Large-Scale Video Event Detection

Guangnan Ye

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Large-Scale Video Event Detection

Guangnan Ye

Because of the rapid growth of large scale video recording and sharing, there is a growing need for robust and scalable solutions for analyzing video content. The ability to detect and recognize video events that capture real-world activities is one of the key and complex problems. This thesis aims at the development of robust and efficient solutions for large scale video event detection systems. In particular, we investigate the problem in two areas: first, event detection with automatically discovered event specific concepts with organized ontology, and second, event detection with multi-modality representations and multi-source fusion.

Existing event detection works use various low-level features with statistical learning models, and achieve promising performance. However, such approaches lack the capability of interpreting the abundant semantic content associated with complex video events. Therefore, mid-level semantic concept representation of complex events has emerged as a promising method for understanding video events. In this area, existing works can be categorized into two groups: those that manually define a specialized concept set for a specific event, and those that apply a general concept lexicon directly borrowed from existing object, scene and action concept libraries. The first approach seems to require tremendous manual efforts, whereas the second approach is often insufficient in capturing the rich semantics contained in video events. In this work, we propose an automatic event-driven concept discovery method, and build a large-scale event and concept library with well-organized ontology, called EventNet. This method is different from past work that applies a generic concept library independent of the target while not requiring tedious manual annotations. Extensive experiments over the zero-shot event retrieval task when no training samples are available show that the proposed EventNet library consistently and significantly outperforms the state-of-the-art methods.

Although concept-based event representation can interpret the semantic content of video events, in order to achieve high accuracy in event detection, we also need to consider and combine vari-

ous features of different modalities and/or across different levels. One one hand, we observe that joint cross-modality patterns (e.g., audio-visual pattern) often exist in videos and provide strong multi-modal cues for detecting video events. We propose a joint audio-visual bi-modal codeword representation, called bi-modal words, to discover cross-modality correlations. On the other hand, combining features from multiple sources often produces performance gains, especially when the features complement with each other. Existing multi-source late fusion methods usually apply direct combination of confidence scores from different sources. This becomes limiting because heterogeneous results from various sources often produce incomparable confidence scores at different scales. This makes direct late fusion inappropriate, thus posing a great challenge. Based upon the above considerations, we propose a robust late fusion method with rank minimization, that not only achieves isotonicity among various scores from different sources, but also recovers a robust prediction score for individual test samples. We experimentally show that the proposed multi-modality representation and multi-source fusion methods achieve promising results compared with other benchmark baselines.

The main contributions of the thesis include the following.

1. Large scale event and concept ontology: a) propose an automatic framework for discovering event-driven concepts; b) build the largest video event ontology, *EventNet*, which includes 500 complex events and $4,490$ event-specific concepts; c) build the first interactive system that allows users to explore high-level events and associated concepts in videos with event browsing, search, and tagging functions.

2. Event detection with multi-modality representations and multi-source fusion: a) propose novel bi-modal codeword construction for discovering multi-modality correlations; b) propose novel robust late fusion with rank minimization method for combining information from multiple sources.

The two parts of the thesis are complimentary. Concept-based event representation provides rich semantic information for video events. Cross-modality features also provide complementary information from multiple sources. The combination of those two parts in a unified framework can offer great potential for advancing state-of-the-art in large-scale event detection.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Professor Shih-Fu Chang, who guided me to the joyful journey in the fields of computer vision and multimedia. In the past five years, I was amazed by his rigorous research attitude, deep research insight, and everlasting research enthusiasm. I'm sure I would be benefit from his advices not only within my PhD program, but also for the rest of my life.

I would like to thank my mentors at AT&T Research Lab, especially Behzad Shahraray, David C. Gibbon, Zhu Liu, Yadong Mu, Eric Zavesky, etc. The internship in 2014 summer is the most memorial research experience I have ever had. Thanks to Dr. Jun Wang at IBM T.J. Watson Research, for the valuable discussions and suggestions on my research and PhD program.

I would like to thank the committee members Behzad Shahraray, Ching-Yung Lin, John Wright, and John Paisley for providing the valuable suggestions on my thesis.

Thanks to all the (former and current) members in DVMM lab. Especially, thanks to Dr. Dong Liu, who taught me from the beginning on how to write my first research paper hand by hand. Thanks to Dr. Yu-Gang Jiang, from whom I learned a lot on experience of research and engineering. Thanks to Dr. I-Hong Jhuo. Without your help, I could not fulfill more achievements. Thanks to Dr. Xinnan Yu, who always provide me support and helps in my PhD program.

Special thanks to my parents, your love will always be my strongest support in my life. Finally, my deepest thanks and appreciation to my wife, Xiang Liu. For your endless understanding, consideration, and encouragement, you deserve everything I have ever achieved!

To my wife Xiang and my parents

Part I

Introduction

Chapter 1

Introduction

1.1 Motivation

The prevalence of video capture devices and the growing practice of video sharing in social media have resulted in an enormous explosion of user-generated videos on the Internet. For example, there are more than 1 billion users on *YouTube*, and 300 hours of video are uploaded every minute to the website. Another media sharing website, *Facebook*, reported recently that the number of videos posted to the platform per person in the U.S. has increased by 94% over the last year.

There is an emerging need to construct intelligent, robust, and efficient search-and-retrieval systems to organize and index those videos. However, most current commercial video search engines rely on textual keyword matching rather than visual content-based indexing. Such keyword-based search engines often produce unsatisfactory performance because of inaccurate and insufficient textural information, as well as the well-known issue of semantic gaps that makes the keyword-based search engines infeasible in real world scenarios. Thanks to recent research in computer vision and multimedia, researchers have attempted to automatically recognize people, objects, scenes, human actions, complex events, etc., and index videos based on the learned semantics in order to better understand and analyze the indexed videos by their semantic meanings. In this thesis, we are especially interested in analyzing and detecting events in videos. The automatic detection of complex events in videos can be formally defined as "detecting a complicated human activity interacting with people and object in a certain scene" [MED, 2010]. Compared with object, scene, or action detection and classification, complex event detection is a more challenging task because it is often combined

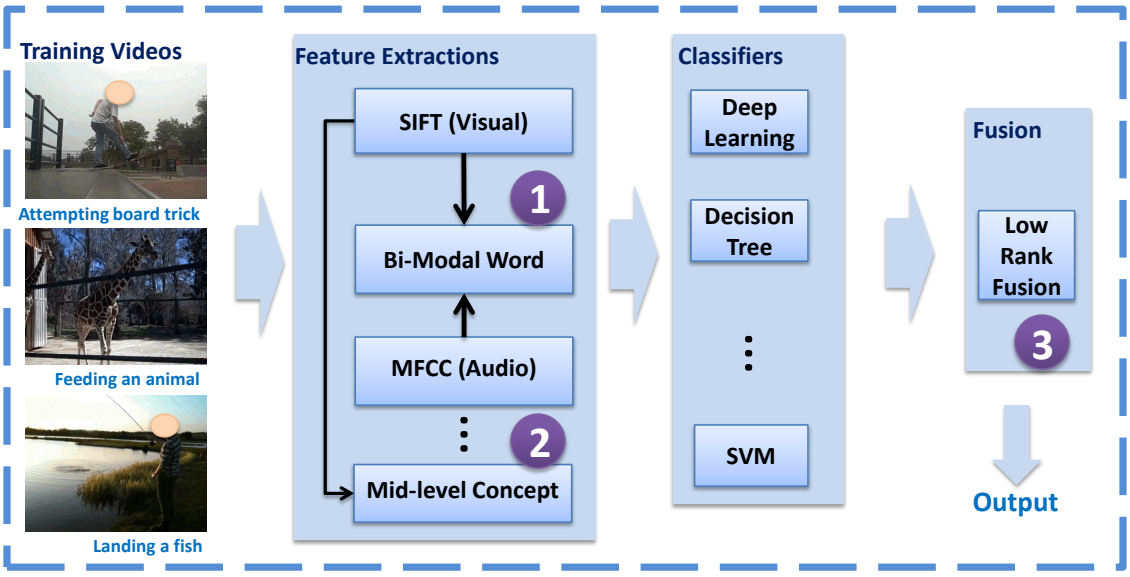


Figure 1.1: Event Detection Framework. Improved components in this thesis are marked accordingly.

with complicated interactions among objects, scenes, and human activities. Complex event detection often provides higher semantic understanding in videos, and thus has great potential for many applications, such as consumer content management, commercial advertisement recommendation, surveillance video analysis, and more.

In general, automatic detection systems, such as the one shown in Figure 1.1, contain three basic components: feature extraction, classifier, and model fusion. Given a set of training videos, state-of-the-art systems often extract various types of features [Y.-G. Jiang and Chang, 2010]. Those features can be manually designed low-level features, e.g., SIFT [Lowe, 2004], Mel-Frequency Cepstral Coefficients (MFCC) [Pols, 1966a], etc., that do not contain any semantic information, or mid-level feature representation where certain concept categories are defined and the probability scores from the trained concept classifiers are considered the concept features. After the feature extraction module, features from multiple modalities are used to train classifiers. Then, fusion approaches [Guangnan Ye and Chang, 2012; A. Rakotomamonjy and Grandvalet, 2009] are applied so that scores from multiple sources are combined to generate detection output. In this thesis, we focus on a few improvements upon this basic framework (e.g., shown as #1, #2, and #3 in

Figure 1.1). In particular, the two major technical components of this thesis, large-scale events and concept ontology construction (shown as #2 in Figure 1.1), and event detection with multi-modality representations (shown as #1 in Figure 1.1) and multi-source fusion (shown as #3 in Figure 1.1), are summarized below.

1.2 Technical Challenges and Proposed Approaches

1.2.1 Large-scale Event and Concept Ontology

Analysis and detection of complex events in videos require a semantic representation of the video content. Concept-based feature representation can not only depict a complex event in an interpretable semantic space that performs better zero-shot event retrieval, but also be considered mid-level features in supervised event modeling. By zero-shot retrieval here, we refer to the scenario in which the retrieval target is novel and thus there are no training videos available for training a machine learning classifier for the specific search target. A key research problem of the semantic representation is how to generate a suitable concept lexicon for events. There are two typical ways for defining concepts for events. The first is event independent concept lexicon that directly applies object, scene, and action concepts borrowed from existing libraryies, e.g., ImageNet [J. Deng and Fei-Fei, 2009], SUN dataset [Patterson and Hays, 2012], UCF 101 [K. Soomro and Shah, 2012], etc. However, because the borrowed concepts are not specifically defined for target events of interest, they are often insufficient and inaccurate for capturing semantic information in event videos. Another approach requires users to pre-define a concept lexicon and manually annotate the presence of those concepts in videos as training samples. This approach seems to involve tremendous manual effort, and it is infeasible for real-world applications.

In order to address these problems, we propose an automatic semantic concept discovery scheme that exploits Internet resources without human labeling effort. To distinguish the work that builds a generic concept library, we propose our approach as an event-driven concept discovery that provides more relevant concepts for events. In order to manage novel unseen events, we propose the construction of a large-scale event-driven concept library that covers as many real-world events and concepts as possible. We resort to the external knowledge base called WikiHow, a collaborative forum that aims to build the world's largest manual for human daily life events. We define *EventNet*,

which contains 500 representative events from the articles of the *WikiHow* website [Wik, 2015], and automatically discover $4,490$ event-specific concepts associated with those events. EventNet ontology is publicly considered the largest event concept library. We experimentally show dramatic performance gain in complex event detection, especially for unseen novel events. We also construct the first interactive system (to the best of our knowledge) that allows users to explore high-level events and associated concepts with certain event browsing, search, and tagging functions.

1.2.2 Multi-modality Representations and Multi-Source Fusion

Note that the state-of-the-art event detection system [Y.-G. Jiang and Chang, 2010] often extracts various types of features from multiple modalities, e.g., low-level visual feature, low-level audio feature, mid-level concept feature, textual feature, etc. In the second part of the thesis, we mainly explore two problems: $1)$ whether there is a cross-modal correlation among different modalities, and $2)$ whether there is a robust fusion method that combines multiple sources effectively.

Joint audio-visual patterns often exist in videos and provide strong multi-modal cues for detecting events. For example, an "explosion" event is best manifested by the transient burst of sound along with visible smoke and flame after the incident. Other examples include strong temporal synchronization (e.g., a horse running with audible footsteps) or loose association (e.g., a runner with cheering sounds in baseball videos). With the assumption that cross-modal correlation tends to be preserved in certain event videos, we propose a joint audio-visual bi-modal representation, called *bi-modal* words. In particular, we build a bipartite graph to model the relationship across the quantized words extracted from the visual and audio modalities. Partitioning over the bipartite graph is then applied to construct the bi-modal words that reveal the joint patterns across modalities. Different pooling strategies are employed to requantize the visual and audio words into the bi-modal words and form bi-modal Bag-of-Word (BoW) representations fed to subsequent event classifiers. Extensive experiments demonstrate the effectiveness of the multi-modality representation, the bi-modal, on video event detection tasks.

Feature combinations from multiple sources are often considered, especially when the features complement each other from heterecious modalities. Here, we focus on the problem of robust late fusion that aims to combine the confidence scores of the models constructed from multiple sources. One challenging problem with the late fusion strategy originates from the possible heterogeneity

among the confidence scores, which produces incomparable numbers at different numeric scales. With the motivation to achieve isotonicity (e.g., scale invariance) among the numeric scores of different sources, while recovering a robust fused prediction score with noise reduction, we propose a robust late fusion method with rank minimization. In particular, we convert each confidence score vector obtained from one source into a pairwise relationship matrix in order to address the scale variance problem. Then we formulate the score fusion problem as seeking a shared rank-two pairwise relationship matrix based on the original score matrix from the individual model that can be decomposed into the common rank-two matrix and sparse deviation errors in order to remove the prediction errors in each source for the individual test sample. We experimentally show that the proposed method can achieve significant performance gains on video event detection. The proposed method is also a general framework for multi-source fusion on other applications.

