

# Selecting Relevant Web Trained Concepts for Automated Event Retrieval

Bharat Singh\*, Xintong Han\*, Zhe Wu, Vlad I. Morariu and Larry S. Davis  
Center For Automation Research, University of Maryland, College Park

{bharat, xintong, zhewu, morariu, lsd}@umiacs.umd.edu

## Abstract

*Complex event retrieval is a challenging research problem, especially when no training videos are available. An alternative to collecting training videos is to train a large semantic concept bank a priori. Given a text description of an event, event retrieval is performed by selecting concepts linguistically related to the event description and fusing the concept responses on unseen videos. However, defining an exhaustive concept lexicon and pre-training it requires vast computational resources. Therefore, recent approaches automate concept discovery and training by leveraging large amounts of weakly annotated web data. Compact visually salient concepts are automatically obtained by the use of concept pairs or, more generally, n-grams. However, not all visually salient n-grams are necessarily useful for an event query—some combinations of concepts may be visually compact but irrelevant—and this drastically affects performance. We propose an event retrieval algorithm that constructs pairs of automatically discovered concepts and then prunes those concepts that are unlikely to be helpful for retrieval. Pruning depends both on the query and on the specific video instance being evaluated. Our approach also addresses calibration and domain adaptation issues that arise when applying concept detectors to unseen videos. We demonstrate large improvements over other vision based systems on the TRECVID MED 13 dataset.*

## 1. Introduction

Complex event retrieval from databases of videos is difficult because in addition to the challenges in modeling the appearance of static visual concepts—e.g., objects, scenes—modeling events also involves modeling temporal variations. In addition to the challenges of representing motion features and time, one particularly pernicious challenge is that the number of potential events is much greater than the number of static visual concepts, amplifying the well-

known long-tail problem associated with object categories. Identifying and collecting training data for a comprehensive set of objects is difficult. For complex events, however, the task of even enumerating a comprehensive set of events is daunting, and collecting curated training video datasets for them is entirely impractical.

Consequently, a recent trend in the event retrieval community is to define a set of simpler visual concepts that are practical to model and then combine these concepts to define and detect complex events. This is often done when no examples of the complex event of interest are available for training. In this setting, training data is still required, but only for the more limited and simpler concepts. For example, [5, 21] discover and model concepts based on single words or short phrases, taking into account how *visual* the concept is. Others model pairs of words or n-grams in order to disambiguate between the multiple visual meanings of a single word [9] and take advantage of co-occurrences present in the visual world [23]. An important aspect of recent work [29, 5] is that concept discovery and training set annotation is performed automatically using weakly annotated web data. Event retrieval is performed by selecting concepts linguistically related to the event description and computing an average of the concept responses as a measure for event detection.

Based on recent advances, we describe a system that ranks videos based on their similarity to a textual description of a complex event, using only web resources and without additional human supervision. In our approach, the textual description is represented by and detected through a set of concepts. Our approach builds on [5] for discovering concepts given a textual description of a complex event, and [9] for automatically replacing the initial concepts with *concept pairs* that are visually salient and capture specific visual meanings.

However, we observe that many visually salient concepts generated from an event description are not useful for detecting the event. In fact, we find that removing certain concepts is a key step that significantly improves event retrieval performance. Some concepts should be removed at training time because they model visually salient concepts that are

\*The first two authors contributed equally to this paper.

