2015

Exploiting Concepts In Videos For Video Event Detection

Ethem Can

EXPLOITING CONCEPTS IN VIDEOS
FOR VIDEO EVENT DETECTION

A Dissertation Presented

by

ETHEM F. CAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2015

College of Information and Computer Sciences

EXPLOITING CONCEPTS IN VIDEOS
FOR VIDEO EVENT DETECTION

A Dissertation Presented

by

ETHEM F. CAN

Approved as to style and content by:

_____

James Allan, Co-chair

_____

R. Manmatha, Co-chair

_____

W. Bruce Croft, Member

_____

Patrick A. Kelly, Member

_____

James Allan, Chair of the Faculty
College of Information and Computer Sciences

Unable are the Loved to die, For Love is Immortality
Hakan
Leyla Gün

ACKNOWLEDGMENTS

ABSTRACT

EXPLOITING CONCEPTS IN VIDEOS
FOR VIDEO EVENT DETECTION

SEPTEMBER 2015

ETHEM F. CAN

B.Sc., BILKENT UNIVERSITY, ANKARA, TURKEY

M.Sc., BILKENT UNIVERSITY, ANKARA, TURKEY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan and Professor R. Manmatha

Video event detection is the task of searching videos for events of interest to a user where an event is a complex activity which is localized in time and space. The video event detection problem has gained more importance as the amount of online video is increasing by more than 300 hours every minute on Youtube alone.

In this thesis, we tackle three major video event detection problems: video event detection with exemplars (VED-ex), where a large number of example videos are associated with queries; video event detection with few exemplars (VED-ex$_{few}$), in which only a small number of example videos are associated with queries; and zero-shot video event detection (VED-zero), where no exemplar videos are associated with queries.

We first define a new way of describing videos concisely, one that is built around using query-independent concepts (e.g., a fixed set of concepts for all queries) with

a space-efficient representation. Using query-independent concepts enables us to learn a retrieval model for any query without requiring a new set of concepts. Our space-efficient representation helps reduce the amount of time required to train/test a retrieval model and the amount of space to store video representations on disk.

When the number of example videos associated with a query decreases, the retrieval accuracy decreases as well. We present a method that incorporates multiple one-exemplar models into video event detection aiming at improving retrieval accuracies when there are few exemplars available. By incorporating multiple one-exemplar models into video event detection with few exemplars, we are able to obtain significant improvements in terms of mean average precision compared to the case of a monolithic model.

Having no exemplar videos associated with queries makes the video event detection problem more challenging as we cannot train a retrieval model using example videos. It is also more realistic since compiling a number of example videos might be costly. We tackle this problem by providing a new and effective zero-shot video event detection model that exploits dependencies of concepts in videos. Our dependency work uses a Markov Random Field (MRF) based retrieval model and assumes three dependency settings: 1) full independence, where each concept is considered independently; 2) spatial dependence, where the co-occurrence of two concepts in the same video frame is treated as important; and 3) temporal dependence, where having concepts co-occur in consecutive frames is treated as important. Our MRF based retrieval model improves retrieval accuracies significantly compared to the common bag-of-concepts approach with an independence assumption.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The number of videos available on the Internet has been increasing rapidly. Video sharing (e.g., through Youtube or social media) has played an important role in this increase (the total number of video views per day on Youtube and Facebook has reached into the billions). In an age where watching videos online is widespread, searching and retrieving videos has become increasingly important.

Video search may be defined as retrieving videos relevant to an information need (i.e., expressed by a query) from a large source of videos such as video sharing web sites. For example, in Youtube, users provide a query about videos they are searching for and then Youtube returns a list of videos that are expected to be relevant to the query. Relevance is often estimated in terms of the similarity between the query and the text associated with videos (e.g., metadata such as title and description). In Figure 1.1, we illustrate the retrieved results when we search for "grooming an animal".

In the figure, the circled texts are the words matching this query. The top three retrievals in the figure are videos of "grooming an animal" (i.e., some "grooming an animal" event is in the video); however, the fourth, fifth, and the sixth retrievals are advertisements about veterinary schools where no "grooming an animal" event exists.

When the retrieval is based on the metadata of videos, the chances of finding a video are aligned with providing the correct terms for a query. Note that the phrase "correct terms" refers to the ones that match with the textual information associated with videos. For this example illustrated on Youtube, video search is only based on

Figure 1.1. Top retrievals when searching for "grooming an animal".

the metadata associated with the videos and some of the retrievals are not relevant to the "grooming an animal" query. This is due to the fact that the text associated with those videos is misleading or insufficient. Understanding the content of videos

will help improve the retrieval and eliminate videos not relevant to the "grooming an animal" query.

For the example, we have assumed that the videos have metadata associated with them. This strong assumption is essential for text-based video search. Where no metadata is available for videos, a text-based approach will not help. Consider the same example (from Figure 1.1) when there is no metadata associated with the videos (see Figure 1.2). In this case, a text-based approach is useless. However, visual cues can still be used for video search purposes. We, as do others (Chen et al. 2014, Dalton et al. 2013, Habibian et al. 2014, Younessian et al. 2012), use visual cues that describe the content of videos for video search. The TREC Video Retrieval Evaluation (TRECVID) multimedia event detection track is devoted to research in video search exploiting the content of videos. The task of searching videos for events is of interest to a user where an event is a complex activity which is localized in time and space. The events are also expected to be observable, involving people interacting with other people or objects (Jiang et al. 2013, NIST 2012, Over et al. 2010). This is our definition of an event and that is aligned with the official definition of an event by NIST(NIST 2012, Over et al. 2010). As Jiang et al. (2013) note that in their survey paper, there is no consensus on defining an event in the literature. The official definition states that an event is a complex activity occurring at a specific time and space (NIST 2012, Over et al. 2010). Observing evidences of this official definition in the queries is difficult. Further, extracting specific "space" and "time" information from a definition of a query is rather complicated (see Section 2.3.1).

In this thesis, we tackle the video event detection (VED) problem by taking the visual content of videos into account. Understanding the visual content of videos is essential whenever there is no metadata associated with videos. Further, they can also be incorporated into text-based approaches (e.g., the ones exploiting the metadata of videos) for a better retrieval, which might help improve the quality of the retrievals.

| grooming an animal | SEARCH |

1.

2.

3.

4.

5.

6.

7.

Figure 1.2. Top retrievals when searching for "grooming an animal" provided without their metadata.

In this thesis, we consider only the visual features (e.g., concepts) to evaluate their effectiveness within the context of VED. We represent videos using their visual content. We make use of concepts (e.g., objects and actions) that provide a high level of abstraction to represent videos. For example, the sample image provided in Figure 1.3 can be described using airplane, clouds, and sky, which are examples of concepts. In order to automatically extract concepts from the sample image (Figure 1.3), detectors specific to these descriptors need to be created. A concept detector should detect or measure the likelihood of that particular concept in an image or video. For instance,

4

a plane detector would give the likelihood of seeing a plane in an image, which should be high for the image in Figure 1.3.

Figure 1.3. A sample image of an airplane above the clouds.

In the video event detection problem, an event query might consist of a text description and a number of relevant example videos associated with the query. We refer to this as video event detection with exemplars (VED-ex). Alternatively, it might consist of just a textual description, and this is called zero-shot video event detection (VED-zero) indicating that there are zero visual examples.

When an event query consists of a textual description as well as a number of exemplars, a retrieval model can be trained on these examples. First, concepts are extracted from videos by running concept detectors against videos so videos are represented using these concepts. We can learn the discriminative concepts for the given query using the exemplars (e.g., concepts common in the exemplars but not in general). Finally, we look for these discriminative concepts in the test videos to identify videos likely to be relevant to the given query.

1.1 Concept Selection

The main issue in tackling the video event detection with exemplars problem is to determine the concepts to be used for an event query. A common approach in VED-ex is to select concepts specific to the event query. In other words, a different set of concepts is determined for each query (Chen et al. 2014). However, crafting a different set of concepts for each query introduces some cost. In the first part of this thesis (Chapter 3), we focus on an alternative approach to query-based concepts within the context of VED-ex.

We hypothesize that we can use a fixed set of concepts for any event query without sacrificing effectiveness. In our work, we compile a fixed set of concepts that are selected without prior knowledge of event queries. This fixed set of 1,000 concepts is collected from a large source of images (Image-Net 2014).

Using a fixed set of concepts enables us to skip the concept selection step. This is important to save a modest amount of time when retrieval models are created and used for event search. Further, it becomes more important when new concepts are needed to be created (e.g., in case the selected concepts are not in the vocabulary) as creating new concepts can be very time consuming and we need more space to store our concept vocabulary.

Our approach stems from the idea of representing a video with a set of predefined concepts. Videos relevant to an event query are expected to be somewhat visually similar to each other. Therefore, we believe that concepts in visually similar videos may be similar as well. Similar approaches have been used and shown to be successful for object recognition. Torresani et al. (2010) focus on the idea of expressing a "novel" object category using a set of predefined categories (the authors call it "classemes"). Cusano et al. (2012) focus on a similar idea. In contrast to Torresani et al. (2010), they do not use the labels of the predefined categories. They call this method "unsupervised

classemes". Our approach of using a fixed set of concepts—selected independently from the queries—is motivationally similar to the latter work.

In order to show that our fixed set of concepts perform as well in accuracy as the ones obtained using query-based concepts, we create event detection models using both methods. Further, we also compare our query-independent concepts with the work of Chen et al. (2014), that is based on query-dependent concepts. The experimental results show that query-based concepts provide comparable retrieval accuracies to our fixed set of concepts—selected independently from queries—within the context of VED-ex (Section 3.4.1). We thus increase the efficiency of retrieval without sacrificing accuracy.

In addition to query dependent concepts, the common video event detection approaches often use a non-sparse representation of videos. The concept detectors are run against videos at the frame or clip level (i.e., a video consists of multiple clips). It is common to pool responses over frames (or clips) for a final representation of videos using concepts in VED-ex. For example, Cheng et al. (2012), Liu et al. (2013a), and, Jiang et al. (2013) use maximum and average pooling of responses over frames. The main issue with these pooling techniques is that the final representation of a video is not sparse. If there are n concepts considered, the final representation of a video becomes a dense histogram of size n where the number of non-zero elements is also n. Keeping dense histograms on the disk is also expensive compared to their sparse counterparts. Sparsity is especially important when machine learning algorithms are involved (e.g., learning a model using exemplars) since most of them work more efficiently with sparse histograms (i.e., the number of non-zero elements is smaller than the actual size of the histogram).

We address these space and time problems by considering only a small number of highly responsive concepts per-frame. Habibian et al. (2013) investigate concept vocabularies for video event detection and conclude that a subset of concepts should

be considered for a better video event detection. Bhattacharya et al. (2014) focus on finding the minimally needed evidence for recognizing events and report that humans can recognize events in videos by watching only a portion of the videos. We follow a similar principle in our work but apply it to machines. Our findings show that our representation of videos is not only efficient but also effective compared to common pooling techniques (e.g., maximum and average) and using all or a large number of concepts (Section 3.4.2).

1.2 One-Exemplar Models

So far, we have focused on how to select concepts and how to represent videos using those concepts. After the representation of videos, we train retrieval models using example videos associated with event queries. Collecting a large number of example videos for an event query is the ideal case in video event detection with exemplars. However, collecting a few exemplars is more realistic than providing a large number of example videos since the annotation task is costly. The major drawback of having a few exemplars to train retrieval models is that they usually provide lower retrieval accuracies compared to models trained on a large number of example videos. In the second part of this thesis (Chapter 4), we focus on video event detection with very few exemplars (VED-ex$_{few}$). We present a method that incorporates multiple one-exemplar models into video event detection aiming at improving retrieval accuracies when there are few exemplars available.

A retrieval model trained on visually different models might carry some amount of noise. For instance, example videos for the "repairing an appliance" query might include contents such as repairing an oven, repairing a refrigerator, and repairing a washing machine. Even though they are example videos for the same query (i.e., repairing an appliance), they are visually different from each other. This situation would presumably have been different if we had a larger set of example videos since

example videos would have covered different variations of the query. In the common video event detection approach, all of the example videos are used to create a single global model (e.g., repairing an oven, repairing a refrigerator, and repairing a washing machine videos are used to create *one* "repairing an appliance" model). Instead, our one-exemplar models are created for each exemplar and are specific to their exemplar (e.g., a repairing an appliance model created using the repairing an oven video, a model using the repairing a refrigerator video, and another model using the repairing a washing machine video). We incorporate the multiple one-exemplar models into the global model to handle both visually different and similar exemplars. The global model works well when the example videos are visually similar. Our one-exemplar models work better when the example videos are visually different. It is very difficult to estimate the variance of the example videos in advance. Therefore, incorporating one-exemplar models into this global model enables us to deal with this difficulty within the context of video event detection with few exemplars.

Our experimental results show that we are able to improve retrieval accuracies using our approach over using a single global model when there are few exemplars available (Section 4.2.1). In addition to our concept-based technique, we investigate the robustness of our method by experimenting on different descriptors. These experiments show that our method is robust to multiple descriptors (Section 4.2.2).

1.3   Zero-Shot Video Event Detection

So far, we have focused on the video event detection with exemplars case. The more challenging part of video event detection is when event queries consist only of a textual description and no exemplars: zero-shot video event detection (VED-zero). In VED-zero the main process is the same as in VED-ex except that of learning a retrieval model. VED-zero is challenging since we cannot train a retrieval model

using exemplars. Therefore, unlike the VED-ex case, ranking of videos relies on the likelihood of observing relevant concepts to a query in videos.

Most existing zero-shot video event detection work focuses on a bag-of-concepts approach that assumes the independence of concepts in videos (e.g., query-likelihood model). This approach assumes the individual occurrences of concepts in videos and ignores their interactions such as their order. For example, consider two cases: 1) three concepts are detected at the beginning of a video, and 2) three concepts are detected at different locations (e.g., one at the beginning, one in the middle, and one at the end). A bag-of-concepts approach treats these two cases the same and ignores the information of their co-occurrence or their order of occurrence.

We hypothesize that the bag-of-concepts approach relies on a weak independence assumption and exploiting concepts and their relationships could enable us to have better retrieval. In the last part of this thesis (Chapter 5), we detail our dependency-based retrieval algorithm within the context of VED-zero and show evidence supporting our hypothesis.

Motivated by the fact that events are complex activities, we believe that considering concepts individually might not be enough to retrieve videos relevant to an event. In other words, considering concepts and their relationships might enable us to have better evidence to recognize videos relevant to an event. For example, in Figure 1.4, we provide two consecutive frames. There are three concepts detected in these frames: "person looking direction," "blowing candles," and "person clapping." The "person looking direction" and "blowing candles" concepts occur in the same space (frame) and these concepts have a temporal relationship with the "person clapping" concepts. Considering these concepts and their spatial and temporal relationships give us more evidence of this video about it being relevant to a "birthday party" event. However, these concepts individually do not provide the same level of confidence. In Figure 1.5, we provide another illustration of concepts in a video. In the first frame, there is

a spatial relationship between the "audience clapping" and "jumping over fence" concepts and these concepts have a temporal relationship with the "judges giving points" concept. When we consider these relationships in addition to the concepts itself, we can identify this video as relevant to the "horse riding competition". On the other hand, considering these concepts independently is not necessarily sufficient for the same detection.

Our dependency work uses a Markov random field (MRF) based approach (Metzler and Croft 2005), a widely accepted algorithm in the information retrieval community. They make use of the occurrences of single terms, ordered phrases, and unordered phrases in their retrieval model. They obtain significant improvements on their dependency-based retrieval model on large web-based collections. Feng and Manmatha (2008) apply a similar approach to *image retrieval*. They propose the use of unigrams for what they call a full independence model and spatial bigrams for a spatial dependence model. In our work, we focus on three dependency assumptions: (1) full independence, (2) spatial dependence, and (3) temporal dependence (Section 5.1.2).

The output of our MRF-based retrieval model is used to rank videos according to their relevance to an event query. In other words, the probability estimation of a video given a query is used for ranking purposes. The output scores of concept detectors are used in the estimations. These scores can be interpreted differently. They can be used as binary values: presence/absence of concepts in videos. Alternatively the detector output scores can be used directly (Section 5.1.3). For example, we have three concepts: "car," "ox," and "barn". We detect the "car" concept with a confidence of 0.4 (the higher the confidence the stronger the detection and confidence values are between zero and one); the "ox" concept with 0.9; and the "barn" concept with 0.5 in a video. When we focus on presence/absence information of concepts, we assume that only concepts with a high detection confidence are present in this videos, and consider

Figure 1.4. Illustration of concepts in a video. Concepts: {("person looking direction"), ("blowing candles"), ("person clapping"); Spatial Relationships: {("person looking direction","blowing candles")}; and Temporal Relationships: { ("person looking direction","person clapping") ,("blowing candles","person clapping")}

these concepts in the estimations. In the alternative version, we consider all of the concepts without ignoring any of them.

Using presence/absence information makes the estimation of probabilities in our formulations easy. Concepts having a low confidence are also eliminated when binary values are used (e.g., the "car" and "barn" concepts can be ignored in the example above). However, when we use binary values, there is a possibility that we might lose some useful information (e.g., the "barn" concept would be useful) As an alternative, we can directly use the scores in our models. In this way, we are not removing any potentially useful information (e.g., we use all three concepts in the estimations in the example above). However, the issue with using outputs directly is that these scores carry some noise. (e.g., the "car" concept might not be relevant). There is a chance

Figure 1.5. Another illustration of concepts in a video. Concepts: {("person clapping"),("jumping over fence"),("judges giving points")}; Spatial Relationships: { ("person clapping","jumping over fence")}; and Temporal Relationships: {("person clapping","judges giving points"), ("jumping over fence","judges giving points")}

that this noise can be carried over into our model. In our MRF-based retrieval model, we focus on both methods, using binary values and using output scores directly. In addition, we also focus on blending these two methods to leverage the advantages of both methods. Evaluation on a large collection shows that our MRF-based retrieval model improves the retrieval accuracies statistically significantly over a retrieval model where dependencies of concepts are not considered (Section 5.3.1). Blending two different choices for concept detector output enables us to further improve the retrieval accuracies (Section 5.3.2).

So far, we have summarized the problems that we tackle in this thesis as well as our solutions to address them. Before formalizing the contributions, we would like to point out the status of current video event detection approaches.

Video event detection is a relatively recent problem. Even though it is new, researchers have shown promising results. However, in most of the cases including this thesis, the biggest goal is to find out the best way to use the content of videos that should be used when videos do not have any metadata. As we are still in early stages of solving the problem, we cannot always retrieve videos relevant to a query very effectively. Indeed, our retrieval accuracy scores are extremely low because we focus on only the visual features and ignore the text modality. Our solutions mainly explore the relative advantages of different ways of using content of videos and that is a promising approach when videos do not contain metadata. We believe that recent progress on the field especially on using deep learning features to create concept detectors will yield increases in accuracy scores.

1.4 Contributions

The research results presented in this dissertation may be summarized by the following three major contributions in video event detection:

1. **We define a new way of describing videos concisely, one that is built around using query-independent concepts with a space-efficient representation.** The recent approaches in video event detection with exemplars focus on selecting a number of concepts based on the query to describe videos (Chen et al. 2014, Cheng et al. 2012, Jiang et al. 2013, Liu et al. 2013a), which becomes costly when done for each of multiple queries.

    (a) Unlike the recent approaches, we compile a set of query-independent concepts and use this set of concepts for any query. The results show that retrieval accuracies obtained using these query-independent concepts are as strong as the ones obtained using the concepts selected specifically each query. In this way, the concept selection process may be skipped and a

retrieval model can be learned directly for any given query without requiring a new set of concepts.

(b) In addition to query-dependent concepts, inefficient techniques for representation of videos using concepts are also employed Cheng et al. (2012), Jiang et al. (2013), Liu et al. (2013a). These techniques use a non-sparse representations of videos, which is space and time costly. In contrast to the space and time costly techniques that use all of the concepts, we consider only a small number of highly responsive concepts to represent videos. In this way, the amount of required *space* to store video representations is *reduced* to one fifth compared to using all concepts. Our space-efficient representation technique also enables us to train retrieval models and run queries against these models *in 1/5 the time* compared to using all concepts.

2. **We present a method that incorporates multiple one-exemplar models into video event detection aiming at improving retrieval accuracies when there are few exemplars available.** The common idea of "the more exemplars we have, the better models can be learned" is followed in video event detection studies while tackling the problem. However, collecting a large number of example videos is usually unrealistic.

(a) We present a method that considers several one-exemplar models each of which is learned using one exemplar and incorporates these models into video event detection with few exemplars. In the common video event detection approach, all of the example videos are used to create a single global model. However, multiple one-exemplar models are created for each exemplar and they are specific to their exemplar. We incorporate several one-exemplar models into the global model.

(b) By incorporating one-exemplar models into video event detection with few exemplars, we are able to obtain 15-35% relative improvements, in average precision compared to the case of not having one-exemplar models in retrieval (i.e., only the global model). One-exemplar models not only enable us to improve retrieval accuracies but also handle the issue of high variance of example videos.

(c) In order to analyze the robustness of our method, we also evaluate it on multiple descriptors. Experimental results show that our method is robust to multiple descriptors.

3. **We provide a new and effective zero-shot video event detection model that exploits dependencies of concepts in videos.** In a real world scenario, a query may consist of only a textual description and no exemplars: the zero-shot video event detection task. Most of the video event detection oriented studies focus on the case where we have exemplars and those that focus on zero exemplar assumes that concepts in videos are independent of each other.

(a) Against this widely accepted assumption, we exploit dependencies of concepts in videos in addition to the independence assumption. Our dependency work uses a Markov random field (MRF) based approach (Feng and Manmatha 2008, Metzler and Croft 2005), a state of the art ranking algorithm from the information retrieval community.

(b) We evaluate three dependency assumptions: (1) full independence, where each concept is considered independently; (2) spatial dependence, where the presence of two concepts in the same video frame is treated as important; and (3) temporal dependence, where having concepts occur in consecutive frames is treated as important.

(c) In our MRF-based retrieval model, we utilize concept outputs in two different ways: 1) converting them into presence/absence information (i.e., Boolean concepts), and 2) using them directly (i.e., scored concepts). In the first approach, we threshold the concept outputs and focus on only the ones that have high confidence. In this way, we are able to remove most of the noise carried over by concept detectors. The major drawback of this approach is that it might also remove some useful information while trimming the concepts that have less confidence. In the latter approach, we make use of the concept outputs directly in our models. In this case, we use all of the available information including some amount of noise. In addition to these two approaches, we also consider a hybrid method that takes advantages of these two approaches together. Using the hybrid approach enables us to improve the retrieval accuracies of Boolean concepts and scored concepts by approximately 30% and 15% (relative improvements) respectively.

(d) Our MRF-based retrieval model improves retrieval accuracies by anywhere from 10% to 150% (relative) in average precision compared to the common independence assumption in a collection of 30 queries and approximately 27,000 videos. In addition to comparing our model with the common bag-of-concepts approach (independence assumption), we also compare our results with the previously reported retrieval accuracies on the same dataset. Our hybrid retrieval model outperforms the previous work by 300% (i.e., 2.2% vs. 9.1%) (Chen et al. 2014), 160% (i.e., 3.5% vs. 9.1%) (Rastegari et al. 2013), 115% (i.e., 4.2% vs. 9.1%) Mazloom et al. (2013a), or 40% (i.e., 6.4% vs. 9.1%) Habibian et al. (2014) in terms of mean average precision (relative improvements).

CHAPTER 2

LITERATURE OVERVIEW

The obvious solution to video event detection is to use textual similarity between query and metadata associated with videos. Most of the commercial video search engines focus on this approach for ranking videos. For example, Youtube is based on this textual similarity.

The major drawback of using textual similarity is that it relies on assuming videos have metadata associated with them and that it accurately and completely describes the video's contents. Therefore, the content of videos also needs to be exploited to address when there is no metadata available. A number of different attributes may be extracted from the content of videos such as concepts (e.g., objects and actions), audio (e.g., speech), and also optical character recognition, OCR, (text in images) associated with frames (e.g., closed captions). Audio and text might not exist in some videos; however, we can extract some concepts from any video. In our approaches, we make use of concepts to tackle the video event detection problem.

In this section, we summarize the previous studies that are related to our work. The majority of the previous papers are on the video event detection with exemplars case. The papers related to zero-shot video event detection are few and recent.

2.1 Video Event Detection with Exemplars

Recent approaches (e.g., Jiang et al. (2010), Natsev et al. (2010), Cao et al. (2011), Liu et al. (2013a), and Oh et al. (2013)) to tackling the video event detection with exemplars problem is to use a number of concepts to represent a video. An event

detection model is often trained for each event query using an off-the-shelf classifier. Test videos are then run against these retrieval models.

2.1.1 Video Representation

In the video event detection with exemplars problem, it is common to utilize example videos associated with queries to learn the distinctive characteristics of queries so that we can leverage this knowledge later to rank videos according to their relevance to queries. For example, we aim to learn the specific characteristics of a "repairing an appliance" query by analyzing the common patterns in example videos of this query (mostly with the help of a machine learning algorithm). Then we focus on the characteristics of "repairing an appliance" while ranking test videos according to their relevance to the "repairing an appliance" query.

Here the issue is to extract some information from videos which helps us learn the characteristics of a query. In other words, we need a way to represent where similar videos have similar patterns. For example, we might represent videos with a bag of concepts (e.g., objects). For the "repairing an appliance" query, we might learn characteristics including "videos relevant to this query contains multiple appliance occurrences". We then use this knowledge to conclude that videos having occurrences of appliances have a higher chance of being relevant to the "repairing an appliance" query.

In order to represent a video using a concept, a detector specific to this concept needs to be trained. For example, a dog detector is needed for the "dog" concept. Similarly, a "walking" detector is needed to measure the likelihood of observing a "walking" action in a video. A common approach to creating a concept detector in video event detection is to train a statistical model using a number of relevant images/videos specific to this concept. The input for these statistical models is often a set of descriptors extracted from examples. Cheng et al. (2012), Liu et al. (2013a)

use motion boundary histograms (MBH), histogram of oriented gradients (HOG), and spatio-temporal interest points (STIP) descriptors to create concepts. Modolo and Snoek (2013) focus on scale invariant feature transform (SIFT) descriptors and its two variants, colorSIFT, and denseSIFT, to create their concepts. Yang and Shah (2012) use MBH, STIP, and SIFT in their study. SIFT and its variants (e.g., denseSIFT) are usually used to create object-based concept detectors (e.g., dog). MBH and HOG exploit motion information in video; therefore, they are employed to create action-based concepts (e.g., walking). For example, Cheng et al. (2012), Liu et al. (2013a), Ma et al. (2013), Mazloom et al. (2013b) create their action-based concepts using these features. In this study we leverage the action based concepts created by SRI International (Cheng et al. 2012, Liu et al. 2013a). Dalton et al. (2013) also make use of their concepts in their work. Off-the-shelf object detectors can also be used as an alternative to creating concept detectors from scratch. ObjectBank (Li et al. 2010) is employed by Althoff et al. (2012), Oh et al. (2014) and Oh et al. (2013) as a source of pre-trained generic object detectors.