1.3 Thesis Outline

The following indicates the organization of the remainder of the thesis. In Chapter 2, we start with a brief literature survey on event detection, especially focusing on the benchmark event detection dataset summary and state-of-the-art system reviews. Part II describes large scale video event and concept ontology, and it contains Chapters 3,4,5. In Chapter 3, we describe an automatic event driven semantic concept discovery method. Chapter 4 describes a large scale structured concept library for complex event detection. In Chapter 5, we describe the EventNet application, where we build an ontology browsing, searching, and event video tagging online system. Part III describes event detection with multi-modality representation and multi-source fusion. In particular, in Part III Chapter 6, we describe the bi-modal codeword construction that discovers joint audio-visual codewords for video event detection, and, in Part III Chapter 7, we describe the robust late fusion method with rank minimization for multi-source fusion. Finally, in Part IV Chapter 8, we conclude the thesis and discuss future work.

Chapter 2

Literature Survey

2.1 Introduction

With the fast growth of video sharing in social media, high-level complex event detection has attracted great interests in computer vision and multimedia areas. Over decades, researchers have made great efforts to collect the large scale event datasets, build up benchmark systems, and propose novel methodologies for the task of video event detection. In this chapter, we briefly review the literatures for event detection task. Specifically, we will provide detailed descriptions in two perspectives including benchmark video event detection datasets, the state-of-the-art event detection system.

2.2 Dataset Summary

By the year of 2010, databases for event detection are still quite limited in the community. Most researchers focus on datasets of human action recognition which captured under constrained environments such as KTH [C. Schuldt and Caputo, 2004], IXMAS [Weinland D and E, 2006], Weizmann [M. Blank and Basri, 2005]. Later, several more realistic action datasets under unconstrained environments were released such as UCF11 [Liu J and M, 2009], UCF Sports [Rodriguez MD and M, 2008], UCF50 [ucf, 2015], UCF101 [K. Soomro and Shah, 2012], Hollywood movie dataset [I. Laptev and Rozenfeld, 2008], and Human Motion Dataset [Kuehne H and T, 2011]. Although those action recognition datasets can not be considered as high-level complex

Figure 2.1: Examples from Columbia Consumer Video database.

event benchmark datasets, the research on action recognition has made important early efforts and explorations for the task of event detection. Next, we describe several popular event detection benchmark datasets.

TRECVID MED dataset [MED, 2010] In order to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation, the annual NIST TRECVID activity defined the most well-known event detection task since the year of 2010, which is called TRECVID multimedia event detection (MED). The MED data contains user-generated content from Internet video and is collected and annotated by the Linguistic Data Consortium. Every year a newly extended event dataset is released for larger scale data evaluations. Over five years' efforts, significant efforts have been made to develop the dataset containing over $200,000$ videos which belong to 48 event categories. Tens of participants attended the evaluation and proposed several novel methods, and promising results have been achieved by the participating teams. MED is believed to be the largest public event detection evaluation worldwide. The dataset is also considered as one of the key benchmark datasets in the following experiments in the later chapters of the thesis.

Columbia Consumer Video (CCV) dataset [Y.-G. Jiang and Loui, 2011] In order to stimulate innovative research on challenging issues in event detection, CCV dataset was released in 2011. The dataset contains $9,317$ YouTube videos spanning over 20 semantic categories, which are "E1:

basketball", "E2: *baseball*", "E3: *soccer*", "E4: *ice skating*", "E5: *skiing*", "E6: *swimming*", "E7: *biking*", "E8: *cat*", "E9: *dog*", "E10: *bird*", "E11: *graduation*", "E12: *birthday*", "E13: *wedding reception*", "E14: *wedding ceremony*", "E15: *wedding dance*", "E16: *music performance*", "E17: *non-music performance*", "E18: *parade*", "E19: *beach*", "E20: *playground*". Figure 2.1 gives an example for each category. The database was collected with extra care to ensure relevance to consumer's interest and originality of video content without post-editing. Class annotations on video level were carefully performed with Amazon MTurk platform.

Stanford Sports-1M dataset [Karpathy *et al.*, 2014] The Sports-1M dataset consists of 1 million YouTube videos annotated with 487 classes. The classes are arranged in a manually-curated taxonomy that contains internal nodes such as Aquatic Sports, Team Sports, Winter Sports, Ball Sports, Combat Sports, Sports with Animals, and generally becomes fine-grained by the leaf level. There are 1000-3000 videos per class and approximately 5% of the videos are annotated with more than one class. The annotations are produced automatically by analyzing the text metadata surrounding the videos. Thus, the data is weakly annotated. Compared with the other datasets, although Sports-1M contains the largest video pool, since the video events are constrained within sports domain, it is not very popular used in general event detection evaluations.

FCVID: Fudan-Columbia Video dataset [Y.-G. Jiang and Chang, 2015] The newly released Fudan-Columbia Video Dataset (FCVID) contains $91,223$ Web videos annotated manually according to 239 categories. The categories in FCVID cover a wide range of topics like social events (e.g., "tailgate party"), procedural events (e.g., "making cake"), objects (e.g., "panda"), scenes (e.g., "beach"), etc. Categories were defined carefully and organized in a hierarchy of 11 high-level groups. In order to minimize subjectivity, multiple people were involved in both the category definition and the manual annotation processes.

Columbia EventNet Dataset [Guangnan Ye and Chang, 2015] In order to build a large scale event specific concept library that covers as many real-world events and their concepts as possible, EventNet was released recently, which contains $95,321$ videos over 500 events with $4,490$ event specific concepts attached to those events. They have built the largest event and concept ontology with well-organized hierarchical structures. Dramatic performance gains were reported by using this library in complex event detection especially for unseen novel events. Details of EventNet will be described in Chapter 4.

2.3 Benchmark Systems

Typical state-of-the-art event detection systems contain three basic module which are feature extraction module, classification module, and fusion module as shown in Figure 1.1. In this section we will summarize the benchmark systems in these three perspectives.

2.3.1 Feature Representation

Feature representation plays an important role in well performed event detection systems [Natarajan, 2011; Y.-G. Jiang and Chang, 2010]. Usually, top-ranked event detection systems used to combine various types of features from multiple modalities. For example, in the top-ranked MED 2010 system [Y.-G. Jiang and Chang, 2010], they extracted low-level features such as SIFT [Lowe, 2004], STIP [Laptev and Lindeberg, 2003], MFCC [Pols, 1966a]. In MED 2015 evaluation, more robust features are extracted, e.g., Opponent SIFT [Baptiste Mazin and Gousseau, 2012], GIST [Baptiste Mazin and Gousseau, 2012], LBP [Baptiste Mazin and Gousseau, 2012], dense trajectories with HOG, HOF, and MBH [H. Wang and Liu, 2011]. However, those low-level features are incapable of providing any interpretation or understanding of semantics presented in complex events. To this point, researches started to explore the mid-level semantic concept features. Briefly speaking, such method first define concept categories related to objects, scenes, actions etc., and train classifiers for each category. Then the confidence score of the presence of a concept can be considered as the mid-level features in supervised event modeling frameworks. Popular concept libraries include Classemes [L. Torresani and Fitzgibbon, 2010], ImageNet [J. Deng and Fei-Fei, 2009], ObjectBank [L.-J. Li and Xing, 2010], ActionBank [Sadanand and Corso, 2012], EventNet [Guangnan Ye and Chang, 2015](shown in Chapter 3), etc. Features from multiple modalities, e.g., MFCC [Pols, 1966a], ASR, OCR, are also considered for complimentary information. Instead of single modality feature, there are also a lot of efforts on constructing multi-modal feature representations. For example, in [W. Jiang and Loui, 2009], the authors proposed a joint audio-visual feature, called audio-visual atom which indicates an image region trajectory associated with both regional visual features and audio features. In [G. Ye and Chang, 2012](shown in Chapter 6), the authors proposed a joint audio-visual bi-modal representation, called bi-modal words, to represent joint audio-visual patterns in event videos. Except for that, features learned from deep learning models, e.g., the last

few layers of deep learning models learned over ImageNet $1K$ or $20K$ [A. Krizhevsky and Hinton, 2012] are considered as strong features for event detection recently.

2.3.2 Classification Method

Given robust feature representations, event recognition can be achieved by various types of classifiers. For example, event detection can be formulated as a one-versus-all manner based on various feature representations, where a two-class SVM is typically trained either with linear setting or kernelized setting (e.g., RBF kernel, Chi-square kernel, etc.). Beyond that, graphical models are proposed in order to deeply analyze the sequential video frames. For example, in [Natarajan P, 2008], an action is modeled by a transition HMM. Conolly proposed modeling and recognition of complex events using CRF [CI, 2007]. Recently, some researchers have borrowed the success of deep learning on the task of large scale image classification [A. Krizhevsky and Hinton, 2012], and applied CNN directly on the task of event detection and achieved promising results [Florian, 2014].

2.3.3 Fusion Method

With various output from multiple sources, a robust fusion method often produces better performance especially when the sources are from heterogenous domain with complimentary information [Y.-G. Jiang and Chang, 2010]. A popular feature combination strategy in computer vision is MKL [Bach *et al.*, 2004], which learns an optimized kernel combination and the associated classifier simultaneously. Varma *et al*. [Varma and Ray, 2007] used MKL to combine multiple features and achieved good results on image classification. Different from this line of research, numerous score late fusion methods in the literature are proposed which work by combining the confidence scores of the models obtained from different features. For example, Jain *et al*. [Jain *et al.*, 2005] transformed the confidence scores of multiple models into a normalized domain, and then combined the scores through a linear weighted combination. SIFT, STIP, and MFCC features were lately fused by Jiang et al. [Y.-G. Jiang and Chang, 2010] in their top ranked TRECVID 2010 MED system. In Chapter 7, we will introduce a novel late fusion method which not only achieves isotonicity but also removes the predictions errors made by the individual models.

Part II

Large Scale Video Event and Concept Ontology

Analysis and detection of complex events in videos require a semantic representation of the video content. Event-specific concepts are the semantic concepts specifically designed for the events of interest, which can be used as a mid-level representation of complex events in videos. Existing video semantic representation methods typically require users to pre-define an exhaustive concept lexicon and manually annotate the presence of the concepts in each video, which is infeasible for real-world video event detection problems. Moreover, such methods that focus only on defining event-specific concepts for a small number of pre-defined events cannot manage novel unseen events.

In this part, we first propose an automatic semantic concept discovery scheme by exploiting Internet images and their associated tags so that users no longer need to annotate the concepts for each video event (shown in Chapter 3). In particular, given a target event and its textual descriptions, we crawl a collection of images and their associated tags by performing a text-based image search using the noun and verb pairs extracted from the event's textual descriptions. The system first identifies the candidate concepts for an event by measuring whether a tag is a meaningful word and visually detectable. Then a concept visual model is built for each candidate concept using an SVM classifier with probabilistic output. Finally, the concept models are applied to generate concept based video representations.

In order to manage unseen events, we apply the automatic event-driven semantic concept discovery scheme to construct a large scale event-specific concept library that covers as many real-world events and their concepts as possible (shown in Chapter 4). In particular, we choose WikiHow, an online forum that contains a large number of how-to articles on human daily life events. We perform a coarse-to-fine event discovery process and discover 500 events from WikiHow articles. Then we use each event name as query to search YouTube and discover event-specific concepts from the tags of returned videos. After an automatic filter process, we end with $95,321$ videos and $4,490$ concepts. We train a *Convolutional Neural Network* (CNN) model on the $95,321$ videos over the 500 events, and use the model to extract deep learning features from the video content. With the learned deep learning feature, we train $4,490$ binary SVM classifiers as the event-specific concept library. The concepts and events are further organized in a hierarchical structure defined by WikiHow, and the resultant concept library is called *EventNet*. Finally, the EventNet concept library is used to generate concept based representation of event videos.

In the last chapter of this part (Chapter 5), we provide some applications for the proposed large scale video event and concept ontology. In particular, we present several novel functions of EventNet: 1) interactive ontology browsing, 2) semantic event search, and 3) tagging of user-loaded videos via open web interfaces. The system is the first (to the best of our knowledge) that allows users to explore rich hierarchical structures among video events, relationships between concepts and events, and automatic detection of events and concepts embedded in user-uploaded videos in a live fashion.

Chapter 3

Event Driven Semantic Concept Discovery

3.1 Introduction

Recognizing complex events from unconstrained videos has received increasing interest in multimedia information retrieval and computer vision research communities [Y.-G. Jiang and Shah, 2013; A. Tamrakar and Sawhney, 2012; J. Revaud and Jégou, 2013]. By definition, an event is a complex activity that involves people interacting with other people and/or objects under certain scene settings. Compared with human action recognition which focuses on simple primitives such as "jumping", "walking" and "running" [Liu and Shah, 2008; L. Laptev and Rozenfeld, 2008; H. Wang and Liu, 2011], event detection is more challenging because it has to manage unconstrained videos that contain various people and/or objects, complicated scenes, and their mutual interaction. For example, a video of "birthday party" could contain several atomic components, including objects such as "cake" and "candle", actions such as "dancing" and "hugging" as well as scenes such as "garden" and "living room".

