6-2016

Event detection with zero example: Select the right and suppress the wrong concepts

Yi-Jie LU

Hao ZHANG

Maaike DE BOER

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts

Yi-Jie Lu[†], Hao Zhang[†], Maaike de Boer[*‡], Chong-Wah Ngo[†]
[†]City University of Hong Kong, [‡]Radboud University
[*]Netherlands Organization for Applied Scientific Research (TNO)
{yijie.lu, hzhang57-c}@my.cityu.edu.hk, maaike.deboer@tno.nl, cscwngo@cityu.edu.hk

ABSTRACT

Complex video event detection without visual examples is a very challenging issue in multimedia retrieval. We present a state-of-the-art framework for event search without any need of exemplar videos and textual metadata in search corpus. To perform event search given only query words, the core of our framework is a large, pre-built bank of concept detectors which can understand the content of a video in the perspective of object, scene, action and activity concepts. Leveraging such knowledge can effectively narrow the semantic gap between textual query and the visual content of videos. Besides the large concept bank, this paper focuses on two challenges that largely affect the retrieval performance when the size of the concept bank increases: (1) How to choose the right concepts in the concept bank to accurately represent the query; (2) if noisy concepts are inevitably chosen, how to minimize their influence. We share our novel insights on these particular problems, which paves the way for a practical system that achieves the best performance in NIST TRECVID 2015.

General Terms

Algorithms, Experimentation, Performance

Keywords

Multimedia Event Detection; Video Search; 0Ex; Concept Selection; Semantic Pooling; Concept Bank

1. INTRODUCTION

Today, almost all real-world solutions of video retrieval, such as YouTube, are mainly based on text matching. This matching requires the availability of textual metadata such as titles and tags provided by video uploader [6]. In contrast to text matching, multimedia research communities have kept promoting video content understanding for years, expecting that machines can search by visual cues as well as high-level semantics, rather than requiring humans to annotate the videos beforehand [31]. Among those practices, in recent years, TRECVID has brought up the problem of multimedia event detection (MED), which is specialized in searching complex multimedia events in a very large video corpus [26]. MED has the following distinctions that cause great difficulty for video retrieval: (1) The query is a description of an everyday event, which is complex and diverse by nature in both of its visual and semantic cues. The diversity is caused by the varieties of appliance types, change of surroundings, illumination, and viewpoints. Figure 1 takes the event "cleaning an appliance" as an example: the definition of an appliance includes large household machines such as air conditioner, dishwasher, refrigerator, kitchen stove, etc. It can also include small devices like coffee makers. The camera position may vary dramatically between videos, causing great difficulty for object recognition. The cleaning operation usually happens in a kitchen or shop, but is not limited to such indoor scenes. For example, people may clean their grill in the backyard, which substantially increases visual diversity. Moreover, although the appearance of a refrigerator in a kitchen is helpful for video screening, key evidence such as spraying cleaning fluid is critical to make a judgement. (2) The task concerned by this paper, namely zero-example event detection (0Ex), focuses on the case that no visual examples are given to train an event classifier. Instead, only textual definition of the event is provided as a information need. Therefore, a search system has to leverage external knowledge to bridge the gap between textual information in the query and visual content in the video corpus. (3) The large video corpus for event search contains 200,000 unconstrained Internet videos with variable length and quality [29], which sets a high demand on both performance and efficiency for a search system.

Figure 1: Visual diversity of the event "cleaning an appliance".