In addition to concepts, the descriptors used to create them are also employed in video event detection studies. For example, SIFT—Scale-Invariant Feature Transform— is a feature descriptor that exploits gradient distributions at corner points (Lowe 1999) and is one of the most commonly used descriptors. SIFT is usually computed at corner points using gray scale intensity values. Variations of the original SIFT implementation such as denseSIFT (in which SIFT descriptors are computed at densely sampled points (e.g., every 5 pixel), colorSIFT where Red-Green-Blue intensity values are used rather than gray-scale intensity values (van de Sande et al. 2010), and motionSIFT which detects interest points and encodes local appearance and models local motion (Chen and Hauptmann 2009) have also been used in event detection. In addition to SIFT and its variations, GIST, a descriptor that represents the dominant spatial structure of a scene (Oliva and Torralba 2001), is also used. SIFT, denseSIFT, colorSIFT, and

GIST are image based features; therefore, they do not exploit the motion information in a video. In order to extract the motion information in a video, MBH—motion boundary histograms (Dalal et al. 2006)— are employed in multiple studies. HOG3D is a volumetric histogram of oriented gradients (HOG) where the third dimension is time (Klaser and Marszalek 2008) and is employed in numerous studies for the same purpose. STIP—Spatio-temporal interest points Laptev (2005)— has also been used commonly. Ballan et al. (2011) briefly summarize descriptors and concepts used in video event detection.

In the literature, the descriptors used to create concepts as used for the video event detection task directly, sometimes the concepts themselves are used for the task, and occasionally they are used in combination. In our work, we use only concepts. However, in a preliminary work of ours, we showed that using our concepts provides higher retrieval accuracies than using the features to form them (Can and Manmatha 2014). In that work, we showed that the retrieval accuracies obtained by SIFT-based (e.g., dense SIFT, and color SIFT) descriptors can be improved by concepts created with these descriptors.

Even though concepts and descriptors provide promising results, fusion of multiple descriptors and concepts provide the highest retrieval accuracies within the context of VED-ex (Cheng et al. 2012, Liu et al. 2013a). For the VED-zero case, descriptors that are used to create concepts cannot be utilized because they do not carry semantic information.

In addition to existing descriptors and concepts in the literature, recent attempts show promising results using deep-learning based features. A recent success, Overfeat (Sermanet et al. 2013), in object detection and localization using deep-learning features catalyzes the use of these features in many fields including video event detection. Gan et al. (2015) propose a flexible deep convolutional neural network (CNN) infrastructure for video event detection and they show promising results. Further,

using Overfeat features also show promising results within the context of VED-ex. Note that we also report some results using Overfeat features in Chapter 4. We believe that further progress and improvements crafted for the video event detection task using deep-learning features will be provided in near future.

So far, we have discussed zero-shot video event detection considering only visual concepts. Other modalities can also be considered to improve the retrievals. Especially the texts recognized in the speech of videos' audio track (automatic speech recognition: ASR) and text extracted from videos (video optical character recognition: VOCR) have shown to be successful in video event detection (Dalton et al. 2013, Habibian et al. 2014, Oh et al. 2013). They usually lead to retrieval results with a high precision and a low recall since not all of the videos have speech or text to be recognized. Further, it is difficult to automatically select the correct terms in ASR/VOCR search as it is uncommon to have exact matches between the query description and the ASR/VOCR text. "Renovating a home" is an example where it is hard to automatically extract the terms from the textual description of a query. Videos relevant to this query might involve conversations about very specific topics such as updating the drainage system in a bathroom and changing the material used in the corner of a living room. For this example, finding a query which yields high recall is a very challenging task.

2.1.2 Concept Selection

Up to now, we have summarized the concepts and the descriptors used to create these concepts in video event detection. In this thesis, we take advantage of exemplar videos to improve the efficiency of video event detection by using query-independent concepts as an alternative to query-based concepts. For query-based concepts, a number of concepts are selected based on a query and video event detection is performed using these concepts. For example, Chen et al. (2014) focus on descriptions of queries to identify candidate concepts to represent videos. As an alternative to this approach,

we make use of concepts that are selected independently from the queries. We show that concepts that are independent of a query (i.e., query-independent concepts) provide as high a retrieval accuracy as the one obtained using concepts that are selected depending upon a query (i.e., query-dependent concepts). To the best of our knowledge, this statement has not been shown yet within the context of video event detection with exemplars.

Similar approaches have been used and shown to be successful in object category recognition. Torresani et al. (2010) focus on the idea of expressing a "novel" object category using a set of predefined categories (the authors call it "classemes"). They use the outputs of predefined categories to create "classemes vectors" which will be used later as an input to their object category classifiers. Cusano et al. (2012) focus on a similar idea. In contrast to Torresani et al. (2010), they do not use the labels of the predefined categories; however, they cluster descriptors extracted from images into k groups each of which is then assumed to be a class/category. They call this method "unsupervised classemes". Next, they train a classifier for each "unsupervised classeme" and the outputs of these classifiers are used to form unsupervised classemes vectors. An unsupervised classemes vector consists of the presence/absence of these unsupervised classemes where the the presence/absence is determined depending upon the outputs of the classifiers.

2.1.3 Representation Efficiency

After selecting concepts, we now can use the selected concepts to represent videos. In video event detection, it is a common practice to detect concepts at different locations in a video as videos consist of a sequence of frames. For example, if a video consists of 10 frames, it is common to detect concepts at each of these *ten* frames. In this way, we can keep track of multiple occurrences of concepts as well as occurrences of concepts at different locations. Even though we detect concepts at multiple locations,

we prefer to have one representation of a video. The raw detection output scores from different locations in a video are pooled over to produce a representation of a video. Traditionally maximum (i.e., considering the maximum output score over video frames) and average (i.e., considering the average output scores over video frames) pooling functions are used for this purpose (Chen et al. 2014, Cheng et al. 2012, Liu et al. 2013a, Natarajan et al. 2011; 2012, Yu et al. 2012).

Using maximum and average pooling leads to a non-sparse representation. Having a sparse representation is a key element to improving the efficiency in video event detection. As an alternative to these commonly used pooling functions, we provide a space-efficient representation that considers only a limited amount of concepts as present at each detection level. Our space-efficient representation assumes presence of the concepts whose signals are the strongest. For example, there are $1,000$ concepts in our dictionary and we only consider the 10 strongest concepts as present at each detection level. This filtering step quantizes the concepts in a sparser way which enables us to have efficient training and testing. Further, storing video representations on a disk becomes less costly as our method reduces the amount of space required to store video representations as well.

Up to this point, we have summarized the previous work for the video event detection with exemplars case in terms of video representation. We have also highlighted the issues that needs to be addressed. Further, we have explained how our methods differ from the previous work as well. In Chapter 3, we detail our method and provide an extensive analysis compared to the previous work that are recent and closest to our work.

2.1.4 Event Modeling

After representing a video using concepts, a classifier is used to create a model for each event. As we mentioned above, the main purpose of learning a retrieval model specific

to a query is to learn distinctive characteristics of that particular query. Leveraging this knowledge, we can retrieve relevant videos to this query. A representation of a video is conventionally a vector of values and they are calculated using concept detection output scores (e.g., quantized as in our case or used directly when pooling functions employed). When we train a retrieval model, we learn a set of weights. These weights help us determine which concepts are important for a query. Recalling the "repairing an appliance" example, the presence of an "appliance" might be strong evidence for videos being relevant to this query. For this example, we expect the weight for the "appliance" concept to be high.

Even though there are several different machine learning algorithms that can be used to train a retrieval model, the variety of the methods that have been employed so far is rather limited. Support vector machines (SVM) are often used and shown to be promising in quite a few studies (Ayari et al. 2011, Cao et al. 2011, Cheng et al. 2012, Natarajan et al. 2012, Natsev et al. 2010, Oh et al. 2013). SVM often utilizes kernels to define similarity between examples. The intersection and χ^2 kernels are the most common kernels used in these studies (Ayari et al. 2011, Cao et al. 2011, Cheng et al. 2012, Natarajan et al. 2012, Natsev et al. 2010, Oh et al. 2013) and the intersection kernel often provides slightly higher retrieval accuracies compared to the χ^2 kernel. These kernels are non-linear kernels and they tend to be slower than their linear counterpart by definition. When efficiency is more important than effectiveness, linear kernels might be employed as well. In Section 3.4.3, we provide a detailed comparison of linear kernel and the intersection kernel in terms of efficiency and effectiveness.

Slight modifications on the SVM classifier have been proposed for a better video event detection. Tang et al. (2012) use a latent structural SVM with a hinge loss function (Felzenszwalb et al. 2010, Yu and Joachims 2009) to learn classifiers for events. Ma et al. (2012) propose a SVM-like approach that fundamentally mines

the correlation between the descriptors and the semantic information using example videos. They compare their approach with a baseline where classifiers are learned using a SVM with Gaussian and χ^2 kernel. The same authors made use of a very similar approach (Ma et al. 2013) where they learn correlations between the video attributes and an event. In the paper, the semantic labels of external videos (e.g., web videos) are used as attributes in contrast to the conventional understanding of attributes where they might be defined as adjectives such as "furry" (e.g., cat) or discriminative phrases "dogs have it but sheep do not" and "has wheel" (Farhadi et al. 2009).

In addition to the SVM classifier, a few other techniques have been employed for better video event detection. Shirahama et al. (2014) present a hidden conditional random field approach to train a retrieval model. They compare their model with SVM with two different pooling techniques (i.e., average and maximum). According to their results, their method outperforms the SVM based retrieval models slightly. Gkalelis and Mezaris (2014) propose a novel nonlinear discriminant analysis method called generalized subclass discriminant analysis as an alternative to kernel based SVMs. They show that their proposed method enables them to improve the retrieval accuracy within the context of video event detection with exemplars. Mazloom et al. (2013b) use a video as a query to retrieve videos similar to the query video. This approach can be summarized as a k-nearest neighbor approach (i.e., finding similar videos for a given video). Alternatively it can be considered "query by example" as in "query by document" where a document is used as a query (Yang et al. 2009b) and "query by image" in which search is based on an image or a visual sketch Snoek and Worring (2008). Bhattacharya et al. (2014) also follow the same setting (i.e., query by video) to discover minimally needed evidence to identify the presence of an event in a video. Mazloom et al. (2014) use tag propagation for video event retrieval in a similar setting.

The SVM classifier and a few variations have been employed in video event detection studies. These studies have a commonality: using all of the available example videos to train a retrieval model. It is often required to do so to generalize a query using example videos. However, it might be insufficient if example videos do not always share common concepts. For example, we might have "repairing an oven", "repairing a refrigerator", and "repairing a dishwasher" as example videos for the "repairing an appliance" query. For this example, we can learn general characteristics of "repairing an appliance" using all available example videos. However, we would be discarding specific characteristics of the individual examples. We address this issue by incorporating one-exemplar models into video event detection. In addition to a global model that uses all available example videos, we create several one-exemplar models each of which is trained using one example. One-exemplar models are specific to their exemplar; therefore, we learn specific characteristics of individual exemplars as well. In Chapter 4, we detail our approach and show that our space efficient representation with query-independent concepts can be improved by using one-exemplar models. Further, we have evaluated our one-exemplar models using other descriptors to show that it is robust to multiple descriptors and not only good for our method.

Malisiewicz et al. (2011) make use of a similar one-exemplar approach for object detection. They create exemplar-based models each of which is trained on one positive example for object recognition. Unlike their work, we focus on the video event detection task and deal with videos. In another study, Can et al. (2014) incorporate query-specific feedback into learning-to-rank models. They point out that models that are trained on a single query provide statistically significant improvements on the retrieval accuracies compared to using a standard learning-to-rank model which is trained on multiple queries.

2.1.5 Ranking

Once the concepts are detected and classifiers are trained, the issue is to rank test videos according to their relevance to a given event query. For this purpose, test videos are run against the classifiers. A classification score is obtained for each video. This classification score is then used for ranking purposes. When multiple descriptors are employed, multiple classification scores are obtained. These scores are fused to have a final score for ranking. This approach is called "late fusion" where the classification scores are combined.

Mazloom et al. (2013b) take the arithmetic mean of the classification scores to fuse multiple scores. Tamrakar et al. (2012) use geometric mean to merge classification scores. They also compare weighted combinations of the scores and point out that tuning weights for each score does not change the final performance significantly. More complicated methods than average mean and geometric mean are also employed in late fusion. Natarajan et al. (2011) fuse classification scores using both Bayesian model combination (BAYCOM) as well as a weighted-average method. Jiang et al. (2012) focus on fusion of descriptors and concepts, that is based on collective classification. They encode concepts into graphs and diffuse the scores on the graph for the final fused prediction. In order to construct the graph, they make use of logarithmic and exponential loss functions and two collective classification techniques: Gibbs sampling and Markov random walk. They theoretically show that their method is scalable. Oh et al. (2013) compare arithmetic mean, geometric mean, MFoW (i.e., Maximal-Figure-of-Merit) (Kim et al. 2012), and Expert Forest (Liu et al. 2012) fusion methods and conclude that geometric mean and expert forest provide higher retrieval accuracies than the others. Oh et al. (2014) focus on a MFoW-based fusion method that is formulated in a linear discriminant function. Myers et al. (2014) evaluate a number of fusion techniques such as arithmetic mean, geometric mean, and weighted fusion methods. They note that arithmetic mean and geometric mean are the best

fusion techniques in terms of retrieval accuracy. Besides, these methods do not require additional steps to learn weights and can be applied directly to the classification scores. Oneata et al. (2012) and Aly et al. (2013) learn the weights for each classification score using part of the training set and their final score consists of a weighted sum of the scores obtained using the classifiers trained on multiple descriptors. Yang and Shah (2012) use sparse coding (Olshausen 2000) to perform late fusion but did not detail their method. Modolo and Snoek (2013) choose arithmetic mean as their fusion method. Lan et al. (2012) describe a fusion schema called double fusion that combines early fusion and late fusion. Their main focus in this work is to mine the example videos to find out the most discriminative features. Their experimental results show that their fusion is better than its precedents. Even though there have been a number of different methods used for late fusion, arithmetic and geometric mean are often reported to be successful and simple methods for late fusion.

As an alternative to late fusion, "early fusion" is also employed (Ayari et al. 2011, Bhattacharya et al. 2014, Natarajan et al. 2011; 2012, Tamrakar et al. 2012). In contrast to late fusion, here fusion is performed before the classification step. In other words, the histograms of the descriptors are concatenated and the merged histograms are fed to the classifier for training. Ayari et al. (2011) indicate that early fusion improves the retrieval accuracies more than late fusion. However, Tamrakar et al. (2012) report that late fusion outperforms early fusion. There is no consensus on early and late fusion within the context of video event detection with exemplars.

To sum up, simple techniques such as average mean and geometric mean seem to be chosen over complicated methods. They do not require any additional steps which might be costly. Further, none of the other methods have been shown to be significantly and constantly better than these mean based techniques. For the comparison of late over early fusion, there is no strong evidence to chose one over the other. As a result they can be assumed equally effective. Keeping these motivations in

mind, we prefer late fusion and arithmetic mean when we evaluate our one-exemplar models in the context of fusion of multiple descriptors.

2.2 The Zero-Shot Case

Zero-shot video event detection is the problem of searching videos for a given event query where no exemplars are available. Therefore, it is a challenging task. In contrast to the video event detection with exemplars case, here we cannot take advantage of the exemplars to train retrieval models. Therefore, we need to focus on a retrieval model that bridges the text description of an event query and the semantic information that we aim to extract from the video content. To this end, concepts are preferred over descriptors since concepts are expected to carry more semantic information.

Unlike the video event detection with exemplars case, there have only been a few papers on zero-shot video event detection. It might be partly because it is challenging and therefore the retrieval accuracies are too low.

Early attempts were based on hand-crafted decision functions that map descriptors to a single concept. For instance, Zhang et al. (1995) mainly focus on detecting "news anchor person" in news videos.

The number of concepts has increased to handle different types of queries in recent studies. For example, Dalton et al. (2013) use multiple modalities to tackle zero-shot event detection problem. They use text extracted from videos, texts recognized in the speech of their audio track, as well as concepts in videos for zero-shot video event detection. They also expand the queries and use relevance feedback for a better retrieval. Similar to Dalton et al. (2013), Jiang et al. (2014) also focus on relevance feedback. They propose a multimodal pseudo relevance feedback method for event search in videos and evaluate their method on different descriptors such as audio speech recognition and video optical character recognition (OCR). Similar to Dalton et al. (2013), Younessian et al. (2012) focus on automatic speech recognition (ASR)

transcripts as well but they are more interested in acoustic concept indexing and ASR then multimodal approach. They propose an adaptive semantic similarity approach to measure textual similarity between ASR transcripts.

Chen et al. (2014) propose an automatic semantic concept discovery scheme by exploiting Internet images and their associated tags. They report that their discovery technique provides promising results on the multimedia event detection task without examples.

Rastegari et al. (2013) focus on bi-concepts for image search. They analyze bi-concepts by searching for concept pairs where a joint classifier is more accurate than their individual counterparts. Mazloom et al. (2013a) use a similar idea in video event detection. They propose bi-concepts where a concept detector is created to cover a pair of concepts. For example, they create a "cat and food" concept that is created using examples having "cat" and "food" together. Concept combinations are practically defined as concept co-occurrences in annotations. Therefore, it is no different than creating yet another concept detector that is tuned to only fire when "cat and food" exist together in a video. The main issue here is that bi-concepts are carved for very specific queries. For example, "cat and food" might provide a strong evidence for the "feeding a cat" query. However, we cannot use this concept for "petting a cat". Further, finding examples to create a "cat and food" detector might be harder than finding examples with "cat" and "food" individually.

In a similar work, Habibian et al. (2014) expand the idea of bi-concepts in video event detection. They make use of both 'and' and 'or' logic operations to combine concepts compared to using only 'and' operator (Mazloom et al. 2013a). They also combine individual concepts after creating the detectors. This is different than the work of Rastegari et al. (2013) and Mazloom et al. (2013a) where joint detectors are created. In their work, they identify concepts to be combined based on training data for each query. This work gets closer to be a video event detection with exemplars

work rather than a zero-shot detection work as concepts to be combined are identified using a number of example videos.

Unlike the previous work, we provide an Markov random field (MRF) based retrieval model that exploits the spatial- and temporal-dependencies in addition to the individual occurrences of concepts that have not been studied previously. Exploiting dependencies of concepts enables us to capture the concept - video relations better. In this way, we provide better coverage for different types of queries. Our dependency work is based on individual concepts (not joint concepts like Rastegari et al. (2013) and Mazloom et al. (2013a)) and does not require example videos to select concepts.

We also provide three different ways to use the concept detection outputs. We first quantize concept detector outputs into presence/absence information and use the frequency of occurrences of present concepts in our estimations. In addition to quantization, we also use the concept detector output scores directly. The first approach might eliminate some noise but might also trim some useful information. Unlike the first approach, the latter one does not trim any useful information. However, it does not remove any noise either. Our final approach considers these approaches together aiming at leveraging advantages of both methods. In Chapter 5, we detail the ways we used in our estimations and show that the final approach improves the retrieval accuracies of the first two approaches.

2.3   Multimedia Event Detection Track

The TRECVID Multimedia Event Detection (MED) Track started in 2010. It aims at searching multimedia recordings (e.g., videos with audio) to satisfy an information need. This need, for this task, is defined to be user-defined events that are based on pre-computed metadata (Over et al. 2010). Multimedia event detection is an extended version of its precedents including video retrieval. These extensions include, but are not limited to, the definition of an information need. While video retrieval

does not shape the limits of a query, the MED Track requires that a query be related to an event, a complex activity, involving people interacting with other people and/or objects, and consisting of several actions/activities (Jiang et al. 2013, NIST 2012, Over et al. 2010). Further, an information need is also defined to contain not only a textual description of a query (e.g., "a birthday party") but also several explanatory metadata items such as definition and explication.

Even though the ultimate goal is to reach the point where we can produce a system which retrieves videos relevant to an event query without supervision, the track has also another subtask: MED with exemplars where we can gain knowledge from the exemplars and use that to provide better ad-hoc MED. In addition to detection of an event, identifying its location in the video (e.g., event recounting) is also targeted in the track. Considering these tasks, extracting concepts (e.g., objects and actions) from videos is highly encouraged as concepts are expected to provide semantically meaningful information about the content of videos. Using semantically meaningful information might help us on ad-hoc MED and event recounting.

Evaluations were required to be provided in terms of missed detection rate by false alarm rate in the beginning of the track. Information retrieval (IR) evaluations including average precision, have been preferred later. In addition to evaluation metrics, test and training sets evolved drastically throughout the track. More videos and queries have been added to the evaluation sets every year, which makes it challenging to settle upon a standardized dataset, forcing researchers to come up with their own settings. It is, therefore, possible to see multiple inconsistent settings in the literature. For example, Mazloom et al. (2013b) focus on a set that consists of approximately 9,000 videos. The same authors in the same year but in a different study focus on a set of approximately 35,000 videos (Mazloom et al. 2013a). Further, for the MED with exemplars case, the number of positive examples is also affected by consistently growing

datasets. Fortunately, a good amount of effort has been dedicated to standardize the evaluation settings for the track.

The highest retrieval accuracies are obtained by blending results of multiple descriptors. Jiang et al. (2010) started to blend results of multiple descriptors within the context of VED. They discover that fusing results of SIFT descriptors, spatio-temporal interest points, and MFCC audio features yield higher retrieval accuracies compared to considering them individually. Many researchers now fuse results of multiple descriptors especially before their formal evaluation submissions.

There has been also a formal evaluation every year. Multiple research teams participated in this evaluation by providing their results on a pre-determined evaluation set and queries by NIST. Then, NIST announces the results. Our work has been part of the research team compiled at University of Massachusetts Amherst mostly partnering with SRI International. Therefore, our work is aligned with the tasks provided by the track. In the following, we will explain the impact of the track on our work.

2.3.1 MED Tasks vs. Our Work

In our experiments, our main focus is on video event detection (VED) rather than multimedia event detection, where we eliminate the audio and textual modalities while ranking videos as our core contributions are centered around the visual modality. As in the MED track, we focus on two major sub problems: video event detection with exemplars (VED-ex) and video event detection with no exemplars (VED-zero). In addition to focusing on the same sub-problems, some of our design choices are also based on the requirements of the track. Below we explain these choices.

Concept Selection: The core aim in the track is to be able to retrieve videos relevant to an event query without any supervision. In order to achieve this, we are encouraged

to extract semantic information from videos. For this purpose, we focus on concepts (e.g., objects and actions) in videos for both sub tasks (i.e., VED-ex and VED-zero).

For VED-ex, we compile our own concepts each of which is created using static images. To do so, we first collect a number of images from a large database of images (Image-Net 2014). After collecting the images, we extract descriptors from those images, which is followed by training classifiers specific to their concepts (i.e., concept detectors). We detail this process in the following chapter (see Section 3.3.2).

The images on ImageNet are human-annotated and they mostly represent an object. "Airplane, umbrella, magnifying glass, pumpkin, leopard, toyshop, and bean" are instances of these concepts. The concepts we use in Chapters 3 and 4 are created using images from this database. We state that we make use of query-independent concepts within the context of VED-ex and the concepts are selected without any prior knowledge to queries and bias. Therefore, it might be difficult to reason using the "magnifying glass" concept to identify videos relevant to a query that is totally unrelated with it. However, in the next chapter we show that labels of concepts have little effect on retrieval accuracy within the context of VED-ex. Note that, these query-independent concepts and sparse representation are part of our submission to the track in 2012 and 2013.

Our claim of using query-independent concepts works for only VED-ex. For the other task, VED-zero, we select a number of concepts for each query. As VED-zero is more challenging compared to VED-ex, we borrow action-based concepts from our partners (SRI International and UCF). They make use of existing action recognition sets: HMDB (Kuehne et al. 2011) and UCF101 (Soomro et al. 2012), while creating their concepts. These concepts are mostly for generic actions such as "running". They created 152 action-based concepts from these sets. In addition to existing sets, as encouraged by the early years of the TRECVID MED track, some of their concepts are inspired by known queries. In other words, they first go over the description of

an event query and determine the concepts to be created for this query. They then create the selected concepts. Therefore, it is not surprising to get a concept labeled as "cutting a tree", which is dedicated to the "felling a tree" event query. These action-based concepts did not have a brief description as we have for the concepts we created using ImageNet. However, human annotators added brief descriptions to some of the action-based concepts. In our work, we create a dictionary of concepts using all of these concepts, that includes these specific concepts as well as a number of generic concepts such as "running". Rather than manually determining the concepts to be used for a query considering our prior knowledge, we automatically detect a subset based on textual similarity of concept labels and query description. We detail this automatic approach in the chapter where we introduce our VED-zero approach (see Section 5.1.1).

Evaluation Settings: In our experiments, we aim at using the standardized sets and queries. To do so, we use the most recent data made available by NIST for the track. The training sets used in this study are EK100 where each query has 100 example videos, and EK10 in which each query has 10 example videos. We make use of EK100 in the evaluation of our QIC and sparse representation. EK10 meets our needs for evaluating one-exemplar models. As the test bed, we focus on a set of approximately 27,000 videos with 30 queries, that is the most recent publicly available test bed for the track. We provide experimental results in terms of average precision (in percent) as it is the most recent decision in the track and standard in IR. More information about our experimental settings is provided in the next chapter while introducing the experimental environment (see Section 3.3).

Space and Time in Event Queries: When we analyze the queries provided by NIST, it sometimes might be difficult to see queries explicitly matching with the official definition (i.e., a complex activity occurring at a specific time and space). For

example, while the definition of the "landing a fish" query states that the "landing a fish" event takes place on or at the shore of a body of water, it does not provide information about its "time".

Extracting "space" and "time" information from a definition of a query is rather complicated. Therefore, using "space" and "time" information in our approaches is also difficult. In our work on dependencies, we treat the "place" constraint as two concepts occurring at the same location at the same time (i.e., same frame). The "time" constraint is treated as concepts co-occurring in an order (e.g., "landing a fish" is expected to be following a "catching a fish"). We detail our spatial and temporal concepts in the last chapter (see Section 5.1.2). Note that our work for VED-zero was used as part of the submission to the track in 2014.

CHAPTER 3

VIDEO EVENT DETECTION
USING QUERY-INDEPENDENT CONCEPTS

This decade has seen an upsurge of interest in video event detection. Motivated by the importance of this task, we identify some shortcomings of recent approaches and provide solutions to address them.

In this chapter, we investigate query-independent concepts as an alternative to selecting concepts based on queries. In particular, we show that query-independent concepts can be used as an alternative to query-dependent concepts without sacrificing the effectiveness of video retrieval. Using query-independent concepts enables us to skip the concept selection step, which is an inevitable step for query-dependent concepts. This is especially important when new concepts are needed to be created (e.g., in case the selected concepts are not in the vocabulary) as creating new concepts can be very time consuming.

Another shortcoming of recent approaches is the non-sparse representation of videos. While using all concepts to represent videos is a common practice in VED-ex, it has a drawback: final representations are dense; therefore, they are costly to store on the disk and expensive when training/testing retrieval models. To address this issue, we provide a space-efficient technique to represent videos using concepts.

In the following, we first detail query-independent concepts and then explain the details of our space-efficient technique. Next, we present the experimental results and discussion, which are followed by an extended analysis of our approach in terms of concept detection, classifier choice, and parameter selection.

3.1 Query-Independent Concepts (QIC)

In the video event detection problem, several concepts are used to represent a video since events are defined to be complex activities. One of the challenges in this problem is the selection of concepts.