Existing event detection works have proposed the use of raw audio-visual features fed into different sophisticated statistical learning frameworks, and have achieved satisfactory performance [Y.-G. Jiang and Chang, 2010; P. Natarajan and Zhuang, 2012]. However, these works are incapable of providing any interpretation or understanding of the abundant semantics present in a complex multimedia event. This hampers high-level event analysis and understanding, es-

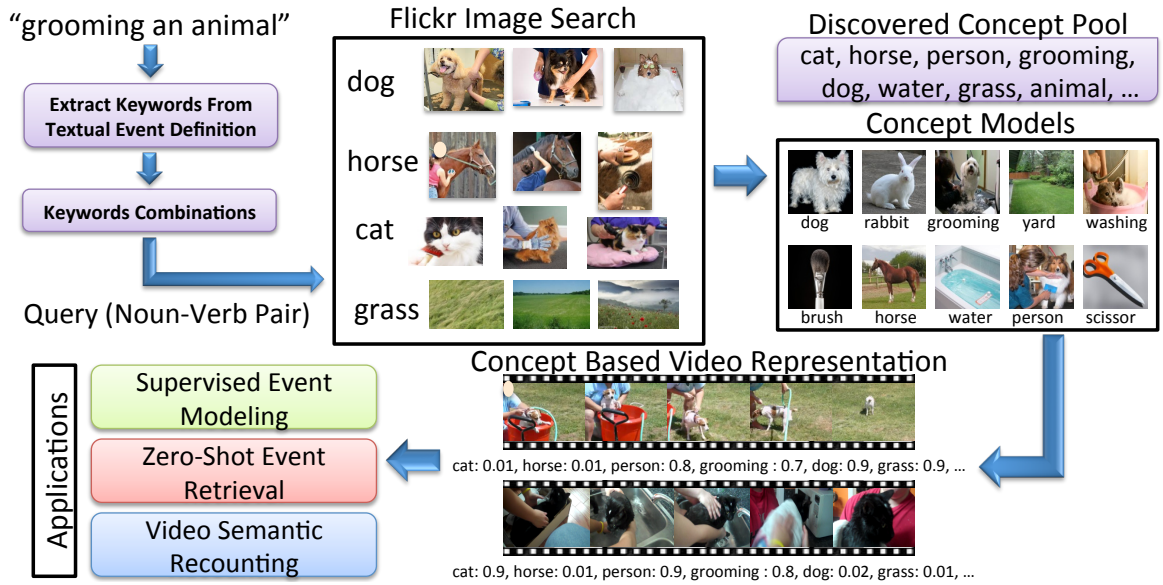


Figure 3.1: The framework of the proposed concept discovery approach. Given a target event (e.g., "grooming an animal") and its textual definition, we extract noun and verb keywords from the textual description of this event and use each combination of a noun and a verb as a textual query to crawl images and their associated tags from Flickr. We then discover potential concepts from the tags by considering their semantic meanings and visual detectabilities. Finally, we train a concept model for each concept based on the Flickr images annotated with the concept. Applying the concept models on the videos will generate concept-based video representations, which can be used in supervised event modeling over concept space, zero-shot event retrieval as well as semantic recounting of video contents.

pecially when the number of training videos is small or non-existent [J. Liu and Sawhney, 2012; M. Merler and Natsev, 2012]. Therefore, a logical and computationally tractable way is to represent a video that depicts a complex event in a semantic space that consists of semantic concepts related to objects, scenes, and actions, where each dimension measures the confidence score of the presence of a concept in the video. Once we have such concept-based video representations, we can use them as middle-level features in supervised event modeling, or directly use the scores of the semantic concepts to perform zero-shot event retrieval [J. Liu and Sawhney, 2012; M. Merler and Natsev, 2012].

One intuitive approach to generate concept-based video representation is to manually define a suitable concept lexicon for each event, followed by annotating the presence/absence of each concept in the videos [M. Mazloom and Snoek, 2013; J. Liu and Sawhney, 2012; M. Merler and Natsev, 2012]. This approach seemingly involves tremendous manual effort, and it is impractical for real-world event detection problems with regard to a huge number of videos. On the other hand, the web is a rich source of information with a huge number of images captured for various events under different conditions, and these images are often annotated with descriptive tags that indicate the semantics of the visual content. Our intuition is that the tags of Internet images related to a target event should reveal certain common semantics that appear in the event, and thus suggest the relevant concepts in the event videos. This stimulates a challenging research problem that has not been well studied, yet(to the best of our knowledge): given a target event (sometimes associated with a textual definition of the event, such as the textual event kits in TRECVID Multimedia Event Detection (MED) task [MED, 2010]), how do we automatically discover the relevant concepts from the tags of Internet images, and construct corresponding concept detection models specifically optimized for the target events? Because we focus on discovering concepts for pre-specified events, we term our approach as **Event-Driven Concept Discovery** in order to distinguish it from the work that builds a generic concept library that is independent of the targets [L. Torresani and Fitzgibbon, 2010]. Figure 3.1 illustrates the overall framework of the proposed system.

There are three main challenges in utilizing Internet images and their associated tags to learn concept models for complex event detection. First, because the tags associated with the Internet images are provided by general Internet users, there are often tags that are meaningless or irrelevant to the target event. To ensure the correctness of concept discovery, our method must choose semantic meaningful tags as the candidate concepts of the target event. To address this task, we perform noisy tag filtering by matching each tag to synsets in WordNet [Miller, 1995].

Moreover, some tags are abstract and not related to visual contents. For example, images with the tags ”economy” and ”science” do not show consistent visual patterns that can be effectively modeled by computer vision techniques. Therefore, we employ a visualness verification procedure to check whether a tag can be visually detected. Only visually related tags are kept as candidate concepts.

The last challenge is that the labels of Internet images are often very noisy. Directly adopting

such images as training samples for a concept can lead to a poor concept model. To solve this problem, we turn to a confidence ranking strategy. Given an image annotated with a concept, we first estimate the posterior probability of the concept's presence based on its visual closeness to other images annotated with the same concept. Then we rank all images based on the probabilities and choose the top ranked ones as the positive training samples of the concept. Such confidence ranking strategy is valuable in reducing the influence of noisy labels because it measures the confidence of the concept's presence in an image from its collective coherence with other images.

We demonstrate both qualitatively and quantitatively that the proposed concept discovery approach can generate accurate concept representations on event videos. By applying the concept scores as concept representations of videos, our method can achieve significant performance gains when evaluated over various semantic-based video understanding tasks including supervised event modeling and zero-shot event retrieval. One major contribution is that the concepts discovered based on the proposed method achieve significant performance gains (228% in zero-shot event retrieval) over concept pools constructed using other well known methods, such as Classemes and ImageNet. We also show that our discovered concepts outperform classic low-level features in supervised event modeling, and can reveal the semantics in a video over the semantic recounting task.

3.2 Related Work

Complex event detection in videos has been investigated in the literature. Duan *et al*. [L. Duan and Luo, 2010] proposed learning cross-domain video event classifiers from the mixture of target event videos and source web videos crawled from YouTube. Tang *et al*. [K. Tang and Koller, 2012] developed a large margin framework to exploit the latent temporal structure in event videos, and achieved good performance on event detection. Natarajan *et al*. [P. Natarajan and Zhuang, 2012] exploited multimodal feature fusion by combining low-level features and available spoken and videotext content associated with event videos. Ma *et al*. [Z. Ma and Hauptmann, 2013a] proposed adapting knowledge from other video resources to overcome the insufficiency of the training samples in small sample video event detection. However, these works focus on modeling events into sophisticated statistical models, and cannot reveal rich semantics in videos.

Some works attempted to accomplish event detection with concept based video representation-

s. Izadinia *et al*. [Izadinia and Shah, 2012] manually annotated a number of concepts on event videos, and proposed a discriminative model that treats the concepts as hidden variables and models the joint relationship among concepts in a concept co-occurrence graph. Liu *at al*. [J. Liu and Sawhney, 2012] observed concepts in event videos and defined a concept ontology that falls into "object", "scene", and "action", through which a number of SVM classifiers are trained as concept detectors to generate concept scores on videos. However, as mentioned before, all these methods require significant manual effort, that are inadequate for real-world event detection tasks with several videos. In [Z. Ma and Hauptmann, 2013b], Ma *et al*. leveraged the concepts contained in other video resources in order to assist detection in event videos, and proposed a joint learning model to learn concept classifier and event detector simultaneously. Mazloom *et al*. [M. Mazloom and Snoek, 2013] first constructed a concept library with $1,346$ concept detectors by mixing 346 manually defined concepts in TRECVID 2011 Semantic Indexing Task [Ayache and Quénot, 2008] and $1,000$ concepts from the ImageNet Large Scale Visual Recognition Challenge 2011 [J. Deng and Fei-Fei, 2009], and then discovered the optimal concept subset using a cross-entropy optimization. Nevertheless, the concepts in other video resources might not be relevant to the content of event videos, and could produce inaccurate semantic descriptions on videos. Yang *et al*. [Yang and Shah, 2012] adopted deep belief nets to learn cluster centers of video clips and treated them as data-driven concepts. Such data-driven concepts do not seem to convey any semantic information, and are not applicable for the semantic representation of videos. Contrary to these methods, we focus on the automatic discovery of semantic concepts in event videos by exploiting Internet images and their tags, which uncovers the semantics in videos without any manual labor.

Berg *et al*. [T. Berg and Shih, 2010] introduced a method for automatically discovering concepts by mining text and image data sampled from the Internet. A text string is recognized as a concept only if the visual recognition accuracy on its associated image is relatively high. Nevertheless, the method merely works on a closed web image set with surrounding text, and cannot be applied in the concept discovery of event videos, none of which does not contain any textual description. Yanai *et al*. [Yanai and Barnard, 2005] adopted a similar idea to discover visual related concepts associated with Internet images. Our work is also related to analyzing videos by leveraging still images. For example, Ikizler-Cinbis *et al*. [N. Ikizler-Cinbis and Sclaroff, 2009] proposed learning actions from the web, which collected images from the Web in order to learn representations of actions, and used

this knowledge to automatically annotate actions in videos. In contrast, we focus on automatically discovering concepts from still images and using them to interpret complex video semantics, which is more challenging than these prior works.

In terms of building a concept bank library, our work is also related to existing concept libraries, such as Object Bank [L.-J. Li and Xing, 2010], Classemes [L. Torresani and Fitzgibbon, 2010], and Action Bank [Sadanand and Corso, 2012]. However, these libraries are designed for generic objects or actions, and hence are not directly relevant to the target event collection at hand. On the contrary, our concept library is designed for a set of pre-specified events, that are more relevant to the target events and could precisely reveal the semantics of the event videos.

3.3 Discovering Candidate Concepts From Tags

In this section, we present a three-step procedure for discovering candidate concepts for each target event as follows:

Step I: Flickr Image Crawling. Given a target event and its textual event description, we can use NLTK [Bird, 2006] to extract the nouns and verbs from the event definition sentence. Then we combine a noun and a verb to form a "noun-verb pair" as a textual query to perform text-based image searches on Flickr. Finally, we download the retrieved images and associated tags for each query and combine them together as the concept discovery pool. Notably, the images retrieved this way have higher relevance to the target event (See Figure 3.6).

Step II. Noisy Concept Filtering. Given a target event, we can crawl a number of images and associated tags that belong to the same event from Flickr. As mentioned before, the tags are typically provided by general Internet users, and there are a sufficient amount of meaningless words that are irrelevant to the target event. To ensure that each tag corresponds to a meaningful concept, a tag filtering process is performed. In particular, we use WordNet [Miller, 1995] as the concept lexicon and look up each tag in it. If a tag is matched successfully to a synset in WordNet, it is regarded as a meaningful concept. Otherwise, it is removed as a noisy word.

Step III. Concept Visualness Verification. After the filtering process, the remaining tags are meaningful concepts. Nevertheless, we notice that some concepts are not visually related. For example, there might be some images associated with concept "economy", but there are no consistent

(a) Birthday party (b) Attempting a bike trick

Figure 3.2: Concept clouds discovered for event “birthday party” and “attempting a bike trick”. The size of each concept indicates its TF-IDF value.

visual patterns within these images because the concept is highly abstract. Involving such concepts can introduce significant distractions to the video representation and degrade the final performance. Therefore, we need to verify the visualness of each concept so that only visually related concepts are included. To accomplish this, we first treat the images associated with a concept as positive training samples and simultaneously choose the same number of images from other concepts as negative training samples. Then we divide all training images into two halves and perform a two-fold cross-validation.

Performance is measured based on Average Precision(AP). Finally, only concepts with high cross validation performance are verified as visually related concepts and retained in the concept library. In order to obtain reliable concept detectors trained with sufficient number of images, we further remove the concepts that contain less than 80 training images. We noticed that each discovered concept is discovered for a specific event, and thus we call the resulting concepts, “event-specific concepts”.

Based on the above strategy, we obtain the initial candidate concepts for each target event. Figure 3.2 shows the concept clouds discovered on two exemplary events, where the concept size indicates its TF-IDF value, and we treat a set of tags in an event as a document.

3.4 Building Concept Models

3.4.1 Training Image Selection for Each Discovered Concept

In this section, we present how to choose reliable training images for each discovered concept, and then introduce how to build the corresponding concept model.