ICMR'16 June 6–9, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912015

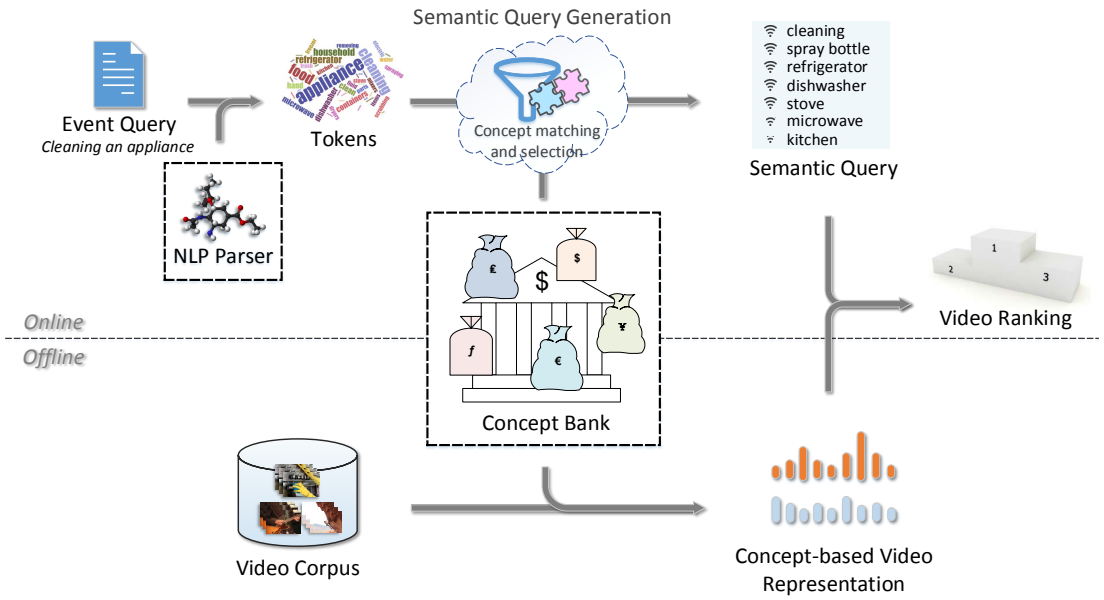

Figure 2: The overall framework for event detection with zero example. The core of the framework is a large concept bank containing detectors of objects, scenes, actions, and activities. Semantic query generation which maps the tokens (noun and verb phrases) to an internal query representation, namely *semantic query*, is essential to system performance.

The state-of-the-art systems for 0Ex are mostly based on the selection of a handful of relevant concepts that can integrally describe an event query, assuming that there is a classifier for each of the selected concepts [15]. We develop our automatic system likewise, which is shown in Figure 2. In the offline phase, all videos are represented by a concept-based representation in which each dimension represents the likelihood of presence of a particular concept. In the online phase, given an event query, e.g. "cleaning an appliance" with detailed description, the noun and verb phrases are first extracted as tokens. Then these tokens are mapped to an internal query representation called *semantic query* by concept matching. In this way, the query is converted to concept-based representation as well. Finally, event search is performed having all the concept-based representations of both queries and videos.

This framework directly raises an important concern: *How to build a large concept bank generalizable for everyday events?* In order to provide good coverage in the first place, we basically mix thousands of concepts in different granularity, i.e. from common objects to activities. Then comes the fundamental problem this paper aims to focus: *How to pick up the right concepts in the concept bank to represent an event query?* Intuitively, one may expect of using more and more relevant concepts for describing an event from different aspects, given that these concepts may help reinforce with each other. This idea, however, turns out to be impractical: Through experiments, we surprisingly find that only selecting a few relevant concepts could already provide a good performance for most events, as long as the selected concepts are accurate. Adding more concepts that are less relevant to the query almost always risks a performance plunge. Hence, a conservative strategy is to simply cap the number of concepts in semantic query by keeping the best matches. Although this strategy generally shows good performance, the optimal number of concepts is hard to determine because it varies among events. In addition, due to imperfection of concept matching, false positives are inevitable even for the top ten relevant concepts. As a result, capping the number of concepts turns to be risky when concept matching becomes unreliable, because a single noisy concept would count in a representation with a small number of concepts. We hereby propose to improve the video representation by pooling the key evidence, rather than pooling all the keyframes. In this new representation, noisy concepts are less likely to have an impact.

The novelty of this paper can be summarized as follows:

- We focus on the problem of choosing the right concepts from a large concept bank for event detection without training examples. We show that, with a large concept bank, concept selection is crucial to the performance. To the best of our knowledge, concept selection in the scope of 0Ex has not been systematically investigated.
- We propose to represent the videos only by their evidential shots to create a system robust to wrong concepts in the semantic query. The detection of the evidential shots is unsupervised, which well fits to the 0Ex pipeline. The representation can lead to more than 50% of improvement without capping the number of concepts in semantic query generation.

The efficiency of the proposed methods was backed by experimental results on a large corpus: According to TRECVID 2015's evaluation on 200,000 videos, addressing the aforementioned problems paved our way to a practical system that achieved the best performance in the 0Ex task of MED.

2. RELATED WORK

Multimedia event detection typically makes use of training examples. Event classifiers are trained by low-level features [9, 25, 33] or semantic features [11, 12, 27] that are directly

extracted from exemplar videos. *Zero-example* multimedia event detection (0Ex) is an emerging topic yet to be explored. In contrast to zero-shot learning which focuses on the recognition of images with unseen labels, 0Ex emphasizes the use of general and external knowledge for textual-to-visual relation. A few pilot studies were proposed very recently [5, 34, 2]. These works built a small concept library, typically hundreds of concepts, for textual-to-visual relation. Mazloom et al. [20] proposed to use tag propagation which propagates tags from a labeled video source to unlabelled videos. For 0Ex retrieval, they found that using tags propagated by concept vector similarity overwhelms bag of visual words similarity, which renders the importance of concepts. More recent work starts to resort to a larger concept library. Ye et al. [35] collected a dataset with 500 events in which more than 4,000 concepts were hierarchically organized. Singh et al. [28] automatically discovered salient visual concepts by web search according to the text query. In event search phase, Habibian et al. [10] indexed concepts by AND/OR composition. Chang et al. [3] proposed a rank aggregation framework that addressed the incomparable scales of scores when merging concepts from different feature spaces. Jiang et al. studied pseudo relevance feedback [14] and self-paced reranking [13] that further improved the performance by reranking. Jiang et al. [15] also systematically investigated 0Ex problem. This work explored the contribution of multiple features including thousands of concepts, as well as the performance of several search models. Our work relies on a large number of concepts as well. But different from previous studies which analyze many components like feature types and search models, we especially focus on the crucial problems raised when the size of the concept library is increased.