Recent approaches focus on selecting concepts based on underlying queries (Chen et al. 2014, Jiang et al. 2012; 2013). For example, Chen et al. (2014) identify a set of possible concepts for a given query then use these concepts for video event detection. In this case, a new query requires revisiting the concept selection process. This process is expensive particularly when new concept detectors need to be created.

As an alternative to query-dependent concepts, we hypothesize that concepts that are selected independently from the queries can be used for video event detection. In this way, higher efficiency can be achieved. We also claim that using query-independent or query-dependent concepts is not a factor for the effectiveness of video retrieval.

To integrate query-independent concepts in video event detection, we create a fixed set of concepts selected without prior knowledge of the queries and use that set of concepts for any query. Using a fixed set of concepts enables us to save time since we do not have to construct a new set of concepts for each query. This is important as it enables us to skip the concept selection process for a new query. For example, while Chen et al. (2014) need to spend time to construct a new set of concepts for every new query, we use the same set of concepts for all queries. Note that when we construct the fixed set of concepts, we do not require that concepts have semantic relationships with queries. Therefore, the selected concepts might be semantically irrelevant to some queries.

Representing a video with a set of predefined concepts is motivationally similar to representing an object with other objects in object recognition (Cusano et al. 2012, Torresani et al. 2010). Further, word embeddings, a recent and fast-paced growing topic, stem from a similar idea: words are mapped to vectors, where distributional

similarities of words are used to describe the target word (Mikolov et al. 2013, Socher et al. 2013, Turian et al. 2010). An older type of word embedding—latent semantic analysis (LSA)—learns semantic word vectors as well (Deerwester et al. 1990, Maas et al. 2011).

In word embeddings, the vectors are often (or at least claimed to be) semantically similar to the target word. In object recognition, Torresani et al. (2010) use a set of objects to describe another object. They claim that a distribution of the outputs of a set of object detectors may be used to describe an object. For example, they report that "helmet," "sports track," "cake pan," "collectible," and "muffin pan" are observed to be important objects while describing a "cowboy hat" in their work. Cusano et al. (2012) embellish their idea, making the claim that the label of a pre-defined category adds nothing to the classifier trained specific to this category. They group the examples of objects considering their visual characteristics rather than their labels. They revisit an important observation from (Torresani et al. 2010): "we work on the assumption that modern category recognizers are essentially quite dumb: so a swimmer recognizer looks mainly for water texture, and the bomber plane recognizer contains some tuning for the 'C' shapes corresponding to the airplane nose and perhaps the 'V' shapes at the wing and tail". Cusano et al. (2012) claim, on top of this observation, that the labels add nothing to the final recognition as the labels do not have direct implications on the recognizers.

Query-independent concepts center around a similar idea. Videos relevant to an event query are expected to be somewhat visually similar to each other. Therefore, the concepts with a high detection confidence on those videos should also be similar to each other independent of their semantic labels. Concept detectors are basically classifiers that we train. A concept detector may fire not only when we have that particular concept but may also have a reasonable response when other concepts are present. This stems from the idea that these detectors are responding to certain visual

features or elements of the picture but not necessarily the semantics. Experimentally, we observe that the distribution of the responses of concept detectors (even those which are not semantically related to the concept) often respond consistently to certain video events.

For example, we learn from the example videos that detecting the c_κ concept is more likely than the c_τ concept in videos relevant to the "horse riding competition" event query. When we rank videos according to their relevance to this query, we put videos that are rich in terms of the c_κ concept on the top ranks, while pushing the ones loaded with the c_τ concept to the bottom ranks. Knowing the semantic labels of c_κ and c_τ adds nothing to the ranking. In our approach, we compile a fixed set of concepts by randomly selecting concepts from a large concept pool. No semantic knowledge from the queries and concepts is involved in this process. Therefore, the concepts are query-independent concepts. The size of this fixed set of query-independent concepts is decided by tuning.

Later in this chapter (Section 3.4.1), we show that our hypothesis holds by showing that we are able to obtain comparable retrieval accuracies using a set of query-independent concepts compared to the ones obtained using their query-dependent counterparts. So far, we have detailed the query-independent concepts, next we explain our space-efficient representation that we use with query-independent concepts.

Figure 3.1. Sample video frames of the "horse riding competition" event.

3.2 Space-Efficient Representation

To detect concepts in videos, we need detectors specific to their concepts. One way of learning a concept detector is to acquire a number of examples relevant to this specific concept and use these examples to learn such a concept detector (Vijayanarasimhan and Grauman 2011). For example, in order to create an *airplane* detector, we can make use of the images provided in Figure 3.2 (see Section 3.3.2 for details of learning concept detectors).

Figure 3.2. Sample airplane images.

A concept detector should detect or measure the likelihood of that particular concept in a video. In other words, the output score of a concept detector indicates the chance of observing that particular concept in a video.

It is common to detect concepts at frame or video clip level in video event detection since concepts capture simple actions, or objects. Videos contain several video clips (i.e., short clips of video) and many frames (a usual video is recorded at 24 frames per second FinalCutPro7-Documentation (2015)). Therefore, multiple scores (e.g., from different parts of a video) of a concept detector are obtained for a video. The chance of observing a concept is then calculated by considering a combination of these multiple scores. This process is called "pooling". For example, Cheng et al. (2012), Jiang et al. (2013), Liu et al. (2013a) use average and maximum pooling of scores over frames.

Assume we have scores as provided in Table 3.1, then the final representation of a video with average and maximum pooling becomes as shown in Table 3.2.

Table 3.1. Concept detector scores Φ_i of the concepts $\{c_1, c_2, ..., c_{10}\}$ at video frames v_1 to v_5.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
v_1	0.13	0.20	0.04	0.04	0.07	0.12	0.19	0.03	0.02	0.15
v_2	0.08	0.20	0.07	0.05	0.06	0.10	0.06	0.07	0.21	0.11
v_3	0.17	0.11	0.10	0.04	0.09	0.02	0.12	0.07	0.17	0.12
v_4	0.15	0.18	0.07	0.07	0.05	0.07	0.21	0.12	0.01	0.07
v_5	0.06	0.10	0.04	0.08	0.16	0.07	0.15	0.09	0.16	0.08

Table 3.2. Final representation calculated by pooling the scores in Table 3.1 over frames.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
Avg. Pooling	0.12	0.16	0.06	0.06	0.09	0.08	0.15	0.08	0.11	0.11
Max. Pooling	0.17	0.20	0.10	0.08	0.16	0.12	0.21	0.12	0.21	0.15

The main issue with these pooling techniques is that the final representation of a video is non-sparse. In the example above, there are ten concepts ($|C| = 10$) and the final representation with average or maximum pooling has ten non-zero values. Sparse representations require less space than their dense counterparts. Further, training a retrieval model using a non-sparse representation takes more time than the sparse ones.

We address these space and time problems by considering only a small number of highly responsive concepts. In this way, the final representation of a video becomes sparser compared to using all concepts. Representation of a video $H = \{H_1, H_2, ..., H_{|C|}\}$ with our space-efficient technique can be computed as follows:

$$H_i = \sum_t \varphi(c_i, v_t, k) \tag{3.1}$$

44

where $\varphi(c_i, v_t, k)$ is defined to be:

$$\varphi(c_i, v_t, k) = \begin{cases} 1, & \Phi^i(t) \geq \Phi_k(t) \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

where $\Phi^i(t)$ is the detector score for the concept c_i at video frame v_t and $\Phi_k(t)$ is the kth maximum score in video frame v_t. By using the equation above, we assume the presence of only k concepts in each frame. Note that these k concepts will change for each frame. However, the final representation does not become a dense representation (see Section 3.4.3.3 for details). k is often tuned to a small number such as 10 or 20 per frame where the total number of possible concepts is a large value (e.g., 1,000). Since only k (where $k << |C|$) concepts are considered for each frame, the representation becomes very sparse which enables us to have efficient training and testing.

When we consider the scores in Table 3.1, our space-efficient representation where $k = 3$ only considers the concepts below in each frame:

$$
\begin{array}{ll}
v_1 \rightarrow & \{c_2, c_7, c_{10}\} \\
v_2 \rightarrow & \{c_2, c_9, c_{10}\} \\
v_3 \rightarrow & \{c_1, c_7, c_9\} \\
v_4 \rightarrow & \{c_1, c_2, c_7\} \\
v_5 \rightarrow & \{c_5, c_7, c_9\}
\end{array}
$$

Using Equations 3.1 and 3.2, the representation of the video, H, becomes:

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
H	2	3	0	0	1	0	4	0	3	2

where each bin corresponds to a concept. For example, the 1st column ($H_1 = 2$) indicates that c_1 occurs two times (above the threshold) over the frames v_1 to v_5, and c_3 does not occur at all ($H_3 = 0$). In the example above, the number of non-zero values

is six, whereas it was ten with average and maximum pooling. Our space-efficient technique provides a sparser representation.

After constructing this representation of the videos, we normalize the representations since the videos have different lengths ranging from 1 second to 4,000 seconds in our data. (The average length of videos is approximately 110 seconds where the standard deviation is approximately 162 seconds. This shows that the variance in the video lengths is very large, motivating our choice to normalize video representations so that comparing them will be easier and more accurate.)

We make use of a normalization technique that is based on the maximum term-frequency normalization used in Information Retrieval (Baeza-Yates et al. 1999), also known as L_∞ normalization. The main motivation stems from the assumption that higher frequencies of concepts are observed in longer videos since longer videos tend to repeat the same patterns over and over. Furthermore, we also take the logarithm of the values due to the large variance in video lengths. The normalization technique used in the space-efficient technique is as follows:

$$H_i' \;\; = \;\; \frac{\log\left(H_i\right)}{\log\left(\max\left(H_1, H_2, ..., H_{|C|}\right)\right)} \tag{3.3}$$

where H' is the normalized representation of H_i.

No spatial layout is enforced for simplicity while detailing our space-efficient representation. Lazebnik et al. (2006) point out that considering spatial layout increases the descriptive ability of descriptors. They state that without considering the spatial layout it is difficult to differentiate an object from its background. They also note that other solutions to this difficult issue—such as generative part models (Fei-Fei et al. 2007, Fergus et al. 2003)—are often computationally expensive. Their approach involves repeatedly subdividing the image and computing histograms of

46

local features at increasingly finer resolutions. Spatial layouts have been shown to be successful (Bosch et al. 2007, Griffin et al. 2007, Lampert et al. 2008, Yang et al. 2009a). In our representation, we also make use of a spatial layout as illustrated in Figure 3.3. We use the video frame itself ($level_0 = \{r_0\}$), the video frame divided into four regions ($level_1 = \{r_1, r_2, r_3, r_4\}$), and the video frame divided into sixteen regions ($level_2 = \{r_5, r_6, r_7, r_8, r_9, r_{10}, r_{11}, r_{12}, r_{13}, r_{14}, r_{15}, r_{16}, r_{17}, r_{18}, r_{19}, r_{20}\}$). For each region we calculate its own representation, each of which are then concatenated for the final representation.
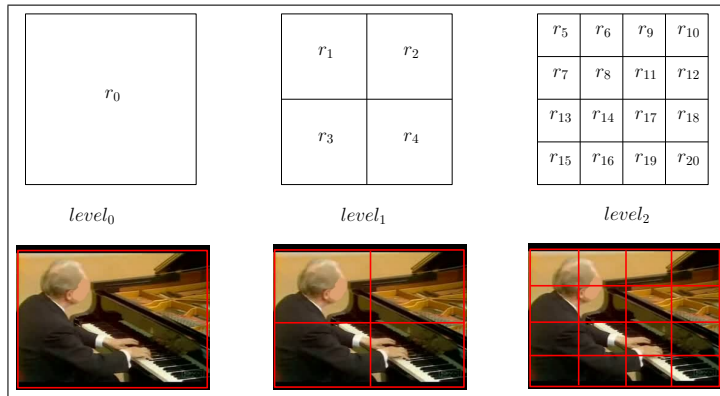
Figure 3.3. Illustration of spatial layout used in our representation.

So far, we have detailed the query-independent concepts with a space-efficient representation for the video event detection with exemplars problem. Next, we introduce our experimental setup. We follow this with the experimental results and discuss them.

3.3   Experimental Setup

Before we provide the experimental results, we introduce our experimental environment starting with the datasets and queries used for evaluation purposes.

3.3.1 Datasets and Event Queries

We evaluated query-independent concepts with a space-efficient representation on a collection of 30 event queries and approximately 27,000 test videos (referred to as MEDTEST, see Table 3.3). Note that 27,000 videos is a larger collection compared to recent action recognition datasets which contains in the range of 5 to 10 thousand videos (e.g., HMDB (Kuehne et al. 2011)). Further, 27,000 videos correspond approximately to 3 million frames/clips.

Retrieval models for the 30 event queries in Table 3.3 are trained using 100 example videos per query and 5,000 non-relevant videos (referred to as EK100). The queries and the datasets are used in the evaluation of NIST's TRECVID Multimedia Event Detection track (Jonathan Fiscus 2014). Note that the event queries from E16 to E20 are officially not released.

Apart from MEDTEST and EK100, a small development set (referred to as TINYSET) is also created for tuning parameters such as the concept vocabulary size $|C| = 1,000$ and the number of concepts to be considered at each frame $k = 10$.

3.3.2 Concept Detection

Representing videos with a set of concepts requires concept detectors. In order to create concept detectors, we first need to extract characteristic features of sample images. For example, we can compute gradient orientation distributions on corner points (e.g., nose, end of wings, joints of body and wings) or alternatively on densely sampled points (e.g., every 5 pixel horizontally and vertically) in the airplane images provided in Figure 3.4. We can make use of well known descriptors such as SIFT (Lowe 1999), SURF (Bay et al. 2008), FREAK (Alahi et al. 2012), and BRIEF (Calonder et al. 2012) for this purpose. In Figure 3.5, we illustrate the gradient orientation distributions of the nose of an airplane. According to the illustration, gradients in the north-east, east, and south-east orientations are the ones having the largest magnitude.

Table 3.3. Id, title, and number of relevant videos of each query in MEDTEST.

Id	Title	# relevant videos
E6	Birthday party	186
E7	Changing a vehicle tire	111
E8	Flash mob gathering	132
E9	Getting a vehicle unstuck	95
E10	Grooming an animal	87
E11	Making a sandwich	140
E12	Parade	234
E13	Parkour	104
E14	Repairing an appliance	78
E15	Working on a sewing project	81
E21	Attempting a bike trick	16
E22	Cleaning an appliance	23
E23	Dog show	20
E24	Giving directions to a location	27
E25	Marriage proposal	33
E26	Renovating a home	33
E27	Rock climbing	18
E28	Town hall meeting	19
E29	Winning a race without a vehicle	22
E30	Working on a metal crafts project	21
E31	Beekeeping	28
E32	Wedding shower	28
E33	Non-motorized vehicle repair	26
E34	Fixing musical instrument	23
E35	Horse riding competition	29
E36	Felling a tree	25
E37	Parking a vehicle	20
E38	Playing fetch	22
E39	Tailgating	27
E40	Tuning musical instrument	26

Then using a machine learning algorithm such as support vector machines (SVM), we can learn the discriminative gradient orientation distributions specific to airplanes so that we will be able to calculate the likelihood of observing an airplane in an image.

In our case, we first extract densely sampled (i.e., every 5 pixels) SIFT—denseSIFT—descriptors from static images. A SIFT descriptor is a 128-dimensional vector indicating the gradient distributions at a point. The vector is normalized; therefore,

Figure 3.4. Sample airplane images.

Figure 3.5. An illustration of gradient orientation distributions of the nose of an airplane.

SIFT is invariant to scale, rotation, and partially to the viewpoint of the camera and illumination (Lowe 2004). We use the implementation of Vedaldi and Fulkerson (Vedaldi and Fulkerson 2008) to extract denseSIFT descriptors. The next step is to quantize the extracted descriptors into visual words. To do so, we randomly choose a large number of denseSIFT descriptors (from static images), each of which is a 128-dimensional vector. We then cluster these descriptors. These clusters are also known as a "codebook" or "visual vocabulary" (Zhang et al. 2010). A denseSIFT descriptor is assigned to the visual word that is the closest cluster centroid to this descriptor (in terms of L_2 distance). Finally, each image is represented by frequencies of these "visual words". This method is known as "bag-of-words" (Fei-Fei and Perona 2005). After representing images by frequencies of visual words, we feed these "bags-of-words"

representations to a SVM classifier for training concept detectors. Next, we run these detectors against the videos to measure the likelihoods of observing these concepts in videos, which will be used to create representation of videos. These representations are used in training/testing a retrieval model.

3.3.3   Training a Retrieval Model

For each query in our collection, we train a retrieval model using example videos as well as non-relevant videos in the EK100 dataset. The retrieval model is trained using a SVM classifier with an intersection kernel. The intersection kernel has been used commonly as a kernel for SVM in computer vision applications ranging from image annotation to action recognition (Can and Manmatha 2013, Wang et al. 2009). It is based on the histogram intersection approach proposed for color indexing with application to object recognition by Swain and Ballard (1991). The intersection kernel for two vector H^A and H^B is defined as follows:

$$K(H^A, H^B) = \sum_i \min\left(H_i^A, H_i^B\right) \tag{3.4}$$

where H_i^A is the ith bin for the vector H^A (in other words i^{th} feature value), and similarly H_i^B for H^B.

Even though a SVM classifier is a classification algorithm, it can also be used for ranking purposes. SVM classifiers are two-class classification algorithms which often produce binary decisions: relevant if $sign(f(x))$ is larger than zero or otherwise non-relevant where $f(x)$ is the decision function. A decision function of a non-linear kernel is as follows (Joachims 1999, Maji et al. 2008):

$$f(x) = \sum_{j=1}^{\#sv} \alpha_j K(x, z_j) + b \tag{3.5}$$

51

where $\#sv$ is the total number of support vectors, K is the mapping function (e.g., map data to another presumably higher dimension), x is a descriptor vector of an example (e.g., representation of a video, H, in our case), z_j is a support vector, and α_j is a set of coefficients for the support vector z_j. In the case of a SVM classifier with a linear kernel, the mapping function is the dot product of x and z_j. In order to rank videos, we, as does the rest of the community, use $f(x)$ for ranking purposes when we use a SVM classifier.

3.3.4 Evaluation

After training a model, we run test videos against the trained model. The output scores of this process are used for ranking purposes. The evaluation of the ranked lists obtained by running our retrieval models is performed by using the relevance judgment released by NIST and the `trec_eval` (http://trec.nist.gov/trec_eval) tool. Average precision (in percent) and mean average precision (in percent) are used as evaluation metrics in our experiments.

Here, we provide our design choices (i.e., concept detectors, classifier choice, and parameters) that we use in our experiments. We also provide an extended analysis of our design choices, after providing results and discussion, detailing how we choose them and why we choose them (see Section 3.4.3).

3.4 Experiments and Discussion

We first show that we are able to obtain comparable retrieval accuracies with query-independent concepts compared to using query-dependent concepts. We then investigate the efficiency and effectiveness of our space-efficient representation comparing it to the techniques that use all concepts.

3.4.1 Query-Independent Concepts (QIC) vs. Query-Dependent Concepts (QDC)

We first compare our results with the results of Chen et al. (2014) since it is a recent and quite successful work on video event detection using query-dependent concepts. To select query-dependent concepts, they first extract nouns and verbs from the event descriptions (an example event description is provided in Figure 3.6) using a natural language processing tool kit (Bird 2006). Then, they form a noun-verb pair using a noun and a verb extracted from the textual description of an event query. Next, they use a number of noun-verb pairs as a textual query to perform a text-based image search on image databases such as Flickr and Image-Net (they mention that the images crafted from Flickr provide better video event detection results compared to images obtained from other sources). Next, they filter the images that might have no visual meaning. They give "economy" as an example and note that this concept is highly abstract, making it difficult to train detectors for it. To remove the image categories having no visual meaning, they measure the accuracy of the concepts using a cross-validation technique.

Event name: Making a sandwich

Definition: Constructing an edible food item from ingredients, often including one or more slices of bread plus fillings.

Explication: Sandwiches are generally made by placing food items on top of a piece of bread, roll or similar item, and placing another piece of bread on top of the food items. Sandwiches with only one slice of bread are less common and are called "open face sandwiches". The food items inserted within the slices of bread are known as "fillings" and often include sliced meat, vegetables (commonly used vegetables include lettuce, tomatoes, onions, bell peppers, bean sprouts, cucumbers, and olives), and sliced or grated cheese. Often, a liquid or semi-liquid "condiment" or "spread" such as oil, mayonnaise, mustard, and/or flavored sauce, is drizzled onto the sandwich or spread with a knife on the bread or top of the sandwich fillers.

Evidential description:
scene: indoors (kitchen or restaurant or cafeteria) or outdoors (a park or backyard)

objects/people: bread of various types; fillings (meat, cheese, vegetables), condiments, knives, plates, other utensils

activities: slicing, toasting bread, spreading condiments on bread, placing fillings on bread, cutting or dishing up fillings.

audio: noises from equipment hitting the work surface; narration of or commentary on the process

---

Figure 3.6. An example textual description of a query.

In Table 3.4, we compare the retrieval accuracies of query-independent concepts (QIC) with our space-efficient representation and the query-dependent method of Chen et al. (2014) (CSI). MEDTEST is used as a test set and EK100 is used for training. Note that Chen et al. (2014) evaluate their method on twenty queries.

Table 3.4. Experimental results of Query-Independent Concepts (QIC) with space-efficient representation and the query-dependent method of Chen et al. (2014), Concepts learned from Selected Images (CSI). Test set: MEDTEST; Training set: EK100. Results are provided in terms of average precision (in percent).

E. Id and Name	CSI(Chen et al. 2014)	QIC
E6 Birthday party	7.5	**13.5**
E7 Changing a vehicle tire	26.0	**26.7**
E8 Flash mob gathering	33.0	**61.7**
E9 Getting a vehicle unstuck	30.0	**30.2**
E10 Grooming an animal	6.0	**17.5**
E11 Making a sandwich	10.0	**14.7**
E12 Parade	17.5	**29.9**
E13 Parkour	20.2	**31.5**
E14 Repairing an appliance	20.3	**35.5**
E15 Working on a sewing project	12.5	**16.5**
E21 Attempting a bike trick	2.5	**7.3**
E22 Cleaning an appliance	2.0	**15.9**
E23 Dog show	**30.0**	28.2
E24 Giving directions to a location	**12.0**	3.6
E25 Marriage proposal	1.0	**5.1**
E26 Renovating a home	7.2	**11.1**
E27 Rock climbing	**11.5**	4.9
E28 Town hall meeting	21.0	**24.9**
E29 Winning a race without a vehicle	12.5	**22.8**
E30 Working on a metal crafts project	7.2	**13.5**
Avg.	14.5	**20.8**

Chen et al. (2014) argue that the query-dependent concepts are better compared to concepts learned from random images. In Table 3.4, we show that we are able to obtain superior effectiveness with query-independent concepts to the ones obtained using query-dependent concepts (see Table 3.5 for sample concepts used in CSI for

the "making a sandwich" query and of query-independent concepts). While we have a retrieval accuracy of 20.8%, query-dependent concepts are able to provide a retrieval accuracy of only 14.5%: that is a 43% relative improvement. This is a statistically significant improvement (p=0.002). Further, on seventeen events out of twenty, our query-independent concepts outperform Chen et al. (2014)'s query-dependent concepts.

Table 3.5. Example concepts of Chen et al. (2014) for the "E11:making a sandwich" event query and our query-independent concepts. Note that there are 1,000 concepts but we only show five for illustration.

Method	E11:making a sandwich
CSI(Chen et al. 2014)	sandwich, food, bread, cooking, cheese
QIC	magnifying glass, pumpkin, leopard, toyshop, bean

Even though our and Chen et al. (2014)'s experimental settings are quite similar, in order to be sure that the effectiveness of query-independent concepts is not due to other reasons such as the source of images, we also compare query-independent concepts with query-dependent concepts within our environment. To do so, we repeat a similar concept selection process of Chen et al. (2014) based on queries and compare it with query-independent concepts. We identify concepts expected to be relevant to events by querying noun phrases in a query description against the textual information of concept. The resulting ranked list consists of the concepts sorted according to their relevance to an event query.

Identification of Query-Dependent Concepts: We collect sample images to train concept detectors from Image-Net (`www.image-net.org`). Each category (i.e., concept) has a brief description and a title (provided by ImageNet). For example, for images titled "microflora", the brief description is "microscopic plants; bacteria are often considered to be microflora." For the query in Figure 3.6, the nouns such as "sandwich," are used to rank the concepts according to their relevance which is

a similar approach to the one used by Chen et al. (2014). In other words, we run these phrases against the concept titles and descriptions (provided by ImageNet) using standard IR approaches of query-likelihood modeling without query expansion to rank the concepts according to their estimated relevance to the "making a sandwich" query. Then we accept the top m concepts as the relevant concepts to the given query. The nouns are extracted using a part-of-speech tagger (Klein and Manning 2003). We provide a few top concept retrievals for the "making a sandwich" query in Table 3.6.

Table 3.6. Top concept retrievals for the "making a sandwich" event query.

ImageNet Title	ImageNet Description
gyro	a Greek sandwich: sliced roast lamb with onion and tomato stuffed into pita bread
garlic bread	French or Italian bread sliced and spread with garlic butter then crisped in the oven
French toast	bread slice dipped in egg and milk and fried; topped with sugar or fruit or syrup
dip	tasty mixture or liquid into which bite-sized foods are dipped
chicken sandwich	a sandwich made with a filling of sliced chicken

In Table 3.7, we compare query-independent concepts with query-dependent concepts within our settings. In the experiments, all the settings are the same except the concept selection process. For query-dependent concepts, a different set of concepts are selected for every query (e.g., gyro, garlic bread, french toast, dip, and chicken sandwich are some sample concepts selected for the "making a sandwich" query). MEDTEST is used for evaluation of both methods and EK100 is used for training. Unlike the previous case, here we consider thirty queries: E6 to E40 since Chen et al. (2014) provide results for only E6 to E30.

Among fourteen events out of thirty, query-dependent concepts outperform their independent counterparts. While we obtain a mean average precision of 19.4% with

Table 3.7. Experimental results of Query-Independent Concepts (QIC) with space-efficient representation and the Query-Dependent concepts (QDC) with space-efficient representation. Test set: MEDTEST; Training set: EK100. Results are provided in terms of average precision (in percent).

E. Id & Name	QDC	QIC
E6 Birthday party	13.4	**13.5**
E7 Changing a vehicle tire	**29.3**	26.7
E8 Flash mob gathering	**62.0**	61.7
E9 Getting a vehicle unstuck	28.2	**30.2**
E10 Grooming an animal	14.3	**17.5**
E11 Making a sandwich	13.7	**14.7**
E12 Parad	**30.6**	29.9
E13 Parkour	31.4	**31.5**
E14 Repairing an appliance	**39.7**	35.5
E15 Working on a sewing project	14.5	**16.5**
E21 Attempting a bike trick	**8.7**	7.3
E22 Cleaning an appliance	**21.4**	15.9
E23 Dog show	**36.0**	28.2
E24 Giving directions to a location	**4.5**	3.6
E25 Marriage proposal	3.9	**5.1**
E26 Renovating a home	9.3	**11.1**
E27 Rock climbing	**5.5**	4.9
E28 Town hall meeting	22.9	**24.9**
E29 Winning a race without a vehicle	14.9	**22.8**
E30 Working on a metal crafts project	**14.8**	13.5
E31 Beekeeping	**22.4**	19.5
E32 Wedding shower	**10.1**	9.2
E33 Non-motorized vehicle repair	26.2	**26.5**
E34 Fixing musical instrument	**21.2**	19.6
E35 Horse riding competition	17.7	**25.3**
E36 Felling a tree	7.3	**10.2**
E37 Parking a vehicle	**17.6**	16.7
E38 Playing fetch	2.0	**2.6**
E39 Tailgating	16.0	**19.7**
E40 Tuning musical instrument	14.0	**17.0**
Avg.	19.1	**19.4**

query-independent concepts, we obtain 19.1% with query-dependent concepts. The

difference is negligibly small. Further, no statically significant difference is observed between the two sets of results (p=0.33).