To eliminate the noisy and outlier images crawled from the Internet, we use a confidence ranking method to choose reliable training images. Given a concept c, we can construct the following two image subsets $\mathcal{X}^+ = \{\mathbf{x}_i\}_{i=1}^{m}$ and $\mathcal{X}^- = \{\mathbf{x}_i\}_{i=m+1}^{m+n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of the i-th image with d being the feature dimensionality, $\mathcal{X}^+$ is a set of m images annotated with concept c, and $\mathcal{X}^-$ contains n images annotated without concept c. We adopt a soft neighbor assignment [J. Goldberger and Salakhutdinov, 2004] in the feature space to estimate the confidence of assigning the given concept to an image. In particular, each image $\mathbf{x}_i$ selects another image $\mathbf{x}_j$ as its neighbor with probability $p(\mathbf{x}_i, \mathbf{x}_j)$ and inherits its label from the image it selects. We define the probability $p(\mathbf{x}_i, \mathbf{x}_j)$ using a softmax operator over the entire image set $\mathcal{X} = \{\mathcal{X}^+, \mathcal{X}^-\}$:

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2)}{\sum_{\mathbf{x}_k \in \mathcal{X} \setminus \{\mathbf{x}_i\}} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2)}, \tag{3.1}$$

where $\|\cdot\|$ denotes the l_2 norm of a vector.

Based upon this stochastic selection rule, we can calculate the probability $p(\mathbf{x}_i)$ of image $\mathbf{x}_i$ being classified as positive with respect to the given concept:

$$p(c|\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathcal{X}^+} p(\mathbf{x}_i, \mathbf{x}_j). \tag{3.2}$$

In the above equation, the confidence score of a concept's presence in an image is measured based on its visual closeness with respect to other images in the same concept category. If an image has a noisy label, it tends to fall apart from the coherent visual pattern of the concept, leading to a small confidence value. On the contrary, images with correct labels always comply with the common visual pattern of the concept, and thus are assigned high confidence values. We use $p(c|\mathbf{x}_i)$ to estimate the confidence of image $\mathbf{x}_i$ belonging to concept c, and select s images with the highest confidence scores as the positive training images for each concept. In this work, we set $s = 200$[1],

[1]If the images annotated with a concept are fewer than 200, we directly utilize all images as positive samples for concept modeling.

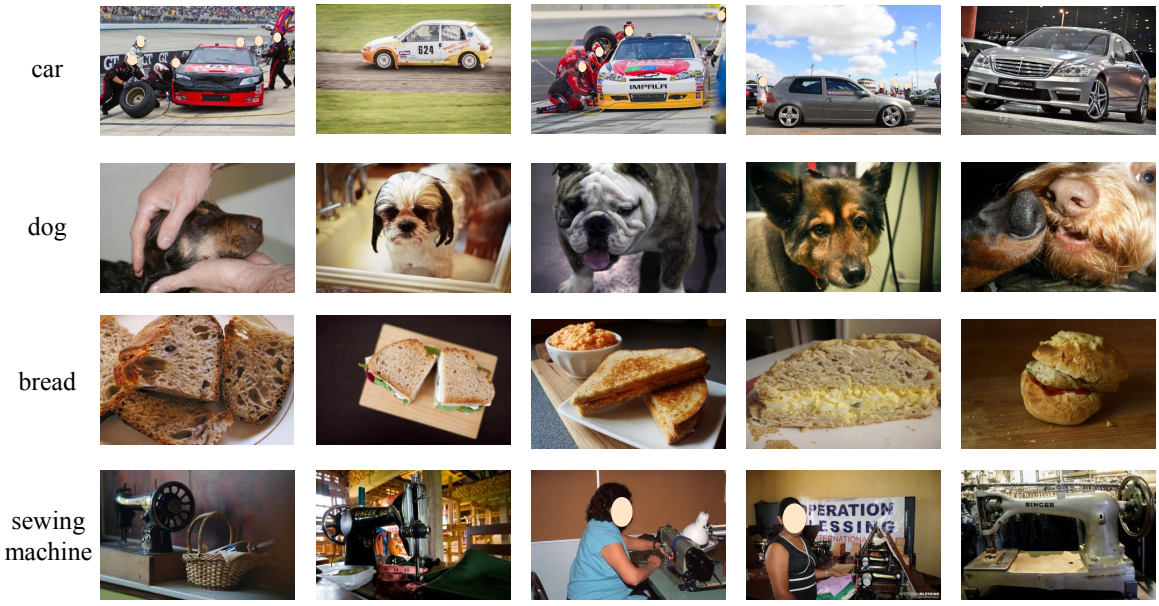


Figure 3.3: The top 5 images ranked by our method for some exemplary concepts.

and choose $t = 1,900$ negative images from other concepts as the negative training images. Figure 3.3 shows the top-5 images ranked by our method for some exemplary concepts. As can be seen, the selected images are highly relevant to the concepts while maintaining reasonable content diversity.

3.4.2 Concept Model Training

Given a concept c discovered for an event, suppose we have an Internet image collection $\mathcal{I} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{s+t}$, where the label $y_i \in \{-1, 1\}$ of each image $\mathbf{x}_i$ is determined by the confidence score ranking method described in Section 3.4.1. In this work, we choose a large margin SVM classifier with RBF kernel as our concept model, where the kernel function is defined as $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-d^2(\mathbf{x}_i, \mathbf{x}_j)/\sigma^2\right)$. Here $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and σ is the mean distance among all images on the training image set. We use the LibSVM library [Chang and Lin, 2011] as the implementation of our SVM concept model, and the optimal tradeoff parameter for SVM is determined via cross-validation.

The overall complexity of concept model training consists of two parts: training image selection and concept model training. In particular, the time complexity for choosing the training images de-

scribed in Section 3.4.1 is $\mathcal{O}\big(d(m+n)^2\big)$, where m and n are, respectively, the number of positive and negative images for each concept and d is the feature dimension. In our experiment implemented on a MATLAB platform on an Intel XeonX5660 workstation with 3.2 GHz CPU and 18 GB memory, a total of 13.58 seconds is required to finish the confidence score calculation when $m = 3000$, $n = 3000$, and $d = 2,659$. On the other hand, the time complexity for the concept model training described in Section 3.4.2 is $\mathcal{O}\big(d(s+t)^2 + (s+t)^3\big)$, which includes operations to compute the kernel and perform matrix inversion [Chang and Lin, 2011]. In our experiment with $s = 200$, $t = 1,900$ and $d = 2,659$, we finish the the training process of a concept model within 2 minutes on average. Considering the efficiency of the concept modeling process, our approach is applicable for constructing a large-scale concept library that consists of a huge number of concept models.

3.5 Video Event Detection with Discovered Concepts

After constructing the models for all concepts in an event, we apply them on the videos and adopt their probabilistic outputs as the concept-based representations, that can be used as an effective representation for semantic event analysis. In more detail, given a video clip in a target event, we can first generate the concept based representation on each video frame and then average them as the final concept representation of the video clip, or we can directly apply the concept models on the averaged feature of the frames in the video. The second approach significantly reduces the concept score generation time, and thus it is adopted as the concept-based video representation generation method in this work. There are typically two use scenarios to apply concept based video representations for complex event detection, as discussed below:

Scenario I: Supervised Event Modeling Over Concept Space. In this scenario, there is usually a number of labeled positive and negative training videos associated with a pre-specified event, and we regard the concept-based video representations as high-level video content descriptors in the concept space for training a classifier of the target event. Therefore, we expect the concept based video representation to be discriminative in order to easily separate the target event from other events. Given a pre-specified event detection task that consists of E events, we choose S concepts with the highest TF-IDF values for each E event from their respective discovered con-

cepts, and concatenate the concept scores into $E \times S$ dimensional feature vector (In this work, we set $E = 20$ and $S = 100$, and generate a $2,000$-dimensional concept-based video representation for this task). With the concept feature representation as input, we can train any supervised model as the event classifier. In this work, we choose binary SVM classifier with χ^2 kernel as our event detection model. In the test stage, we adopt SVM probabilistic output as the event detection score on each test video, through which the video retrieval list can be generated.

Scenario II: Zero-Shot Event Retrieval. In this scenario, we do not have any training videos of the target event, but only directly use the event name to retrieve relevant videos from the large video archive. Under this setting, the only available information is the concept scores on the test videos. We call this task zero-shot event retrieval because the procedure is purely semantic based. Given that each concept has different levels of semantic relevance with respect to the query event, we use a weighted summation strategy to calculate the detection score of each test video. Given an event name e comprised of multiple words, we use WordNet [Miller, 1995], a large lexical database of English words, to estimate the semantic similarity of two words. The semantic relevance $r(e, c)$ between event e and concept c is determined as the maximum semantic similarity between concept c and all words that appear in event name e. We use NLTK [Bird, 2006], a Python API for WordNet, to calculate the Wu-Palmer Similarity [Wu and Palmer, 1994] of two words. For an event and a test video with concept representation $\{s_1, \ldots, s_T\}$, where each concept score s_i corresponds to the event-specific concept c_i, the detection score of this video with respect to the target event can be estimated as $\sum_{i=1}^{T} r(e, c_i)s_i$. Finally, the event retrieval result can be generated by ranking the videos with these weighted summation scores.

3.6 Experiment

Our experiment aims to verify the effectiveness of our discovered concepts over the complex video event detection dataset. We begin with a description of the dataset, and then perform experiments that evaluate different aspects of our method.

3.6.1 Dataset and Feature Extraction

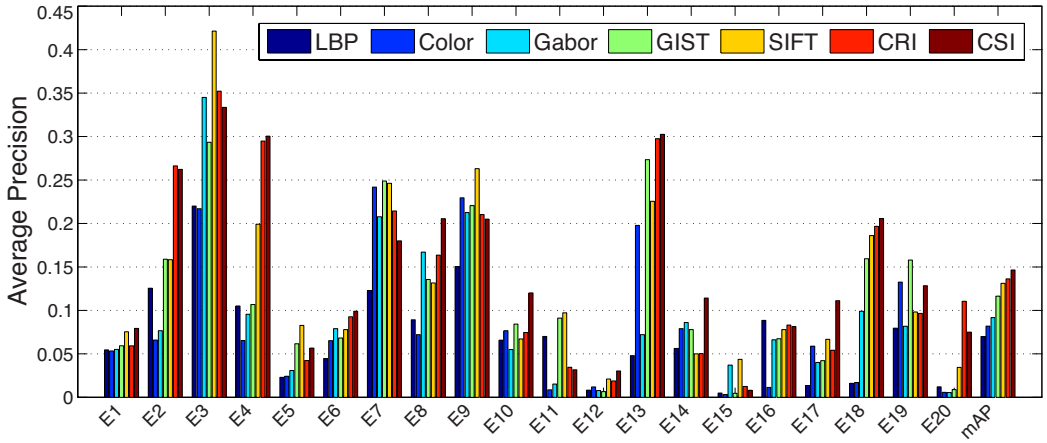
Event Video Set. The TRECVID MED 2013 pre-specified task consists of a collection of Internet videos collected by Linguistic Data Consortium from various Internet video hosting sites [MED, 2010]. The dataset contains $32,744$ videos that fall into 20 event categories and the background category. The names of these pre-specified 20 events are respectively "E1: *birthday party*", "E2: *changing a vehicle tire*", "E3: *flash mob gathering*", "E4: *getting a vehicle unstuck*", "E5: *grooming an animal*", "E6: *making a sandwich*", "E7: *parade*", "E8: *parkour*", "E9: *repair an appliance*", "E10: *working on a sewing project*", "E11: *attempting a bike trick*", "E12: *cleaning an appliance*", "E13: *dog show*", "E14: *giving directions to a location*", "E15: *marriage proposal*", "E16: *renovating a home*", "E17: *rock climbing*", "E18: *town hall meeting*", "E19: *winning a race without a vehicle*", and "E20: *working on a metal crafts project*". On this pre-specified event detection task in TRECVID MED 2013, each event is associated with a textual event kit that specifies the key concepts and detailed process of this event.

Internet Image Set. To collect Internet images, we utilize query words related to each event in order to perform image search on Flickr. For each event, we extract the keywords in the event textual kit, and then attempt different combinations of any two keywords (typically, one noun and one verb) to perform keyword-based image searches on *Flickr.com*, where we combine all images from different queries as the Internet images for this event. This way, we download $20,000$ images and their associated tags as the pool for concept discovery in each event. In particular, we perform the candidate concept discovery described in Section 3.3, and choose 100 potential concepts from the crawled image set for each event.

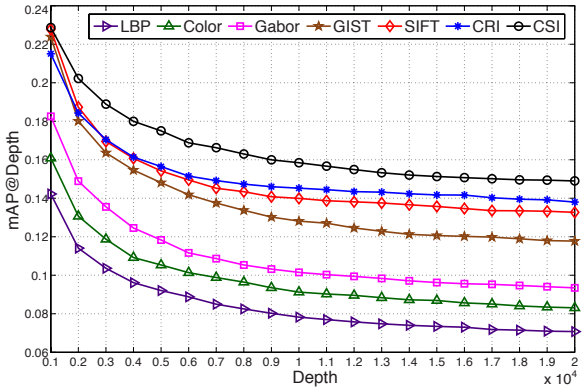
Feature Extraction. We extract the $2,659$-dimensional Classemes feature [L. Torresani and Fitzgibbon, 2010] as the feature representation of both Internet images and video frames. Given a video clip, we simply aggregate all frames of the Classemes feature as the video-level feature representation. Other codebook based features, such as SIFT BoW, can also be used as the alternatives in our work.