Regarding the retrieval performance of an 0Ex system, semantic query generation, which translates event query to an internal query representation, is the crux. With a large concept bank, while trying to avoid relating queries to wrong concepts on the one hand, an alternative to improve performance is to somehow tolerate the wrong concepts in event search. We draw inspiration from semantic pooling [36]. Rather than simply averaging the frame-level concept responses for the whole video [19, 21], recent studies suggest to only pool the evidential parts that are semantically important to an event query [22, 36]. The idea is based on the underlying assumption that important evidence is sparsely scattered in a video's timeline, thus aggregating all keyframes like average pooling would collect a bunch of junk information. Yu et al. [36] learned the concept importance of a small concept set and pooled the low-level features according to the importance of their related concepts. Mettes et al. [22] clustered the keyframes into fragment proposals and learned the importance of each proposal. Lately, more complex work tends to optimize the event detection by jointly discovering the evidence, given that a good recounting should assist detection in the first place [4, 30], rather than only interpret detection [8]. All the above works need training examples to discover discriminative segments for pooling. Furthermore, Bhattacharya et al. [1] conducted intensive user study and found that a human can recognize most events by merely looking into very few sample segments of a video. Inspired by this finding, we take query words to locate key evidence in a video for pooling. Unlike previous works, the key evidence is proposed without any need of training videos. We show that event search is robust to noisy concepts even by exploiting as few as three pieces of key evidence for video representation.

3. SEMANTIC QUERY GENERATION

Semantic query generation, as shown in Figure 2, maps the user generated *event query* to an internal, concept-based representation, a.k.a. *semantic query*. This query translation has two goals in general: First, the semantic query can be directly used for event search by the system. Second, the semantic query is able to reflect the essentials of event query for event detection purpose. In addition, the readability, i.e. how concise the semantic query is, is an optional goal if humans are considered to interactively refine the semantic query. We investigate two major steps that have significant impact on retrieval performance in semantic query generation.

3.1 Concept matching

Matching concepts in concept bank requires to measure how close each concept is to the query. We design a flexible phrase matching method that can easily combine with word similarities of WordNet [23], and weighting policies, such as TF-IDF and word specificity.

The concept matching process can be formalized as a mapping from query tokens $\boldsymbol{q}$ to the concepts $\boldsymbol{c}$, denoted as $map(\boldsymbol{q}, \boldsymbol{c})$. Then, we denote the name of one of the concepts as $\boldsymbol{c}_i$, a noun or verb phrase from the query tokens as $\boldsymbol{q}_j$, and the set of common words[1] between them as $\boldsymbol{U}$. Obviously, $\boldsymbol{U} = \boldsymbol{T}(\boldsymbol{c}_i) \cap \boldsymbol{T}(\boldsymbol{q}_j)$, where $\boldsymbol{T}(\boldsymbol{c}_i)$ stands for the set of words in the concept name $\boldsymbol{c}_i$, and $\boldsymbol{T}(\boldsymbol{q}_j)$ likewise. The concept matching is solved by first calculating the similarities for each $(\boldsymbol{c}_i, \boldsymbol{q}_j)$ pair, then selecting the concept with the maximum similarity for each phrase $\boldsymbol{q}_j$, thus forming a set of event-related concepts.

Specifically, a similarity matrix $\boldsymbol{S}$ is first obtained by calculating the similarity for each $(\boldsymbol{c}_i, \boldsymbol{q}_j)$ pair via word-to-word matching, given

$$sim(\boldsymbol{c}_i, \boldsymbol{q}_j) = \frac{|\boldsymbol{U}|}{|\boldsymbol{c}_i|} \cdot \max(\boldsymbol{t}_{\boldsymbol{U}} \odot \log \boldsymbol{s}_{\boldsymbol{U}}) \tag{1}$$

where $\boldsymbol{t}_{\boldsymbol{U}}$ is the TF-IDF weights, in which TF (term frequency) represents the frequency of appearance in event query for the words in $\boldsymbol{U}$, and IDF (inverse document frequency) is estimated based on a collection of Wikipedia pages downloaded from the Web. $\boldsymbol{s}_{\boldsymbol{U}}$ is the word specificity vector defined by the minimum depths of word for the words in $\boldsymbol{U}$ based on WordNet hierarchy. $\odot$ stands for element-wise product. $\max(\cdot)$ operation only picks up the maximum value. $|\boldsymbol{U}|$ means the number of common words, and $|\boldsymbol{c}_i|$ likewise. $\frac{|\boldsymbol{U}|}{|\boldsymbol{c}_i|}$ gives credits to an exact match of the concept name compared to a partial match. Note that for the concepts which do not have any overlapped words with token phrases, the corresponding similarities are set to 0. Then, we define the max-filtered similarity matrix $\boldsymbol{S}^{\boldsymbol{F}}$ as

$$[\boldsymbol{S}^{\boldsymbol{F}}_{ij}] = \begin{cases} sim(\boldsymbol{c}_i, \boldsymbol{q}_j) & \text{if } sim(\boldsymbol{c}_i, \boldsymbol{q}_j) = \max\limits_k sim(\boldsymbol{c}_k, \boldsymbol{q}_j); \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

[1]We find it helpful to consider synonyms as common words as a supplement to the overlapped words. For example, "baby" and "infant" can be regarded as common words.