The similarity in retrieval accuracies obtained with query-dependent and query-independent concepts stems from the idea that a concept detector may fire not only for its specific concept but may also provide strong signals when other concepts are present. Further, concept detectors often respond consistently to certain video events and we learn these detectors (distributions of these detector responses practically) while training a retrieval model. Therefore, we obtain good retrievals even when we use query-independent concepts.

Even though the results of query-dependent concepts are improved within our settings (compared to Chen et al. (2014)'s query-dependent concepts), they still cannot outperform the results obtained with query-independent concepts. This finding supports that our hypothesis holds: we can obtain comparable retrieval accuracies with query-independent concepts to the ones obtained using query-dependent concepts. Our findings also align with the observations of Cusano et al. (2012): the semantic labels do not contribute to the final accuracy. In our case, semantic labels of the concepts added nothing to the retrieval accuracies as well.

3.4.2 Space-Efficient Representation

After showing that we are able to obtain comparable retrieval accuracies using query-independent concepts to the ones obtained using query-dependent concepts, we focus on evaluating our space-efficient representation in terms of first efficiency and then effectiveness.

As our space-efficient representation provides a sparser representation compared to using all concepts with a pooling technique such as maximum and average, we need less space to store our representations. This is also important while training and testing retrieval models. The sparser the representations are, the faster the retrieval

models are trained. In Table 3.8, we provide the amount of space required to store our space-efficient representations and representations created using all concepts. Our space-efficient representation enables us to store the representations created using all concepts with a 1-to-5 ratio of efficiency.

Table 3.8. Amount of space required to store our space-efficient representations and the representations created using all concepts with a pooling technique. Numbers are provided in terms of megabytes (MB)

	Our Space-Efficient Rep.	Average Pooling	Maximum Pooling
Training	268 MB	1,492 MB	1,615 MB
Testing	1,452 MB	6,340 MB	7,911 MB

The sparsity of the representations directly affects the running time for training a retrieval model and testing test examples against it as well. In Table 3.9, we provide the running time required to train and test a retrieval model using the representations created by our space-efficient technique as well as the techniques using all concepts. Similar to the space situation, our efficient technique enables us to train and test retrieval models five times faster than the techniques that use all concepts. Note that testing requires more time than training in all of the cases since the MEDTEST dataset (used for testing) is five times larger than the EK100 set.

Table 3.9. Running time to train/test a retrieval model using our space-efficient representations and the representations created using all concepts with a pooling technique. Numbers are provided in terms of seconds (s).

	Our Space-Efficient Rep.	Average Pooling	Maximum Pooling
Training	245 s	1,005 s	1,504 s
Testing	1,217 s	5,012 s	7,341 s

So far, we have analyzed the efficiency of our technique. We investigate its effectiveness as well. In Table 3.10 we compare the effectiveness of our approach with the techniques using all concepts, namely average and maximum pooling.

Table 3.10. Experimental results of our space-efficient technique (QIC) compared with the techniques using all concepts: average pooling (QIC-AVG) and maximum pooling (QIC-MAX) Test set: MEDTEST; Training set: EK100. Results are provided in terms of average precision (in percent).

E. Id	QIC	QIC-AVG	QIC-MAX
E6	**13.5**	8.5	3.0
E7	**26.7**	9.3	1.5
E8	**61.7**	39.6	25.9
E9	**30.2**	25.1	6.0
E10	**17.5**	4.6	1.9
E11	**14.7**	10.4	6.4
E12	**29.9**	23.9	10.1
E13	**31.5**	11.5	4.1
E14	**35.5**	23.8	15.0
E15	**16.5**	10.6	7.1
E21	**7.3**	6.2	1.1
E22	**15.9**	1.3	5.4
E23	**28.2**	25.4	13.4
E24	**3.6**	1.1	1.4
E25	5.1	**6.9**	3.5
E26	**11.1**	7.8	1.9
E27	4.9	**5.2**	1.1
E28	**24.9**	11.2	11.4
E29	**22.8**	12.7	5.5
E30	**13.5**	10.9	0.5
E31	**19.5**	6.4	1.2
E32	**9.2**	3.2	7.2
E33	**26.5**	13.0	3.3
E34	**19.6**	9.9	3.2
E35	**25.3**	5.1	3.2
E36	**10.2**	2.1	1.9
E37	**16.7**	9.7	0.6
E38	**2.6**	0.7	0.4
E39	**19.7**	8.3	1.7
E40	**17.0**	9.0	1.6
Average	**19.4**	10.8	5.0

Our space-efficient technique significantly outperforms the techniques using all concepts with average pooling and maximum pooling. While a mean average precision

of 10.8% is obtained with the technique using all concepts with average pooling, this score drops to 5% with maximum pooling. This huge difference can be explained with the number of concepts used to represent videos. In our case, we focus on the highly responsive concepts. The other approaches consider all concepts. Using a more informative subset of concepts than using all concepts is a better choice in terms of efficiency as well as effectiveness within the context of video event detection. Previous observations support our findings as well. Habibian et al. (2013) investigate concept vocabularies for video event detection and conclude that a subset of concepts should be considered for a better video event detection. Merler et al. (2012) and Mazloom et al. (2013a) focus on finding the optimal concept dictionary. They report that using a subset of the concepts performs better than using all of the concepts and some subset of concepts are more informative than others. We follow a similar principle in our work.

Note that blending results of different approaches for higher scores is generally a common practice in video event detection. However, using all of the concepts is a terrible choice within the context of VED-ex. Therefore, we do not consider blending our space-efficient representation with the techniques using all of the concepts.

3.4.3 Further Analysis

3.4.3.1 Concept Detection

It is common to create individual concept detectors within the context of video event detection (Chen et al. 2014, Cheng et al. 2012, Habibian et al. 2014, Liu et al. 2013a, Oh et al. 2013). For example, in order to train a "dog" concept, a number of images having "dog" in them are collected. Then, SVM is employed to train a dog detector using descriptors extracted from the sample dog images. In this way, each concept detector is trained independently.

As an alternative to individual concept detectors, multiple concept detectors (MCD) can be considered as well. An MCD learns multiple concepts together rather than learning them independently. For example, if there are 3 concepts: car, bike, and mountain, we create an MCD based on these three concepts using a multi-class classifier (see illustration in Figure 3.7). An MCD produces a car-bike-mountain detector (first row in the figure), whereas the other approach considers a car detector, a bike detector, and a mountain detector, each of which is trained separately using a two-class (e.g., binary) classifier (last row in the figure).
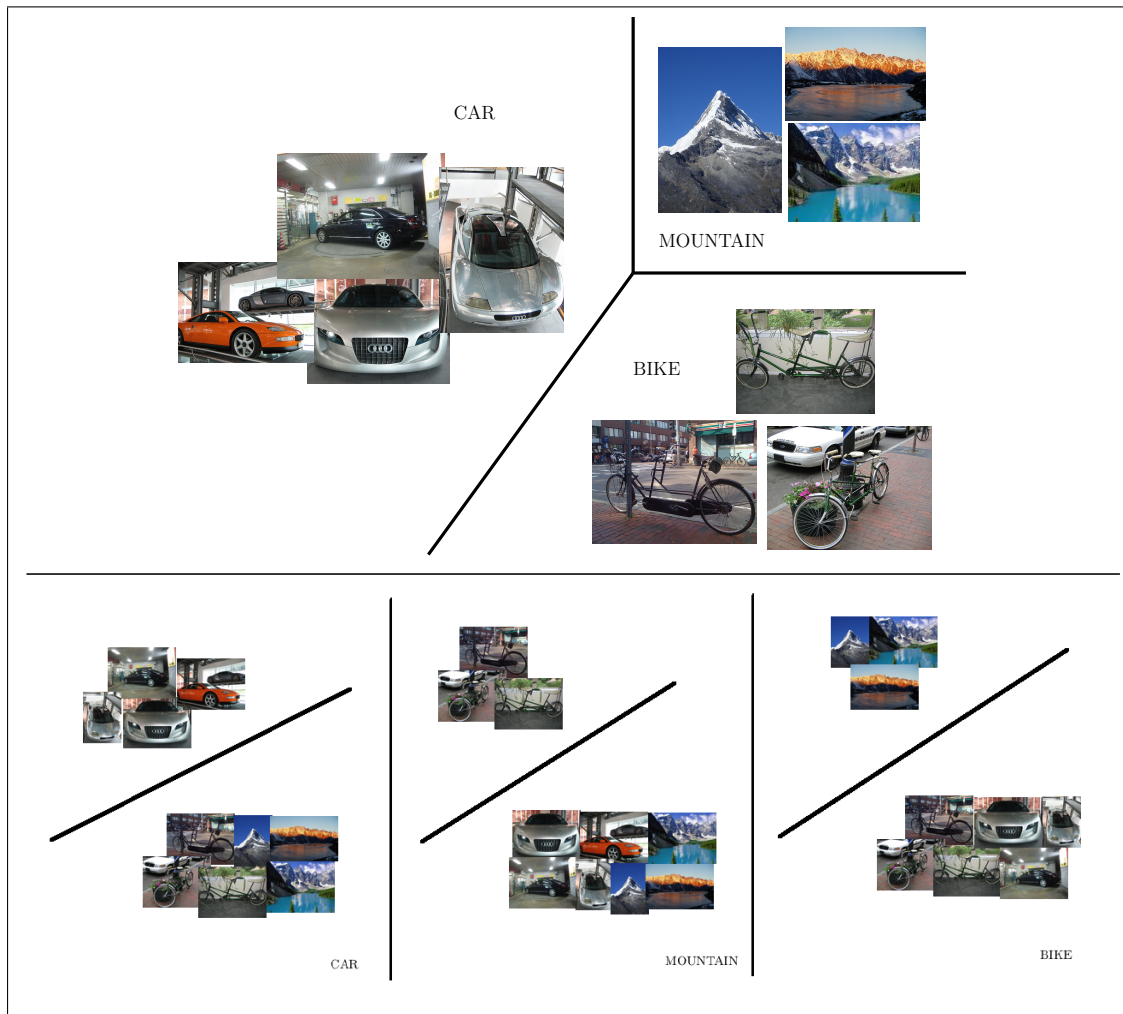
Figure 3.7. Illustration of a multiple concept detector (first row) and individual concept detectors (second row) .

We make use of a linear multi-class SVM Joachims (2014) to train an MCD. The implementation used in this study is a one-vs-rest multi-class SVM. In other words, it creates k separating hyperplanes for k classes. Therefore, an MCD outputs k scores, each of which represents a score of a concept. By employing an MCD, we are able to obtain k scores from the same model which enables us to have a better comparison of the concept detector scores to identify the ones that are most likely to be in a video compared to the scores of individual concept detectors. After creating an MCD, we run test videos against it to measure the likelihood of observing concepts in videos. In order to run videos against an MCD, we need to extract denseSIFT descriptors from videos as well as static images. We first sub-sample videos into video frames at a rate of 3 frames per second. Then we follow the same procedures for video frames that we did for images. We can run video frames against the MCD after completing extraction of denseSIFT descriptors from videos. Figure 3.7 summarizes how video frames are run against an MCD.

In our case, we make use of a MCD to obtain detector scores. We would like to have concept detector scores calculated using the same model, which enables us to have more reliable comparison of scores. To validate our choice, we compare these two methods within the context of video event detection with exemplars. In Table 3.11, we provide the retrieval accuracies of both methods: using individually created concept detectors (QIC-IND) and using a multiple concept detector.

Video event detection results when using a multiple-concept detector are better than the ones obtained using individually created concept detectors. Retrieval accuracies with MCD on 20 of 30 queries are higher than the individually created concepts. There is a 10% relative difference in mean average precision, which is a statistically significant improvement where p=0.01. These results verify our choice for concept detectors. Even though individually created concept detectors can be used, a MCD fits better in our formulation.

Table 3.11. Comparison of using individually created concepts (QIC-IND) and multiple-concept detector. Test set: MEDTEST; Training set: EK100. Results are provided in terms of average precision (in percent).

E. Id	E. Name	QIC-IND	QIC
E6	Birthday party	**14.1**	13.5
E7	Changing a vehicle tire	**27.8**	26.7
E8	Flash mob gathering	58.5	**61.7**
E9	Getting a vehicle unstuck	19.3	**30.2**
E10	Grooming an animal	12.8	**17.5**
E11	Making a sandwich	**16.2**	14.7
E12	Parade	27.9	**29.9**
E13	Parkour	24.4	**31.5**
E14	Repairing an appliance	34.8	**35.5**
E15	Working on a sewing project	16.0	**16.5**
E21	Attempting a bike trick	**7.6**	7.3
E22	Cleaning an appliance	**17.9**	15.9
E23	Dog show	26.6	**28.2**
E24	Giving directions to a location	3.2	**3.6**
E25	Marriage proposal	4.7	**5.1**
E26	Renovating a home	9.5	**11.1**
E27	Rock climbing	3.9	**4.9**
E28	Town hall meeting	24.6	**24.9**
E29	Winning a race without a vehicle	12.1	**22.8**
E30	Working on a metal crafts project	9.3	**13.5**
E31	Beekeeping	**21.7**	19.5
E32	Wedding shower	**11.8**	9.2
E33	Non-motorized vehicle repair	**27.8**	26.5
E34	Fixing musical instrument	**24.6**	19.6
E35	Horse riding competition	15.7	**25.3**
E36	Felling a tree	8.1	**10.2**
E37	Parking a vehicle	9.8	**16.7**
E38	Playing fetch	1.0	**2.6**
E39	Tailgating	17.2	**19.7**
E40	Tuning musical instrument	**18.7**	17.0
Avg.		17.6	**19.4**

3.4.3.2 Classifier Choice

So far, the results that have been reported in this chapter are obtained using a SVM classifier with an intersection kernel. It has been used commonly as a kernel for SVM in

computer vision applications ranging from image annotation to action recognition (Can and Manmatha 2013, Wang et al. 2009). It is a non-linear kernel which requires more memory and time for training and testing compared to linear kernels. Approximations are proposed to reduce the memory and time requirements in the testing phase (Maji et al. 2008). As an alternative to the intersection kernel, we also provide our results using linear kernels considering the faster training and classification speeds with significantly less memory requirements compared to a non-linear kernel.

SVM-rank learns a model based on pair-wise comparisons of the training examples in contrast to point-wise (or example-wise) comparisons in SVM classification methods. Joachims (2002) showed that a ranking problem can be formulated by maximizing the number of following inequalities satisfied:

$$\forall(v_i, v_j) : wx_i > wx_j \tag{3.6}$$

where x_i is a descriptor vector (e.g., representation of a video using concepts, H, in our case), w is the weight vector, v_i is a relevant example, and v_j is a non-relevant example. Non-negative slack variables are employed to solve the optimization problem (focusing on ROC-area) by transforming inequalities to equalities in a similar way to SVM classification. SVM-rank produces prediction scores which are then used for ranking purposes.

Here, we investigate whether similar retrieval accuracies can be achieved with faster ranking algorithms. For this purpose, we compare retrieval accuracies when retrieval models are trained using SVM-rank with a linear kernel.

In Table 3.12, we provide the retrieval accuracies using a SVM classifier with an intersection kernel (SVM IK.) and SVM-rank with a linear kernel (SVM-rank Linear K.). The results are calculated when the EK100 dataset is used for training and the MEDTEST dataset for testing on thirty events (E6-E40).

Table 3.12. Comparison of using SVM classifier with an intersection kernel (QIC SVM IK.) with SVM-rank with a linear kernel (QIC SVM-rank Linear K.) in video event detection with exemplars. Test set: MEDTEST; Training set: EK100. Results are provided in terms of average precision (in percent).

E. Id	QIC SVM-rank Linear K.	QIC SVM IK.
E6	11.8	**13.5**
E7	25.5	**26.7**
E8	59.7	**61.7**
E9	29.5	**30.2**
E10	13.2	**17.5**
E11	11.7	**14.7**
E12	**29.9**	**29.9**
E13	27.9	**31.5**
E14	34.3	**35.5**
E15	14.4	**16.5**
E21	4.4	**7.3**
E22	**19.1**	15.9
E23	27.1	**28.2**
E24	2.7	**3.6**
E25	4.2	**5.1**
E26	**11.7**	11.1
E27	**4.9**	**4.9**
E28	**25.0**	24.9
E29	18.4	**22.8**
E30	9.4	**13.5**
E31	19.2	**19.5**
E32	5.9	**9.2**
E33	22.3	**26.5**
E34	19.2	**19.6**
E35	19.9	**25.3**
E36	**11.4**	10.2
E37	**19.1**	16.7
E38	**2.6**	**2.6**
E39	19.3	**19.7**
E40	14.4	**17.0**
Avg.	17.9	**19.4**

Results obtained using a SVM classifier with an intersection kernel outperform the results obtained using SVM-rank with a linear kernel, as expected. However,

there are eight event queries where SVM-rank with a linear kernel does better than SVM IK. In Table 3.13, we provide the required times (in terms of seconds and averaged over thirty events) for training and testing a retrieval model using 1) SVM classifier with an intersection kernel (SVM IK.), and 2) SVM-rank with a linear kernel (SVM-rank Linear K.). Note that the fast version of the intersection kernel (Maji et al. 2008) is used in all our computations. The table shows that—considering the time requirements of a linear kernel and an intersection kernel— a linear kernel might be preferred over an intersection kernel when efficiency is most important with a small loss in performance. Note that testing requires more time than training in all of the cases since the MEDTEST dataset (used for testing) is five times larger than the EK100 set.

Table 3.13. Amount of time (in seconds) required to train and test an event detection model using SVM classifier with an intersection kernel (SVM IK.), and SVM-rank with a linear kernel (SVM-rank Linear K.). Test set: MEDTEST; Training set: EK100.

	SVM IK.	SVM-rank Linear K.
Train	245 s	27 s
Test	1,217 s	103 s

3.4.3.3 Parameter Selection

So far, we have provided results of our experiments. Now we detail our parameter choices in our algorithm. Apart from MEDTEST and EK100, a small development set (referred to as TINYSET) is also created for tuning parameters. There are two parameters that we have tuned in our settings: 1) the number of concepts considered at each frame, $k = 10$ and 2) the total number of concepts, $|C| = 1,000$. We also tune the regularization parameter of the SVM classifiers on the TINYSET dataset.

Tuning-$|C|$: We tune the $|C|$ parameter on the TINYSET. In Figure 3.8, we provide the retrieval accuracies for different values of concept vocabulary size $|C|$. The retrieval

accuracy increases when $|C|$ gets larger untill $|C|=1,000$. For $|C|>1,000$ it becomes steady and do not change or changes slightly. Therefore, we set $|C|$ to 1,000.
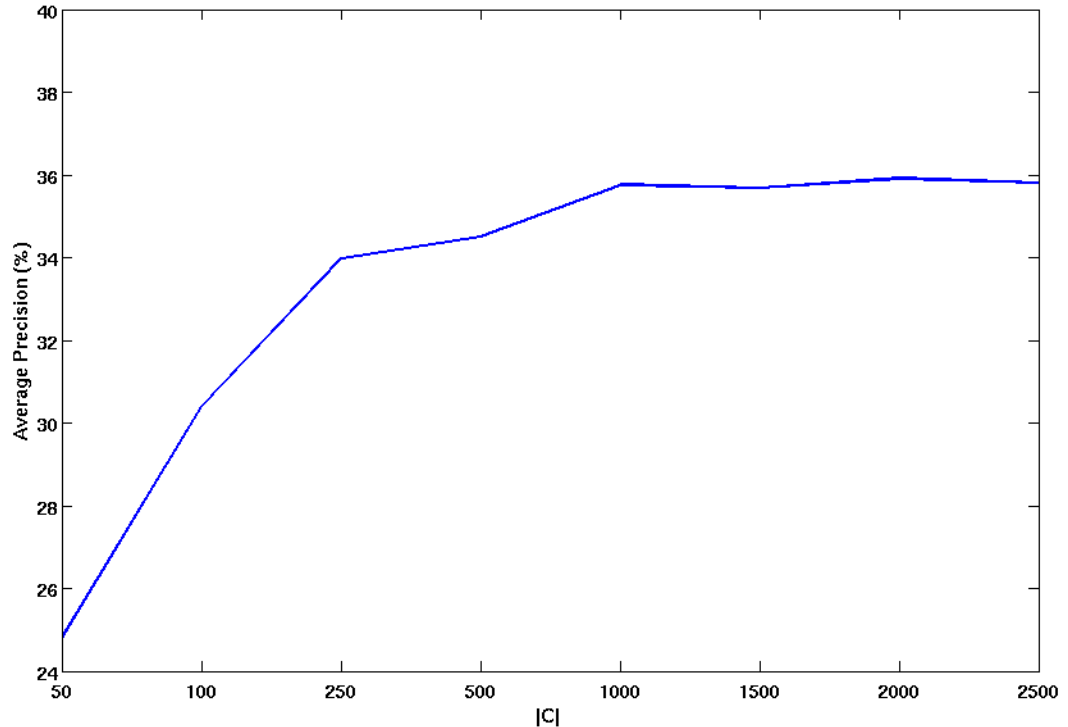
Figure 3.8. Illustration of tuning the concept vocabulary size, $|C|$, on the TINYSET dataset.

Tuning-k: We tune the $k = 10$ parameter on the TINYSET. In Figure 3.9, we illustrate the retrieval accuracies obtained for different values of k. While the retrieval accuracies for small values of (i.e., 5 to 50) are comparable to each other, they drastically drop for values larger than 50.

k **vs. # of non-zero values in** H In Figure 3.10, we provide the ratio of the number of concepts considered in H (i.e., the non-zero values in H) to the total number of available concepts for different values of k.

Figure 3.10 shows that when k becomes larger, the non-zero values in the video representation, H, decrease. The values provided in the figure are averaged over the
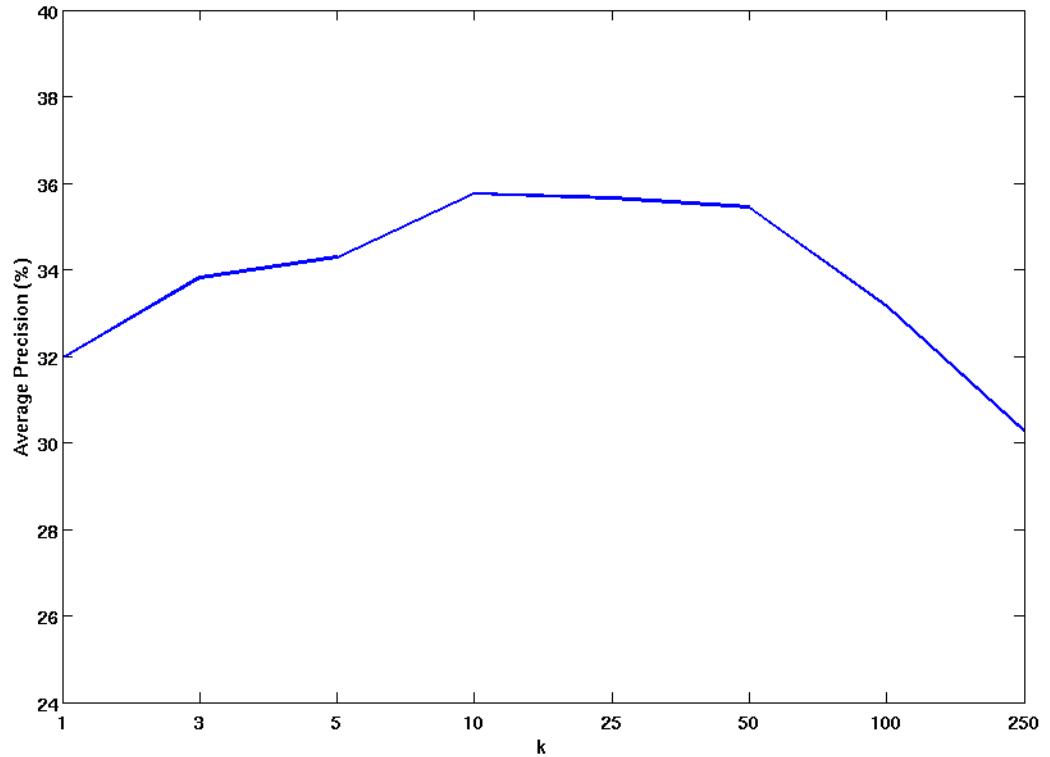
Figure 3.9. Illustration of tuning the k parameter on the TINYSET dataset.

videos in the TINYSET. Besides, for large values of k we need more time for training and testing an event detection model according to the figure.

Selection of a Concept Vocabulary In our experiments, we have selected a concept vocabulary of $|C| = 1,000$ and we use the same fixed set for all experiments in this chapter.

In order to gather the concept vocabulary we make use of a large dataset of images: ImageNet (Image-Net 2014). Using the same set in our experiments enables us to be sure that the changes in the retrieval accuracies is not due to different concepts. However, here we would like to investigate if using different concept vocabularies of the same size would yield significant changes on the retrieval accuracies.
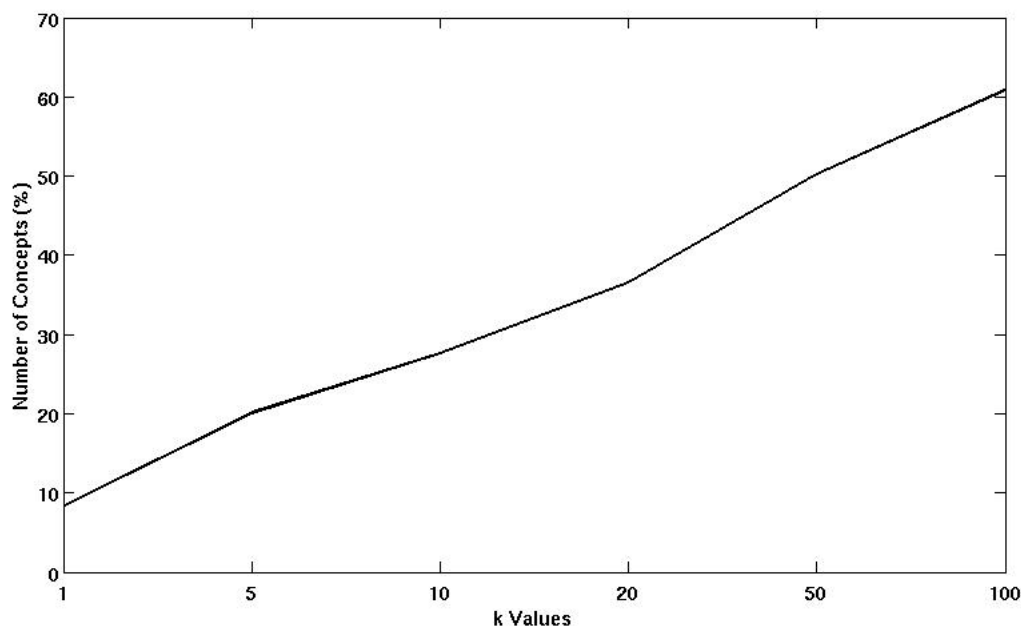
Figure 3.10. The ratio of the number of concepts considered in H to the total number of available concepts (non-zero values in H) for different values of k.