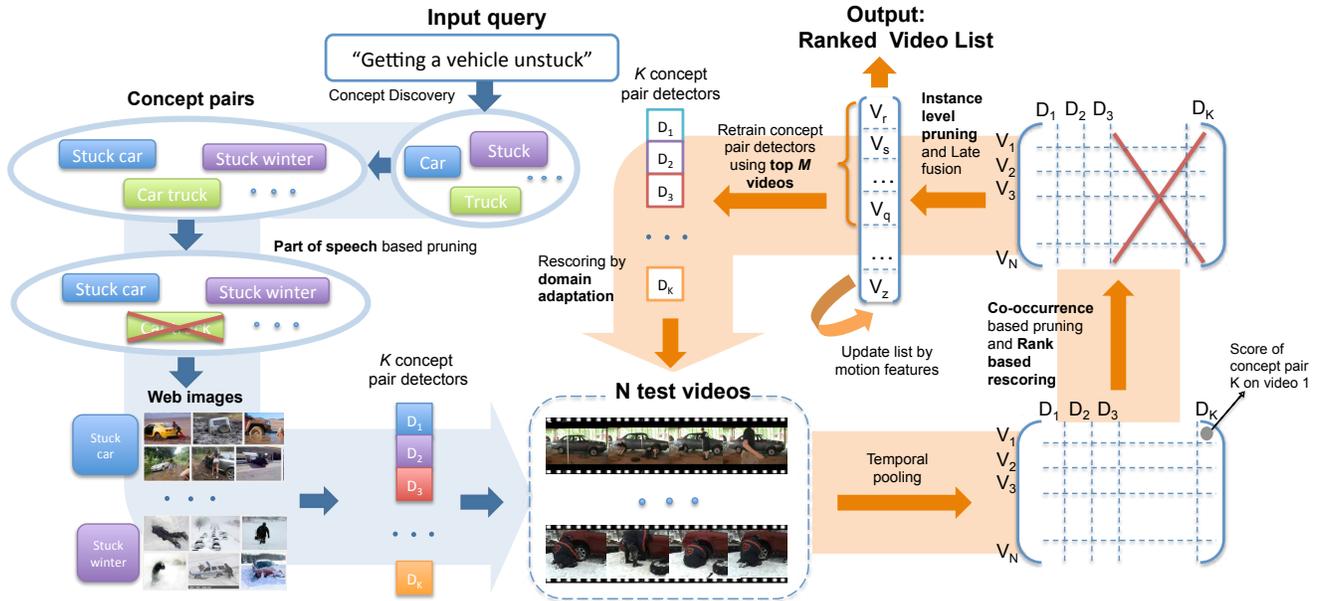


Figure 1. Framework overview. An initial set of concepts is discovered from the web and transformed to concept pairs using an action centric part of speech (grammar) model. These concept pairs are used as Google Image search text queries, and detectors are trained on the search results. Based on the detector scores on the test videos, co-occurrence based pruning removes concepts that are likely to be outliers. Detectors are calibrated using a rank based re-scoring method. An instance level pruning method determines how many concepts are likely to be observed in a video and discards the lowest scoring concepts. The scores of remaining concepts are fused to score each video. Motion features of the top ranked videos are used to train a SVM and update the video list. Finally, the initial detectors are re-trained using the top ranked videos of this video list, and the process of co-occurrence based pruning, instance level pruning and rank based calibration is repeated to re-score the videos.

not likely to be meaningful based on linguistic considerations. Others should be removed if an analysis of video co-occurrences and activation patterns indicates that a concept is likely to be irrelevant or not among the subset of concepts that occur in a video instance. These problems are further confounded by the fact that concept detectors are initially trained on weakly supervised web images<sup>1</sup>, so there is a domain shift to video, and detector responses are not properly calibrated.

Our contribution is a fully automatic algorithm that discovers concepts that are not only visually salient, but are also likely to predict complex events by exploiting co-occurrence statistics and activation patterns of concepts. We address domain adaptation and calibration issues in addition to modelling the temporal properties. Evaluations are conducted using the TRECVID EK0 dataset, where our system outperforms state-of-the-art methods based on visual information.

<sup>1</sup>We prefer to use web images for concept training because a web search is a weak form of supervision which provides no spatial or temporal localization. This means that if we search for video examples of a concept, we do not know how many and which frames contain the concept (a temporal localization issue), while an image result is much more likely to contain the concept of interest (the spatial localization still remains).

## 2. Related Work

Large scale video retrieval commonly employs a concept-based video representation (CBRE) [1, 22, 24, 30], especially when only few or no training examples of the events are available. In this setting, complex events are represented in terms of a large set of concepts that are either event-driven (generated once the event description is known) [5, 13, 21] or pre-defined [29, 7, 8]. A test query description is mapped to a set of concepts whose detectors are then applied to videos to perform retrieval. However, methods based on pre-defined concepts need to train an exhaustive set of concept detectors a priori or the semantic gap between the query description and the concept database might be too large. This is computationally expensive and currently infeasible for real-world video retrieval systems. Instead, in this paper, given the textual description of the event to be retrieved, our approach leverages web image data to discover event-driven concepts and train detectors that are relevant to this specific event.

Recently, web (Internet) data has been widely used for knowledge discovery [11, 2, 9, 29, 10, 14]. Chen et al. [6] use web data to weakly label images, learn and exploit common sense relationships. Berg et al. [2] automatically discover attributes from unlabeled Internet images and

their associated textual descriptions. Duan et al. [11] describe a system that uses a large amount of weakly labeled web videos for visual event recognition by measuring the distance between two videos and a new transfer learning method. Habibian et al. [14] obtain textual descriptions of videos from the web and learn a multimedia embedding for few-example event recognition. For concept training, given a list of concepts, each corresponding to a word or short phrase, web search is commonly used to construct weakly annotated training sets [5, 29, 9]. We use the concept name as a query to a search engine, and train the concept detector based on the returned images.

Moreover, retrieval performance depends on high quality concept detectors. While the performance of a concept detector can be estimated (e.g., by cross-validation [9]), ambiguity remains in associating linguistic concepts to visual concepts. For example, *groom* in *grooming an animal* and *groom* in *wedding ceremony* are totally different, and while two separate detectors might be capable of modeling both types of *groom* separately, a single *groom* detector would likely perform poorly. Similarly, tire images from the web are different from frames containing tires in a video about *changing a vehicle tire*, since there are often people and cars in these frames. To solve this problem, [9, 19] use an n-gram model to differentiate between multiple senses of a word. Habibian et al. [13] instead leverage logical relationships (e.g., “OR”, “AND”, “XOR”) between two concepts. Mensink et al. [23] exploit label co-occurrence statistics to address zero-shot image classification. However, it is not sufficient to discover visually distinctive concepts, since not all concepts are equally informative for modeling events. We present a pruning process to discover visually distinctive *and useful* concepts by a pruning process.