Type	*Concepts*
Relevant	Rock climbing, bouldering, sport climbing, artificial rock wall
Non-discriminative	Rope climbing, climbing, rock
False positive	Rock fishing, rock band performance
Different context	Stone wall, grabbing rock

Table 1: Concept examples for manual concept selection regarding the event *rock climbing*.

in which we only retain the concept with the maximum similarity given a phrase $\boldsymbol{q}_j$. Finally, we map the text queries $\boldsymbol{q}$ to concepts $\boldsymbol{c}$ by

$$\boldsymbol{w} = map(\boldsymbol{q}, \boldsymbol{c}) = \boldsymbol{S^F} \cdot \mathbb{1} \tag{3}$$

which results in a sparse weight vector $\boldsymbol{w}$, with each dimension representing the importance of a concept.

Given the concept detector responses of all the videos $\boldsymbol{R}$, the event search now can simply be performed by calculating the product of the videos and query:

$$\boldsymbol{h} = \boldsymbol{R}\boldsymbol{w} \tag{4}$$

3.2 Concept selection

Given an event query, concept matching can basically match to dozens of concepts with a concept bank sized about 3,000. Using all of the matches for event search would eventually result in poor performance. On the one hand, this is because concepts with lower importance are less discriminative to the event. For instance, although the concept *water* frequently appears in the event *distinguishing a fire*, water is too common to be seen in many other events such as *competitive swimming*, *bungee jumping*, and *cleaning an appliance*. On the other hand, a concept with low importance is more likely to be a false positive example. Therefore, it makes sense to deploy a concept selection module aiming to filter out irrelevant concepts.

Automatic concept selection simply rank the concepts by importance given the weight vector $\boldsymbol{w}$, then select the top k concepts for semantic query generation.

Manual concept selection employs human subjects to perform concept screening. Along this process, all noisy concepts are expected to be removed, leaving only relevant and discriminative concepts. We basically follow the criteria below for manual concept screening:

- *Relevancy*: Remove false positives by looking at the names of concepts;
- *Context relatedness*: Remove concepts for which training videos appear in different context based on human's common sense;
- *Discrimination*: Only include concepts that are discriminative to this event if a higher-level activity/event detector that matches the event is found.

These criteria are exploited based on the results of automatic concept selection. Take the event *rock climbing* as an example, semantic query generation can pick up roughly 50 concepts by automatic concept matching. A human judge then needs to quickly screen these concepts and remove irrelevant ones. Table 1 shows different types of irrelevant concepts, which considers relevancy, discrimination, and context relatedness. For example, *Rope climbing* is not discriminative because it can also appear in the event *climbing a tree*; *grabbing rock* is removed as it may present in different context, e.g. alongside a river. It's worth to mention that *artificial rock wall* can be regarded as a relevant concept because, based on common sense, artificial rock wall is only used for practicing rock climbing. As a rule of thumb, it's practically effective to only represent an event by relevant concepts. The non-discriminative concepts and concepts that come out in different context are still helpful if no relevant concepts can be found.

Algorithm 1 Evidential pooling for one video

input: weight vector $\boldsymbol{w}$, keyframe-level concept detector responses $\boldsymbol{J}$, number of concepts restricted for evidence proposal k_e, number of evidential shots m
output: video-level concept detector responses $\boldsymbol{r}$

For concepts not ranking in the top k_e of $\boldsymbol{w}$, set their weights in $\boldsymbol{w}$ to 0, getting $\boldsymbol{w_e}$;
Calculate $\boldsymbol{J}\boldsymbol{w_e}$, the importance scores for all the keyframes;
Rank all the keyframes by their importance scores;
$n \leftarrow 0$, $G \leftarrow \varnothing$;
while $n < m$ **and** not all the keyframes are processed **do**
  get the next important keyframe y in the rank list;
  if $G = \varnothing$ **or** y is not adjacent to any of the keyframes in G **then**
    add y to a new set P;
    add set P to G;
    $n \leftarrow n + 1$;
  else
    add y to the existing set P in G where P has a keyframe adjacent to y;
  end if
  if $n > 1$ **then**
    for each $(P_i, P_j) \in G$ **do**
      if P_i has a keyframe adjacent to P_j **then**
        merge P_i and P_j into one set;
        $n \leftarrow n - 1$;
      end if
    end for
  end if
end while
$\boldsymbol{r} \leftarrow$ Average pooling J for the selected keyframes in G only;
return $\boldsymbol{r}$;

4. EVIDENTIAL POOLING

Automatic concept selection simply caps the number of concepts used in semantic query in order to avoid the involvement of irrelevant concepts. Although this method is generally simple and efficient, it has two major drawbacks. First, it is difficult to determine a threshold k because the number of concepts that achieves the best performance varies between events; second, using fewer number of concepts for event representation is more sensitive to noisy concepts, given that each concept would count much. We would like to have a best of both worlds approach that, it can both benefit from a representation with larger number of concepts and resist the negative impact from less relevant or even irrelevant concepts.