In order to show that the concept vocabulary is not a significant factor for the retrieval accuracies, we randomly selected ten more concept vocabularies having the same size. We then perform the same experiments that we have done in this chapter using different concept vocabularies. We run these experiments on the TINYSET as well.

In Figure 3.11, we provide the retrieval accuracies obtained using different concept vocabularies. In the experiments, we fix the concept vocabulary size to 1,000 and k to 10. The results show that using different concept vocabularies does not change the final accuracies. Even though retrieval accuracies of individual queries are different, their mean average precision are very close to each other. This finding supports that the retrieval accuracies we obtained in our experiments are not stemmed from using a "good" concept vocabulary.
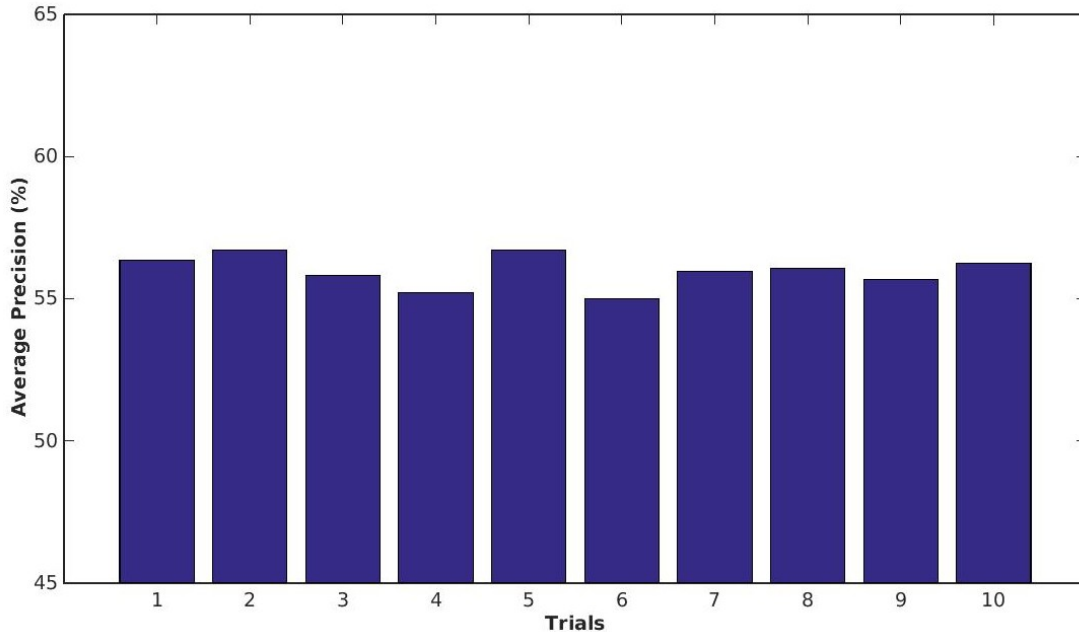
Figure 3.11. Illustration of the results using different concept vocabularies on the TINYSET dataset.

3.5 Summary of the chapter

Here we show that for event detection query-independent concepts provide as good retrievals as query-dependent concepts (when concepts are selected based on queries). This argues for using the same set of concepts for any query, which enables us to save time since we do not have to revisit the concept selection process multiple times.

In our experiments, we empirically show that query-independent concepts can be used as an alternative to using concepts selected for each query. The comparisons are performed within our settings as well as with a previous work using query-dependent concepts. The results supports our claim.

In addition to query-independent concepts, we provide a sparse way of representing videos using concepts. Existing techniques to represent videos using concepts is costly as they use all concepts. As an alternative, we provide a space-efficient representation

of videos. We show that our sparse representation is efficient as well as effective compared to using all concepts.

Finally, we discuss our design choices. For example, we discuss the reasons to choose using multiple concept detector (MCD) as an alternative to individual concept detectors. Further, we also analyze the efficiency and effectiveness of using a linear kernel instead of an intersection kernel (i.e., a non-linear kernel) within the context of VED-ex. At last, we explain our parameter selection experiments to show how we select the parameters (e.g., k and $|C|$) used in our experiments.

CHAPTER 4

INCORPORATING ONE-EXEMPLAR MODELS
INTO VIDEO EVENT DETECTION

Recent approaches (Chen et al. 2014, Jiang et al. 2012; 2013, Ma et al. 2013) in video event detection with exemplars (VED-ex) use a large number of example videos for a query. The retrieval models trained on a large collection of example videos most likely provide better retrieval accuracies than the models created using a few exemplars. However, collecting a large number of exemplars is often either difficult or unrealistic. Here, we present a method that incorporates multiple one-exemplar models into video event detection aiming at improving retrieval accuracies when there are few exemplars available.

A single retrieval model is usually trained using all available example videos in VED-ex. While it might be sufficient for the queries where example videos are visually similar, it is problematic for learning stronger characteristics of the queries when example videos are visually different from each other. For example, exemplars of the "repairing an appliance" query contain events such as repairing an oven, repairing a refrigerator, and repairing a washing machine. In Figure 4.1, we show sample frames of "repairing an appliance" videos. Conventionally, a single global model is trained using these examples, implicitly assuming that these exemplars are all visually similar to each other. The resulting retrieval model then becomes weak and overly-generic (Malisiewicz et al. 2011). This is a major problem especially when we have very few exemplars. In contrast to creating one single global model, we create multiple one-exemplars models, each of which is trained using one exemplar. We

present a method that incorporates these models into video event detection aiming at improving retrieval accuracies by addressing the issue discussed above.
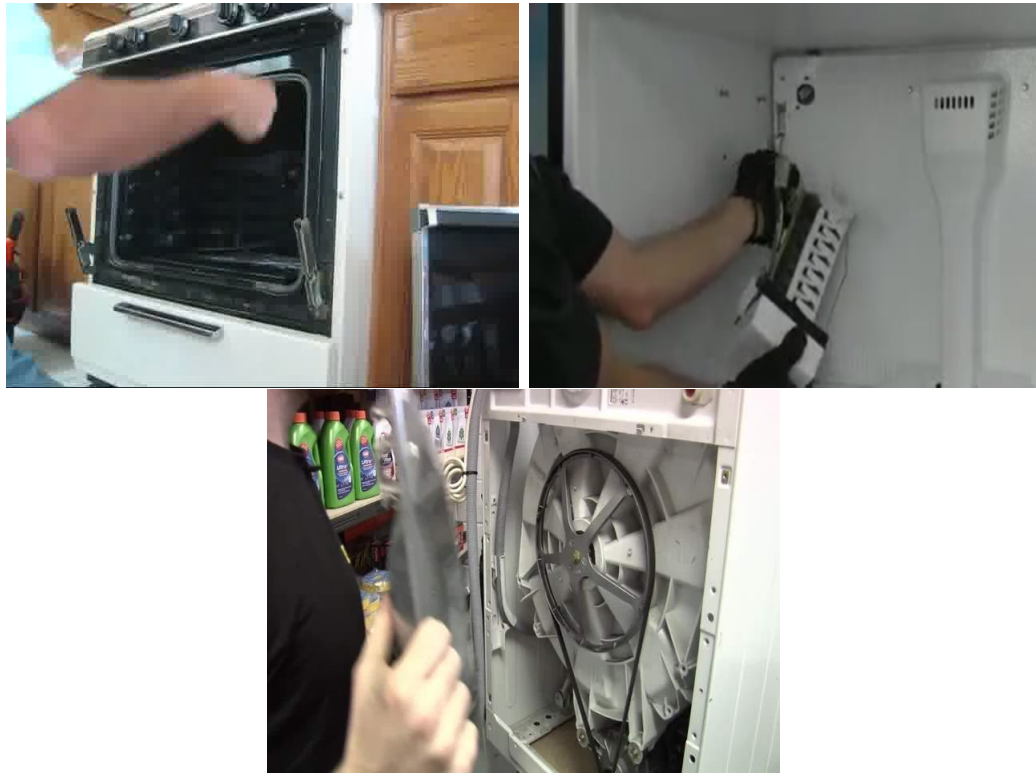
Figure 4.1. Sample frames of example videos of the "repairing an appliance" query.

Our one-exemplar models are created for each exemplar and are specific to their exemplars. For example, three one-exemplar models are created for the "repairing an appliance" query (see Figure 4.1), in contrast to creating only one single global model.

The global model works well when the example videos are visually similar. Our one-exemplar models work better when the example videos are visually different (see Figure 4.1). It is very difficult and expensive to estimate the variance of the example videos in advance. Therefore, incorporating one-exemplar models into the global model enables us to deal with this difficulty within the context of video event detection with very few exemplars (VED-ex$_{few}$).

Similar approaches to one-exemplar models have been considered to solve various problems ranging from object detection to learning-to-rank (Can et al. 2014, Malisiewicz et al. 2011). Our approach is related in spirit to Malisiewicz et al. (2011). They create a number of object detectors each of which consists of a positive example and a number of negative ones. They conclude that the results are very promising. In another work, McCallum et al. (2000) point out that the reduction in the computational cost obtained by dividing the data into overlapping subsets—called canopies—in the context of efficient clustering can be performed without any performance loss.

In the following, we first detail our one-exemplar models within the context of VED-ex$_{few}$. We then present the experimental results and discussion which are followed by analysis of the extension of our approach to tackle the query specific relevance feedback problem in the learning-to-rank framework.

4.1   One-Exemplar Models

Consider a set of video clips $T = \{V_1, V_2, ...V_\ell\}$ associated with event query E, and $T = T_p \cup T_n$ where T_p consists of example videos of E and T_n consists of clips non-relevant to E. As there are $|T_p| = k$ exemplars, we create k one-exemplar models (M_i), each of which considers one exemplar $V_i \in T_p$ and multiple negative examples $\forall j, V_j \in T_n$. Consider an example to illustrate one-exemplar models and the global model. Assume that there are eight videos: three positive exemplars, T_p, and the rest are not-relevant videos, T_n, (illustrated in Figure 4.2). We first train a global model (M) using all the videos (as illustrated in the left side of the figure). We then train three one-exemplar models $(M_1, M_2, $ and $M_3)$ using one exemplar per model (as illustrated on the right side of the figure).

After creating one-exemplar models (M_i) and the global model (M), we run the test videos against these models for ranking purposes. Figure 4.3 illustrates the testing
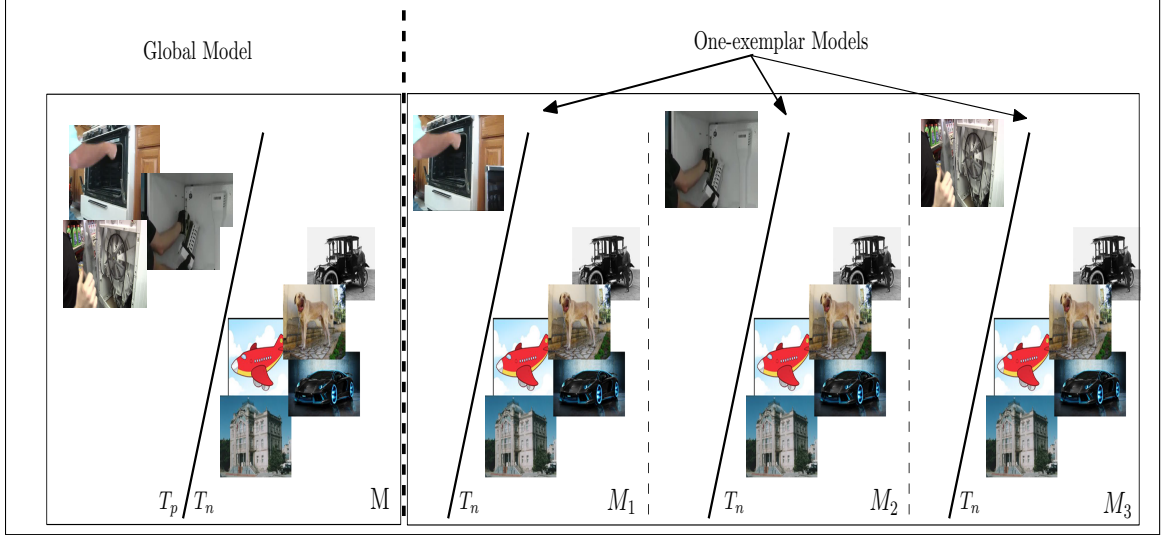
Figure 4.2. Illustration of one-exemplar models (M_1, M_2, M_3) and the global model (M).

phase of our approach. Test videos—$V_j \in Z$—are run against M as well as the models M_i, and their combination provides the final ranking of the videos.

Each test video gets a score for an event query E indicating the likelihood of that video being relevant to the event query E. Given that there are k one-exemplar models and a global model, there will be $k+1$ outputs—one from each model. (One-exemplar models are trained independent of each other and independent of the global model.) To improve the retrievals by incorporating these models, we jointly consider a global model and one-exemplar models. In our approach, we aim to estimate a probability, $P(E|V_j)$, of a test video V_j being relevant to a query, E, considering models M and M_i. Our sample space is a collection of disjoint one-exemplar models (e.g., M_i) and a global model (e.g., M). We estimate $P(E|V_j)$ considering multiple one-exemplar models and one global model, from the law of total probability, and assuming the disjointness of the models as follows:
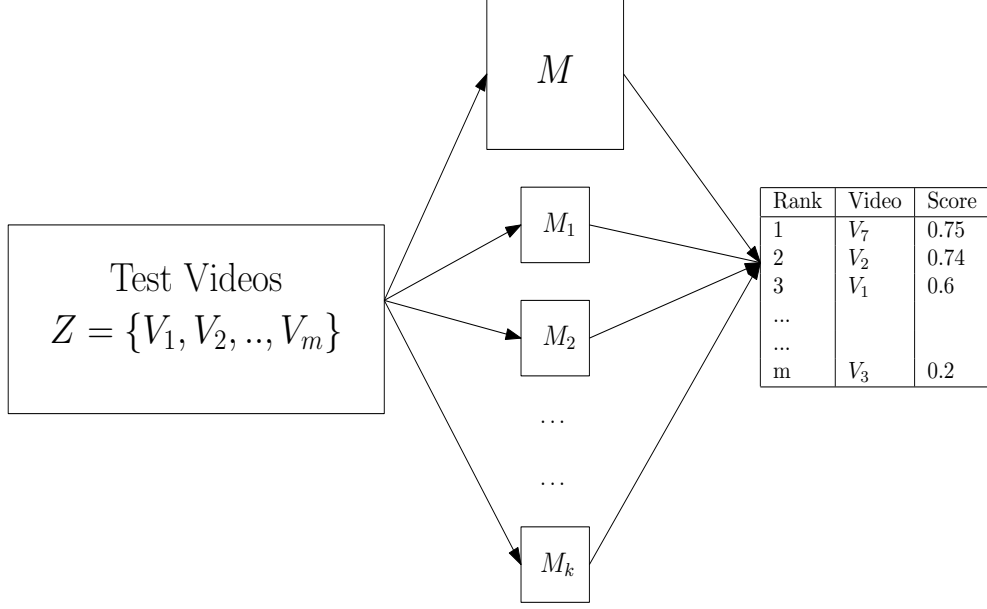
Figure 4.3. Illustration of testing videos against the standard model M and exemplar-based models M_i.

$$P(E|V_j) \quad = \quad \sum_i P(E_M, E_{M_i}|V_j) \tag{4.1}$$

assuming E_M is conditionally independent of E_{M_i} given V_j (i.e., $E_M \perp E_{M_i} \mid V_j$)

$$P(E|V_j) \quad = \quad \sum_i^k P(E_M|V_j)P(E_{M_i}|V_j) \tag{4.2}$$

where $P(E_M|V_j)$ is the probability of a video V_j being relevant to a query E calculated using the model M, and similarly $P(E_{M_i}|V_j)$ indicates the probability of a video V_j being relevant to E calculated using the model M_i. We use this final estimation of a test video V_j being relevant to an event E for ranking purposes. A posterior relevance probability is approximated for each V_j as proposed by Platt et al. (1999). As proposed in Platt et al. (1999)'s work, a sigmoid function is used for this purpose (Chang and Lin 2011, Platt et al. 1999). Note that we take logarithms not to deal with very small

numbers during the calculation (a sigmoid function does not approximate to zero probability by definition).

4.2   Experiments and Discussion

So far, we have explained our one-exemplar approach within the context of VED-ex$_{few}$. Here, we provide empirical evidence to show that our approach improves retrieval accuracy when there are very few exemplars available.

In the previous chapter (Chapter 3), we made use of query-independent concepts with a space-efficient representation. Here, we first evaluate our approach using the same approach and settings except the training set. Fortunately, NIST provides a collection of videos where the event queries have only ten example videos (Jonathan Fiscus 2014). This collection aligns well with our evaluation purposes. The collection is referred to as EK10 where each of thirty event queries has ten example videos and approximately 5,000 non-relevant videos. We also consider the cases when there are less than ten example videos available to see if our approach also works with much fewer exemplars. We also analyze the robustness of our approach in terms of different descriptors. To do so, we also evaluate our approach on multiple descriptors used in video event detection.

4.2.1   Experiments comparing with and without one-exemplar models

In Section 3.3, we introduced the event queries and the test collection (i.e., MEDTEST) used for evaluation purposes. Here, we use the same settings except the training set is different as we use a dataset with 10 examples rather than 100 for evaluation purposes. In Table 4.1, we compare our approach, one-exemplar models combined with the global model are incorporated into video event detection (w/ OX), with the case where only a global model is considered and our one-exemplar models are not

involved in the detection (w/o OX). Bold face indicates a higher average precision over its counterparts.

Table 4.1. Experimental results of the case when one-exemplar models are incorporated into video event detection (w/ OX) with the case when they are not involved in the detection (w/o OX). Test set: MEDTEST; Training set: EK10. Results are provided in terms of average precision (in percent).

E. Id and E. Name	w/o OX	w/ OX
E6 Birthday Party	2.8	**3.3**
E7 Changing a vehicle tire	4.8	**5.8**
E8 Flash mob gathering	33.0	**34.1**
E9 Getting a vehicle unstuck	5.5	**6.4**
E10 Grooming an animal	3.8	**4.2**
E11 Making a sandwich	6.2	**6.4**
E12 Parade	11.0	**12.0**
E13 Parkour	16.5	**16.7**
E14 Repairing an appliance	9.1	**14.0**
E15 Working on a sewing project	**4.1**	3.7
E21 Attempting a bike trick	1.1	**2.4**
E22 Cleaning an appliance	1.9	1.9
E23 Dog show	1.8	1.8
E24 Giving directions to a location	0.5	0.5
E25 Marriage proposal	0.3	**0.4**
E26 Renovating a home	0.4	0.4
E27 Rock climbing	**6.1**	5.4
E28 Town hall meeting	8.0	**12.1**
E29 Winning a race without a vehicle	2.4	**2.9**
E30 Working on a metal crafts project	4.0	**6.8**
E31 Beekeeping	3.6	3.6
E32 Wedding shower	3.3	**3.6**
E33 Non-motorized vehicle repair	**7.6**	6.6
E34 Fixing musical instrument	1.3	**1.6**
E35 Horse riding competition	**11.0**	10.1
E36 Felling a tree	3.7	**3.9**
E37 Parking a vehicle	2.9	**3.2**
E38 Playing fetch	0.6	**0.7**
E39 Tailgating	**13.0**	12.9
E40 Tuning musical instrument	3.0	**4.2**
Avg.	5.8	**6.5**

According to Table 4.1, our approach enables us to improve the retrieval accuracies to 6.5% from 5.8% within the context of VED-ex$_{few}$. It is a statistically significant improvement (p<0.008, calculated using paired t-test). Further, there are 20 events where our approach outperforms its counterpart. The average relative improvements on these events are approximately 30%.

One-exemplar models works well when the example videos are visually different and a single global model works well in the other case. The experimental results also validate this claim. For example, our approach boosts retrieval accuracies for several events such as "repairing an appliance" where there are video clips related to repairing an oven, a refrigerator, and a washing machine. Even the videos related to "repairing a refrigerator" are visually different than each other: one is about repairing the ice-maker in the freezer, whereas the other one is about fixing a refrigerator shelf. Similarly, example videos of the "working on a metal craft project" query are all different from each other (see Figure 4.4). However, for the "non-motorized vehicle repair" event, a single global model—without one-exemplar models— works better than our approach. When we analyze the example videos for this event, we find out that nine out of ten videos are related to fixing bikes. Only one video is different from the others and it is about fixing a ski. In Figure 4.5, we provide sample video frames from these example videos. The first nine (top three row) frames are about repairing a bike and the last one (the fourth row) is about fixing a ski.

Motivated by the sample frames in Figure 4.5, we perform an experiment where we remove the outlier example (i.e., the last frame, fourth row, in the figure) from the training set and train the retrieval model (a global model) with the remaining nine exemplars. The retrieval accuracy grows to 8.9% from 7.6%, that is obtained using the retrieval model trained using ten exemplars. This shows that having a training set consisting of visually similar examples enables us to have a stronger retrieval model.
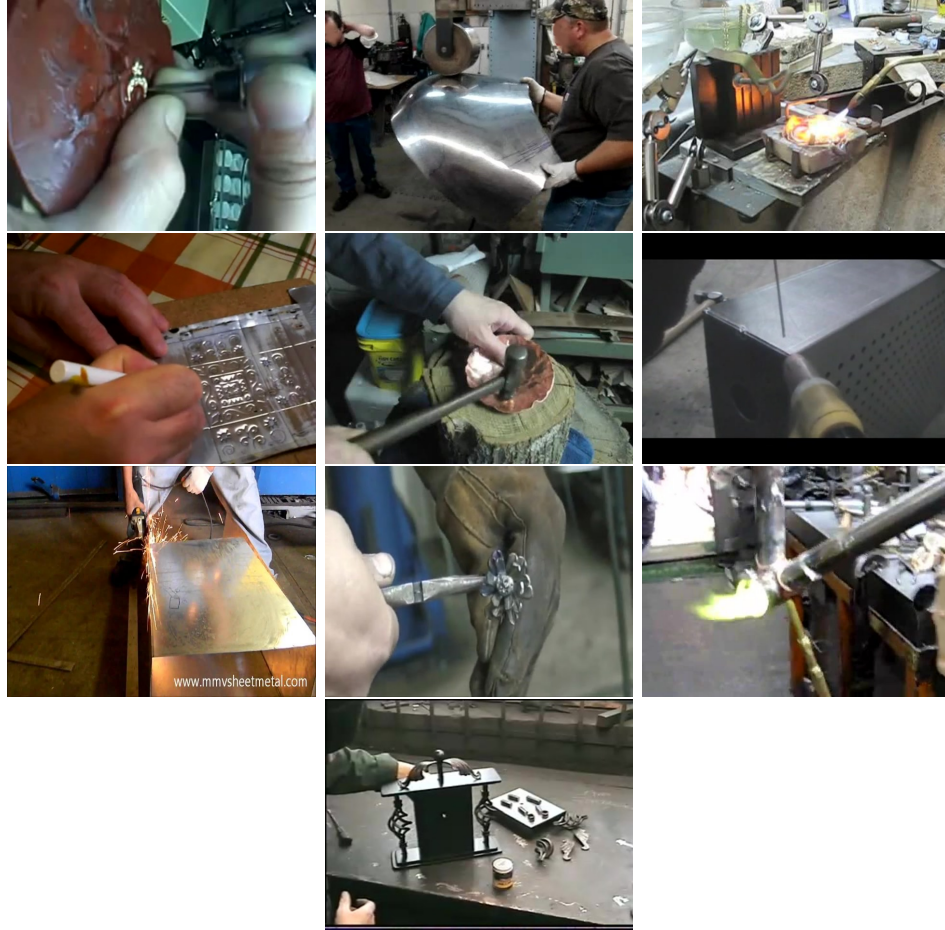
Figure 4.4. Sample video frames extracted from example "working on a metal craft project" videos.

