event E. $v_c(k)$ is the *video-level* concept detection score of video k, it is computed by average pooling of *shot-level* concept detection scores. Higher score indicates this video is more likely to be match the event query.

3.3 Event Recounting with Diversity

Different from event detection, event recounting aims at selecting a subset of relevant and diverse video segments for users quick grasping the hints. Similar to event detection, we compute the relevance score Rel_i^E of video segment i for event E as the sum of *shot-level* detection scores from the selected concepts as $Rel_i^E = \sum_{c \in \mathcal{C}_E} f_c(i)$, where $\mathcal{C}_E$ is the set of selected concepts for event E, $f_c(i)$ is the concept detection score of segment i. Rel_i^E terms from the same videos are normalized to $[0, 1]$.

One can directly use the relevance scores for recounting, by selecting the video segments with highest relevance scores. However, it is intuitive that it may be beneficial for the system to display a diverse selection of video segments, while preserving the event relevance for the selected segments. To address this issue, we introduce a diversity term for segment selection. It is measured by the semantic distances between two video segments i and j. Let $\mathbf{c}_i = [f_{C_1}(i),\ f_{C_2}(i),\ ...,\ f_{C_K}(i)]$ as the concept representation for segment i, we define the diversity score between segments i, j as $Diff_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|_2$.

Putting the two terms together, the objective function of event recounting for event E is to select a subset of video segments $\mathcal{I}$ to maximized $\sum_{i \in \mathcal{I}} Rel_i^E + \sum_{i \in \mathcal{I}, j \in \mathcal{I}} Diff_{ij}$.

Denote $s_i \in \{1, 0\}$ as a binary indicator of whether segment i is selected or not, it can be transformed into:

$$\sum_{i=1}^{T} Rel_i^E \cdot s_i + \sum_{i,j} Diff_{ij} \cdot s_i \cdot s_j \tag{1}$$

We introduce an auxiliary variable $s_{ij} \in \{0, 1\}$. It takes the value of 1 only if both segments i and j are selected. Define L as the maximum number of selected segments per video, the resulting formulation has the form, which is an integer linear programming problem:

$$\begin{aligned} \text{Maximize:} \quad & \sum_{i=1}^{T} Rel_i^E \cdot s_i + \sum_{i,j} Diff_{ij} \cdot s_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{T} s_i \leq L, \\ & s_{ij} \leq s_i, \quad s_{ij} \leq s_j \quad \forall i, j, \\ & s_i + s_j - s_{ij} \leq 1 \quad \forall i, j, \\ & s_i \in \{0, 1\} \quad \forall i, \\ & s_{ij} \in \{0, 1\} \quad \forall i, j. \end{aligned} \tag{2}$$

where T is the total number of segments in a video.

ILP is NP-Hard. However, we found that relaxing the problem into linear programming already produces competitive results, and can be solved efficiently using off-the-shelf solvers in MATLAB: we allow the binary variables to take any value from 0 to 1, and round the outputs to either 0 or 1 afterwards. Generating event recounts for a video with around 40 shots takes only a few seconds.

After the problem is solved, we use the video segments with $s_i > 0$ to generate the final event recounting, each video segment is accompanied by a text description generated from the selected concept with highest response.

Discussion: There are alternative formulations (*e.g.* diverse ranking (Radlinski, Kleinberg, and Joachims 2008)) which may also provide diversity, yet we choose our formulation for its simplicity and effective performance.

By maximizing the objective function, the video recounting results are not only relevant to the events, but are also diverse. Although a balance term can be added in Equation 1 to weight relevance and diversity, we fix it to 1 as: (1) the two terms have been scaled to $[0, 1]$; (2) it is difficult to select the weighting parameter under zero-shot setting.

4 Experiments

We present the dataset, experimental settings, evaluation criteria and experimental results in this section.

4.1 MED-recounting Dataset

So far, no existing datasets are available for automatic quantitative evaluation of recounting results. Previous work on recounting relies on humans to watch the program outputs and rate their qualities, without explicitly defining the recounting ground truth (Sun et al. 2014). To fill this gap, we introduce a new video dataset **MED-Recounting** and provide temporal annotations of the evidence locations within the videos. We also design an automatic evaluation metric.