Our idea is majorly enlightened by Bhattacharya et al.[1] that, for event recognition, humans can make a quick decision in most cases by just looking into several shots of a video. This fact indicates that merely a few shots can provide sufficient information to reach a decision. We further hypothesis that using few sample shots of a video is more discriminative than using all of its content. Specifically, consider the case of a video only represented by the most evidential shot: a large portion of the video is thus stripped

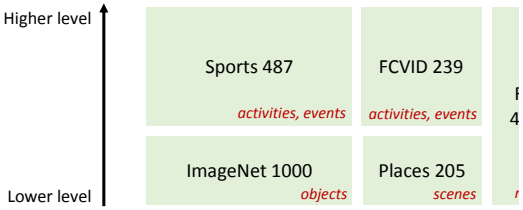

Figure 3: Concept bank overview. 2,774 concept off-the-shelf datasets are collected. *ImageNet 1(* cialized in objects, *Places 205* in scenes. *Sport FCVID 239* mainly contain higher-level semanti activities and events. *Research Collection (RC SIN 346* are mixed with concepts of different gra

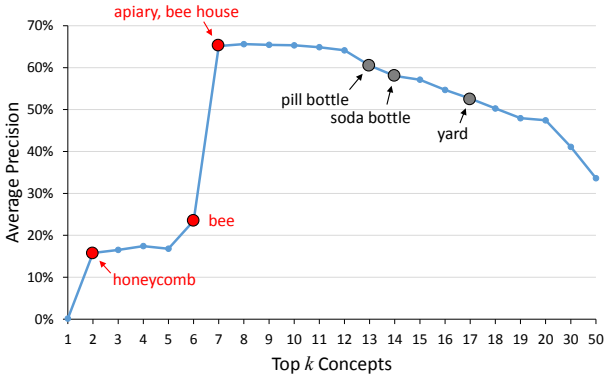

Figure 4: For the exemplar event query *beekeeping*, the performance drops significantly when adding more concepts to the semantic query. Concepts having a large impact are marked in the figure.

off, making the representation sparse and "clean" a neat outcome of this clean representation – *it is relatively robust against the noisy concepts in the semantic query.* For instance, a typical video clip of *bike trick* might happen on a street. A prolonged clip would capture many objects on the street, such as cars, buses and traffic lights. If a given event query *parking a car* also expects the presence of such street objects, the clip of *bike trick* is likely to be detected as a false positive, simply because it happily responds to the semantic query that expects these non-discriminative street objects. But the clean representation which takes only the most evidential shot is less likely to suffer from this problem, as the representation is more concentrated. Rare concepts such as buses and traffic lights are naturally filtered out.

Algorithm 1 details the evidential pooling. The pivot is finding m evidential shots. The algorithm takes the semantic query as information need to rank keyframes in a video (steps 1–3). Evidential keyframes are sequentially selected based on the rank (steps 5–22) while aggregating them into evidential shots (steps 15–20). Video representation is formed by pooling through m evidential shots only, rather than all the keyframes of the video clip (step 23). It is important to note that the ranking of keyframes is based on $\boldsymbol{w_e}$, where the number of concepts is restricted to k_e (step 1). Basically, k_e is set to a small number concerning that the evidential shots should be proposed by the most relevant concepts selected by k_e. We actually find that k_e can be chosen from a wide range that does not affect the performance. It is also worth noting that k_e is different from k which is used for automatic concept selection. With evidential pooling, the event search should be tolerant to noise, i.e., less sensitive to the value of k in the automatic concept selection phase.

5. EXPERIMENTS

5.1 Setups

The experiments are conducted on the TRECVID Multimedia Event Detection (MED) datasets. We use the event kits that contain 20 event queries from E021 to E040. The test set, named MED14Test, includes around 25,000 testing videos with no textual metadata. The performance is evaluated by the standard metric Mean Average Precision (MAP). The ground truth of MED14Test set is officially provided for self-evaluation [26]. As for 0Ex, no training videos are used throughout the experiments. Besides, we also show the official results on MED15Eval released by NIST. MED15Eval set has two divisions: MED15EvalFull contains 200,000 videos without textual metadata; MED15EvalSub is a random subset of MED15EvalFull, which contains around 32,000 videos. No ground truth is provided for MED15Eval.

5.2 Concept bank

We build a large concept bank containing a total of 2,774 semantic concepts with varied granularity. They are sourced from off-the-shelf datasets with concept types covering objects, scenes, actions and activities. The six datasets and their settings are listed below:

- ***ImageNet 1000:*** We use the same DCNN architecture proposed in [18]. Specifically, the DCNN architecture can be denoted as $Image - C48 - P - N - C128 - P - N - C192 - C192 - C128 - P - F4096 - F4096 - F1000$, in which C are the convolutional layers followed by the number of filters, F are the fully-connected layers, P are the max-pooling layers and N are the local contrast normalization layers. The parameters of DCNN are learnt on ILSVRC 2012 [7]. The neural responses of the eight layer (F1000) are extracted for each keyframe of a test video.
- ***SIN 346:*** A set of 346 concept detectors fine-tuned with AlexNet DCNN structure on the TRECVID SIN 2014 dataset [37]. The concept responses are extracted for each keyframe of a test video.
- ***Research Collection 497:*** Similar to [24], we select 497 frequent concepts from the MED'14 Research Collection dataset [29]. At most 200 positive keyframes are manually annotated for each concept. We fine-tune 497 concept detectors using the AlexNet DCNN architecture. As with previous methods, we extract the responses of the concept detectors for each keyframe of a test video.
- ***Places 205:*** Multimedia events usually take place in notable scenes, e.g., *bike trick* is played in a park or on a street, whereas *cleaning an appliance* happens in a kitchen or shop. To capture the scene information, we fine-tune 205 scene categories on the MIT Places dataset [38] using the AlexNet DCNN architecture. The concept responses are extracted for each keyframe of a test video.