3.6.2 Supervised Event Modeling Over Concept Space

In this task, we treat the concept-based video representations as feature descriptors for supervised event modeling. We follow the pre-defined training ($7,787$ videos) and test ($24,957$ videos) data



(a) AP on each event

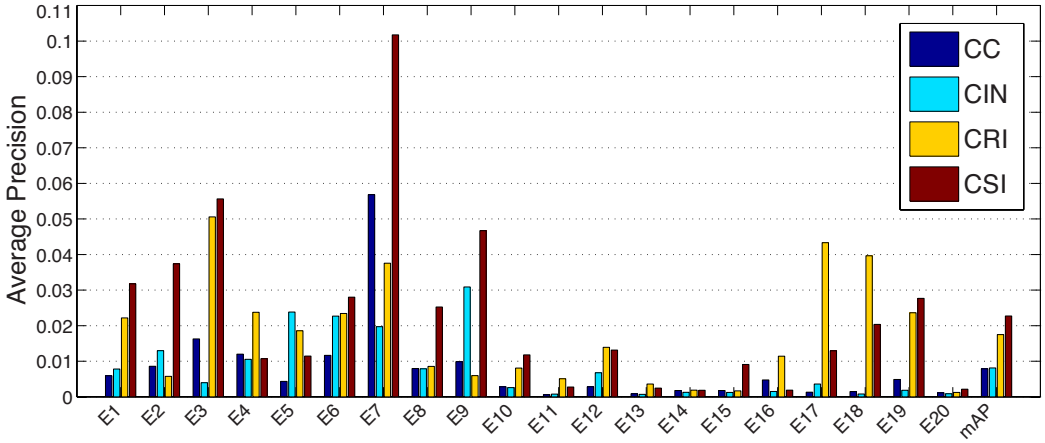


(b) mAP at variant depths

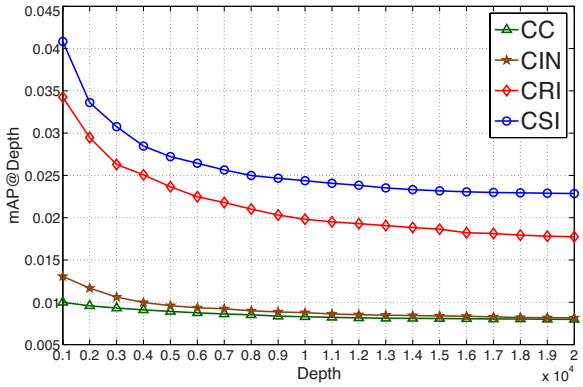
Figure 3.4: Performance of different methods on supervised event modeling task. CC: Classemes, CIN: ImageNet, CRI: proposed method without training image selection, CSI: proposed method with image selection. This figure is best viewed in color.

divided in the pre-specified EK100 task in TRECVID MED 2013 [MED, 2010] in our experiment. There are 100 positive training videos and approximately 50 negative training videos in each event category. Moreover, both training and test sets contain a significant number of background videos that do not belong to any target category, thus making the detection task very challenging. On each event, AP, which approximates the area under the precision/recall curve, is adopted as evaluation metric for event detection. Finally, we further calculate mean Average Precision (mAP) across all 20 events as the overall evaluation metric on the entire dataset.

To evaluate the effectiveness of our discovered concept features in supervised event modeling,



(a) AP on each event



(b) mAP at variant depths

Figure 3.5: Performance of different concept-based classifiers for zero-shot event detection task. This figure is best viewed in color.

we compare the following middle or low level feature representations: (1) **SIFT** [Lowe, 2004] BoW, (2) **GIST** [Oliva and Torralba, 2001] BoW, (3) **Gabor** [D. Field, 1987] BoW, (4) **LBP** [T. Ojala and Maenpaa, 2002] BoW, and (5) Transformed **Color** Distribution [K. Van De Sande and Snoek, 2010] BoW. All the above five descriptors are densely extracted on grids of 20×20 pixels with 50% overlap from images. For each type of extracted descriptor, we train a codebook with 400 codewords, and partition each image into 1×1 and 2×2 blocks for spatial pyramid matching [S. Lazebnik and Ponce, 2006]. Finally, we adopt soft quantization [J. van Gemert and Geusebroek, 2010] to represent each image as a $2,000$-dimensional histogram, with the same dimension as our concept-based video representation to ensure a fair comparison. (6) Concepts learned from Random

Images (**CRI**). Concept models are learned using a randomly chosen subset of images associated with the concept tag directly without content consistency filtering, as described in Section 3.4.1. (7) Our proposed $2,000$ Concepts learned from Selected Images (**CSI**). We use our method to select reliable training images for concept modeling.

Figure 3.4 illustrates the performance of all the methods on this task quantitatively. From the results, we obtain the following observations: (1) the concepts generated by our CSI method consistently ourperform the other methods by a large margin, which demonstrates its effectiveness in concept-based video representation. On some events where our method is inferior to other baselines, we observe significant domain difference between Flickr images and MED videos. We will address this issue by exploring cross domain adaptation methods in our future work. (2) The CSI method outperforms the five types of low-level features, which implies that our discovered concepts can not only reveal the semantic concepts, but also be utilized as an effective feature description for event discrimination. (3) The CSI method performs significantly better than the CRI method. This is because the former leverages more reliable training images than the random images utilized in the latter. This verifies the soundness of our confidence ranking method in selecting clean training images for concept modeling. Figure 3.4(b) shows the mAP comparisons at varied returned depths (i.e., the number of top ranked test videos included in the result evaluation). From the results, we can see that our method achieves significant and consistent mAP improvements over the other methods at varied returned depths.

3.6.3 Zero-Shot Event Retrieval

In this task, we directly apply the concept scores to rank the videos in the archive without leveraging any training video samples. Similar to Section 3.6.2, we adopt AP and mAP as our evaluation metrics. The focus of this experiment is to reveal the effectiveness of the discovered semantic concepts compared with those discovered from other methods. To this end, the following concepts generated from different methods are compared: (1) Classemes Concepts (**CC**). We extract the $2,659$-dimensional CC feature from the videos. Given each event name, we use the WordNet Wu-Palmer semantic similarity between event and concept names (see Section 3.5) in Classemes to find the 100 most relevant CC for this event. (2) Concepts discovered from ImageNet (**CIN**) [J. Deng and Fei-Fei, 2009]. In this method, we want to discover concepts for each event from all the concepts

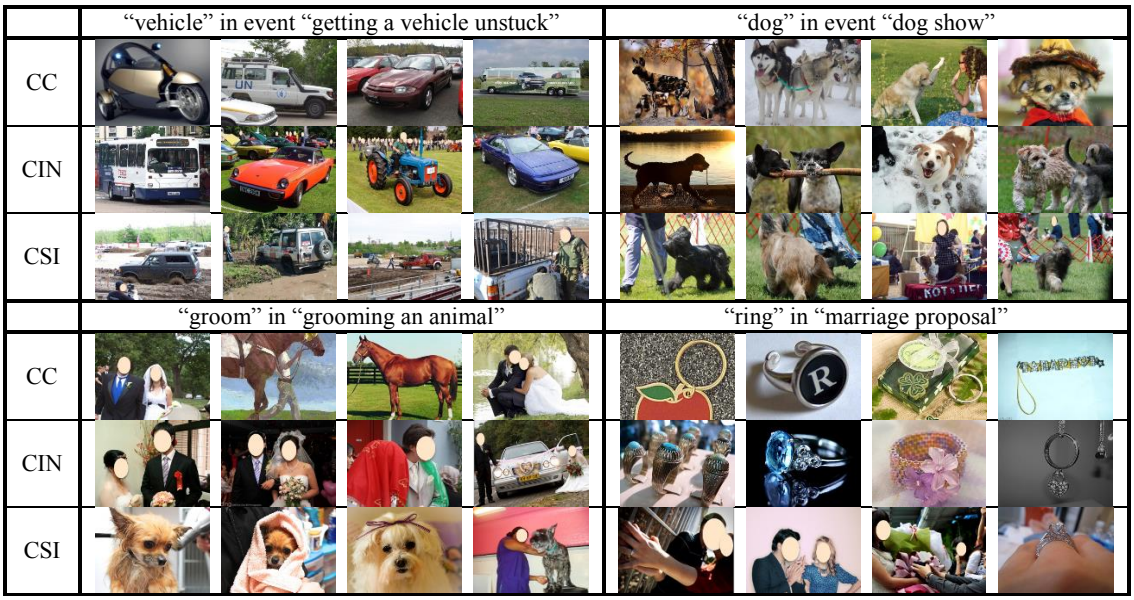


Figure 3.6: Concept training images for different concept generation methods. Note that the training images utilized in our method (CSI) not only contain the concept but also convey context information about the event.

in ImageNet. For each event name, we also adopt WordNet to calculate the semantic similarity between its event and concept names in ImageNet, and choose the same number of concepts (100 for each event) from ImageNet . For each concept, we choose 200 images from its ImageNet synset as the positive training images, and $1,900$ images from other discovered ImageNet concepts as the negative training images. These images are then fed into an SVM classifier as the concept model. (3) Concepts learned from Random Images (**CRI**), where we randomly select the same number of images as were used in our method for training images for concept modeling. (4) Our proposed Concepts learned from Selected Images (**CSI**).

Figure 3.5(a) shows the per-event performance of all methods. In Figure 3.5(b), we further plot mAP at different returned depths for different comparison methods. From the results, we obtain the following observations: (1) our CSI method achieves the best performance (with a relative performance gain as high as 228%). over most events. Becuase the task is purely semantic-based, the results clearly verify that the concepts discovered by our method are applicable to semantic-based event retrieval. (2) The zero-shot event retrieval performance is worse than the supervised event modeling. This is because the latter uses training videos to obtain a more sophisticated event

E4 Getting A Vehicle Unstuck: blizzard, snow, road, stuck, truck

E6 Making A Sandwich: recipe, cheese, bread, baking, chocolate

E8 Parkour: jump, slacklining, free run, acrobatic, gap

E11 Attempting A Bike Trick: boy, park, BMX, ride, trick

E19 Wining A Race Without A Vehicle: speed, hurdle, athlete, track, stadium

Figure 3.7: Semantic recounting examples on videos from some exemplary events in TRECVID MED 2013: each of the 5 rows shows evenly subsampled frames of an example video and the top 5 relevant concepts detected in the video.

model, whereas the former is merely based on concept score aggregation. (3) Our CSI method performs much better than Classemes, becuase most concepts in Classemes are irrelevant to the events. (4) The CSI method clearly outperforms the CIN method. This shows that our concept discovery method can obtain more accurate concept representations for the events than the concept representations discovered from other lexicon such as ImageNet. The reason might be two-fold: first, the concepts discovered from Flickr images are more relevant to real-world events than the general concepts in ImageNet. In fact, our discovered concepts are the most semantically relevant to the events compared with the concepts discovered from Classemes and ImageNet, as indicated in Table 3.1. Second, the Flickr images used to train our concept model contain visual clues of the event, whereas the images in ImageNet usually contain clean objects without event backgrounds. Figure 3.6 illustrates the training images from Flickr and ImageNet for some concepts. We can see that for the concept "groom" in event "grooming an animal", because of the ambiguity of the word "groom", the training images from CC and CIN generally refer to "bridegroom" whereas our concept refers to the action of "grooming an animal". Another example is the concept "ring" in event "marriage proposal", where the training images from CC and CIN are general "ring" with simple backgrounds and do not contain any context information about marriage proposals.

3.6.4 Human Evaluation

In addition to the performance comparison described in the previous sections, we also design an experiment to ask human judges to evaluate the quality of concepts discovered from different sources including Classemes, ImageNet and Flickr. The details of the experiment are illustrated in Figure 3.8.

First, we randomly select one event from the 20 pre-specified events in TRECVID MED 2013, and generate 100 concepts from each of the three sources (Flickr, ImageNet, Classemes) using the approach described in Section 3.6.3. In this step, we rank the 100 concepts from each source in descending order based on the WordNet semantic similarity between concept and event name. Second, from each of the three ranked concept lists, we randomly choose five concepts at the same rank positions, and form three concept subgroups, each of which contains the sampled five concepts. Third, we randomly select two concept subgroups from the three, and obtain a pairwise concept component that is, sent to a user in order to determine which five concepts are more relevant to the

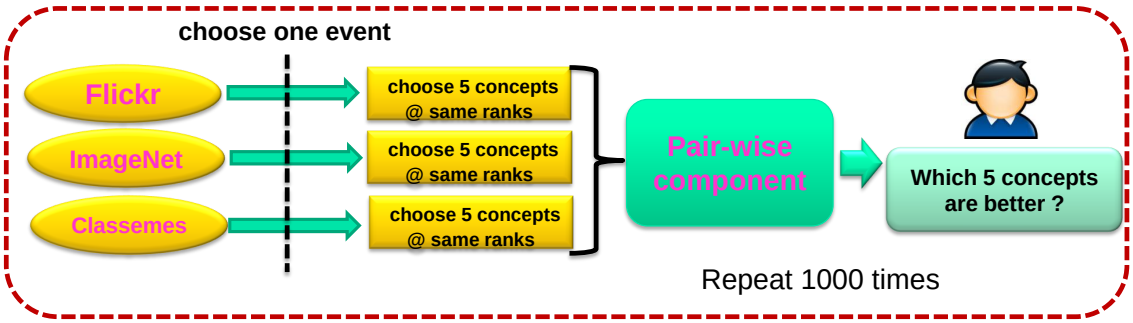


Figure 3.8: Human Evaluation on Concept Relevance.