Video Data: We use the videos in the challenging NIST TRECVID Multimedia Event Detection 2014 dataset (MED'14) to evaluate the recounting performance. It has 20 event categories. Videos in MED'14 have large variations in length, quality and resolution. The average length of the videos is over 2 minutes. For the purpose of zero-shot event recounting evaluation, we select 10 videos per event from the MED'14. Most of the videos have a duration from 1 to 5 minutes. The total number of videos used for recounting evaluation is 200.

Annotation Protocol and Evaluation Metric: We segment video into shots using the algorithm described in Section 3.1. For every shot in the same video, we first ask the users "Does the shot contain supporting evidence for event *A*?". The possible answers are "Yes" or "No". For those shots marked as "yes", we ask them to group the shots that they believe offer the same type of evidence. We use majority vote rule to combine annotations from different annotators. The final annotation is in the form of integer labels for all shots in a video, where each positive number stands for a different evidence category, and -1 stands for no evidence. Typically, each video contains about 40 video segments, and 3 key evidence categories are marked in each video. To evaluate recounting quality (RQ) for each video event, assume that the total number of selected shots is up to L, we use the percentages of evidence categories have been hit as evaluation metric, which is defined as:

$$RQ = \frac{\#evidence_{hit}}{\#evidence_{total}} \tag{3}$$

where $\#evidence_{hit}$ represents the number of key evidence categories has been covered in the recounting result, and $\#evidence_{total}$ the total number of key evidence categories within the test video. When the number of selected shots are fixed, a higher RQ score indicates that more of the evidence categories are covered by the recounting result for the video.

As selecting video shots from the same evidence category does not increase $\#evidence_{hit}$, our evaluation metric favors video recounting that is both relevant and diverse.

4.2 Zero-shot Event Detection

Experiment setup: We conduct experiments on TRECVID MEDTest 2013 and 2014 datasets. Each includes around 25,000 test videos with per-video ground truth annotations for 20 event categories, all officially provided by NIST. Since we focus on zero-shot settings, we just take event names as inputs. To evaluate the results, we apply the official metric: average precision (AP) per event, and mean Average Precision (mAP) by averaging all 20 events.

ZeroMED Method	mAP (%)
Concept Discovery (Chen et al. 2014)	2.3
Bi-concept (Habibian, Mensink, and Snoek 2014a)	6.0
Composite-concept (Habibian, Mensink, and Snoek 2014a)	6.4
SPaR (Jiang et al. 2014)	12.9
EventNet (Ye et al. 2015)	8.9
Weak concept (Wu et al. 2014)	12.7
Singh et al. (Singh et al. 2015)	11.6
Semantic embedding (Elhoseiny et al. 2016)	13.5
DECK (Ours)	**17.8**

Table 1: Comparisons with state-of-the-art on MED13.

To evaluate whether DECK helps in concept selection for ZeroMED, we first compare with skip-gram based concept selection approach under the same features on MEDTest 2014 dataset for fair comparisons. We also conduct experiments on MEDTest 2013 to compare against state-of-arts, since most ZeroMED systems have reported results on it.

Comparison with state-of-the-art: First, we compare the DECK approach with recent state-of-the-art methods. We report results on MED'13 which was used by published methods. Among them, *Bi-concept*, *Composite-concept*, *EventNet* use web videos to train concept detectors; Singh et al. uses web images to train concepts and re-train the event detectors with top retrieved test videos; *Weak concept*, *SPaR*, *Semantic embedding* and *MMPRF* use similar pre-trained concept banks as ours, but select concepts from manually generated event descriptions with semantic similarity. From Table 1, we observe that systems using pre-trained concepts generally have higher performance than webly-trained concepts. Our system outperforms all the other approaches significantly, which indicates that DECK is able to select *better* concepts *automatically*.