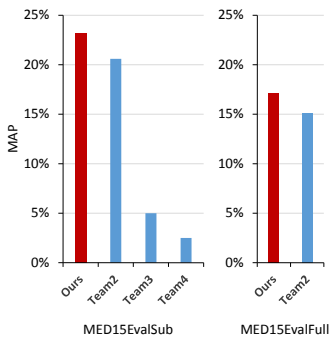

Figure 5: The official results of 0Ex MAP performance on MED15Eval released by NIST.

Dataset	*#Concepts*	*Optimum k*	*MAP*
Sports	487	10	0.103
FCVID	239	1	0.071
Research Collection	497	2	0.053
ImageNet	1,000	3	0.049
Places	205	2	0.020
SIN	346	5	0.014
Concept Bank	2,774	9	0.129
AutoSQGSys [15]	4,043	-	0.115

Table 2: The optimum of concepts k per dataset using automatic concept selection. The MAP is reported on MED14Test.

- ***FCVID 239:*** To capture the high level activities of multimedia events, 239 concept detectors are trained with SVMs on FCVID [16] dataset. Specifically, as concepts in this dataset are mainly annotated at video level, we first extract the AlexNet Layer-7 [18] responses for each keyframe in a training video. Then, the training features for a video are generated by average pooling the responses of all the keyframes. The concept detectors are trained by these video-level training features. A similar pipeline can be applied to test videos to extract video-level testing features. Finally, the responses of 205 concept detectors can be directly extracted for test videos.
- ***Sports 487:*** The 487 concepts of general sports are trained with the 3D-CNN architecture [32] on the Sport-1M dataset [17] that contains one million videos. The concept responses are extracted for each of 15-frame segments of a test video.

Figure 3 summarizes the granularity of concepts for the six datasets.

5.3 Concept selection

Automatic concept selection To demonstrate the importance of concept selection, we illustrate an event *beekeeping* in Figure 4. By concept matching, 34 concepts are chosen as candidates for this event. We then employ automatic concept selection to only retain the top k concepts. As the number k increases, the performance increases at first, but decreases dramatically when $k > 12$, dropping almost a half at the end. From the figure, we see three concepts that largely contribute to the performance increase: *honeycomb*, *bee* and *apiary (bee house)*. These concepts are relevant and discriminative. In contrast, as k further increases, more

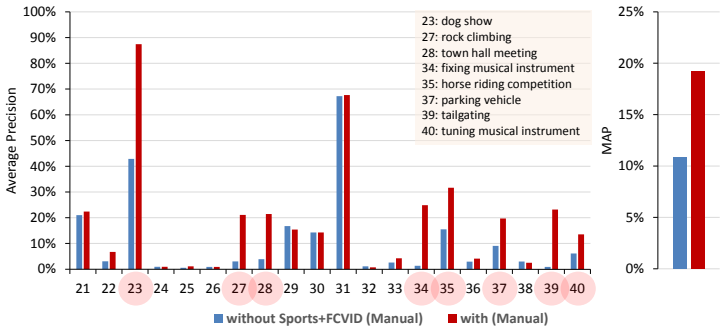

Figure 6: The contribution of concepts from Sports 487 and FCVID 239 on MED14Test. Both the Average Precision (AP) per event and MAP are shown. Concepts from these two datasets have a large contribution on 8 of the 20 events, leading to a boost of MAP.

k_e	2	4	8	16
MAP	0.0725	0.0771	0.0775	0.0770

Table 3: Evidential pooling is not sensitive to the number of concepts k_e used for evidence proposal. For this test, all the matched concepts are used for event search. The MAP is based on MED14Test.

concepts that are less relevant are involved. For example, a *bottle* as beekeeping container is different from a *pill bottle*. Although there is a partial match in their names, the context of these two concepts is different. *Yard* is another example which turns down the performance. One might intuitively think that *yard* can be helpful in recognizing beekeeping, given that beekeeping usually takes place in a yard. However, due to the presence of more accurate and discriminative concepts like *honeycomb* and *apiary*, high performance has already been achieved. Adding a less discriminative concept *yard* would decrease the performance in this case, since *yard* is commonly seen in many other activities, which tends to introduce noise.

As the performance is sensitive to the number of concepts, a good practice for automatic concept selection is to stop at a small k. For the beekeeping example shown above, the optimal number k is 8. We find that the optimum k is bound to both event query and concept bank. Table 2 lists the optimum k for each individual dataset as well as the merged concept bank tested on MED14Test. As you can see, many datasets have their best performance by less than 4 concepts. Note that the performance of our concept bank with automatic concept selection has already exceeded the state-of-the-art full system AutoSQGSys as described by Jiang et al. [15] even without combining the ASR and OCR features.