In Figure 4.6, we illustrate pair-wise similarities between example videos of the "E30: working on a metal crafts project" (on the left V_1^{30}, V_{10}^{30}) and "E33: non-motorized vehicle repair" (on the right $V_1^{33}, ..., V_{10}^{33}$). In the Figure, darker colors show higher similarity between pairs. For E30, the colors are lighter compared to E33, which indicates that the pairwise similarities of exemplars of E33 are higher than the ones of E30. As the similarities are higher for the E33 case, one global retrieval model works better than one-exemplar models. Similarly, for the E30 case, one-exemplar models work better since the exemplars are dissimilar to each other. Note that, in our illustrations in the figure, we embed the weights trained for the retrieval model on top
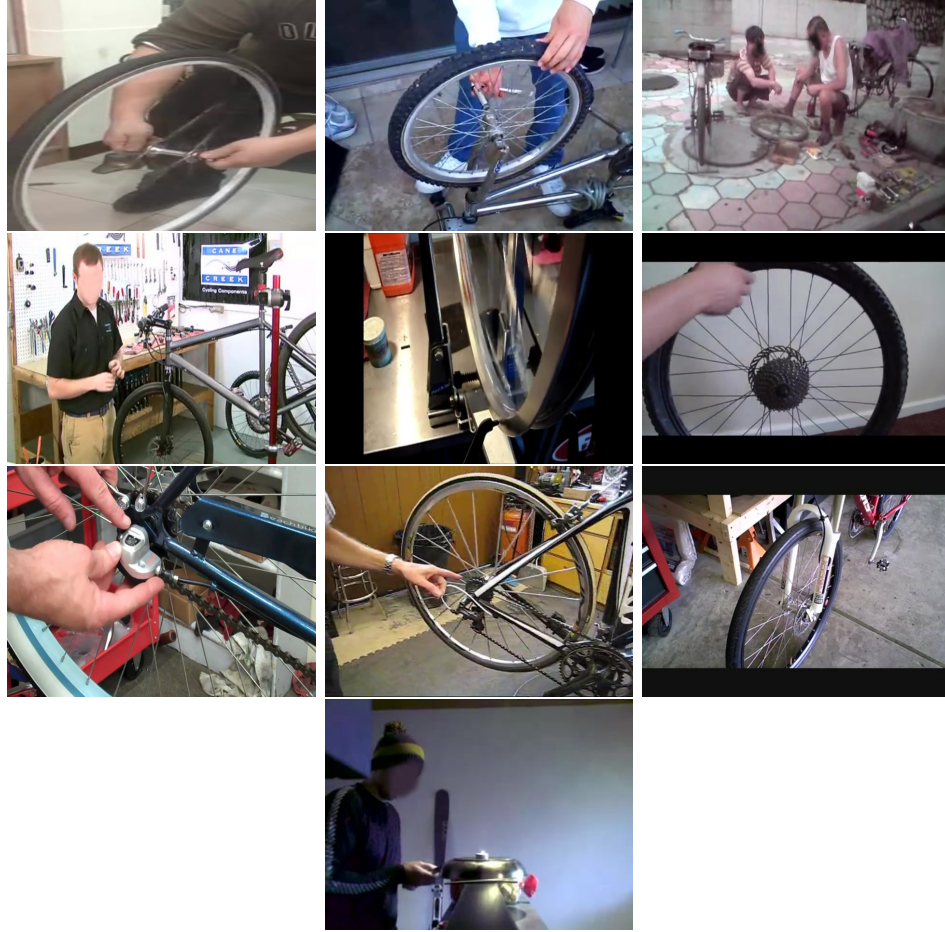
Figure 4.5. Sample video frames extracted from example "non-motorized vehicle repair" videos.

of the descriptors to visualize the similarities. It is rather difficult to identify visually different examples automatically from the rest only considering their descriptors and without any learning stage. This suggests us to use stronger descriptors as well as compiling a subset of concepts based on the query in the zero-shot video event detection (VED-zero) case. In VED-zero there is no training phase and the problem gets more difficult. In the next Section (Section 5), we detail how to address this issue.
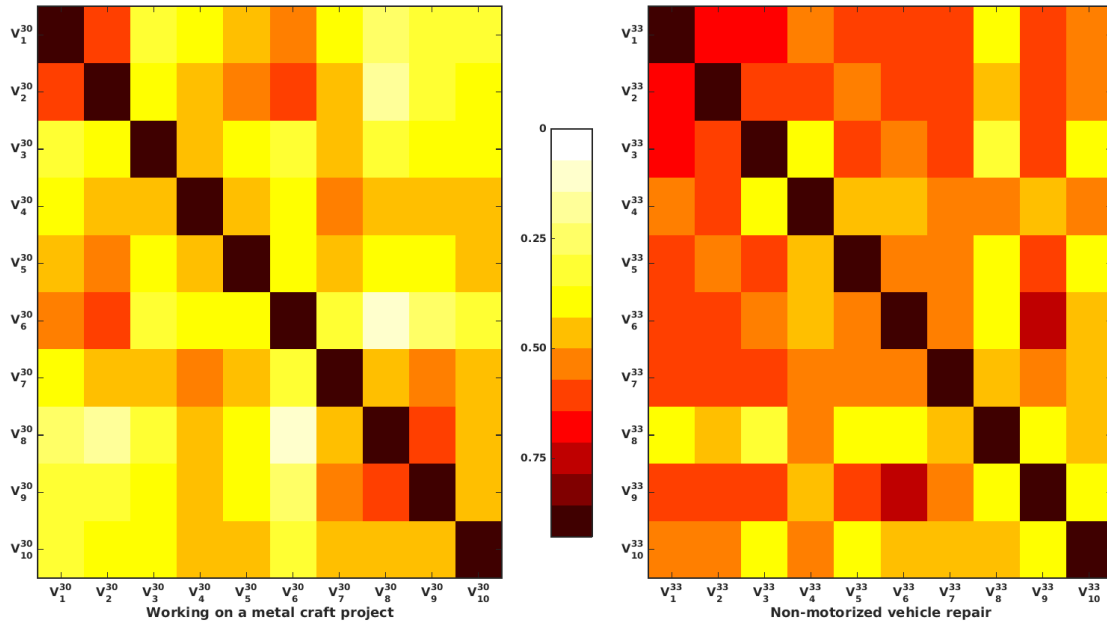
Figure 4.6. Illisturation of the similarities between example videos of the "working on a metal crafts project" (on the left) and "non-motorized vehicle repair" (on the right).

4.2.1.1 One-exemplar models considering different number of positive examples

So far, we have focused on the case where we have ten example videos. We might end up with fewer example videos. Here, we investigate this case and analyze our approach when we have fewer example videos: five exemplars (EK5) and two exemplars (EK2) per event query. To do so, we randomly select two and five example videos for each event query. We repeat this process multiple times to avoid biased selections. We then report the average values calculated over these multiple iterations.

In Table 4.2, we provide the retrieval accuracies of our approach (w/ OX) when there are fewer example videos (e.g., five:EK5 and two:EK2) as well as the case when one-exemplar models are not involved in the detection (w/o OX). According to the results, our approach enables us to improve the retrieval accuracies even when there

Table 4.2. Experimental results of the case when one-exemplar models are incorporated into video event detection (w/ OX) with the case when they are not involved in the detection (w/o OX). Test set: MEDTEST; Training set: EK5 (five example videos per event query) and EK2 (two example videos per event query). Results are provided in terms of average precision (in percent).

E. Id	EK5 w/oOX	EK5 w/OX	EK2 w/oOX	EK2 w/OX
E6	2.8	**3.1**	2.3	**2.5**
E7	3.3	**4.3**	1.5	**1.9**
E8	24.7	**25.8**	13.1	**13.2**
E9	4.0	**4.7**	1.2	**2.4**
E10	1.5	**2.7**	**1.8**	1.6
E11	**4.2**	3.9	1.5	**1.8**
E12	9.1	**9.5**	5.0	**5.7**
E13	10.1	**11.4**	4.3	**4.4**
E14	4.8	**6.8**	1.4	**1.6**
E15	1.3	**1.6**	1.6	1.6
E21	0.8	**0.9**	0.1	**0.2**
E22	1.2	1.2	0.5	**0.7**
E23	3.5	**3.9**	1.5	**1.6**
E24	0.3	0.3	0.2	**0.3**
E25	0.3	0.3	0.2	0.2
E26	0.3	**0.5**	0.3	0.3
E27	3.8	**3.9**	2.0	**2.4**
E28	3.8	**5.0**	2.6	**3.0**
E29	**3.1**	2.6	2.3	**2.6**
E30	3.3	**3.6**	1.4	**2.0**
E31	4.5	**4.8**	1.6	1.6
E32	2.0	**2.4**	1.3	**2.1**
E33	4.4	4.3	**2.2**	2.1
E34	1.0	**1.1**	**0.5**	0.4
E35	3.7	**3.8**	1.6	**2.0**
E36	2.2	**3.1**	2.1	**2.2**
E37	1.8	**1.9**	0.7	**0.9**
E38	0.2	**0.9**	0.2	0.2
E39	6.4	**7.1**	2.3	**2.4**
E40	3.0	3.0	**1.0**	0.8
Avg.	3.8	**4.3**	1.9	**2.2**

are five and two example videos per query. This shows that our approach is robust to different number of example videos.

For the EK5 case, we are able to improve the retrieval accuracies to 4.3% from 3.8%, which is a statistically significant improvement (p<0.00008). There are 24 event queries where our approach outperforms its counterpart. The average relative improvement on those events is approximately 30%. For the EK2 case, we are able to improve the retrieval accuracies to 2.2% from 1.9%, which is also a statistically significant improvement (p<0.0001). There are 21 event queries where our approach does better than the case where one-exemplar models are not involved in the detection. The average relative improvements on those events is similar to the EK5 case. However, we obtain higher statistical significance in the EK5 case compared to the EK2 case. In the EK2 case, the models are created using only two exemplars rather than five example videos.

Recalling our running example, for the "non-motorized vehicle repair" event, the difference in retrieval accuracies shrink to almost zero considering the case where one-exemplar models are involved in detection (w/ OX) or not (w/o OX). However, this situation was different when we have more exemplars (i.e., 7.6% w/o OX and 6.6% w/ OX in EK10). For this event, the exemplars are visually similar to each other and splitting the exemplars into smaller subsets causes weaker retrieval models. On the other hand, for the "working on a metal craft project" event, incorporating one-exemplar models into the detection enables us to improve retrieval accuracies for the EK10 case as well as for the EK5 and EK2 cases since the exemplars are visually different from each other.

Up to that point, we have discussed one-exemplar models when we have few exemplars including ten, five, and two. We also create as many one-exemplar models as the number of exemplars. For example, if we have *five* example videos in our training set, then we create *five* one-exemplar models. When we have more example

videos in our training set, we might consider a different approach. There are two reasons we might follow a different approach: efficiency and effectiveness. Creating few one-exemplar models and running videos against them are not costly steps. However, many one-exemplar models might require additional considerations for efficiency purposes such as using multiple cores. One exemplar models work well when exemplars are different than each other. When we have many example videos, the chances of observing similar videos are higher. For example, the chance of having two "repairing an oven" is larger when we have many example videos compared to having few exemplars. Splitting similar examples is usually not a desired outcome. Perhaps, a method that considers grouping example videos and creating few-exemplar models might be a solution. However, our preliminary experiments on an approach that aims at grouping similar exemplars in a training set show that grouping similar exemplars without any supervision is quite challenging. As an alternative, we can select a subset of exemplars and create one-exemplar models using only this subset. We also observe that selecting this subset without any supervision/feedback is challenging. However, in another work of ours (Can et al. 2014), we show that we can use explicit relevance feedback to select such a subset of exemplars. In the same work, we also observe that optimal size of a subset (based on effectiveness) would be half of the total number of positive examples in the training set.

4.2.2 Robustness of one-exemplar models to different descriptors

So far, we have focused on our query-independent concepts and space-efficient representation in our experiments to valuate one-exemplar models. Here we investigate the case where we have different descriptors used to represent videos.

A broad range of visual cues can be employed to describe content of videos in addition to our query-independent concepts. The visual cues might be similar QIC where detectors are built on top of other descriptors. Alternatively, the descriptors

that are used to built concept detectors might be used directly as well (e.g., Red, Green, and Blue distributions of an image/video). Edge orientation histograms help to describe the gradient distributions of the regions in an image/video which can later be used in the problems such as for tracking, object identification, and object recognition Lowe (2004). Similarly, trajectory based features focus on the magnitude and shape of the motion in a video.

We focus on three more descriptors to show that our approach is not only good for query-independent concepts (QIC) that we have introduced in the previous chapter but for other descriptors as well. The first descriptor we focus on is the histogram of oriented gradients calculated on densely extracted trajectories, HOG,(Dalal and Triggs 2005). The second descriptor is motion boundary histograms calculated on densely extracted trajectories, MBH, (Dalal et al. 2006) as well. The last one is called Overfeat and calculated using convolution neural networks that are learned on images (Sermanet et al. 2013). The first two descriptors exploit motion information in videos. The last one as well as QIC are image-based concepts. Overfeat uses other descriptors (i.e., deep learning features) to built detectors as QIC. In this way, we evaluate our approach on descriptors structurally and motivationally different.

In Table 4.3, we provide the results of the experiments to evaluate our approach on different descriptors. According to the table, our approach is robust to multiple descriptors, where we obtain statistically significant improvements on HOG, MBH, and Overfeat ($p < 0.03$, $p < 0.008$, and $p < 0.002$ respectively).

When one-exemplar models are incorporated into the detection, we are able to improve the retrieval accuracies to 5.4% from 4.7% with HOG. For 23 out of 30 events our approach provides higher retrieval accuracies. Further, the average relative improvements on these events are approximately 20%. We observe similar improvements considering the MBH descriptor. There are 20 event queries, where our approach enables us to improve retrieval accuracies. Considering the Overfeat

Table 4.3. Experimental results of different descriptors (HOG, MBH, Overfeat) for the case when one-exemplar models are incorporated into video event detection (w/ OX) with the case when they are not involved in the detection (w/o OX). Test set: MEDTEST; Training set: EK10. Results are provided in terms of average precision (in percent).

E. Id	HOG		MBH		Overfeat	
	w/o OX	w/ OX	w/o OX	w/ OX	w/o OX	w/ OX
E6	3.4	**3.7**	4.0	**5.1**	4.4	**5.0**
E7	3.2	**3.3**	1.2	**1.3**	26.2	**28.2**
E8	18.9	**20.0**	31.0	**31.5**	35.2	**37.0**
E9	3.7	**3.8**	3.6	**6.5**	26.7	**29.3**
E10	3.0	**4.2**	5.0	**5.3**	8.4	**9.7**
E11	**4.3**	3.9	**6.2**	6.1	5.5	**5.8**
E12	12.5	**22.2**	**22.5**	22.3	14.2	**14.7**
E13	14.8	**15.1**	**50.2**	50.2	17.9	**19.9**
E14	5.7	**8.7**	4.0	**4.7**	17.3	**19.0**
E15	2.3	2.3	7.4	**8.0**	2.5	**2.7**
E21	2.5	**3.6**	1.2	**2.3**	**10.1**	6.7
E22	0.5	**0.7**	1.7	**2.4**	1.5	**2.1**
E23	1.6	1.6	4.0	**4.2**	6.8	**8.3**
E24	0.4	0.4	**0.5**	0.4	0.5	0.5
E25	0.1	**0.2**	0.4	0.4	0.1	**0.2**
E26	1.8	**3.4**	**5.3**	5.2	5.6	**6.7**
E27	1.5	**1.6**	10.1	10.1	6.4	**6.9**
E28	8.2	**8.7**	12.3	**12.6**	2.6	**2.8**
E29	2.3	**2.5**	17.8	**19.1**	5.5	**6.6**
E30	7.4	**8.0**	1.7	**2.0**	**9.1**	8.0
E31	2.1	**2.3**	**5.2**	3.3	25.5	**26.4**
E32	2.9	**3.0**	1.4	**1.5**	1.7	**1.8**
E33	**4.4**	4.3	1.2	1.2	9.4	**9.6**
E34	0.6	0.6	0.7	**0.8**	7.3	**10.6**
E35	16.2	**16.3**	21.3	**23.0**	7.5	**7.8**
E36	1.9	**2.1**	4.9	**5.9**	7.1	**7.4**
E37	4.4	**4.5**	8.0	**8.8**	8.5	**10.1**
E38	0.4	**0.5**	0.9	0.9	0.5	**0.7**
E39	7.6	**8.5**	3.0	**3.8**	7.1	**7.8**
E40	**1.4**	1.3	8.2	**8.5**	**2.8**	2.6
Avg.	4.7	**5.4**	8.2	**8.6**	9.5	**10.2**

descriptor, we observe a relative improvement of approximately 20% on 26 event queries where our approach outperforms its counterpart.

In addition to individual descriptors, we also evaluate our method on their fusion. Fusing a number of descriptors is a common practice in video event detection Cheng et al. (2012), Jiang et al. (2010), Liu et al. (2013a), Natarajan et al. (2012). We blend the resulting ranked list from different descriptors for a final list. We consider the HOG, MBH, and Overfeat descriptors in addition to our query-independent concepts. In Table 4.4, we provide the results of blending four descriptors. Consdering the blend of these four descriptors, we observe similar improvements in the retrieval. There are 23 out of 30 events where incorporating one-example models into the global detection model improves the retrievals. The improvements are statistically significant as well (p<0.003).

4.3   Summary of the chapter

Here, we show that our one-exemplar models enables us to improve retrieval accuracies statistically significantly when we have few exemplars. One-exemplar models are specific to its example and video event detection models generalize over the example videos. Our approach leverage from both approaches.

Experimental results show that our approach is robust to the number of example videos. We investigate our approach when there are five and two example videos per query. Similar to the case where we have ten example videos, our approach provides statistically significant improvements compared to the case when one-example models are not considered in the detection.

In our detailed analysis of our experiments, we observe that one-exemplar models work better when example videos of a query are dissimilar to each other. On the other hand, a global retrieval model works better when example videos of a query are related to each other. For example, for the "working on a metal craft project"

Table 4.4. Experimental results of the case when four descriptors are blended. Test set: MEDTEST; Training set: EK10. Results are provided in terms of average precision (in percent).

E. Id & Name	Fusion of HOG, MBH, Overfeat, and QIC	
	w/o OX	w/ OX
E6 Birthday party	5.9	**7.2**
E7 Changing a vehicle tire	18.3	**20.6**
E8 Flash mob gathering	44.8	**46.9**
E9 Getting a vehicle unstuck	26.9	**30.4**
E10 Grooming an animal	10	**11.8**
E11 Making a sandwich	6.4	**8.2**
E12 Parade	18.5	**19.4**
E13 Parkour	**47.2**	44.8
E14 Repairing an appliance	18.7	**24.6**
E15 Working on a sewing project	6.7	**7.7**
E21 Attempting a bike trick	11.1	**12.4**
E22 Cleaning an appliance	**3.3**	3.0
E23 Dog show	5.7	**7.4**
E24 Giving directions to a location	1.0	**1.1**
E25 Marriage proposal	0.3	0.3
E26 Renovating a home	6.3	**6.6**
E27 Rock climbing	**11.8**	11.7
E28 Town hall meeting	12.1	**12.3**
E29 Winning a race without a vehicle	7.4	**9.7**
E30 Working on a metal crafts project	8.7	**11.4**
E31 Beekeeping	**21.2**	18.2
E32 Wedding shower	3.4	**3.6**
E33 Non-motorized vehicle repair	12.8	**12.7**
E34 Fixing musical instrument	2.0	**3.2**
E35 Horse riding competition	**22.5**	18.5
E36 Felling a tree	7.7	**10.8**
E37 Parking a vehicle	10.5	**12.8**
E38 Playing fetch	1.0	1.0
E39 Tailgating	18.4	**19.4**
E40 Tuning musical instrument	6.1	**6.9**
Avg.	12.6	**13.5**

query, one-exemplar models work better compared to a single global model since the examples of this query are dissimilar to each other.

We also investigate the robustness of our approach to different type of descriptors as well as their fusion. To do so, we evaluate our approach using multiple descriptors. The results show that our approach is robust to multiple descriptors and their fusion within the context of video event detection with very few exemplars.

CHAPTER 5

ZERO-SHOT VIDEO EVENT DETECTION

So far, we have focused on the video event detection with exemplars case (VED-ex and VED-ex_{few}). There is yet another case of video event detection where an event query consists of only textual description and no example videos. This is a more realistic case of video event detection compared to the previous cases. However, as no exemplars are associated with event queries, we cannot benefit from machine learning algorithms to learn a retrieval model while determining the relevance of videos to an event query, which makes the problem more challenging. Here, we tackle this more realistic and also more challenging video event detection problem: *zero-shot video event detection (VED-zero).*

Events are complex activities localized in time and space (Jiang et al. 2013, NIST 2012, Over et al. 2010). Most of the recent zero-shot video event detection studies (Chen et al. 2014, Jiang et al. 2014, Liu et al. 2013b, Wu et al. 2014) consider the bag-of-concepts approach that assumes that concepts in videos are independent of each other and that order does not matter which is, we believe, an insufficient assumption. Videos are sequence of frames and the ordering of frames is not utilized in the bag-of-concepts approach. Assume that two concepts are detected in a video and consider these two scenarios: 1) one of these concepts is detected at the beginning of video and the other is detected at the end, 2) they are detected at consecutive frames. We cannot differentiate these two scenarios when we use bag-of-concepts. However, concepts detected at consecutive frames would get greater share of evidence than those detected far away from each other (similar ideas have been used in different

problems such as positional language model (Lv and Zhai 2009), and proximity-based named entity retrieval (Petkova and Croft 2007).

We believe that considering concepts and their relationships might enable us to have better evidence to recognize videos relevant to an event and, hence, we tackle the problem of exploiting concepts dependencies in videos. The idea is that if important concepts show dependencies in a video, that would be stronger evidence for relevance of this video to an event query.

Our dependency work uses a MRF based approach (Metzler and Croft 2005), a widely accepted algorithm in the information retrieval community. Feng and Manmatha (2008) use a similar approach to tackle the image retrieval problem. In our work, we focus on three dependency assumptions: (1) full independence, (2) spatial dependence, and (3) temporal dependence.

In the following, we first detail our approach to tackle the zero-shot video event detection problem. Next, we present the experimental results and discussion.

5.1 Our Approach

The main task in VED-zero is similar to the VED-ex cases: ranking videos according to their relevance to an event query. The major difference is that in VED-zero there are no example videos associated with a query; therefore, no retrieval model can be trained. Unlike the previous cases, we cannot leverage the same set of concepts for every query. We need to modify our pipeline for this problem. Below, we start by discussing concept detection and concept detector output scores indexing issues. We then detail concept selection and ranking.

As in VED-ex, we first detect concepts in videos. This process consists of running concept detectors against videos. We then need to store the detector output scores in a way so that we can efficiently run queries. In Chapter 3, we showed that we can use the same set of and query-independent concepts for retrieval. This holds when example

videos are associated with a query. Leveraging example videos enables us to learn a retrieval model specific to its query using even the same set of query-independent concepts. However, we cannot apply the same idea for VED-zero since we rely on only the detector outputs and no learning is involved in this case. Therefore, for VED-zero we identify a number of concepts that are expected to be relevant for an event query (Section 5.1.1).

As we select a different set of concepts for each query, we need a different structure to store the detector output scores than in VED-ex. Therefore, we create concept indexes, similar to the inverted indexes in information retrieval. In this way, we can efficiently search the videos for the selected concepts. Further, searching different concepts (e.g., for a different query) does not require traversing the whole collection of videos but, is rather, limited to traversing the disk space of the selected concepts. Fortunately, concept detection and indexing can be performed offline.

When concepts are detected in videos and indexing the output scores is completed, we can run a query against the videos in our collection. In Figure 5.1, we summarize the steps to run an event query against videos. For an event query (i.e., textual description of a query), we first identify a number of concepts that are expected to be relevant to this query. Then, we use these selected concepts to rank videos according to their relevance to the given query. In what follows we explain these steps in detail.
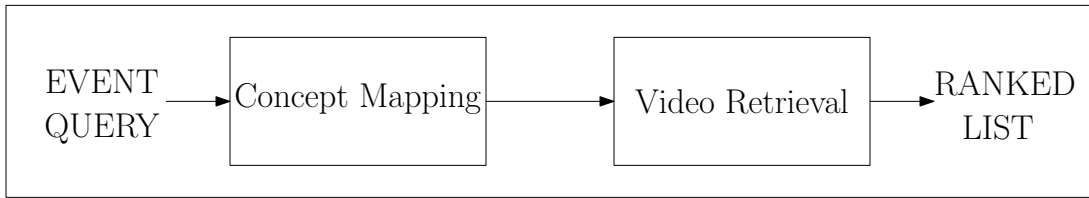
Figure 5.1. Illustration of the steps to run an event query against videos.

5.1.1 Concept Mapping

Information need can be expressed in infinitely many different ways. Creating a concept detector to cover each query that a person can use to express his information need is infeasible as it would require crafting an extensive amount of resources. Therefore, we have a limited vocabulary of concepts and we map an event query to a subset of the concepts in our vocabulary that are expected to be more related to this query than others.

As we do not have one-to-one matching between the textual description of a query and the set of concepts, we consider a mapping that focuses on identifying concepts that are expected to be relevant to a query. Let $C = \{c_1, c_2, ..., c_n\}$ be the set of concepts in our vocabulary and consider a mapping from an event query E to a subset of concepts C':

$$E \rightarrow C' \tag{5.1}$$

where $C' \subseteq C$. We follow a similar approach when we select concepts based on queries in Chapter 3. Here we apply the same approach to identify concepts that are expected to be relevant to a query. We first extract noun phrases from the textual description of a query (an example textual description of a query is illustrated in Figure 5.2). Then we run these noun phrases against the descriptions of the concepts. The resulting list is a list of concepts ranked according to their relevance to an event query (see Table 5.1). We then focus on the top m concepts (C') in this list while retrieving videos. The concept selection process is performed automatically without any human involvement. In addition to our work, Chen et al. (2014) and Dalton et al. (2013) also employ a similar approach in their work. Chen et al. (2014) focus on Flickr tags to identify possible matches with query descriptions. They also filter some candidates such as "economy" since it seems to be an abstract concept which is difficult to illustrate in images. Dalton et al. (2013) use a sequential dependence model by adjusting different weights to different fields (e.g., definition and explication Figure 5.2) in the query

description. Against this automatic approach, alternative approaches including using a training set to identify a set of concepts for each query (Habibian et al. 2014), have been used in VED-zero. Habibian et al. (2013) determine the concepts to be used for a query by mining a training set.

Event name: Felling a tree

Definition: One or more people fell a tree.

Explication: Felling is the process of cutting down an individual tree transforming its position from vertical to horizontal. Felling a tree can be done by hand or with a motorized machine. If done by hand, it usually involves a tool such as a saw, chainsaw, or axe. A tree-felling machine, known as a feller buncher, can also be used. Felling is part of the logging process, but can also be done to single trees in non-logging contexts. possibly climbing the tree or accessing upper parts of the tree from a cherry-picker bucket and then cutting branches from the tree before felling it, possibly cutting a horizontal wedge from the tree's trunk to cause the tree to fall in a desired direction, cutting horizontally through the trunk of the tree with saw(s) or ax(es), using wedges or rope(s) to prevent the tree from falling in some particular direction (such as onto a house).

Evidential description:
scene: outdoors, with one or more trees

objects/people: persons in work clothing, hand saws or chain saws, axes, metal wedges, tree-felling machines

activities: sawing, chopping, operating tree felling machine

audio: chainsaw motor, sounds of chopping, sawing, tree falling

---

Figure 5.2. An example textual description of a query.

Table 5.1. Example of concepts retrieved in top ranks for the query provided in Figure 5.2

Concepts
cutting a tree
chopping a tree
tree falling
sawing a tree
putting down an object on the floor
person sawing
people marching on street
cutting fabric
falling
machine sawing

5.1.2 Video Retrieval

After selecting a subset of concepts that are expected to be relevant to a query, we now can run this query against videos in our collection by focusing on only these concepts.

Our dependency work uses a MRF based retrieval model and makes three dependency assumptions: (1) full independence, (2) spatial dependence, and (3) temporal dependence. The idea is that when important concepts show dependencies in a video, there is a stronger evidence for relevance of this video to a query. In Figure 5.3, we illustrate these dependency assumptions (v_t is a video frame at time t, v_{t+1} is a video frame at time $t+1$, c_i, and c_j are concepts). The left-most figure illustrates the independence assumption. The graph in the middle shows the spatial dependency between concepts ($\{(v_t, c_i, c_j), (v_{t+1}, c_i, c_j)\}$), where the presence of two concepts in the same video frame is treated as important. The right-most graph illustrates the temporal dependency between concepts($\{(v_t, v_{t+1}, c_i, c_j)\}$, where having concepts occur in consecutive frames is treated as important.
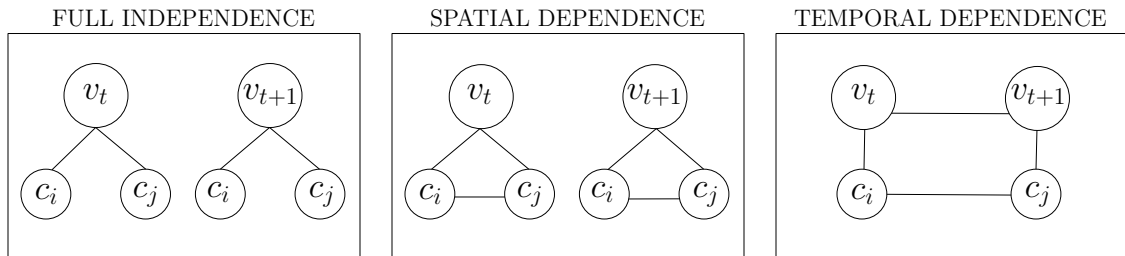
Figure 5.3. Illustration of dependency assumptions in our MRF based retrieval model.