Recent work has also explored multiple modalities—e.g., automatic speech recognition (ASR), optical character recognition (OCR), audio, and vision—for event detection [16, 17, 29] to achieve better performance over vision alone. Jiang et al. [17] propose MultiModel Pseudo Relevance Feedback (MMPRF), which selects several feedback videos for each modality to train a joint model. Applied to test videos, the model yields a new ranked video list that is used as feedback to retrain the model. Wu et al. [29] represent a video by using a large concept bank, speech information, and video text. These features are projected to a high-dimensional concept space, where event/video similarity scores are computed to rank videos. While multi-modal techniques achieve good performance, their visual components alone significantly under-perform the system as a whole.

All these methods suffer from calibration and domain adaptation issues, since CBRE methods fuse multiple concept detector responses and are usually trained and tested on different domains. To deal with calibration issues, most

related work uses SVMs with probabilistic outputs [20]. However, the domain shift between web training data and test videos is usually not addressed by calibration alone. To reduce this effect, some ranking-based re-scoring schemes [16, 17] replace raw detector confidences with the confidence rank in a list of videos. To further adapt to new domains (e.g., from images to videos), easy samples have been used to update detector models [27, 16]. Similar to these approaches, we use a rank based re-scoring scheme to address calibration issues and update models using the most confident detections to adapt to new domains.

### 3. Overview

The framework of our algorithm is shown in Fig. 1. Given an event defined as a text query, our algorithm retrieves and ranks videos by relevance. The algorithm first constructs a bank of concepts by the approach of [5] and transforms it into concept pairs. These concept pairs are then pruned by a part of speech model. Each remaining concept pair is used as a text query in a search engine (Google Images), and the returned images are used to train detectors, which are then applied to the test videos. Based on detector responses on test videos, co-occurrence based pruning removes concept pairs that are likely to be outliers. Detectors are then calibrated using a rank based re-scoring method. An instance level pruning method determines how many concept pairs should be observed in a video from the class, discarding the lowest scoring concepts. The scores of the remaining concept pairs are fused to rank the videos. Motion features of the top ranked videos are then used to train a SVM and re-rank the video list. Finally, the top ranked videos are used to re-train the concept detectors, and we use these detectors to re-score the videos.

The following sections describe each part of our approach in detail.

### 4. Concept Discovery

The concept discovery method of [5] exploits weakly tagged web images and yields an initial list of concepts for an event. Most of these visual concepts correspond to single words, so they may suffer from ambiguity between linguistic and visual concepts. Consequently, we follow [9] by using n-grams to model specific visual meanings of linguistic concepts and [23] by using co-occurrences. From the top  $P$  concepts in the list provided by [5], we combine single-word concepts into pairs and retain the phrase concepts to form a new set of concepts. The resulting concepts reduce visual ambiguity and are more informative. We refer to the concepts trained on pairs of words as *pair-concepts*.

Fig. 2 shows the frames ranked highest by the proposed pair-concept detectors, the original concept detectors for single words, and the sum of two independently trained con-

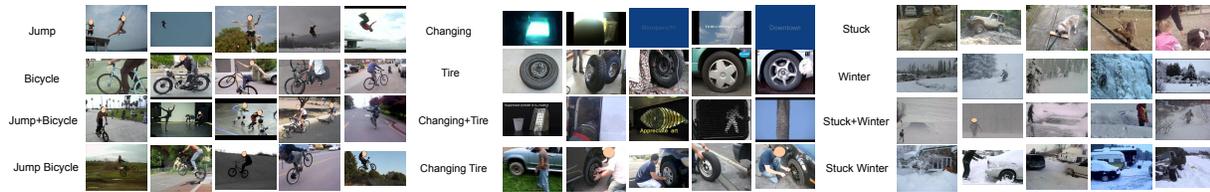


Figure 2. Top five ranked videos by different concept detectors trained using web images for three events: (a) *attempting a bike trick*, (b) *changing a vehicle tire*, (c) *getting a vehicle unstuck*. The first and second rows show the results of running unary concepts on test videos. The third row combines two unary concept detectors by adding their scores. The fourth row shows the results of our proposed pair-concept detectors. Pair-concepts are more effective at discovering frames that are more semantically relevant to the event.

cept detectors on the words constituting the pair-concept. Pair-concept detectors are more relevant to the event than the unary detectors or the sum of two detectors. For example, in Fig. 2, the event query is *attempting a bike trick*, and two related concepts are *jump* and *bicycle*. The *jump* detector can only detect a few instances of jumping, none of which are typical of a bike trick. The *bicycle* detector successfully detects bicycles, but most detections are of people riding bicycles instead of performing a bike trick. If the two detectors are combined by adding their scores, some frames with bikes and jump actions are obtained, but they are still not relevant to *bike trick*. However, the *jump bicycle* detections are much more relevant to *attempting bike trick*—people riding a bicycle are jumping off the ground.