Manual concept selection By concept screening according to the rules in Section 3.2, we refine the automatically matched concepts for all the 20 events. Note that manual concept screening is allowed in TRECVID and used by many teams [15, 26]. The manually refined semantic queries, combined with ASR and OCR features, provide us clear advantage against other teams (Figure 5), based on the 0Ex evaluation in TRECVID 2015.

We further investigate the contribution of concepts according to their source datasets. Figure 6 shows a large

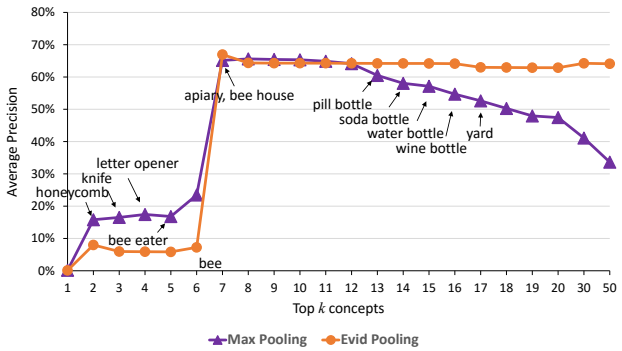

Figure 7: Evidential pooling vs. max pooling regarding the event query *beekeeping*.

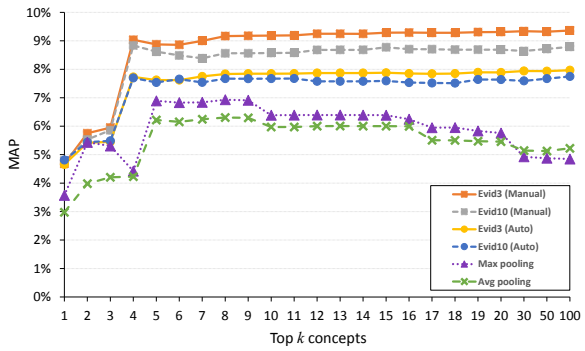

Figure 8: Evidential pooling with various settings for the 20 events in MED14Test. The MAP with regard to the top k concepts is shown for each pooling method.

Pooling Method	*MAP*
3 shots proposed by manually refined SQ	0.0936
10 shots proposed by manually refined SQ	0.0879
3 shots proposed by automatic SQ	0.0797
10 shots proposed by automatic SQ	0.0775
Max pooling	0.0485
Average pooling	0.0522

Table 4: The performance summarization of Figure 8 when all the matched concepts are used for event search. The settings are exactly the same as in Figure 8. SQ stands for Semantic Query. The k_e in automatic SQ is set to 8.

gap of MAP between systems with and without the higher-level concepts from Sports-1M and FCVID. Semantic queries for both systems are manually refined. As seen in the figure, the MAP drops more than 40% by removing these two datasets. We therefore conclude that both lower and higher level concepts devote to a good overall performance. Higher-level concept detectors, such as the activity detectors from FCVID, are accurate and discriminative. However, they require exact matches to the information need, otherwise they are not applicable. For instance, the activity *"professional American football"* would only satisfy queries of American football game. In contrast, lower-level concepts such as objects are componential and complementary. Although they are less discriminative, they offer help in case no exact match can be found among the higher-level concepts. Figure 6 also gives such insights when looking into the comparison of each individual event: Almost half of the events are the events with exact matches in Sports-1M or FCVID, hence their performance boosts. On the opposite side, for the events in which exact matches are not found, the contributions of Sports-1M and FCVID are relatively low.

5.4 Evidential pooling

Algorithm 1 tries to localize key evidence in a video according to a given semantic query. To generate a semantic query suitable for evidence proposal, we first cap the number of concepts k_e to a small value. This makes sense because, for localizing key evidence, we only need to focus on precision regardless of recall. We hence expect the semantic query to be as precise as possible. Table 3 indeed proves that it is acceptable to choose from a wide range of k_e. Even a very risky setting, e.g. $k_e = 2$, can have a satisfactory retrieval performance. We simply choose $k_e = 8$ and keep it fixed for all the automatic runs. Then, for each video clip, we locate the evidential shots regarding the semantic query. It is worth mentioning that the evidential shots can be generated for arbitrary videos, but in this case the content of these shots is likely to be arbitrary too.