In order to rank videos for a given query, we calculate the probability of a video given a query $P(V|E)$ which, by Bayes Rule can be formulated as follows:

$$P(V|E) = \frac{P(E|V)P(V)}{P(E)} \tag{5.2}$$

where $P(E)$ is the same for all videos and can be ignored for ranking purposes. The prior probability is commonly assumed as uniform across all videos so we ignore it as well. Then the equation becomes:

$$P(V|E) \stackrel{rank}{=} P(E|V) \tag{5.3}$$

When we consider the spatial and temporal dependence of concepts, we need to modify our function. We apply a similar MRF framework that is provided for term dependencies by Metzler and Croft (2005) and Feng and Manmatha (2008) for image retrieval. Then, we modify the posterior for ranking purposes as follows:

$$P_\Lambda(V|E) \stackrel{rank}{=} \sum_{\ell \in L(G)} log(\psi(\ell; \Lambda)) \tag{5.4}$$

where $L(G)$ is the set of cliques in graph G, and each non-negative potential function (ψ) over cliques is parametrized by Λ. Potential functions are often parametrized:

$$\psi(\ell; \Lambda) = e^{\lambda_\ell f(\ell)} \tag{5.5}$$

where $f(\ell)$ is a feature function over cliques and λ_ℓ is the coefficient to $f(\ell)$. Then the ranking function becomes:

$$P_\Lambda(V|E) \stackrel{rank}{=} \sum_{\ell \in L(G)} \lambda_\ell f(\ell) \tag{5.6}$$

The potential functions for different variants (illustrated in Figure 5.3): full independence, spatial dependence, and temporal dependence might be defined as illustrated in Figure 5.4.

FULL INDEPENDENCE   SPATIAL DEPENDENCE   TEMPORAL DEPENDENCE

$$\log \psi_I = \lambda_I \log P(c_i|V) \qquad log\psi_S = \lambda_S \log P_S(c_i, c_j|V) \qquad \log \psi_T = \lambda_T \log P_T(c_i, c_j|V)$$
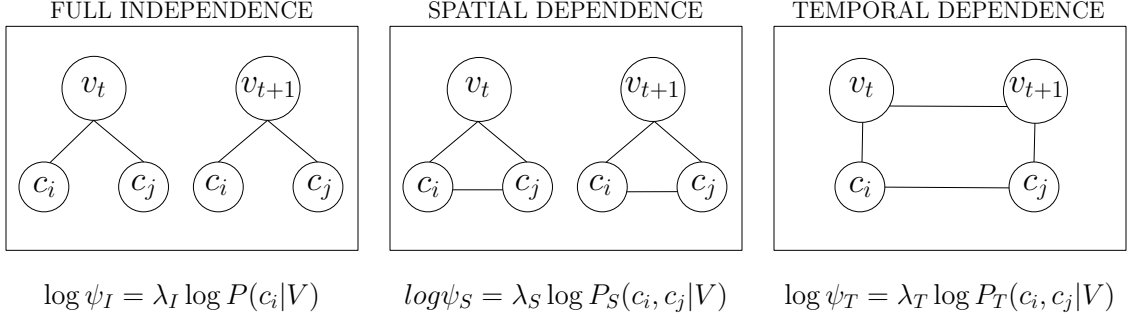
Figure 5.4. Illustration of dependency assumptions in our MRF based retrieval model.

where ψ_I is the potential function for the full independence case and in this case the cliques are defined to be between a concept (e.g., c_i) and a video frame (e.g., v_t). In the spatial dependence case, cliques are defined to be between two concepts (e.g., c_i and c_j) and a video frame (e.g., v_t) and ψ_S is the potential function for this case. The potential function for the temporal dependence case is ψ_T. In this case, the cliques are defined to be between two concepts (e.g., c_i and c_j) and consecutive frames (e.g., v_t and v_{t+1}). Replacing the potential functions our formulation becomes:

$$\sum_{c_i \in C'} \lambda_I \, log \, P(c_i|V) + \sum_{c_i, c_j \in C'} \lambda_S \, log \, P_S(c_i, c_j|V) + \sum_{c_i, c_j \in C'} \lambda_T \, log \, P_T(c_i, c_j, t|V) \quad (5.7)$$

where $C' \subseteq C$ is the selected concepts for this particular event query, E, and the coefficients (λ) can be set to $\lambda_I = \lambda_S = \lambda_T$ or alternatively they can be tuned as well.

For a given query, we employ the ranking function above to rank videos in terms of their relevance. The probabilities for the full independence might be estimated (considering a smoothed language model) as follows:

$$log \, P_I(c_i|V) = \log \left[(1 - \alpha) \frac{f(c_i, V)}{\sum_j f(c_j, V)} + \alpha \frac{f(c_i, W)}{\sum_j f(c_j, W)} \right] \quad (5.8)$$

where $f(c_i, V)$ is the frequency of c_i in video V (it can also be interpreted as the total number of frames in video V where c_i concept is present) and $f(c_i, W)$ is the

total frequency of c_i in video collection W. We leverage linear smoothing (Jelinek and Mercer 1980) in our estimations not to encounter zero probabilities. α acts as the smoothing parameter, which can be tuned and λ_I is the coefficient for the full independence case.

As we mention earlier in this chapter, the concept detector output scores can be interpreted differently. In this work, we use these output scores in two different ways. In the following, we explain them in detail.

5.1.3 Estimation of Probabilities

We focus on two different ways of interpreting the output scores of concept detectors. Concept detectors measure the likelihood of observing a concept in videos. The output scores can be unbounded real numbers. However, for simplicity we map them between zero and one, where an output score close to one indicates a higher confidence than an output score close to zero.

The first interpretation we consider in our work is the presence/absence of concepts (referred to as Boolean concepts) in videos. In Chapter 3, we showed that using a number of concepts that have the highest output score at each frame enables us to have successful retrieval. Here, we employ the same idea to convert the output scores into presence/absence of concepts in videos. We consider concepts having the k highest output scores as present in a frame, and assume the rest are not detected in this particular frame.

When we use Boolean concepts, we can directly use the probability estimations used in (Metzler and Croft 2005) since the Boolean concepts are countable. Recall the example in Table 3.1 where the concept output scores are as follows:

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
v_1	0.13	0.20	0.04	0.04	0.07	0.12	0.19	0.03	0.02	0.15
v_2	0.08	0.20	0.07	0.05	0.06	0.10	0.06	0.07	0.21	0.11
v_3	0.17	0.11	0.10	0.04	0.09	0.02	0.12	0.07	0.17	0.12
v_4	0.15	0.18	0.07	0.07	0.05	0.07	0.21	0.12	0.01	0.07
v_5	0.06	0.10	0.04	0.08	0.16	0.07	0.15	0.09	0.16	0.08

and when we assume that the concepts having the three highest output score as present in this video then it becomes:

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
v_1	0	1	0	0	0	0	1	0	0	1
v_2	0	1	0	0	0	0	0	0	1	1
v_3	1	0	0	0	0	0	0	0	1	1
v_4	1	1	0	0	0	0	1	0	0	0
v_5	0	0	0	0	1	0	1	0	1	0

When we have Boolean concepts as in the example above, we can directly use the frequencies of concepts while estimating the probabilities. In the example, concept pairs as spatial dependence includes: $\{(c_1, c_2), (c_1, c_7), (c_2, c_7)\}$ for v_4 and $\{(c_5, c_7),$ $(c_5, c_9), (c_7, c_9)\}$ for v_5. The concept pairs considering temporal dependence of concepts in v_4 and v_5 are: $\{(c_1^{v_4}, c_5^{v_5}), (c_1^{v_4}, c_7^{v_5}), (c_1^{v_4}, c_9^{v_5}), (c_2^{v_4}, c_5^{v_5}), (c_2^{v_4}, c_7^{v_5}), (c_2^{v_4}, c_9^{v_5}), (c_7^{v_4}, c_5^{v_5})$ $(c_7^{v_4}, c_9^{v_5})\}$.

Analogous to the full independence case, the spatial dependence term in Equation 5.7, considering a linearly smoothed language model, can be estimated as follows:

$$\log P_S(c_i, c_j|V) = \log[(1 - \alpha)\frac{f_S(c_i, c_j, V)}{\sum_{a,b} f_S(c_a, c_b, V)} + \alpha\frac{f_S(c_i, c_j, W)}{\sum_{a,b} f_S(c_a, c_b, W)}] \qquad (5.9)$$

where $f_S(c_i, c_j, V)$ is the total number of frames in which c_i and c_j occur together in video V, $f_s(c_i, c_j, W)$ is the total number of frames in which c_i and c_j co-occur in video

collection W. $\sum_{a,b} f_S(c_a, c_b, V)$ is the total number of frames including co-occurrences of all concepts pairs (c_a, c_b) where $\forall_{a,b}\, c_a \in C', c_b \in C'(a \neq b)$ in video V.

The temporal dependence case is mostly similar to the spatial dependence case; but we consider the concepts occurring in a window of *consecutive* frames rather than concepts occurring in the *same* frame. The temporal dependence term in Equation 5.7 can be, considering a linearly smoothed language model, estimated as follows:

$$\log P_T(c_i, c_j, t|V) = \log[(1 - \alpha)\frac{f_T(c_i, c_j, V)}{\sum_{a,b} f_T(c_a, c_b, V)} + \alpha\frac{f_T(c_i, c_j, W)}{\sum_{a,b} f_T(c_a, c_b, W)}] \quad (5.10)$$

where $f_T(c_i, c_j, W)$ is the total number of times in which c_i and c_j co-occur in a window of consecutive frames in video V and $f_T(c_i, c_j, W)$ is the total number of times where c_i and c_j occur together in window of consecutive frames.

Using Boolean concepts enables us to calculate frequencies without any modification to the common language model estimations. Boolean concepts are strong and ignoring a number of concepts (e.g., concepts having low confidence are not assumed to be present) might help reducing noisy concept detection. However, while reducing the noise we might also lose some valuable information. As an alternative to Boolean concepts, we also employ the output scores directly in our formulations to address this issue.

When we use the output scores of concept detectors directly (referred to as scored concepts) we cannot use the same equations that we have used for the Boolean concepts. In order to address this issue, we modify Equations 5.8, 5.9, and 5.10. For the full independence case, we are able to use the output scores as frequencies and consider maximum likelihood estimation as we did for the Boolean concepts. The estimation for the full independence model then becomes:

$$(1 - \alpha)\frac{f'(c_i, V)}{\sum_j f'(c_j, V)} + \alpha\frac{f'(c_i, W)}{\sum_j f'(c_j, W)} \quad (5.11)$$

102

where $f'(c_i, V)$ is the sum of c_i's output scores in video V and defined to be: $f'(c_i, V) = \sum_{t=1}^{|V|} \Phi^i(t)$ in which $|V|$ is the total number of frames and $\Phi^i(t)$ is output score of c_i at frame t.

Similarly, $f'(c_i, W)$ is the sum of c_i's output scores in video collection W.

The main challenge is to calculate the co-occurrence counts using scored concepts. The output scores usually do not follow a specific distribution. This might stem from their unconstrained nature. Therefore, estimating co-occurrences based on formal distributions becomes difficult.

In order to address this challenge, we focus on an approximation method that estimates the co-occurrence counts of the concepts in videos. Our approximation method aims at maximizing the co-occurrence values as much as possible considering the maximum output score of concepts efficiently. Similar ad-hoc approximations are shown to be successful and efficient by Srikanth and Srihari (2002) as well as Lin and Och (2004). The estimation below is used for the spatial dependence case:

$$(1 - \alpha)\frac{f'_S(c_i, c_j, V)}{\sum_{a,b} f'_S(c_a, c_b, V)} + \alpha\frac{f'_S(c_i, c_j, W)}{\sum_{a,b} f'_S(c_a, c_b, W)} \tag{5.12}$$

where $f'_S(c_i, c_j, V)$ is the approximation of co-occurrences of concepts c_i and c_j in video V and defined to be: $f'_S(c_i, c_j, V) = \sum_{t=1}^{|V|} \max(\Phi^i(t), \Phi^j(t))$ in which $\Phi^i(t)$ and $\Phi^j(t)$ are output scores of c_i and c_j at frame t. $f'_S(c_i, c_j, W)$ is calculated as follows: $\sum_{v \in W} f'_S(c_i, c_j, v)$. We consider only the cases where the following constraints are satisfied: $\Phi^i(t) > 0$ and $\Phi^j(t) > 0$. The temporal dependence case is analogous to the spatial case except and f'_T is defined to be as follows: $f'_T(c_i, c_j, V) = \sum_{t=1}^{|V|-1} \max(\Phi^i(t), \Phi^j(t+1))$. Similar to the spatial case, we consider only the cases there the following constraints are satisfied: $\Phi^i(t) > 0$ and $\Phi^j(t+1) > 0$.

So far, we have explained our approach to tackle the zero-shot video event detection problem starting from concept detection / indexing to ranking models. Next, we detail

103

the experimental environment. We then provide experimental results and discuss them.

5.2 Experimental Setup

In order to evaluate our MRF based retrieval, we again focus on MEDTEST, a collection of videos with thirty event queries (E6-E40), as our test collection. Here, we do not need any training data; therefore, the sets EK10 and EK100 are not involved in the evaluation of our approach for zero-shot video event detection.

The concept detectors used in this chapter are different than those used for VED-ex experiments. Previously our concept detectors were created on top of static images with densely sample SIFT features (DSIFT). As the zero-shot event detection case is much more challenging due to non-existence of exemplars, here we employ stronger detectors. For this purpose, we focus on the detectors created using three descriptors: DSIFT, histogram of oriented gradients (HOG), and motion boundary histograms (MBH). The last two descriptors are calculated on top of densely extracted trajectories and they exploit the motion information in video. Therefore, they are calculated on video clips (i.e., a small part of a video consisting of a sequence of frames: 270 frames in our case) formed using a sliding window approach where the step size is 90. In total there are 676 concepts used in this work. Ten of them are selected for each query.

For the Boolean concepts, we focus on the concepts having the top ten highest output scores in each frame. The smoothing parameter is set to 0.1 and is fixed for all experiments. The co-efficients for the dependence terms in our formulation is set to 1.0) and not changed for all experiments. For the temporal dependence case we not only look for two consecutive frames but we also consider three more consecutive frames as it is a common practice in information retrieval (Metzler and Croft 2005).

The evaluation of the ranked lists obtained by running our retrieval models is performed by using the relevance judgement released by NIST and the `trec_eval` [1] tool. Average precision (percent) and mean average precision (percent) are used as evaluation measures in our experiments.

5.3   Experiments and Discussion

We compare our MRF based retrieval model with a baseline where a bag-of-concepts approach is used. In the baseline, concepts are assumed to be independent from each other. We focus on the Boolean concepts in the first part of our evaluation. We then provide the experimental results focusing on scored concepts. Finally, we provide the results of the case where we combine these two approaches.

In the last part of the evaluation of our approach, we compare our results with previously reported results (Chen et al. 2014, Habibian et al. 2014, Mazloom et al. 2013a, Rastegari et al. 2013). Recent studies on zero-shot video event detection show promising improvements. Chen et al. (2014) provide a mean average precision of 2.2% (on twenty events E6-E30). There are also other results reported such as 3.5% by Rastegari et al. (2013) and 4.2% by Mazloom et al. (2013a), which are higher than Chen et al. (2014)'s results. Recently Habibian et al. (2014) improve these results and set the bar to 6.4%.

5.3.1   MRF based Retrieval Model

In Table 5.2, we provide the results of our MRF based model as well as the baseline: a bag-of-concepts approach with independence of concepts assumption.

For *Boolean concepts*, our MRF based model improves the baseline from 5.5% to 6.2%, which is a statistically significant improvement where $p < 0.0005$ calculated using a paired t-test. This shows that our model not only provides improvements on

---

[1]`http://trec.nist.gov/trec_eval`

Table 5.2. Results of our MRF based model as well as the baseline. Test set: MEDTEST

Id & Name	Boolean concepts		Scored concepts	
	Baseline	Our Approach	Baseline	Our Approach
E6 Birthday party	13.4	**13.6**	**10.3**	9.8
E7 Changing a vehicle tire	10.2	**10.4**	**9.9**	9.5
E8 Flash mob gathering	9.7	**10.2**	10.3	**15.1**
E9 Getting a vehicle unstuck	2.7	**3.1**	3.7	**4.4**
E10 Grooming an animal	5.8	**6.1**	4.1	**5.3**
E11 Making a sandwich	8.6	**9.5**	**9.7**	9.2
E12 Parade	27.6	27.6	**28.2**	25.5
E13 Parkour	19.5	**22.4**	14.3	**26.0**
E14 Repairing an appliance	6.0	**6.9**	5.4	**5.5**
E15 Working on a sewing project	7.5	**8.5**	9.6	**12.3**
E21 Attempting a bike trick	1.3	**2.0**	1.2	**2.0**
E22 Cleaning an appliance	0.7	0.7	0.5	**0.6**
E23 Dog show	0.3	**0.4**	0.5	**0.7**
E24 Giving directions to a location	1.7	**2.6**	1.4	1.4
E25 Marriage proposal	0.8	**1.9**	1.7	**2.9**
E26 Renovating a home	**0.8**	0.7	**1.3**	1.2
E27 Rock climbing	4.4	**4.8**	5.5	**6.9**
E28 Town hall meeting	0.3	0.3	0.5	0.5
E29 Winning race without a vehicle	1.4	**1.5**	8.9	**9.1**
E30 Working on metal crafts project	9.1	**10.2**	3.5	**7.7**
E31 Beekeeping	5.4	5.4	3.8	**5.7**
E32 Wedding shower	1.8	**5.1**	2.0	**2.7**
E33 Non-motorized vehicle repair	1.8	**2.0**	1.8	**2.6**
E34 Fixing musical instrument	**2.6**	1.9	1.0	**1.3**
E35 Horse riding competition	8.1	**11.0**	14.1	**15.4**
E36 Felling a tree	4.9	**5.6**	6.3	**15.5**
E37 Parking a vehicle	2.7	**2.9**	1.8	**3.1**
E38 Playing fetch	4.7	4.7	0.4	0.4
E39 Tailgating	0.6	0.6	1.0	**1.2**
E40 Tuning musical instrument	1.8	**2.3**	5.2	5.2
Average	5.5	**6.2**	5.6	**7.0**

only one event but we obtain improvements on many events. We observe that our MRF based model outperforms the baseline on 22 out of 30 queries.

For *Scored concepts*, our MRF based retrieval model enables us to boost the retrieval accuracy from 5.6% to 7%, which is a statistically significant improvement ($p < 0.008$). On 21 out of 30 events, our MRF based retrieval model outperforms the baseline, where the average improvements on those events are approximately 45%.

When we compare these two approaches, we observe that the improvements with *Boolean concepts* over the baseline is larger compared to the improvements with *Scored concepts* over the baseline based on retrieval accuracies. Dependencies of concepts would be a very strong evidence for relevance and when we exploit these dependencies using scored concepts we obtain 25% relative improvements over the baseline. However, this figure is approximately 15% for the *Boolean concepts* case. It might stem from the idea that some of the concepts (and so are their dependencies) are ignored due to their weak signals.

Further, there a number of queries where either Boolean concepts or the other approach provides higher retrieval accuracies. This discrepancy between Boolean and scored concepts makes sense upon further investigation. We believe that this is due to the amount of noise happening during concept detection (e.g., selected concepts might not be detected with high confidence). We can rephrase the same idea as: the Boolean concepts approach usually provides better retrieval when the concept detectors work really well, because we trim most of the noise and only the concepts having a higher confidence remain. For example, for the "E24:Giving directions to a location" query, Boolean concepts outperforms scored concepts. When we analyze the top ranked videos for this event, we observe that the concepts selected for this query align with the concepts detected in these videos and therefore they are strong signals. For this query, the presence of especially the "talking" and "pointing directions" concepts helps pull the related videos to top ranks.

The selected concepts for the "E36:Felling a tree" query are not good matches with the concepts detected with high confidence in videos. In this case, scored concepts

outperforms the Boolean concepts. The main advantage of this approach is that it still works even if the detectors are very noisy. Here, our latter approach is favorable.

5.3.2   MRF based Retrieval Model with Blend of Two Approaches

We know motivationally and observe empirically that *Boolean concepts* and *scored concepts* outperform each other for different cases. In other words, both methods have advantages when they are used in our retrieval model. We can leverage the advantages of both methods by considering them together in a hybrid approach. Therefore, we also focus on their fusion, where we create an ensemble of the results of Boolean and scored concepts by taking the average of the final estimations of both methods.

In Table 5.3, we provide the experimental results of our MRF based retrieval model as well as the baseline when Boolean concepts and scored concepts are considered together.

Scores in the hybrid approach is calculated by averaging the final estimations of both methods. Averaging of the scores is an efficient approach which does not require additional efforts and shown to be successful in action recognition (Can and Manmatha 2013, Can et al. 2015). For the hybrid approach, we are able to obtain a mean average precision of 8% which improves the baseline 25% relatively and is a statistically significant improvement ($p <0.0001$). The p value here is less than the previous cases. Blending multiple results yield improvements when the results are different than each other. In other words, blending two different, and preferably equally accurate, ranked lists would yield better results than blending two similar ranked lists. Having a smaller p value here might indicate that we obtain a more accurate ranked list when we fuse the results of our MRF based model compared to fusion of the baseline. The number of event queries showing improvement also aligns with this finding. The hybrid approach of our MRF based retrieval model enables us to outperform the hybrid approach of the baseline on 26 out of 30 queries. The

Table 5.3. Results of our MRF based model as well as the baseline when blending Boolean concepts and scored concepts. Test set: MEDTEST

Id & Name	Blended Baseline	Blended Our Approach
E6 Birthday party	13.6	**13.7**
E7 Changing a vehicle tire	12.0	**14.6**
E8 Flash mob gathering	10.4	**12.8**
E9 Getting a vehicle unstuck	3.2	**4.5**
E10 Grooming an animal	6.8	**8.1**
E11 Making a sandwich	9.3	**11.2**
E12 Parade	29.2	**32.9**
E13 Parkour	21.6	**29.1**
E14 Repairing an appliance	7.3	**8.7**
E15 Working on a sewing project	10.2	**13.7**
E21 Attempting a bike trick	1.4	**2.4**
E22 Cleaning an appliance	0.7	**0.8**
E23 Dog show	0.5	**0.7**
E24 Giving directions to a location	2.0	**2.7**
E25 Marriage proposal	1.4	**5.1**
E26 Renovating a home	0.9	**1.0**
E27 Rock climbing	5.2	**6.3**
E28 Town hall meeting	0.4	**0.5**
E29 Winning a race without a vehicle	2.1	**2.9**
E30 Working on a metal crafts project	**10.7**	9.6
E31 Beekeeping	7.1	**8.2**
E32 Wedding shower	2.1	**4.0**
E33 Non-motorized vehicle repair	2.4	**3.1**
E34 Fixing musical instrument	**2.5**	2.3
E35 Horse riding competition	12.2	**14.0**
E36 Felling a tree	7.3	**12.6**
E37 Parking a vehicle	2.7	2.7
E38 Playing fetch	4.7	4.7
E39 Tailgating	0.8	**1.1**
E40 Tuning musical instrument	3.3	**5.5**
Avg.	6.5	**8.0**

number in the previous cases was 21 for *scored concepts* and 22 for *Boolean concepts*.

Recall that the average performance from Table 5.2 was 6.2% for *Boolean concepts* and 7.0% for *scored concepts*.

5.3.3 Comparing Our Results with Previously Reported Numbers

So far, we have evaluated our model by comparing it with a baseline that considers a bag-of-concepts approach with an independence assumption. Here, we also compare our results with the previously reported results on the same dataset.

In Table 5.4, we compare our MRF based retrieval model (hybrid approach) with four previously reported results on the same dataset. Note that the comparison is provided on only twenty event queries since Chen et al. (2014), Habibian et al. (2014), Mazloom et al. (2013a), Rastegari et al. (2013) provide their results on only twenty events.

According to the results, our approach outperforms the results of Chen et al. (2014) by 300%, Mazloom et al. (2013a) by 115%, Rastegari et al. (2013) by 160%, and Habibian et al. (2014) by 40%. On 14 out of 20 events, our approach provides higher retrieval accuracy compared to these results in the literature. Note that these comparisons are performed on 20 events as most of the previous work only provided their results on these 20 event queries.

In addition to the mean average precision, our method outperforms the previous work on 70% of the queries. Among these majority, there are some event queries, our approach did really well. For example, for the "E13: Parkour" event query, our method outperforms the previous work drastically. This stems from the concept detectors. In other words, the concepts we use for this particular event query work really well and enable us to obtain videos relevant to this query on the top ranks. On the other hand, there are some event queries where our method does not do well compared to the previous work. The "E29: Winning a race without a vehicle" event query is an instance where Habibian et al. (2014)'s method provide quite high retrieval accuracy compared to the other methods. This might also stem from using different concept detectors. Having different experimental environments reduces our confidence in comparison with our results with the previously reported ones. To address this,

Table 5.4. Results of our MRF based hybrid approach as well as the previously reported results. Test set: MEDTEST

Id	Semantic E. Desc. (Chen et al. 2014)	Select. C. (Mazloom et al. 2013a)	Bi Concepts (Rastegari et al. 2013)	Composite C. (Habibian et al. 2014)	Our Approach
E6	3.1	4.9	4.7	7.6	**13.7**
E7	3.8	1.0	1.8	1.8	**14.6**
E8	5.5	23.0	9.0	**37.3**	12.8
E9	1.0	0.4	3.1	5.5	**4.5**
E10	1.1	0.9	0.9	0.9	**8.1**
E11	2.8	7.7	7.4	7.9	**11.2**
E12	10.5	21.9	19.3	22.4	**32.9**
E13	2.5	0.5	0.9	2.2	**29.1**
E14	4.6	1.2	0.9	2.5	**8.7**
E15	1.1	1.4	1.4	1.5	**13.7**
E21	0.2	1.1	0.6	2.2	**2.4**
E22	**1.2**	0.5	0.5	0.8	0.8
E23	0.2	0.1	0.3	0.1	**0.7**
E24	0.1	0.6	0.6	2.3	**2.7**
E25	0.9	0.1	0.2	0.2	**5.1**
E26	0.1	0.6	0.6	**2.3**	1.0
E27	1.2	14.2	13.9	**14.7**	6.3
E28	**2.0**	1.0	0.6	1.5	0.5
E29	2.8	3.1	3.1	**13.6**	2.9
E30	0.2	0.4	0.5	0.6	**9.6**
Avg.	2.2	4.2	3.5	6.4	**9.1**

we simulate a similar environment to the one defined in (Habibian et al. 2014) and re-calculate their results using our concepts. In this way, we can compare different methods by eliminating the effect of using different detectors. We provide the results of these experiments in Table 5.5.