Impact of concept selection methods: We compare concepts selected by DECK with web images against those selected by semantic similarity as defined by word2vec embeddings. In particular, we evaluate the following settings:

- **word2vec (I)**: use event names to select 3 image concepts with top word2vec similarities.
- **word2vec (I + V)**: use event names to select 3 image and 3 video concepts with top word2vec similarities.
- **DECK (I)**: use DECK to select 3 image concepts.
- **DECK (I + V)**: use DECK to select 3 image and 3 video concepts.

Table 2 lists the ZeroMED performance with different concept selection methods. We can see that DECK has better mAP than word2vec. We also observed that the difference in AP is larger when the event names are more abstract and event composition knowledge is non-trivial to infer (*e.g. playing fetch*), or when word2vec fails to retrieve semantic similar concepts (*e.g. rock crab* for *rock climbing*).

Discover knowledge or build event classifiers? The web images collected for DECK can also be used to train event

Method	Image classifiers	word2vec (I)	DECK (I)	word2vec (I + V)	DECK (I + V)
MAP	6.4	5.4	9.1	12.4	**16.3**

Table 2: Comparison of different concept selection methods for ZeroMED on MED14 dataset.

Dataset	word2vec	DECK
MED14 recounting	0.534	0.648

Table 3: Comparison of mean RQ on zero-shot recounting task for different concept selection methods.

Method	mean RQ
Random	0.251
Uniform	0.279
Clustering	0.362
DECK w/o diversity	0.535
DECK	**0.642**

Table 4: Event recounting results comparing with baseline approaches. Higher scores indicates better performance.

classifiers directly. The *Image classifiers* column in Table 2 shows the APs for this baseline. The image-based event classifiers were trained with VGG-19 CNN features and SVM classifiers from 90 web images per event. We can see that its performance is still far behind from our best DECK system. This indicates that transferring event composition knowledge from web images requires less data than directly training event classifiers. For some events (*e.g. bike trick*), although video events and web images share similar relevant concepts, their appearances differ a lot.

4.3 Zero-shot Event Recounting

We evaluate ZeroMER performance on our MED-recounting dataset. Web images used for DECK are the same as used in ZeroMED. Table 3 shows the mean RQ for different concept selection methods. We can see that DECK outperforms word2vec. In the following experiments, we use the concepts selected by DECK.

Comparison with baseline: To demonstrate the effectiveness of our proposed ZeroMER framework, we compare our framework against four baseline systems:

- **Random**: randomly select L shots.
- **Uniform**: divide the video into L parts uniformly, and choose one shot from each part randomly.
- **Clustering**: for each video, cluster the concept features of the video segments into L clusters, and use each centers.
- **DECK w/o diversity**: select the top L shots with highest relevance scores.

To compute relevance scores, we use DECK to select image and video concepts. We set L to 3 and compare RQ on event level as the average number of evidence categories is around 3. From Table 4 we can see that by choosing segments with relevant concepts, both *DECK w/o diversity* and *DECK* outperform the other systems significantly.

Human evaluation: We asked 10 human evaluators to compare the recounting results generated by *DECK* and *DECK w/o diversity*. They are students with knowledge in Computer Vision. We used all 200 videos in MED-recounting for evaluation and fixed $L = 3$. For each video, we provided the human evaluators the ground truth event name and the description. We then showed the three key frames from recounting generated by the two systems on each side of the screen respectively (with randomly order), and asked the evaluators to choose from the following: *1st is better*, *2nd is better*, *equally good or bad* as suggested in (Liu, Mei, and Chen 2016). We aggregated the evaluation results using majority vote. On average, 78.5% of the evaluators agreed on their votes for specific videos.

Better	Worse	Similar
71.5%	18.5%	10.0%

Table 5: Human comparison of the recounting results generated by DECK against DECK w/o diversity.

According to the evaluators (Table 5), *DECK* generates better recounting results than *DECK w/o diversity* in 71.5% of all videos. It has similar performance in 10% of the videos and is worse in 18.5%, possibly due to irrelevant segmented selected to achieve diversity. This indicates that the RQ evaluation metric agrees well with humans.