Event Name		Concepts Discovered via Different Methods
getting a vehicle unstuck	CC	air transportation vehicle, all terrain vehicle, amphibious vehicle, armed person, armored fighting vehicle, armored recovery vehicle, armored vehicle, armored vehicle heavy, armored vehicle light, command vehicle
	CIN	vehicle, bumper car, craft, military vehicle, rocket, skibob, sled, steamroller, wheeled vehicle, conveyance
	CSI	**tire, car, snow, stick, stuck, winter, vehicle, truck, night, blizzard**
grooming an animal	CC	adult animal, animal, animal activity, animal blo, animal body part, animal body region, animal cage, animal container, animal pen, animal shelter
	CIN	groom, animal, invertebrate, homeotherm, work animal, darter, range animal, creepy-crawly, domestic animal, molter
	CSI	**dog, pet, grooming, cat, animal, bath, cute, canine, puppy, water**
making a sandwich	CC	baking dish, cafe place, classroom setting, collection display setting, cutting device, dish drying rack, food utensil, hair cutting razor, hdtv set, hole making tool
	CIN	sandwich, open-face sandwich, butty, reuben, ham sandwich, gyro, chicken sandwich, hotdog, club sandwich, wrap
	CSI	**sandwich, food, bread, cooking, cheese, spice, baking, pan, kitchen, breakfast**
working on a sewing project	CC	clothes iron, landing craft, laundry room, living room, missile armed craft, multi room unit, work environment, work station, steel mill worker, carpentry tool
	CIN	sport, outdoor game, rowing, funambulism, judo, blood sport, gymnastics, water sport, track and field, outdoor sport
	CSI	**sewing, handmade, embroidery, craft, quilt, fabric, hand, sewing machine, textile, thread**
cleaning an appliance	CC	action on object, animal container, armed person, art object, back yard, bag, bilateral object, box the container, butcher shop, capsule container
	CIN	appliance, gadgetry, gimbal, injector, mod con, device, musical instrument, acoustic device, adapter, afterburn
	CSI	**kitchen, furniture, washing, bed, sink, divan, spring bed, cleaning, stove, dishwasher**
rock climbing	CC	astronomical observatory building, attached body part, auto part, bar building, body movement event, body of water, building, building cluster, building security system, cavity with walls
	CIN	rock, uphill, outcrop, whinstone, xenolith, tor, slope, ptyalith, kidney stone, urolith
	CSI	**climbing, rock climbing, bouldering, mountain, sport, hiking, climber, landscape, peak, rope**

Table 3.1: Top concepts for different concept discovery methods.

test event name.

The above procedure is repeated $1,000$ times, through which we obtain the following statistics about concept quality. (1) Flickr is better than others with 81.29% probability. (2) ImageNet is better than others with 34.19% probability. (3) Classemes is better than others with a 32.46% probability. From these results, we can see that our discovered concepts from Flickr comply with human knowledge, which further verifies the effectiveness of our concept discovery method.

3.6.5 Video Semantic Recounting

For each video of a target event, we rank all the concepts discovered for the event based on their confidence scores, and treat the top-ranked concepts as the semantic description of the video content. Such a procedure can reveal the semantic information contained in a video, and it is thus called video semantic recounting. Figure 3.7 shows the recounting results on videos from some exemplary events in TRECVID MED 2013, where the top-5 ranked concepts generated by our method are selected as concepts for each video. As can be seen, these concepts reveal the semantics contained in the videos, which verifies the effectiveness of our discovered concepts in representing video semantics.

3.7 Summary and Discussion

In this chapter, we introduced an automatic event driven concept discovery method for semantic based video event detection. Given a target event, we crawl a collection of Flickr images and their associated tags related to this event as the concept discovery knowledge pool. Our method first estimates the candidate concepts present in the event by measuring the visualness of each concept and its semantic meaning. Then a concept model is constructed for each candidate concept based on a large margin SVM classifier. Finally, the individual concept models are applied on the event videos to generate concept-based video representations. We tested our discovered concepts over two video event detection tasks including supervised event modeling over concept space and zero-shot event retrieval, and the promising experiment results demonstrated the effectiveness of the proposed event-driven concept discovery method.

Although the proposed event-driven concept discovery method achieved good performance, it can only manage a small number of target events whose definitions are known in advance. When a novel unseen event emerges, it is no longer applicable because of the lack of relevant concepts for the unseen event. In the next chapter, we apply a similar event-driven concept discovery method and target to manage novel unseen events. We build a large scale event specific concept library that covers as many real-world events and their concepts as possible, and call it *EventNet*. Then, when novel unseen events occur, we can predict them with sufficient event-driven concepts properly.

Chapter 4

Large-Scale Structured Concept Library for Complex Event Detection in Video

4.1 Introduction

In the previous chapter, we introduced an automatic event driven concept discovery method for semantic based video event detection. Although the proposed approach achieves promising performance, as mentioned previously, the method can only manage a small number of target events whose definitions are known in advance. When a novel unseen event emerges, it is no longer applicable because of the lack of relevant concepts for the unseen event.

To address this problem, this chapter proposes the construction of a large-scale event-driven concept library that covers as many real-world events and concepts as possible. Figure 4.1 illustrates the overall framework of the proposed method, where we highlight the two main challenges addressed in this thesis. The first is how to define events and their relevant concepts in order to construct a comprehensive concept library. To achieve this goal, we resort to the external knowledge base called WikiHow [Wik, 2015], a collaborative forum that aims to build the world's largest manual for human daily life events. We define 500 events from the articles of the WikiHow website. For each event, we use its name as query keywords to perform text-based video search on YouTube, and apply our automatic concept discovery method to discover event-specific concepts from the tags of the returned videos. Then we crawl videos associated with each discovered concept as the training videos to learn deep learning video features using CNN and event-specific concept models. This

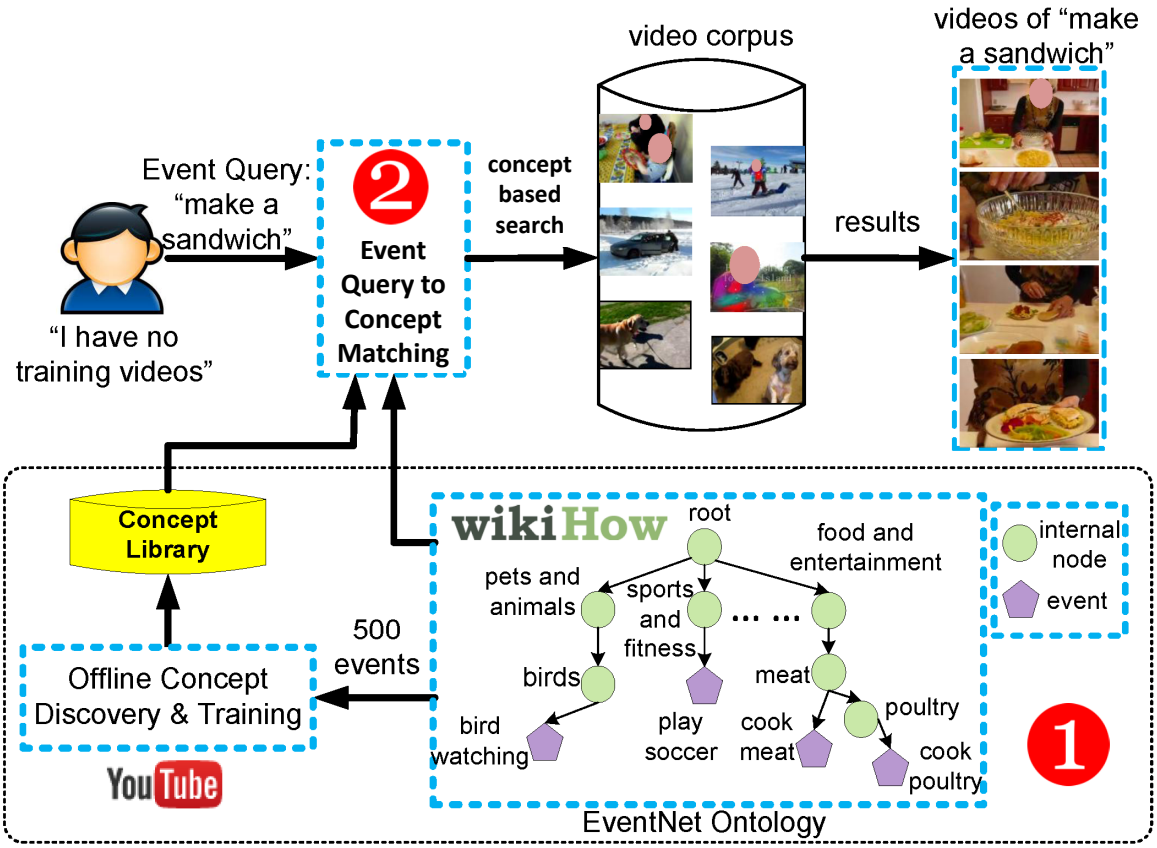


Figure 4.1: Concept based event retrieval by the proposed large scale structured concept library *EventNet*. We propose two unique contributions: (1) A large scale structural event ontology. (2) Effective event-to-concept mapping via the ontology.

leads to an event-specific concept library composed of $4,490$ concept models trained over $95,321$ YouTube videos. We further organize all events and their associated event-specific concepts into a hierarchical structure defined by WikiHow, and call the resulting concept library *EventNet*.

The second challenge is how to find semantically relevant concepts from EventNet that can be used to search video corpus to answer a new event query. The existing methods address this by calculating the semantic similarity between the event query and candidate concepts, and then select the top ranked concepts with the highest similarities [J. Chen and Chang, 2014; Y. Cui and Chang, 2014; S. Wu and Natarajan, 2014]. However, considering that our concepts are event-specific, each concept is associated with a specific event that can be used as contextual information in measuring the similarity between the query event and the concept. Moreover, because of the short text of event

names that contain only very few text words, direct measurement of semantic similarity might not be able to accurately estimate semantic relevance, and the concept matching results could become quite unsatisfactory even when EventNet library does contain relevant events and concepts. To solve these issues, we propose a cascaded concept matching method that first matches relevant events and then finds relevant concepts specific to the matched events. For the queries that cannot be well answered by automatic semantic similarity calculation, we propose to leverage the hierarchical structure of EventNet and allow users to manually specify the appropriate high-level category[1] in the EventNet tree structure, and then only perform concept matching under the specified category (cf. Section 4.7)

We demonstrate that the proposed EventNet concept library leads to dramatic performance gains in concept-based event detection over various benchmark video event datasets. In particular, it outperforms the $20K$ concepts generated from the state-of-the-art deep learning system trained on ImageNet [A. Krizhevsky and Hinton, 2012] by 207% in zero-shot event retrieval. We also show that EventNet can detect and recount the semantic cues that indicate the occurrence of an event video. Finally, the video corpus in EventNet can be used as a comprehensive benchmark video event dataset. The browser of the EventNet ontology and the downloading information of the models and video data can be found at `http://eventnet.ee.columbia.edu`.

We summarize our major contributions as follows: (1) a systematic framework for discovering several events related to human events (Section 4.3). (2) Construction of the largest ontology, including 500 complex events and $4,490$ event-specific concepts (Section 4.4 and 4.6). (3) Rigorous analysis of the properties of the constructed ontology (Section 4.5). (4) Dramatic performance gains in complex event detection especially for unseen novel events (Task I in Section 4.8). (5) The benefit of the proposed ontology structure in semantic recounting (Task II in Section 4.8) and concept matching (Task III in Section 4.8). (6) A benchmark event video dataset for advancing large scale event detection (Task IV in Section 4.8).

[1]As shown in Figure 4.1, the category nodes in EventNet are high-level categories in WikiHow used to organize articles into a hierarchy, such as “pets and animals”, “sports and fitness”, and more.

4.2 Related Work

There are some recent works that focus on detecting video events using concept-based representations. For example, Wu *et al*. [S. Wu and Natarajan, 2014] mined concepts from the free-form text descriptions of TRECVID research video set, and applied them as weak concepts of the events in TRECVID MED task. As mentioned earlier, these concepts are not specifically designed for events, and may not capture well the semantics of event videos.

Recent research also attempted to define event-driven concepts for event detection. Liu *et al*. [J. Liu and Sawhney, 2012] proposed to manually annotate related concepts in event videos, and build concept models with the annotated video frames. Chen *et al*. [J. Chen and Chang, 2014] proposed discovering event-driven concepts from the tags of Flickr images crawled using keywords of the events of interest. This method can find relevant concepts for each event and achieves good performance in various event detection tasks. Despite such promising properties, it relies heavily on prior knowledge about the target events, and therefore cannot manage novel unknown events that might emerge at a later time. Our EventNet library attempts to address this deficiency by exploring a large number of events and their related concepts from external knowledge resources, WikiHow and YouTube. A related prior work [Y. Cui and Chang, 2014] tried to define several events and discover concepts using the tags of Flickr images. However, as our later experiment shows, concept models trained with Flickr images cannot generalize well to event videos because of the well-known cross-domain data variation [K. Saenko and Darrell, 2010]. In contrast, our method discovers concepts and trains models based on YouTube videos, which more accurately capture the semantic concepts that underlie the content of user generated videos.