Concepts which do not result in good visual models (e.g., *cute water*, *dancing blood*) can be identified [9, 5]. But, even when concepts lead to good visual models, they might still not be informative (e.g., *car truck*, and *puppy dog*). Moreover, even if concepts are visual and informative, videos do not always exhibit all concepts related to an event, so expecting all concepts to be observed will reduce retrieval precision. For these reasons, it is not only necessary to select concepts that can be modeled visually, but also to identify subsets of them that are useful to the event retrieval task. We propose three concept pruning schemes to remove bad concepts: pruning based on grammatical parts of speech, pruning based on co-occurrence on test videos, and instance level pruning. The first two schemes remove concepts that are unlikely to be informative, while the last identifies a subset of relevant concepts for each video instance.

#### 4.1. Part of speech based pruning

Action centric concepts are effective for video recognition, as shown in [3, 26]. Based on this, we require that a pair-concept contain one of three types of action centric words: 1) Nouns that are events, e.g., party, parade; 2) Nouns that are actions, like celebration, trick; 3) Verbs, e.g., dancing, cooking, running. Word types are determined by their lexical information and frequency counts provided by WordNet [25]. Then, action centric concepts are paired with other concepts that are not action centric to yield the final

set of pair-concepts.

Table 1 shows the pair-concepts discovered for an event. Qualitatively, these concepts are more semantically relevant to events than the *single* word concepts from [5]. An improvement would be to learn the types of pair-concepts that lead to good event models, based on their parts of speech. However, as our qualitative and quantitative results show, the proposed action-centric pruning rule leads to significant improvements over using all pairs, so we leave data-driven learning for future work.

Pair-concept detectors are trained automatically using web images. For each concept, 200 images are chosen as positive examples, downloaded by using the concept as the textual query for image search on Google Images. Then, 500 negative examples are randomly chosen from the images of other concepts from all events. Based on the deep features [15] of these examples, the detectors are trained using a RBF kernel SVM using LibSVM [4] with the default parameters.

#### 4.2. Co-occurrence based pruning

Not all action-centric pair-concepts will be useful, for a number of reasons. First, the process of generating unary-concepts from an event description is uncertain [5], and might generate irrelevant ones. Second, even if both unary concepts are relevant individually, they may lead to non-sensical pairs. And finally, even if both unary concepts are relevant, web search sometimes returns irrelevant images which can pollute the training of concept detectors.

To reduce the influence of visually unrelated and noisy concepts, we search for co-occurrences between detector responses and keep only pair-concepts whose detector outputs co-occur with other pair-concepts at the video level. The intuition is that co-occurrences between good concepts will be more frequent than coincidental co-occurrences between bad concepts. One reason for this is that if two pair-concepts are both relevant to the same complex event, they are more likely to fire in a video of that event. Another reason is that detectors are formed from pairs of concepts, so many pair-concepts will share a unary concept and so are likely to be semantically similar to some extent. For example *cleaning kitchen* and *washing kitchen* share *kitchen*.

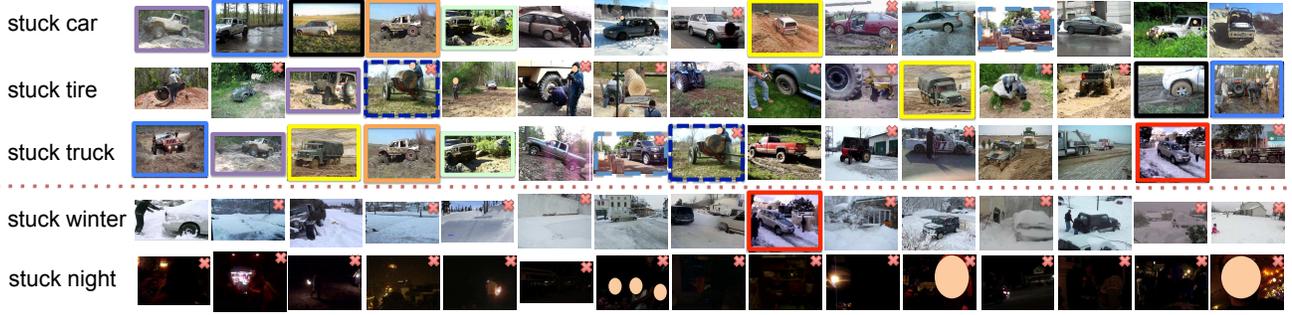


Figure 3. Example of co-occurrence based concept pruning. The five rows correspond to the top 15 videos retrieved by five concept detectors (*stuck car*, *stuck tire*, *stuck truck*, *stuck winter*, *stuck night*) for detecting the event *getting a vehicle unstuck*. Frames from the same videos are marked with bounding boxes of the same color, and repeating colors across concept detectors denote co-occurrences. For example, the yellow border in rows corresponding to *stuck car*, *stuck tire*, and *stuck truck* signifies that all three concept detectors co-occur in the video represented by the yellow color. The solid boxes denote the positive videos, while the dashed ones are negatives. *Stuck night* and *stuck winter* do not co-occur often with other concepts, so they are discarded. Note that the negatives are all marked with red cross in the upper-right corner.