We compare the evidential pooling to the conventional max/average pooling[2]. First, we review the *beekeeping* event in Figure 7. The evidential pooling shows clear advantage over max pooling: As the number of concepts k increases, max pooling tends to suffer much from less relevant and noisy concepts. In contrast, evidential pooling is considerably robust. By using all of the 34 matched concepts, max pooling drops to an AP of 0.336, while evidential pooling stays above 0.64, which is almost twice the AP of max pooling. To investigate whether this observation is generalizable, we further conduct experiments on all the 20 events in MED14Test. Figure 8 shows the performance of evidential pooling with various settings. The corresponding MAP scores are summarized in Table 4. The benefit of evidential pooling is clearly seen over the 20 events. First, both max and average pooling are sensitive to k, whereas evidential pooling is robust to a wide range of k. Second, using all the matched concepts without concept selection, evidential pooling clearly overwhelms max and average pooling by more than 50% of MAP. Third, merely exploiting three evidential shots for pooling can achieve a great performance. Fourth, although proposing evidence by manually refined semantic query performs better, automatic semantic query is not far behind. Most importantly, the event detection using evidential pooling with automatic semantic query is not sensitive to the parameter k_e. Finally, we see that evidential pooling is not only robust, the MAP it reaches by using all the concepts is but also higher than the peak MAP of max pooling by picking the top 8 concepts.

[2]These experiments use a smaller concept bank that does not include FCVID and Sports-1M datasets as lacking of keyframe-based concept responses.

6. CONCLUSION

We presented a framework that practically achieved good performance in 0Ex. Because query representation relying on a large concept bank is sensitive to noisy concepts, restricting the number of concepts and manually removing the weakly relevant concepts are crucial to the performance. Evidential pooling, nevertheless, can involve more concepts

including that of weakly relevant without sacrificing performance. From the experimental results we arrive at the following conclusions:

- An effective practice for generating the semantic query is to keep the concepts precise and discriminative. Using fewer but precise concepts performs better than more but less precise concepts.
- For manual concept screening, we recommend to remove concepts based on their relevancy, context relatedness and discrimination.
- Pooling from the evidential shots rather than all the keyframes can be robust to the noise in the semantic query as potentially non-evidential information is excluded from the search process.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213).

7. REFERENCES

[1] S. Bhattacharya, F. X. Yu, and S.-F. Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *ICMR*, pages 105–112, New York, NY, USA, 2014.

[2] M. Boer, K. Schutte, and W. Kraaij. Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications*, pages 1–19, 2015.

[3] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015.

[4] X. Chang, Y.-L. Yu, Y. Yang, and A. G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *23rd ACM MM*, pages 581–590, New York, NY, USA, 2015. ACM.

[5] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *22Nd CIKM*, pages 1857–1860, New York, NY, USA, 2013. ACM.

[6] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The youtube video recommendation system. In *4th ACM RecSys*, pages 293–296, New York, NY, USA, 2010.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[8] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Video event recounting using mixture subclass discriminant analysis. In *ICIP*, pages 4372–4376, Sept 2013.

[9] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *22Nd ACM MM*, pages 17–26, New York, NY, USA, 2014.

[10] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, pages 17–24, New York, NY, USA, 2014.

[11] A. Habibian, K. E. van de Sande, and C. G. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, pages 89–96, New York, 2013.

[12] L. Jiang, A. G. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *20th ACM MM*, pages 449–458, New York, NY, USA, 2012.

[13] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *22Nd ACM MM*, pages 547–556, New York, 2014.

[14] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, pages 297–304, New York, NY, USA, 2014.

[15] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *5th ICMR*, pages 27–34, New York, NY, USA, 2015.

[16] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv:1502.07209*, 2015.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732. IEEE, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[19] M. Mazloom, A. Habibian, and C. G. Snoek. Querying for video events by semantic signatures from few examples. In *21st ACM MM*, pages 609–612, New York, NY, USA, 2013.

[20] M. Mazloom, X. Li, and C. G. M. Snoek. Few-example video event retrieval using tag propagation. In *ICMR*, pages 459–462, New York, NY, USA, 2014.

[21] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, Feb 2012.

[22] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, and C. G. Snoek. Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting. In *5th ICMR*, pages 427–434, New York, NY, USA, 2015.

[23] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[24] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo. BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems. In *NIST TRECVID Workshop*, 2013.

[25] S. Oh, S. Mccloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. G. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Mach. Vision Appl.*, 25(1):49–69, Jan. 2014.

[26] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Queenot, and R. Ordelman. TRECVID 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[27] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *ICMR*, pages 265–272, New York, NY, USA, 2014.

[28] B. Singh, X. Han, Z. Wu, V. I. Morariu, and L. S. Davis. Selecting relevant web trained concepts for automated event retrieval. In *ICCV*, pages 4561–4569, Dec 2015.

[29] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous audio visual internet collection. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, Istanbul, Turkey, may 2012. ELRA.

[30] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, pages 2569–2576, June 2014.

[31] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003-2009). *J. Am. Soc. Inf. Sci. Technol.*, 62(4):613–627, Apr. 2011.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[33] F. Wang, Z. Sun, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and feature selection. *IEEE Transactions on Multimedia*, 16(5):1303–1315, Aug 2014.

[34] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, pages 2665–2672, June 2014.

[35] 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 *23rd ACM MM*, pages 471–480, New York, NY, USA, 2015. ACM.

[36] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Semantic pooling for complex event detection. In *21st ACM MM*, pages 733–736, New York, NY, USA, 2013.

[37] W. Zhang, H. Zhang, T. Yao, Y. Lu, J. Chen, and C.-W. Ngo. VIREO @ TRECVID 2014: instance search and semantic indexing. In *NIST TRECVID Workshop*, 2014.

[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.