The proposed EventNet also introduces a benchmark video dataset for large scale video event detection. Current event detection benchmarks typically contain only a small number of events. For example, in the well known TRECVID MED task [MED, 2010], significant effort has been made to develop an event video dataset that contains 48 events. Columbia Consumer Video (CCV) dataset [Y.-G. Jiang and Loui, 2011] contains $9,317$ videos of 20 events. Such event categories might also suffer from data bias, and thus fail to provide general models applicable to unconstrained real-world events. In contrast, EventNet contains 500 event categories and $95K$ videos, which covers different aspects of human daily life and is believed to be the largest event dataset currently. Another recent effort also attempts to build a large scale structured event video dataset that con-

Dataset	EM	PM	RE	NM	Total Class #
MED 10-14 [MED, 2010]	16	17	15	0	48
CCV [Y.-G. Jiang and Loui, 2011]	6	5	8	1	20
Hollywood [L. Laptev and Rozenfeld, 2008]	6	0	1	0	7
KTH [Laptev and Lindeberg, 2003]	4	1	1	0	6
UCF101 [K. Soomro and Shah, 2012]	58	11	20	12	101
Matched Class #	90	34	45	13	182

Table 4.1: The matching results between WikiHow articles and event classes in the popular event video datasets, where "EM" denotes "Exact Match", "PM" denotes "Partial Match", "RE" denotes "Relevant" and "NM" denotes "No Match".

tains 239 events [Y.-G. Jiang and Chang, 2015]. However, it does not provide semantic concepts associated with specific events, such as those defined in EventNet.

4.3 Choosing WikiHow as EventNet Ontology

A key issue in constructing a large-scale event-driven concept library is to define an ontology that covers as many real-world events as possible. For this, we resort to the Internet knowledge bases constructed from crowd intelligence as our ontology definition resources. In particular, WikiHow is an online forum that contains several how-to manuals on every aspect of human daily life events, where a user can submit an article that describes how to accomplish given tasks such as "how to bake sweet potatoes", "how to remove tree stumps", and more. We choose WikiHow as our event ontology definition resource for the following reasons:

Coverage of WikiHow Articles. WikiHow has a good coverage over different aspects of human daily life events. As of February 2015, it included over $300K$ how-to articles [Wik, 2015], among which some are well-defined video events[2] that can be detected by computer vision techniques, whereas others such as "how to think" or "how to apply for a passport", do not have suitable corresponding video events. We expect a comprehensive coverage of video events from such a

[2]We define an event as a video event when it satisfies the event definition in NIST TRECVID MED evaluation, i.e., a complicated human activity that interacts with people/object in a certain scene.

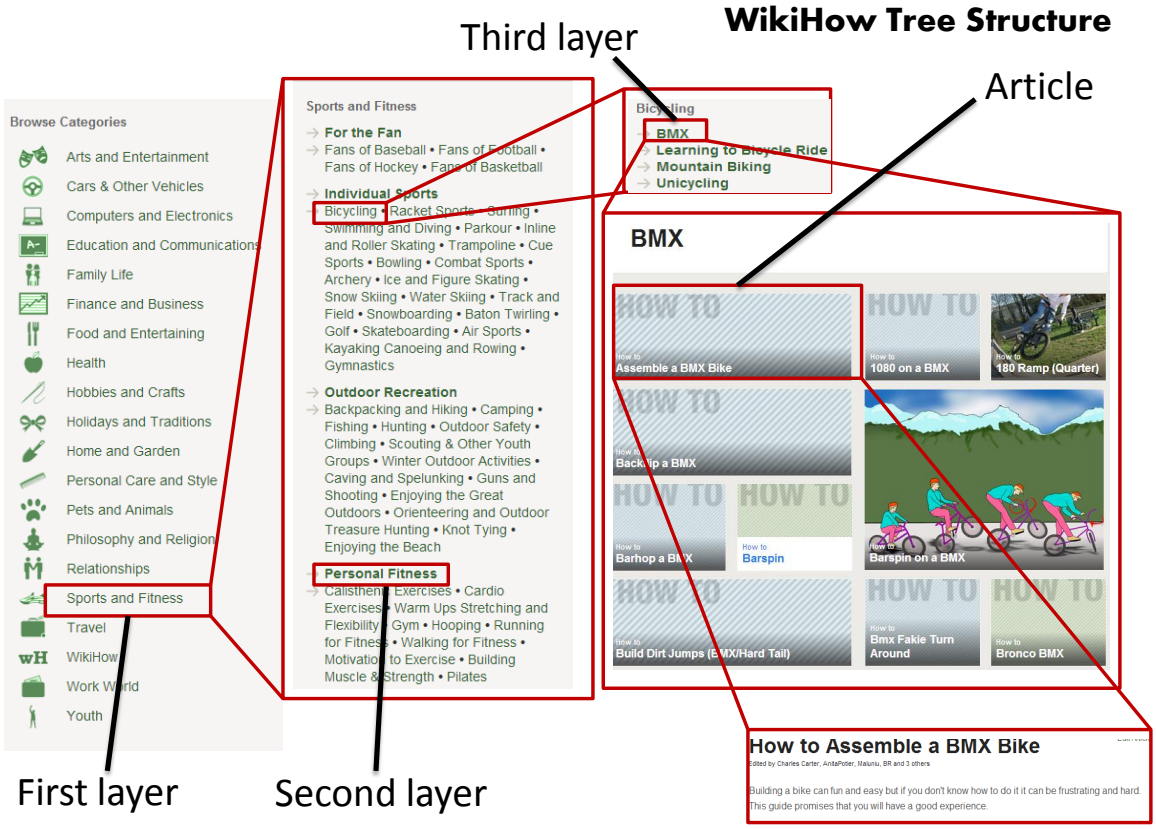


Figure 4.2: The hierarchial structure of WikiHow.

massive number of articles created by the crowdsourcing knowledge from Internet users.

To verify that WikiHow articles have a good coverage of video events, we conduct a study to test whether WikiHow articles contain events in the existing popular event video datasets in the computer vision and multimedia fields. To this end, we choose the event classes in the following datasets: TRECVID MED 2010-2014 (48 classes) [MED, 2010], CCV (20 classes) [Y.-G. Jiang and Loui, 2011], UCF 101 (101 classes) [K. Soomro and Shah, 2012], Hollywood movies (7 classes) [L. Laptev and Rozenfeld, 2008], KTH (6 classes) [Laptev and Lindeberg, 2003]. Then, we use each event class name as a text query to search WikiHow and examine the top-10 returned articles, from which we manually select the most relevant article title as the matching result. We define four matching levels to measure the matching quality. The first is *exact matching*, where the matched article title and event query are exactly matched (e.g., "clap hands" as a matched result to the query "hand clapping"). The second is *partial match*, where the matched article discusses a certain aspect

of the query (e.g., "make a chocolate cake" as a result to the query "make a cake"). The third case is *relevant*, where the matched article is semantically relevant to the query (e.g., "get your car out of the snow" as a result to the query "getting a vehicle unstuck"). The fourth case is *no match*, where we cannot find any relevant articles about the query. The matching statistics are listed in Table 4.1. If we count the first three types of matching as successful cases, the coverage rate of WikiHow over these event classes is as high as $169/182 = 93\%$, which confirms the potential of discovering video events from WikiHow articles.

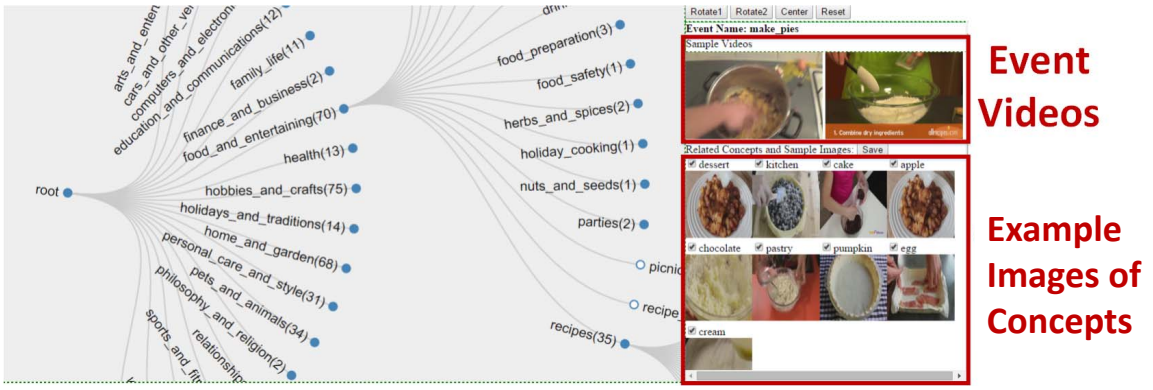


Figure 4.3: Event and concept browser for the proposed EventNet ontology. The hierarchical structure is shown on the left and the example videos and relevant concepts of each specific event are shown to the right.

Hierarchical Structure of WikiHow. WikiHow categorizes all its articles into $2,803$ categories and further organizes all categories into a hierarchical tree structure. Each category contains a number of articles that discuss different aspects of the category, and is associated with a node in the WikiHow hierarchy. As shown in Figure 4.2 of the WikiHow hierarchy, the first layer contains 19 high-level nodes that range from "arts and entertainment", "sports and fitness" to "pets and animal". Each node is further divided into a number of children nodes that are subclasses or facets of the parent node, with the deepest path from the root to the leaf node containing seven levels. Although such a hierarchy is not based on lexical knowledge, it summarizes humans' common practice of organizing daily life events. Typically, a parent category node includes articles that are more generic than those in its children nodes. Therefore, the events that reside along similar path in the WikiHow tree hierarchy are highly relevant (cf. Section 4.4). Such hierarchical structure helps users quickly localize the potential search area in the hierarchy for a specific query in which he/she is interest-

ed, and thus improves concept matching accuracy (cf. Section 4.7). In addition, such hierarchical structure also enhances event detection performance by leveraging the detection result of an event in a parent node to boost detection of the events in its children nodes, and vice versa. Finally, such hierarchical structure also allows us to develop an intuitive browsing interface for event navigation and event detection result visualization [H. Xu and Chang, 2015], as shown in Figure 4.3.

4.4 Constructing EventNet

In this section, we describe the procedure used to construct EventNet, including how to define video events from WikiHow articles and discover event specific concepts for each event from the tags of YouTube videos.

4.4.1 Discovering Events

First we aim to discover potential video events from WikiHow articles. Intuitively, this can be done by crawling videos using each article title and then applying the automatic verification technique proposed in [T. Berg and Shih, 2010; J. Chen and Chang, 2014] to determine whether an article corresponds to a video event. However, considering that there are $300K$ articles on WikiHow, this requires a massive amount of data crawling and video processing, thus making it computationally infeasible. For this, we propose a coarse-to-fine event selection approach. The basic idea is to first prune WikiHow categories that do not correspond to video events, and then select one representative event from the article titles within each of the remaining categories. In the following, we describe the event selection procedure in detail.

Step I: WikiHow Category Pruning. Recall that WikiHow contains $2,803$ categories, each of which contains a number of articles about the category. We observe that many of the categories refer to personal experiences and suggestions, that do not correspond to video events. For example, the articles in the category "Living Overseas" refer to how to improve the living experience in a foreign country, and do not satisfy the definition of video event. Therefore, we want to find such event irrelevant categories and directly filter their articles, in order to significantly prune the number of articles to be verified in the next stage. To this end, we analyze $2,803$ WikiHow categories and manually remove those that are irrelevant to video events. A category is deemed as event

irrelevant when it cannot be visually described by a video, and none of its articles contain any video events. For example, "Living Overseas" is an event-irrelevant category because "Living Overseas" is not visually observable in videos and none of its articles are events. On the other hand, although the category "Science" cannot be visually described in a video because of its abstract meaning, it contains some instructional articles that correspond to video events, such as "Make Hot Ice", and "Use a Microscope". As a result, in our manual pruning procedure, we first find a to-be-pruned category name and then carefully review their articles before deciding to remove the category.

Step II: Category-based Event Selection. After category pruning, only event relevant categories and their articles remain. Under each category, there are still several articles that do not correspond to events. Our final goal is to find all video events from these articles and include them into our event collection, which is a long-term goal of the EventNet project. In the current version, EventNet only includes one representative video event from each category of WikiHow ontology. An article title is considered to be a video event when it satisfies the following four conditions: (1) it defines an event that involves a human activity interacting with people/objects in a certain scene. (2) It has concrete non-subjective meanings. For example, "decorating a romantic bedroom" is too subjective because different users have a different interpretation of "romantic". (3) It has consistent observable visual characteristics. For example, a simple method is to use the candidate event name to search YouTube and check whether there are consistent visual tags found in the top returned videos. Tags may be approximately considered visual if they can be found in existing image ontology, such as ImageNet. (4) It is generic, not too detailed. If many article titles under a category share the same verb and direct object, they can be formed to a generic event name. After this, we end with 500 event categories as the current event collection in EventNet.