In other cases, pair-concepts may share visual properties as they are derived for a specific event, for example *stuck car* and *stuck tire* can co-occur because a tire can be detected along with a car in a frame or in a video.

Let  $\mathcal{V} = \{V_1, V_2, \dots, V_N\}$  denote the videos in the test dataset, where  $V_i$  contains  $N_i$  frames  $\{V_{i1}, V_{i2}, \dots, V_{iN_i}\}$  (sampled from the video for computational reasons). Given  $K$  concept detectors  $\{D_1, D_2, \dots, D_K\}$  trained on web-images, each concept detector is applied to  $\mathcal{V}$ . For each video  $V_i$ , an  $N_i \times K$  response matrix  $\mathbf{S}_i$  is obtained. Each element in  $\mathbf{S}_i$  is the confidence of a concept for a frame. After employing a hierarchical temporal pooling strategy (described in section 5) on the response matrix, we obtain a confidence score  $s_{ik}$  for the detector  $D_k$  applied to video  $V_i$ . Then for each concept detector  $D_k$ , we rank the  $N$  videos in the test set based on the score  $s_{ik}$ . Let  $L_k$  denote the top  $M$  ranked videos in the ranking list. We construct a co-occurrence matrix  $\mathbf{C}$  as follows:

$$C_{ij} = \begin{cases} |L_i \cap L_j|/M & 1 \leq i, j \leq K, i \neq j \\ 0 & i = j, \end{cases} \quad (1)$$

where  $|L_i \cap L_j|$  is the number of videos common to  $L_i$  and  $L_j$ . A concept detector  $D_i$  is said to co-occur with another detector  $D_j$  if  $C_{ij} > t$ , where  $t$  is between 0 and 1. A concept is discarded if it does not co-occur with  $c$  other concepts.

An example is shown in Fig. 3. Here, the top 15 ranked videos retrieved by five different concept detectors for the event *getting a vehicle unstuck* are shown. The *stuck winter* detector co-occurs with other detectors in only one of the top 15 videos, the *stuck night* detector does not co-occur with any other detector, so these two detectors are discarded. Also, fewer positive examples of the complex event are retrieved by the two discarded detectors than the other three, suggesting that the co-occurrence based pruning strategy is effective in removing concepts which are

outliers. After pruning some concepts using co-occurrence statistics, we fuse the scores of good concepts by taking the mean score of these concepts and rank the videos in the test set using this score.

### 4.3. Instance Level Pruning

Although many concepts may be relevant to an event, it is not likely that all concepts will occur in a single video. This is because not all complex event instances exhibit all related concepts, and not all concept instances are detected even if they are present (due to computer vision errors). Therefore, computing the mean score of all concept detectors for ranking is not a good solution. So, we need to predict an event when only a subset of these concepts is observed. However, the subset is video instance specific and knowing all possible subsets a priori is not feasible with no training samples. Even though these subsets cannot be determined, we can estimate the average cardinality of the set based on the detector responses observed for the top  $M$  ranked videos after computing the mean score of detectors. For each event, the number of relevant concepts is estimated as:

$$N_r = K - \min\left(\left\lceil \frac{\sum_{k=1}^K \sum_{i=1}^M \mathbf{1}(s_{ik} < T)}{M} \right\rceil, \lambda\right) \quad (2)$$

where  $\mathbf{1}(\cdot)$  is the indicator function—it will be 1 if the confidence score of concept  $k$  present in video  $V_i$  is less than a detection threshold  $T$  (i.e., detector  $D_k$  does not detect the concept  $k$  in the video  $V_i$ ) and 0 otherwise.  $\lceil \cdot \rceil$  is the ceiling function, and  $\lambda$  is a regularizer to control the maximum number of concepts to be pruned for an event. This equation computes the average number of detected concepts in the top ranked videos. When combining the concept scores, we keep only the top  $N_r$  responses and discard the rest.

Table 1. Concepts discovered after different pruning strategies

Event	Discovered Concepts	
<i>Working on a metal crafts project</i>	Initial Concepts	art, bridge, iron, metal, new york, new york city, united state, work, worker
	After part of speech based pruning	iron art, iron bridge, iron craft, metal art, metal bridge, metal craft, new york, new york city, united state, work iron, work metal, work worker, worker art, worker bridge, worker craft
	After co-occurrence based pruning	iron art, iron craft, metal art, metal bridge, metal craft, work iron, work metal
<i>Dog show</i>	Initial Concepts	animal, breed, car, cat, dog, dog show, flower, pet, puppy, show
	After part of speech based pruning	animal pet, breed animal, breed car, breed cat, breed dog, breed flower, breed puppy, car pet, cat pet, dog pet, dog show, flower pet, puppy pet, show animal, show car, show cat, show dog, show flower, show puppy
	After co-occurrence based pruning	animal pet, breed animal, breed car, breed cat, breed dog, breed puppy, cat pet, dog pet, dog show, puppy pet, show cat, show dog, show puppy
<i>Parade</i>	Initial Concepts	city, gay, gay pride, gay pride parade, new york, new york city, nyc event, parade, people, pride
	After part of speech based pruning	city parade, gay city, gay people, gay pride, gay pride parade, new york, new york city, nyc event, people parade, pride parade
	After co-occurrence based pruning	city parade, gay pride, gay pride parade, people parade, pride parade