Our approach still outperforms the results of Habibian et al. (2014) using our concepts. While the mean average precision on twenty events (see Table 5.4) is 6.7%, here it slightly increases and becomes 7% (on the same twenty event queries). Changing the source of concepts led to approximately 5% difference on the results. We

Table 5.5. Results of our MRF based model as well as the results of Habibian et al. (2014) using our concepts. Test set: MEDTEST

Id & Name	Habibian et al. (2014)	Our hybrid approach
E6 Birthday party	9.8	**13.7**
E7 Changing a vehicle tire	7.4	**14.6**
E8 Flash mob gathering	**31.5**	12.8
E9 Getting a vehicle unstuck	1.8	**4.5**
E10 Grooming an animal	3.8	**8.1**
E11 Making a sandwich	2.8	**11.2**
E12 Parade	27.6	**32.9**
E13 Parkour	**29.3**	29.1
E14 Repairing an appliance	1.3	**8.7**
E15 Working on a sewing project	5.3	**13.7**
E21 Attempting a bike trick	0.2	**2.4**
E22 Cleaning an appliance	0.5	**0.8**
E23 Dog show	**3.0**	0.7
E24 Giving directions to a location	0.4	**2.7**
E25 Marriage proposal	0.3	**5.1**
E26 Renovating a home	0.8	**1.0**
E27 Rock climbing	**6.4**	6.3
E28 Town hall meeting	**0.9**	0.5
E29 Winning race without a vehicle	**6.0**	2.9
E30 Working on metal crafts project	0.4	**9.6**
E31 Beekeeping	6.1	**8.2**
E32 Wedding shower	0.8	**4.0**
E33 Non-motorized vehicle repair	**10.1**	3.1
E34 Fixing musical instrument	0.7	**2.3**
E35 Horse riding competition	5.2	**14.0**
E36 Felling a tree	3.8	**12.6**
E37 Parking a vehicle	0.6	**2.7**
E38 Playing fetch	0.7	**4.7**
E39 Tailgating	**1.4**	1.1
E40 Tuning musical instrument	1.1	**5.5**
Avg.	5.7	**8.0**

believe that the main issue is not the source of concepts but the way to select the pair of concepts in their work. The process is based on performances of concepts calculated on a training set per event query. In other words, concepts providing better retrieval on a training set are favored for a particular query. Even though this does not fit in the

context of zero-shot video event detection, we also investigate the performance of our approach using the concepts selected in their way. When we use their concepts in our blended MRF based retrieval model, we are able to obtain a mean average precision of 10.4%, while it was 8% with our concepts. The improvement on the retrieval accuracies are not surprising since selecting the concepts on *training data* usually yield better retrievals. Observing these improvements also leads us to investigate the performance of our approach when we have exemplars. In other words, we use training examples to train weights for the tuples (e.g., concept-frame pairs: full independence, concept-concept-frame triples: spatial dependence, and concept-concept-frame-frame quadruples: temporal dependence).

5.3.4 Applying Zero-shot to VED-ex$_{few}$

So far, we have assumed that each tuple has the same weight in the model. Different weights for each tuple can be trained by leveraging the example videos of a query. We employ ten example videos per query (using the EK10 set as training set) and train a retrieval model using these examples. Then we run test videos against these models. The final retrieval accuracy for ten queries becomes 11.5% for our blended MRF based retrieval model, whereas the baseline (bag-of-concepts) turns out to be 9.5%. This statistically significant improvement ($p < 0.0001$ and improvements on 25 out of 30 event queries) shows that our approach can also be used when there are example videos associated with queries.

In Chapter 4, by incorporating one-exemplar models into video event detection, we were able to improve, on the average (e.g., multiple descriptors), 22 event queries out of 30. In this case, when we use the same training set, we observe improvements on 25 out of 30 events. The p value here is less than the ones in the previous chapter (i.e., 0.008, 0.002, 0.008, and 0.03 for multiple descriptors). We observe that exploiting

dependencies of concepts improves retrieval slightly more than the one-exemplar approach.

5.4   Summary of the chapter

In this chapter, we focused on zero-shot video event detection, which is the video event detection case when no example videos are associated with queries. As no exemplars are available, we cannot leverage them to learn a retrieval model. Therefore, it is a more challenging and realistic task.

We tackle this problem considering the dependencies of concepts in videos. Events often involve multiple concepts and their interactions. Most of the recent attempts on this problem consider the bag-of-concepts approach with an independence assumption. We believe that it is a weak assumption and we have shown empirically that using dependencies of concepts provides higher retrieval accuracy than the independence assumption.

We focus on a MRF based retrieval model with three dependency assumptions: (1) full independence, (2) spatial dependence, and (3) temporal dependence. The idea is that if important concepts show dependencies in a video, that would be a stronger evidence for relevance of this video to a query.

In our formulations we interpret the concept detector outputs in two different ways: 1) assuming concepts having higher confidence are present and the rest are not present in a video (i.e., Boolean concepts) and 2) using the output scores directly without any selection (i.e., scored concepts). Both approaches have their advantages and we leverage their advantages by also considering a hybrid approach. Experimental results show that we are able to outperform a bag-of-concepts with independence assumption baseline in all of the cases. We also compare our results with previously reported results on the same evaluation datasets. We show that our MRF based retrieval model outperforms previous figures by 40% (i.e., 6.4% vs. 9.1%) to 300% (2.2% vs. 9.1%).

We further investigate the effectiveness of our approach in the context of VED-ex. Experimental results show that, exploiting dependencies improves the retrieval accuracy when there are also example videos associated with event queries. The improvements obtained by exploiting dependencies seem to be slightly higher than the ones obtained with incorporating one-exemplar models into video event detection.

CHAPTER 6

CONCLUSION

Searching videos has become increasingly important as huge number of videos are available online and people spend plenty of time watching videos online. In this thesis, we tackled the video event detection problem, that is the task of searching videos for events of interest to a user. The problem can be seen as retrieving videos that are expected to be relevant to a query.

Most of the recent commercial video search engines (e.g., Youtube) focus on the text similarity between a query and metadata associated with videos for retrieval. However, considering only textual similarity might be misleading or insufficient. Further, it is implausible to use text-based when there is no metadata associated with videos. Text-based approaches can be used or can be improved with the help of visual features (i.e., concepts). However, in this thesis, we mainly investigate the relative advantages of different ways of using visual content of videos. It is an promising approach when no metadata associated with videos. Using only visual features and ignoring the text modality is the major factor of low accuracy scores. Even though VED using only visual features has not reached its "product ready" state, we believe that recent progress, especially on deep learning oriented features, will help researchers to reach that point faster.

In order to include the content of videos into retrieval, we extract semantic information from videos. For this purpose, we create object-based concept detectors and leverage action-based concept detectors. We then run these detectors against

videos to measure the likelihood of observing them in videos. In this way, we have an understanding of the concepts (e.g., objects and actions) in videos.

Creating a concept detector involves labor- and resource-intensive steps and therefore it is an expensive process. As a consequence, our concept dictionary is limited in size; therefore, we need to build a mapping between queries and our dictionary. One way of doing this is to look for the semantic (e.g., textual) similarities between concepts and query descriptions. For example, for a "horse grazing" query, we would be looking for concepts such as "horse" and "grass". Alternatively, we can compile a fixed set of concepts and use them for any query within the context of video event detection with exemplars (VED-ex). We hypothesized that using query-independent (i.e., using a fixed and the same set of concepts for any query) or query-dependent (i.e., selecting concepts based on queries—and perhaps a different set of concepts for different queries) concepts would not change the retrieval accuracies significantly. In Chapter 3, we showed that we are able to obtain similar retrieval accuracies by using query-independent concepts compared to using the ones obtained using query-dependent concepts within the context of VED-ex.

In addition to the query-independent concepts, we also provided a space-efficient way of representing videos, which enables us to reduce the resources required to process the representations. We focused on only concepts that we detect with high confidence at each frame and ignore the rest. As a result, the final representation of a video become sparser. Having a sparse representation enables us to save space to store them on disk and time to train/test them (1-to-5 ratio in both cases). Further, we agreed with Mazloom et al. (2013a), Merler et al. (2012), as well as Habibian et al. (2013) and empirically showed that using a subset of concepts is advantageous compared to using all them.

Up to that point, we have assumed that several example videos are associated with an event query. The major parameter in the video event detection problem is

the number of example videos associated with queries. Having a large number of example videos is ideal; however, it is costly. As the number of exemplars decreases, the quality of the retrieval models—trained on them—decreases as well. We presented a method that incorporates multiple one-exemplar models into video event detection aiming at improving retrieval accuracies when there are few exemplars available. One-exemplar models are created using only one example video and therefore they are specific to their exemplars. For instance, example videos for the "repairing an appliance" query might include contents of repairing different appliances such as oven, refrigerator, and washing machine. In the common video event detection approach, a single global model is trained using all of the example videos. Against this practice, we create several one-exemplars each of which is trained using one exemplar video and incorporated them into this global model. Our approach enabled us to improve the retrieval accuracies anywhere from 15 to 35% when there are few exemplars available.

In this thesis, we also focus on the video event detection problem when no example videos are associated with queries (i.e., zero-shot video event detection). When example videos are unavailable, we cannot train a retrieval model using them, which makes the problem more challenging.

We tackled the zero-shot video event detection problem by exploiting dependencies of concepts. Events are complex activities which are localized in space and time. We hypothesized that considering concepts individually might be insufficient and considering concepts and their relationships might enable us to have a stronger evidence to retrieve videos relevant to an event query.

Our dependency work employs a Markov random field (MRF) based retrieval model that is a broadly accepted algorithm in information retrieval community. We focus on three dependency assumptions: 1) full independence, concepts being independent from each other; 2) spatial dependence, co-occurrences of concepts in the same frame

are treated as important, and 3) temporal-dependence, co-occurrences of concepts in consecutive frames are treated as important.

The concept detectors measure the likelihood of observing concepts in videos. We considered two different interpretations of output scores in our formulations. On the one hand, we treated only the concepts having high confidence as present in videos and fed these Boolean concepts into our retrieval model. On the other hand, we fed the output scores directly into the model. The first approach reduces noise due to low quality detectors. However, it might also trim some useful information. The latter approach keeps the potential useful information along with some amount of noise. Both approaches have their own advantages. We considered both approaches individually as well as a hybrid approach that benefits from both of them.

Evaluation of our approach showed that our MRF based retrieval model statistically significantly improves the common bag-of-concepts approach with independence assumption ranging from 10% to 150%. Comparison with the previously reported results on the same dataset showed that our approach outperforms the previously reported figures by 40% (i.e., 6.4% vs. 9.1%) to 300% (2.2% vs. 9.1%).

6.1 Future Work

So far, we have explained and discussed the approaches that we have provided to tackle the video event detection problem. Here, we present our thoughts for the next directions based on this work.

6.1.1 Determining the Numbers of Concepts to Consider at Each Frame

In Chapter 3 and 5, we focus on only a subset of concepts at each frame depending upon their output scores. In other words, we assume that only k concepts of those having the highest output scores are present in a frame.

The k parameter is tuned on a validation set and based on the maximization of the mean average precision figure. In the analysis of individual queries, we observe that queries have their peak results at different values of k. For example, while we obtain the peak retrieval accuracy for a number of queries when $k = 25$, for other queries this value becomes 5 or 10. We use $k = 10$ as it is the optimal parameter validated based on the mean average precision across multiple queries.

Developing an efficient way to estimate different values of k for different queries is a promising direction. Relationships between k and the inputs (e.g., query and data) can be exposed with further analysis.

6.1.2 Richer Dependencies

In our MRF based retrieval model, we consider three dependency assumptions: full independence, spatial dependence, and temporal dependence. The spatial dependence case exploits the co-occurrence of concepts in the same frame. We consider the co-occurrences of concepts in a window of consecutive frames in the temporal dependence case.

The window size in our experiments was fixed. Developing an efficient way of covering co-occurrence of concepts in different window sizes and selecting one or more of them is a promising direction. In this way, we would not miss temporal dependencies of concepts co-occurring at different distances. For example, for the "horse riding competition" event query, the "horse riding" and "jumping over fence" concepts might co-occur in a short window of frames, whereas for the "felling a tree" query, we might need a larger window for the co-occurrence of "sawing a tree" and "tree falling" as "sawing a tree" usually takes a considerable amount of time.

In addition to temporal dependence, we can improve spatial dependence as well. If we have higher quality detection, we can also focus on co-occurrences of concepts in a finer level. For example, the co-occurrence of "horse" and "person" in the same

frame is good evidence for the "horse-riding competition" query. However, "person" *on top of* "horse" is stronger evidence for the same query.

6.1.3 Creating a Larger Concept Dictionary Efficiently

One of the major challenges to create a concept detector is to collect positive examples related to a particular concept. As a results of this issue, the total number of concept detectors stays in the range of few thousands.

We believe that having a richer concept dictionary would enable us to have better retrievals. We can use publicly available datasets to create more detectors efficiently. ImageNet (`http://www.image-net.org/`) is one of the datasets that provides a number of annotated images. ImageNet has been partially used so far; however, we also need to consider larger datasets, which might contain videos as well. For example, Yahoo Labs has recently announced another dataset, which consists of approximately 100 million annotated images as well as about 700,000 annotated videos (`http://labs.yahoo.com/news/yfcc100m/`).

In addition to existing datasets, we can also make use of social multimedia. In other words, we can use the videos/images posted online since they already have comments or descriptions associated with them. Even though information on social media is too noisy, we still believe that it is one of the most efficient ways to address the lack of annotated data issue.

6.1.4 Handling Concept Selection Errors

In Section 5.1.1, we have explained how to select concepts based on a query within the context of VED-zero. Our approach is based on identifying concepts that are expected to be relevant to a query. The identification process relies on similarity between textual description of queries and concepts and highly dependent upon matching words.

In the concept dictionary we used in VED-zero, several concepts are created based on queries as our evaluation environment is dedicated to test the correctness of our

methods. Further, we make use of the full description of a query including name, definition, and explication. Therefore, we did not face the problem of having no concepts identified as relevant to a query.

Having a larger dictionary of concepts might solve this problem or decrease the chance of facing with this problem. However, when we use only the title of a query, we might need expand the query to catch concepts that might be relevant to this query. Query expansion techniques might be employed to address this issue. For example, Allan et al. (2013) show that it is possible to use only titles of queries. They run query titles against the Wikipedia, and in this way more terms are compiled for a query. This expansion process increases the chance of identifying relevant concepts to a query.

6.1.5 Towards Higher Quality Detectors

Even though the major parameter in the approaches tackling video event detection is the detectors, we researchers do not pay much attention to the quality of the detectors. The main reason stems from the urge for creating a very large dictionary of concepts. Therefore, we focused on the quantity more than the quality. In some cases, we take the quality of concepts for granted.

For example, the HMDB dataset was released for evaluating approaches for the action recognition problem (Kuehne et al. 2011). When it was first released (in 2011), the state-of-the-art for that dataset was around 20% (Kuehne et al. 2011). Then, researchers improved the state-of-the-art to 65%. We transfer the knowledge we have gained from HMDB and apply to video event detection. However, we then observe that recognizing the "running" action with 80% accuracy in HMDB does not necessarily mean that we can detect the "running" actions in videos with the same quality.

A number of solutions seem promising to address this issue. HMDB is hand-crafted (not only HMDB but most of the existing datasets also suffer from the same

issue). It is a simple dataset (e.g., videos having similar lengths, only one action, and similar quality) compared to videos we have in video event detection (e.g., different in size as well as quality, and may contain multiple actions). First, we can focus on unconstrained and not-hand-crafted sets while creating our detectors. We can also try to focus on robust descriptors which might work fairly similar in different datasets. Recent progress on deep learning oriented features shows promising and robust results in object detection and recognition.

Increasing quality of detectors relies on accepting the fact that detectors are key to video event detection. Therefore, we need to focus on the quality of detectors in addition to the quantity of them.

Recent progress on deep learning approaches has shown that high quality descriptors can be created to understand the content of images. For example, a new deep learning oriented descriptor named Overfeat has shown promising results on object detection and recognition (Sermanet et al. 2013). In Section 4.2.2, we also observe that concepts that are created using Overfeat descriptors show promising retrieval accuracies. Perhaps, a deeper analysis in this direction might yield higher quality detectors and we can take the advantage of them by replacing them with low-quality detectors.

We believe that increasing the quality of the detectors will result in higher accuracy scores for both query-dependent concepts (QDC) and query-independent concepts (QIC). QDC might show relatively larger improvements compared to QIC. The success of QIC is correlated with quality of the detectors. In other words, the current detectors are not good enough to detect concepts very effectively, where QDC cannot make a difference compared to using their query independent counterparts. Using higher quality detectors, especially the ones created with deep learning features, we might be able to take better advantage of QDC. A further analysis on comparing query-dependent and -independent concepts might be yet another direction for future studies on VED.

For the one-exemplar case, we believe that our approach will still work using higher quality detectors. When we evaluated our approach, we also investigated its robustness on multiple detectors. We also evaluated one-exemplar models using a deep learning oriented feature (i.e., Overfeat (Sermanet et al. 2013)) and the results showed that our approach works on this descriptor as well.

Similar to VED-ex, we believe that higher quality detectors will yield decent improvements on accuracy scores within the context of VED-zero. While we can take the advantage of example videos in VED-ex, our models rely on only concept outputs in VED-zero. Therefore, any change in the quality of the detectors will have a relatively larger impact on accuracy scores in VED-zero compared to VED-ex.

BIBLIOGRAPHY

Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–517, 2012.

James Allan, Jeffrey Dalton, John Foley, R. Manmatha, Venkatesh Murthy, and David Wemhoener. Short text queries for video retrieval. In *NIST TRECVID Workshop*, 2013.

Tim Althoff, Hyun Oh Song, and Trevor Darrell. Detection bank: An object detection based video representation for multimedia event recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1065–1068, 2012.

Robin Aly, Relja Arandjelovi and Ken Chatfield, Matthijs Douze, Basura Fernando, Zaid Harchaoui, Kevin McGuinness, and et al. Axes at trecvid 2013. In *NIST TRECVID Workshop*, 2013.

Mohamed Ayari, Jonathan Delhumeau, Matthijs Douze, Herv Jgou, Danila Potapov, Jrme Revaud, Cordelia Schmid, and Jiangbo Yuan. Inria 2011 trecvid: Copy detection and multimedia event detection. In *NIST TRECVID Workshop*, 2011.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern Information Retrieval*, volume 463. ACM press New York, 1999.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, 2011.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.

Subhabrata Bhattacharya, Felix X Yu, and Shih-Fu Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, page 105, 2014.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, 2006.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pages 401–408, 2007.

M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua. BRIEF: Computing a Local Binary Descriptor Very Fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, 2012.

Ethem F. Can and R Manmatha. Formulating action recognition as a ranking problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 251–256, 2013.

Ethem F. Can and R Manmatha. Modeling concept dependencies for event detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 289–297, 2014.

Ethem F. Can, W. Bruce Croft, and R. Manmatha. Incorporating query-specific feedback into learning-to-rank models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014.

Ethem F. Can, James Allan, and R. Manmatha. The simpler the better: An action recognition framework extended with deep learning. Technical report, University of Massachusetts Amherst, Collage of Information and Computer Sciences, 03 2015.

Liangliang Cao, Shih-Fu Chang, Noel Codella, Courtenay Cotton, Dan Ellis, Leiguang Gong, Matthew Hill, Gang Hua, John Kender, Michele Merler, et al. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system. In *NIST TRECVID Workshop*, 2011.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

Ming-yu Chen and Alexander Hauptmann. Mosift: Recognizing human actions in surveillance videos, 2009.

H Cheng, J Liu, S Ali, O Javed, Q Yu, A Tamrakar, A Divakaran, HS Sawhney, R Manmatha, J Allan, et al. Sri-sarnoff aurora system at trecvid 2012: Multimedia event detection and recounting. In *NIST TRECVID Workshop*, 2012.

Claudio Cusano, Riccardo Satta, and Simone Santini. Unsupervised classemes. In *European Conference on Computer Vision (ECCV) Workshops and Demonstrations*, pages 406–415, 2012.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision (ECCV)*, pages 428–441, 2006.

Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1857–1860, 2013.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785, 2009.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531. IEEE, 2005.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

Shaolei Feng and Raghavan Manmatha. A discrete direct retrieval model for image and video retrieval. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pages 427–436, 2008.

Robert Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition (2003)*, volume 2, pages II–264, 2003.

FinalCutPro7-Documentation. Apple final cut pro 7 documentation. https://documentation.apple.com/en/finalcutpro/usermanual/index. html{\#}chapter=D%26section=3%26tasks=true, 2015.

Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

Nikolaos Gkalelis and Vasileios Mezaris. Video event detection using generalized subclass discriminant analysis and linear support vector machines. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, page 25, 2014.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

Amirhossein Habibian, Koen EA van de Sande, and Cees GM Snoek. Recommendations for video event recognition using concept vocabularies. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 89–96, 2013.

Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Composite concept discovery for zero-shot video event detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, page 17, 2014.

Image-Net. Image-net: An image database, 2014. URL \url{http://www.image-net.org}.

Fredrick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the W. on Pat. Rec. in Prac.*, 1980.

Lu Jiang, Alexander G Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 449–458, 2012.

Lu Jiang, Teruko Mitamura, Shoou-I Yu, and Alexander G Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, page 297, 2014.

Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

Thorsten Joachims. Making large scale svm learning practical. In *Advances in Kernel Methods*. MIT Press, 1999.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

Thorsten Joachims. Svm-light and svm-rank and multi-class svm, 2014. URL `http://www.cs.cornell.edu/People/tj/`.

David Joy Paul Over Vladimir Dreyvitser Jonathan Fiscus, Greg Sanders. Trecvid 2014 multimedia event detection task. In *NIST TRECVID Workshop*, 2014.

Ilseo Kim, Sangmin Oh, Byungki Byun, AG Amitha Perera, and Chin-Hui Lee. Explicit performance metric optimization for fusion-based video retrieval. In *Europen Conference on Computer Vision (ECCV) Workshops and Demonstrations*, pages 395–405, 2012.

Alexander Klaser and Marcin Marszalek. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference (BMVC)*, pages 275–1, 2008.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL) Volume 1*, pages 423–430, 2003.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

Christoph H Lampert, Matthew B Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Double fusion for multimedia event detection. In *Advances in Multimedia Modeling*, pages 173–185, 2012.

Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006.

Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386, 2010.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics, 2004.

Jingchen Liu, Scott McCloskey, and Yanxi Liu. Local expert forest of score fusion for video event classification. In *European Conference on Computer Vision (ECCV)*, pages 397–410, 2012.

Jingen Liu, Hui Cheng, Omar Javed, Qian Yu, Ishani Chakraborty, Weiyu Zhang, Ajay Divakaran, and et al. Sri-sarnoff aurora system at trecvid 2013 - multimedia event detection and recounting. In *NIST TRECVID Workshop*, 2013a.

Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Video event recognition using concept attributes. In *Workshop on Applications of Computer Vision at WACV*, pages 339–346, 2013b.

David G Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Yuanhua Lv and ChengXiang Zhai. Positional language models for information retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2009.

Zhigang Ma, Yi Yang, Yang Cai, Nicu Sebe, and Alexander G Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *Proceedings of the ACM International Conference on Multimedia*, pages 469–478, 2012.

Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, and Alexander G Hauptmann. Complex event detection via multi-source video attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2627–2633, 2013.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150, 2011.

Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *IEEE Internation Conference on Computer Vision (ICCV)*, pages 89–96, 2011.

Masoud Mazloom, Efstratios Gavves, Koen van de Sande, and Cees Snoek. Searching informative concept banks for video event detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 255–262, 2013a.

Masoud Mazloom, Amirhossein Habibian, and Cees GM Snoek. Querying for video events by semantic signatures from few examples. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 609–612, 2013b.

Masoud Mazloom, Xirong Li, and Cees GM Snoek. Few-example video event retrieval using tag propagation. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, page 459, 2014.

Andrew McCallum, Kamal Nigam, and Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178, 2000.

Michele Merler, Bert Huang, Lexing Xie, Gang Hua, and Apostol Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, 2012.

Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 472–479, 2005.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.

Davide Modolo and Cees GM Snoek. Can object detectors aid internet video event retrieval? In *IS&T/SPIE Electronic Imaging*, 2013.

Gregory K Myers, Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ramakant Nevatia, Chen Sun, Amirhossein Habibian, Dennis C Koelma, Koen EA van de Sande, Arnold WM Smeulders, et al. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 25(1):17–32, 2014.

Pradeep Natarajan, Prem Natarajan, Vasant Manohar, Shuang Wu, Stavros Tsakalidis, Shiv N Vitaladevuni, Xiaodan Zhuang, Rohit Prasad, Guangnan Ye, Dong Liu, et al. Bbn viser trecvid 2011 multimedia event detection system. In *NIST TRECVID Workshop*, 2011.

Pradeep Natarajan, Prem Natarajan, Shuang Wu, Xiaodan Zhuang, Amelio Vazquez-reina, Shiv N Vitaladevuni, Carl Andersen, Rohit Prasad, Guangnan Ye, Dong Liu, et al. Bbn viser trecvid 2012 multimedia event detection and multimedia event recounting systems. In *NIST TRECVID Workshop*, 2012.

Apostol Natsev, John R Smith, Matthew Hill, Gang Hua, Bert Huang, Michele Merler, Lexing Xie, Hua Ouyang, and Mingyuan Zhou. Ibm research trecvid-2010 video copy detection and multimedia event detection system. In *NIST TRECVID Workshop*, 2010.

NIST. 2011 trecvid multimedia event detection track, 2012. URL `http://www.nist.gov/itl/iad/mig/med11.cfm`.

Sangmin Oh, AG Amitha Perera, Ilseo Kim, Megha Pandey, Kevin Cannons, Hossein Hajimirsadeghi, Arash Vahdat, Greg Mori, Ben Miller, Scott McCloskey, et al. Trecvid 2013 genie: Multimedia event detection and recounting. In *NIST TRECVID Workshop*, 2013.

Sangmin Oh, Scott McCloskey, Ilseo Kim, Arash Vahdat, Kevin J Cannons, Hossein Hajimirsadeghi, Greg Mori, AG Amitha Perera, Megha Pandey, and Jason J Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, 25(1):49–69, 2014.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

Bruno A Olshausen. Sparse coding of time-varying natural images. In *Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation*, pages 603–608, 2000.

Dan Oneata, Matthijs Douze, Jrme Revaud, Jochen Schwenninger, Danila Potapov, Heng Wang, Zaid Harchaoui, Jakob Verbeek, and Cordelia Schmid. Axes at trecvid 2012: Kis, ins, and med. In *NIST TRECVID Workshop*, 2012.

P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, and et al. Trecvid 2010 - an overview. In *NIST TRECVID Workshop*, 2010.

Desislava Petkova and W Bruce Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 731–740, 2007.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Mohammad Rastegari, Ali Diba, Devi Parikh, and Ali Farhadi. Multi-attribute queries: To merge or not to merge? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3310–3317. IEEE, 2013.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

Kimiaki Shirahama, Marcin Grzegorzek, and Kuniaki Uehara. Multimedia event detection using hidden conditional random fields. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, page 9, 2014.

Cees GM Snoek and Marcel Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2008.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Munirathnam Srikanth and Rohini Srihari. Biterm language models for document retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–426, 2002.

Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3681–3688, 2012.

Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1250–1257, 2012.

Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, pages 776–789, 2010.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–394, 2010.

K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1449–1456, 2011.

Gang Wang, Derek Hoiem, and David Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *IEEE International Conference on Computer Vision (ICCV)*, pages 428–435, 2009.

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2665–2672. IEEE, 2014.

Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009a.

Yang Yang and Mubarak Shah. Complex events detection using data-driven concepts. In *European Conference on Computer Vision (ECCV)*, pages 722–735, 2012.

Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. Query by document. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 34–43, 2009b.

Ehsan Younessian, Teruko Mitamura, and Alexander Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, page 51, 2012.

Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proceedings of the International Conference on Machine Learning*, pages 1169–1176, 2009.

S. Yu, Z. Zu, D. Ding, W. Sze, F. Vicente, Z. Lan, Y. Chai, S. Rawat, et al. Informedia @ trecvid 2012. In *NIST TRECVID Workshop*, 2012.

HongJiang Zhang, Shuang Yeo Tan, Stephen W Smoliar, and Gong Yihong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(6):256–266, 1995.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.