From Table 4, we have three key observations. First, we find that that the proposed method achieves promising recounting results, above half of the evidence within the video could be identified by directly transferring knowledge from Wikipedia and web image search engine. Secondly, we find that the introduced diversity term could further improve the recounting quality. Thirdly, the diversity term may fail to work well for some events, *e.g giving directions to a location*, due to no relevant concept match this event. This limitation could be addressed by more semantic concept coverage.

5 Conclusion

We introduce a novel problem of zero-shot multimedia event recounting (ZeroMER). It aims at providing persuasive evidence for the events, without using training videos. We present the DECK algorithm to select relevant concepts for zero-shot event detection and recounting fully automatically, and use them to select video segments that are relevant, diverse and compact. Experimental results based on automatic and human evaluations show that the DECK framework achieves promising results for both event recounting and event detection tasks.

Acknowledgement: Chuang Gan is supported in part by the National Natural Science Foundation of China Grant 61033001 and 61361136003.

References

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM TIST*.

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*.

Chen, X.; Shrivastava, A.; and Gupta, A. 2013. NEIL: Extracting visual knowledge from web data. In *ICCV*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Divvala, S. K.; Farhadi, A.; and Guestrin, C. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*.

Elhoseiny, M.; Liu, J.; Cheng, H.; Sawhney, H. S.; and Elgammal, A. M. 2016. Zero-shot event detection by multimodal distributional semantic embedding of videos. In *AAAI*.

Frome, A.; Corrado, G.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Gan, C.; Lin, M.; Yang, Y.; Zhuang, Y.; and Hauptmann, A. G. 2015a. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In *AAAI*.

Gan, C.; Wang, N.; Yang, Y.; Yeung, D.; and Hauptmann, A. G. 2015b. DevNet: A deep event network for multimedia event detection and evidence recounting.

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

Habibian, A.; Mensink, T.; and Snoek, C. G. 2014b. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*.

Jain, M.; van Gemert, J. C.; Mensink, T.; and Snoek, C. G. M. 2015. Objects2action: Classifying and localizing actions without any video example. *CoRR* abs/1510.06939.

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2014. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM MM*.

Jiang, L.; Yu, S.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2015. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*.

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.-T.; Liu, D.; Chen, M.-S.; and Chang, S.-F. 2014a. Recognizing complex events in videos by learning key static-dynamic evidences. In *ECCV*.

Lai, K.-T.; Liu, D.; Chen, M.-S.; and Chang, S.-F. 2014b. Video event detection by inferring temporal instance labels. In *CVPR*.

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

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

Liu, Y.; Mei, T.; and Chen, C. W. 2016. Automatic suggestion of presentation image for storytelling. In *ICME*, 1–6.

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*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.; and Dean, J. 2014. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*.

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

Radlinski, F.; Kleinberg, R.; and Joachims, T. 2008. Learning diverse rankings with multi-armed bandits. In *ICML*.

Rohrbach, M.; Stark, M.; Szarvas, G.; Gurevych, I.; and Schiele, B. 2010. What helps where - and why? semantic relatedness for knowledge transfer. In *CVPR*.

Rohrbach, M.; Stark, M.; and Schiele, B. 2011. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *CVPR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Singh, B.; Han, X.; Wu, Z.; Morariu, V. I.; and Davis, L. S. 2015. Selecting relevant web trained concepts for automated event retrieval. *ICCV*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sun, C., and Nevatia, R. 2013. Large-scale web video event classification by use of fisher vectors. In *WACV*.

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

Sun, C.; Burns, B.; Nevatia, R.; Snoek, C.; Bolles, B.; Myers, G.; Wang, W.; and Yeh, E. 2014. ISOMER: Informative segment observations for multimedia event recounting. In *ICMR*.

Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *ICCV*.

Wu, S.; Bondugula, S.; Luisier, F.; Zhuang, X.; and Natarajan, P. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*.

Ye, G.; Li, Y.; Xu, H.; Liu, D.; and Chang, S.-F. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*.

Yu, S.-I.; Jiang, L.; Mao, Z.; Chang, X.; Du, X.; Gan, C.; Lan, Z.; Xu, Z.; Li, X.; Cai, Y.; et al. Informedia@ trecvid 2014 MED and MER.