## 5. Hierarchical Temporal Pooling

Our concept detectors are frame-based, so we need a strategy to model the temporal properties of videos. A common strategy is to treat the video as a bag, pooling all responses by the average or max operator. However max-pooling tends to amplify false positives; on the other extreme, average pooling would be robust against spurious detections, but expecting a concept to be detected in many frames of a video is not realistic and would lead to false negatives. As a compromise, we propose hierarchical temporal pooling, where we perform max pooling within sub-clips and average over sub-clips over a range of scales. Note that the top level of this hierarchy corresponds to max pooling, the bottom level corresponds to average pooling, and the remaining levels correspond to something in-between. The score for a concept  $k$  in video  $V_i$  is computed as follows,

$$s_{ik} = \sum_{n=1}^l \sum_{j=1}^n \frac{m_{nj}}{n} \quad (3)$$

where,  $l$  is the maximum number of parts into which a video is partitioned (a scale at which the video is analyzed),  $m_{nj}$  is the max pooling score of the detector in part  $j$  of the video partitioned into  $n$  equal parts. Temporal pooling has been widely used in action recognition [18] for representing space-time features. In contrast, we perform temporal pooling over SVM scores, instead of pooling low level features.

## 6. Domain Adaptation

**Score Calibration.** The detectors are trained on web-images, so their scores are not reliable because of the domain shift between the web and video domains. In addition,

each detector may have different response characteristics on videos, e.g., one detector is generic and has a high response for many videos, while another detector is specific and has a high response only for a few videos. Thus we calibrate their responses before fusion as follows :

$$s'_{ik} = \frac{1}{1 + \exp(\frac{R_k(s_{ik}) - u}{u})} \quad (4)$$

where,  $s'$  is the calibrated score,  $R_k$  is the rank of video  $V_i$  when generating the rank list only using concept detector  $D_k$ , and  $u$  controls the decay factor in the exponential. This re-scoring function not only calibrates raw detector scores, but it also gives much higher score to highly ranked samples while ignoring the lower ranked ones, which is appropriate for retrieval.

**Detector Retraining.** Based on the domain adaptation approach of [17], we use pseudo-relevance from top-ranked videos to improve performance. Since web-detectors only capture static scene/object cues in a video, it is beneficial to extract Fisher Vectors (FV) on Improved Dense Trajectory (IDT) features [28] to capture the motion cues. Based on the rank list obtained from concept detectors, we train a linear SVM using LIBLINEAR [12] on the top ranked videos using the extracted Fisher Vectors. The lowest ranked videos are used as negative samples. These detectors are applied again on the test videos. Finally, we use late fusion to combine the detection scores obtained using motion features with web-detectors.

We further adapt the concept detectors to the video domain by retraining them on frames from top-ranked videos. For each detector, we obtain frames with the highest response in the top ranked videos (after fusion with motion features) to train a concept detector (with the constraint that

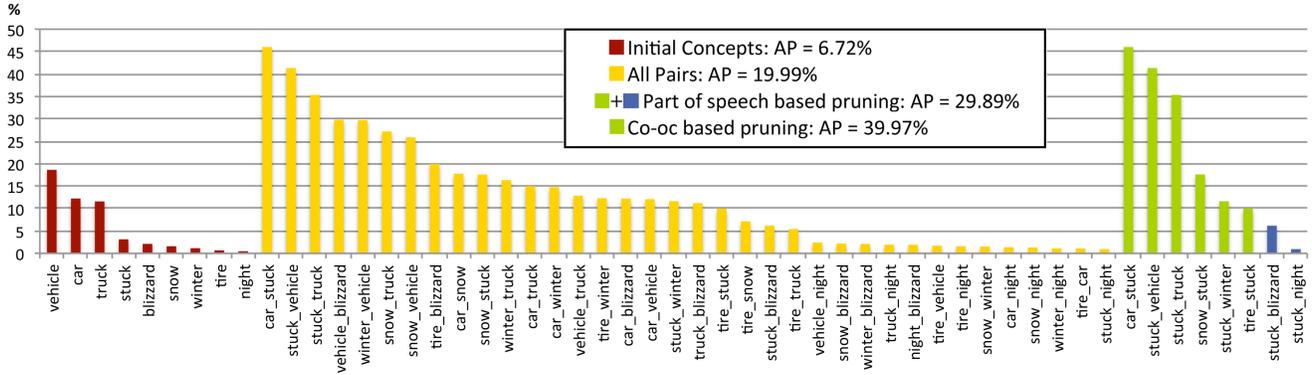


Figure 4. Average Precision (AP) scores of initial concepts, all pair-concepts, the concepts after part of speech based and co-occurrence based pruning are shown for the event "Getting a vehicle unstuck". AP after combining the concepts is also reported. Note that part of speech pruning helps in removing many pair-concepts with low AP. Moreover, co-occurrence based pruning removes the two lowest performing pair-concepts and improves the AP after part of speech pruning significantly.

similar frames should not be selected twice to encourage diversity). We then repeat the process of co-occurrence based pruning, instance level pruning and rank based calibration to fuse the scores for the new concept detectors. Finally, the video scores are updated by summing the fused scores (original concept detectors + IDT) and the scores of adapted concept detectors.

## 7. Experiments and Results

We perform experiments on the challenging TRECVID Multimedia Event Detection (MED) 2013 dataset. We first verify the effectiveness of each component of our approach, and then show the improvement on the EK0 dataset by comparing with state-of-the-art methods.

### 7.1. Dataset and Implementation Details

The TRECVID MED 2013 EK0 dataset consists of unconstrained Internet videos collected by the Linguistic Data Consortium from various Internet video hosting sites. Each video contains only one complex event or content not related to any event. There are 20 complex events in total in this dataset, with ids 6-15 and 21-30. These event videos together with background videos (around 23,000 videos), form a test set of 24,957 videos. In the EK0 setting, no ground-truth positives training videos are available. We apply our algorithm on the test videos, and mAP score is calculated based on the video ranking.

For each event in EK0 dataset, we choose the top 10 concepts (i.e.,  $P = 10$ ) in the list provided by [5] and transform them into pair-concepts. The web image data on which concept detectors are trained is obtained by image search on Images using each pair-concept as a query. The *Type* option is set to *Photo* to filter out irrelevant cartoon images. We downloaded around 200 images for each concept pair query. We sample each video every two seconds to obtain a set of frames. Then, we use Caffe [15] to extract deep

features on all the frames and web images, by using the model pre-trained on ImageNet. We used the fc7 layer after dropout, which generates a 4,096 dimensional feature for each video frame or image. The hyper-parameters to determine if a concept co-occurs with another  $t$ , length of the intersection list  $M$ , regularization constant  $\lambda$ , detector threshold  $N_r$  are selected based on leave one-event out cross validation, since they should be validated with event categories different from the one being retrieved. We found hyper-parameters to be robust after doing sensitivity analysis. The number of levels  $l$  in hierarchical temporal pooling was set to 5. The Fisher Vectors of the top 50 ranked videos and the bottom 5,000 ranked videos are used to train a linear SVM.

Table 2. Comparative results

Pre-Defined	Method	mAP
No	Concept Discovery [5]	2.3%
Yes	SIN/DCNN [17]	2.5%
Yes	CD+WSC [29]	6.12%
Yes	Composite Concept [13]	6.39%
No	<i>Initial</i> concepts	4.91%
	<i>All</i> Pair-concepts	7.54%
	+Part of speech pruning	8.61%
	+Cooc & inst pruning	10.85%
	<b>+Adaptation</b>	<b>11.81%</b>

### 7.2. Evaluation on MED 13 EK0

Table 1 shows the initial list of concepts, the concepts that remain after part of speech based pruning, and the concepts that remain after co-occurrence based pruning for three different events. Although the initial concepts are related to the event, web queries corresponding to them would provide very generic search results. Since we have 10 unary concepts per event, there are 45 unique pair-concept detectors for each event. Approximately 10-20 pair-concepts remain after part of speech based pruning. This helps to re-

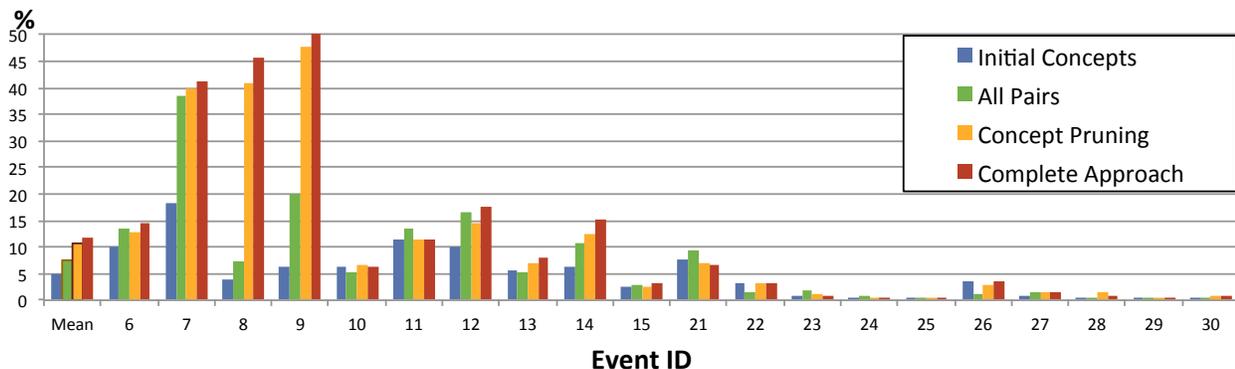


Figure 5. Mean average precision (mAP) scores for the events on the MED13 EK0 dataset. By pruning concepts that are not useful for retrieving the complex event of interest, our approach progressively improves the utility of the remaining.

duce the computational burden significantly and also prunes away noisy pairs. Finally, co-occurrence based pruning discards additional outliers in the remaining pair-concepts.

Table 2 shows the results of our method on the TRECVID EK0 dataset. We observe significant performance gains (5.4% - 11.81% vs 6.39%) over other vision based methods which do not use any training samples. Our performance is almost 2-5 times their mAP. Note that the methods based on pre-defined concepts must bridge the semantic gap between the query specification and the pre-defined concept set. On the other hand, we leverage the web to discover concepts. Our approach follows the same protocol as [5] which performs the same task. Using the same initial concepts as [5], our method obtains 5 times the mAP as that of [5]. Fig. 5 shows the effect of each stage in the pipeline. Replacing the initial set of concepts by action based pair-concepts provides the maximum gain in performance of  $\sim 3.7\%$  (4.9% to 8.61%). Next, co-occurrence based pruning improves the mAP by 1.8% (8.61% to 10.4%). Calibration of detectors and instance level pruning improves the mAP score to 10.85%. Finally, adapting each detector on the test dataset and using motion information allows us reach a mAP of 11.81%. The performance is low for events 21 to 30 because there are only  $\sim 25$  videos for these events while events 6-15 have around 150 videos each in the test set.

To illustrate that the proposed pruning methods remove concepts with low AP, in Fig 4 we plot AP scores of initial unary concepts, all pair-concepts, part of speech based concepts and the concepts after co-occurrence based pruning. Note that almost 50% of pair-concepts had an average precision below 10% before pruning. After part of speech and co-occurrence based pruning, our approach is able to remove all these low scoring concepts in this example.

We would note that Hierarchical Temporal Pooling provides significant improvement in performance for this task. In Table 3, we show mAP scores for different pooling meth-

ods for initial, pair-concepts and after concept pruning (before Detector Retraining). It is clear that Hierarchical Temporal Pooling improves performance in all three cases. We also observe that concepts after pruning have best performance across all pooling methods .

Table 3. Pooling Results

	Initial	All Pairs	After Pruning
Avg. Pooling	2.84%	4.54%	5.94%
Max. Pooling	4.45%	6.87%	9.01%
Hierarchical	4.91%	7.54%	10.85%

## 8. Conclusion

We demonstrated that carefully pruning concepts can significantly improve performance for event retrieval when no training instances of an event are available, because even if concepts are visually salient, they may not be relevant to a specific event or video. Our approach does not require manual annotation, as it obtains weakly annotated data through web search, and is able to automatically calibrate and adapt trained concepts to new domains.

## Acknowledgement

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not with standing any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. The authors would like to thank Yin Cui for providing the initial concepts. The authors acknowledge the University of Maryland supercomputing resources <http://www.it.umd.edu/hpcc> made available for conducting the research reported in this paper.

## References

- [1] S. M. Assari, A. R. Zamir, and M. Shah. Video classification using semantic concept co-occurrences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. 2
- [2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision—ECCV 2010*. Springer. 2
- [3] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. 4
- [4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 4
- [5] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of International Conference on Multimedia Retrieval*, page 1. ACM, 2014. 1, 2, 3, 4, 7, 8
- [6] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *IEEE International Conference on Computer Vision (ICCV), 2013*. 2
- [7] Y. Cui, D. Liu, J. Chen, and S.-F. Chang. Building a large concept bank for representing events in video. *arXiv preprint arXiv:1403.7591*, 2014. 2
- [8] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013. 2
- [9] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. 1, 2, 3, 4
- [10] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. 2
- [11] L. Duan, D. Xu, I.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1667–1680, 2012. 2, 3
- [12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*. 6
- [13] 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*, page 17. ACM, 2014. 2, 3, 7
- [14] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014. 2, 3
- [15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4, 7
- [16] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM MM*, 2014. 3
- [17] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014. 3, 6, 7
- [18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008*. 6
- [19] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011. 3
- [20] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platts probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007. 3
- [21] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In *Applications of Computer Vision (WACV), 2013*. IEEE. 1, 2
- [22] M. Mazloom, A. Habibian, and C. G. Snoek. Querying for video events by semantic signatures from few examples. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 609–612. ACM, 2013. 2
- [23] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. 1, 3
- [24] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *Multimedia, IEEE Transactions on*, 14(1):88–101, 2012. 2
- [25] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4
- [26] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. 4
- [27] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, 2012. 3
- [28] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV), 2013*. 6
- [29] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. 1, 2, 3, 7
- [30] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Multimedia event recounting with concept based representation. In *Proceedings of the 20th ACM international conference on Multimedia*, 2012. 2