4.4.2 Mining Event Specific Concepts

We apply the concept discovery method developed in our prior work [J. Chen and Chang, 2014] to discover event-driven concepts from the tags of YouTube videos. For each of the 500 events, we use the event name as query keywords to search YouTube. We check the top $1,000$ returned videos and collect the ten most frequent words that appear in the titles or tags of these videos. Then we further filter the $1,000$ videos to only include those videos that contain at least three of the frequent words. This step helps us remove many irrelevant videos from the searching results. Using this

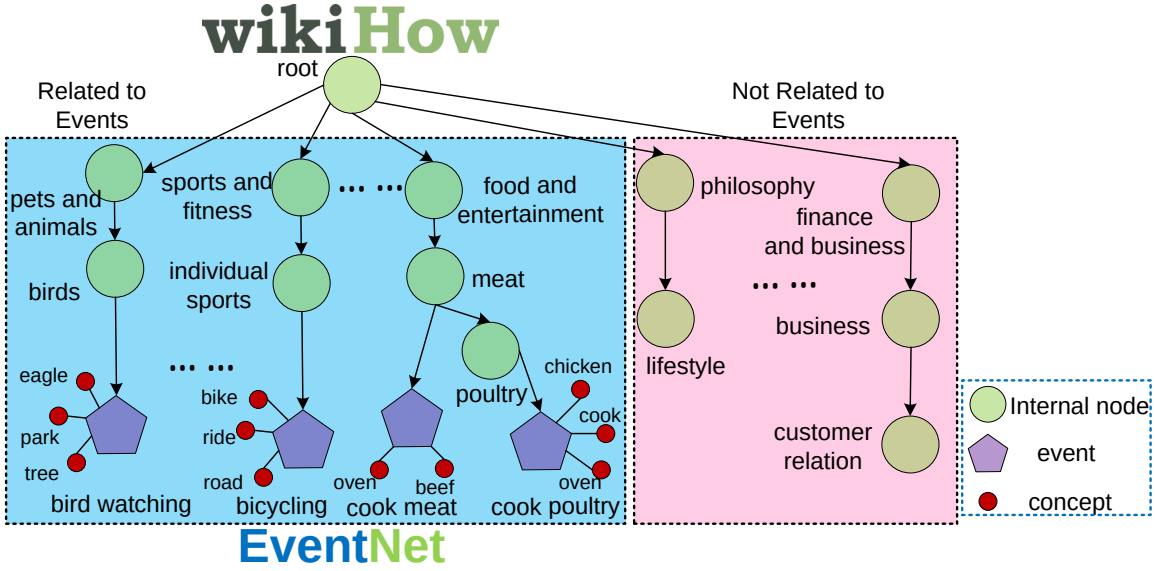


Figure 4.4: A snapshot of EventNet constructed from WikiHow.

approach, we crawl approximately 190 videos and their tag lists as concept discovery resource for each event, ending with $95,321$ videos for 500 events. We discover event-specific concepts from the tags of the crawled videos. To ensure the visual detectability of the discovered concepts, we match each tag to the classes of the existing object (ImageNet [J. Deng and Fei-Fei, 2009]), scene (SUN [Patterson and Hays, 2012]) and action (Action Bank [Sadanand and Corso, 2012]) libraries, and only keep the matched words as the candidate concepts. After going through the process, we end with approximately nine concepts per event, and a total of $4,490$ concepts for the entire set of 500 events. Finally, we adopt the hierarchical structure of WikiHow categories and attach each discovered event and its concepts to the corresponding category node. The final event concept ontology is called EventNet, as illustrated in Figure 4.4.

One could argue that the construction of EventNet ontology depends heavily on subjective evaluation. In fact, we can replace such subjective evaluation with automatic methods from computer vision and natural language processing techniques. For example, we can use concept visual verification to measure the visual detectability of concepts [J. Chen and Chang, 2014], and use text based event extraction to determine whether each article title is an event [A. Ritter and Clark, 2012]. However, as the accuracy of such automatic methods is still being improved, currently we focus on the design of principled criteria for event discovery and defer the incorporation of automatic discovery

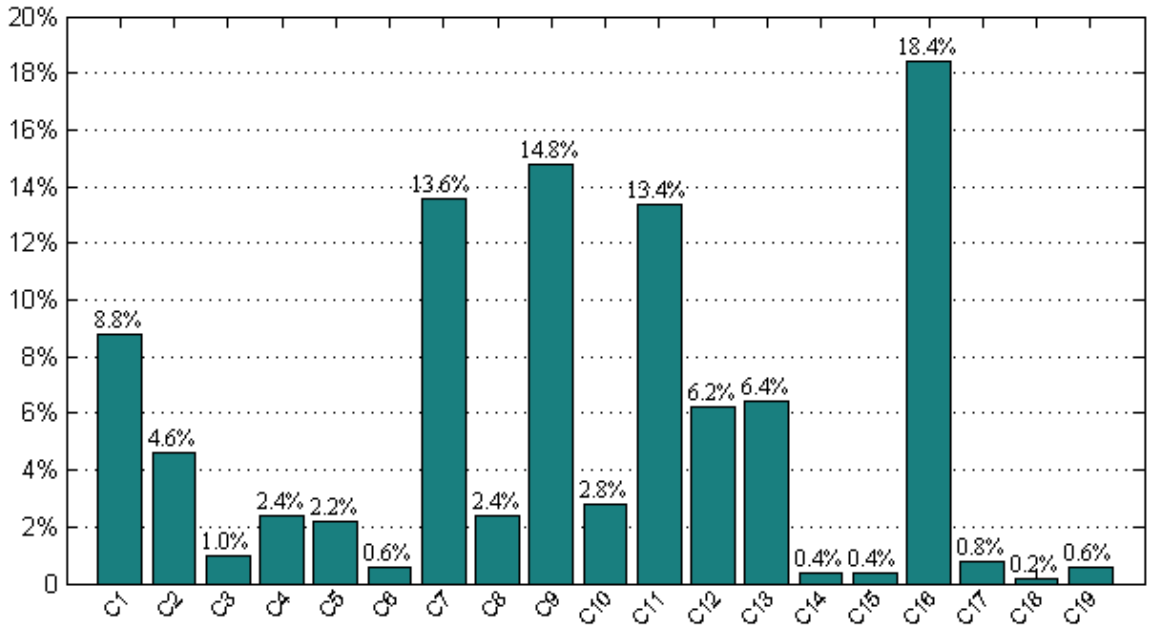


Figure 4.5: Event distribution over the top-19 categories of EventNet, where C1 to C19 are "arts and entertainment", "cars and other vehicles", "computers and electronics", "education and communications", "family life", "finance and business", "food and entertaining", "health", "hobbies and crafts", "holidays and traditions", "home and garden", "personal care and style", "pets and animals", "philosophy and religion", "relationships", "sports and fitness", "travel", "work world", and "youth".

processes until future improvement.

4.5 Properties of EventNet

In this section, we provide a detailed analysis on the properties of EventNet ontology, including basic statistics about the ontology, event distribution over coarse categories, and event redundancy.

EventNet Statistics. EventNet ontology contains 682 WikiHow category nodes, 500 event nodes and $4,490$ concept nodes organized in a tree structure, where the deepest depth from the root node to the leaf node (the event node) is eight. Each non-leaf category node has four child category nodes on average. Regarding the video statistics in EventNet, the average number of videos per event is 190, and the number of videos per concept is 21. EventNet has $95,321$ videos with an average duration of approximately 277 seconds ($7,334$ hours in total).

Event Distribution. We show the percentage of the number of events distributed over the top-19 category nodes of EventNet, and the results are shown in Figure 4.5. As can be seen, the top four popular categories that include the most number of events are "sports and fitness", "hobbies and craft", "food and entertainment", and "home and garden", whereas the least four populated categories are " work world", "relationships", "philosophy and religion" and "youth", which are abstract and cannot be described in videos. A further glimpse of the event distributions tells us that the most popular categories reflects the users' common interests in video content creation. For example, most event videos captured in human daily life refer to their life styles reflected in food, fitness, and hobbies. Therefore, we believe that the events included in EventNet have the potential to be used as an event concept library to detect popular events in human daily life.

Event Redundancy. We also conduct an analysis on the redundancy among the 500 events in EventNet. To this end, we use each event name as a text query, and find its most semantically similar events from other events located at different branches from the query event. In particular, given a query event e_q, we first localize its category node C_q in the EventNet tree structure, and then exclude all events attached under the parent and children nodes of node C_q. The events attached to other nodes are treated as the search base to find similar events of the query based on the semantic similarity described in Section 4.7. The reason for excluding events in the same branch of the query event is that those events that reside in the parent and children category nodes manifest hierarchical relationships such as "cook meat" and "cook poultry". We treat such hierarchical event pairs as a desired property of the EventNet library, and therefore do not involve them into the redundancy analysis. From the top-5 ranked events for a given query, we ask human annotators to determine whether there is a redundant event that refers to the same event as the query. After applying all 500 events as queries, we find zero redundancy among query event and all other events that reside in different branches of the EventNet structure.

4.6 Learning Concept Models from Deep Learning Video Features

In this section, we introduce the procedure for learning concept classifiers for the EventNet concept library. Our learning framework leverages the recent powerful CNN model to extract deep learning features from the video content, while employing one-vs-all linear SVM trained on top of the

Event Query	without EventNet structure	with EventNet structure
landing a fish	landing a plane cook fish	fishing hunt an animal
wedding shower	wedding ceremony take a shower	wedding ceremony make a wedding veil
woodworking project	working out using a rowing machine running	make wood projects make a crochet project

Table 4.2: Top-2 matched events of some event queries without (2nd column) and with (3rd column) leveraging EventNet structure.

features as concept models.

4.6.1 Deep Feature Learning with CNN

We adopt the CNN architecture in [A. Krizhevsky and Hinton, 2012] as the deep learning model to perform deep feature learning from video content. The network takes the RGB video frame as input and outputs the score distribution over the 500 events in EventNet. The network has five successive convolutional layers followed by two fully connected layers. Detailed information about the network architecture can be found in [A. Krizhevsky and Hinton, 2012]. In this work, we apply *Caffe* [Jia, 2013] as the implementation of the CNN model described by [A. Krizhevsky and Hinton, 2012].

For training of the EventNet CNN model, we evenly sample 40 frames from each video, and end with 4 million frames over all 500 events as the training set. For each of the 500 events, we treat the frames sampled from its videos as the positive training samples of this event. We define the set of 500 events as $E = \{0, 1, ..., 499\}$. Then the prediction probability of the k-th event for input sample n is defined as:

$$p_{nk} = \frac{\exp(x_{nk})}{\sum_{k' \in E} \exp(x_{nk'})}, \tag{4.1}$$

where x_{nk} is the k-th node's output of the n-th input sample from CNN's last layer. The loss function L is defined as a multinomial logistic loss of the softmax which is $L = \frac{-1}{N} \sum_{n=1}^{N} \log(p_{n,l_n})$, where $l_n \in E$ indicates the correct class label for input sample n, and N is the total number of inputs. Our

CNN model is trained on NVIDIA Tesla K20 GPU, and it requires approximately 7 days to finish $450K$ iterations of training. After CNN training, we extract the $4,096$-dimensional feature vector from the second to the last layer of the CNN architecture, and further perform ℓ_2 normalization on the feature vector as the deep learning feature descriptor of each video frame.

4.6.2 Concept Model Training

Given a concept discovered for an event, we treat the videos associated with this concept as positive training data, and randomly sample the same number of videos from concepts in other events as negative training data. This obviously has the risk of generating false negatives (videos without a certain concept label does not necessarily mean it is negative for the concept). However, in view of the prohibitive cost of annotating all videos over all concepts, we follow this common practice used in other image ontologies such as ImageNet [J. Deng and Fei-Fei, 2009]. We directly treat frames in positive videos as positive and frames in negative videos as negative to train a linear SVM classifier as the concept model. This is a simplified approach and there are emerging works [K.-T. Lai and Chang, 2014] for selecting more precise temporal segments or frames in videos as positive samples.

To generate concept scores on a given video, we first uniformly sample frames from it and extract the $4,096$-dimensional CNN features from each frame. Then we apply the $4,490$ concept models on each frame, and use all $4,490$ concept scores as the concept representation of this frame. Finally, we average the score vectors across all frames and adopt the average score vector as the video level concept representation.

4.7 Leveraging EventNet Structure for Concept Matching

In concept-based event detection, the first step is to find some semantically relevant concepts that are applicable for detecting videos with respect to the event query. This procedure is called *concept matching* in the literature [J. Chen and Chang, 2014; S. Wu and Natarajan, 2014]. To accomplish this task, the existing approaches typically calculate the semantic similarity between the query event and each concept in the library based on external semantic knowledge bases such as WordNet [Miller, 1995] or ConceptNet [Liu and Singh, 2004], and then select the top ranked concepts as the relevant concepts for event detection. However, these approaches might not be applicable to our EventNet

concept library because the involved concepts are event-specific and depend on their associative events. For example, the concept "dog" under "feed a dog" and "groom a dog" are treated as two different concepts because of the different event context. Therefore, concept matching in EventNet needs to consider event contextual information.

To this end, we propose a multi-step concept matching approach that first finds relevant events and then chooses those from the concepts associated with the matched events. In particular, given an event query e_q and an event e in the EventNet library, we use the textual phrase similarity calculation function developed in [L. Han and Weese, 2013] to estimate their semantic similarity. The reason for adopting such semantic similarity function is that both event query and candidate events in the EventNet library are textual phrases, that need a sophisticated phrase level similarity calculation that supports the word sequence alignment and strong generalization ability achieved by machine learning. However, these properties cannot be achieved using the standard similarity computation methods based on WordNet or ConceptNet alone. Our empirical studies confirm that the phrase based semantic similarity can obtain better event matching results.

However, because of word sense ambiguity and the limited amount of text information in event names, the phrase similarity-based matching approach can also generate wrong matching results. For example, given the query "wedding shower ", the event "take a shower" in EventNet receives a high similarity value because "shower" has an ambiguous meaning, and it is mistakenly matched as a relevant event. Likewise, the best matching results for the query "landing a fish" are "landing an airplane" and "cook fish" rather than "fishing" which is the most relevant. To address these problems, we propose exploiting the structure of the EventNet ontology to find relevant events for such difficult query events. In particular, given the query event, users can manually specify the suitable categories in the top level of the EventNet structure. For instance, users can easily specify that the suitable categories for the event "wedding shower" is "Family Life", while choosing "Sports and Fitness" and "Hobbies and Crafts" as suitable categories for "landing a fish". After the user's specification, subsequent event matching only needs to be conducted over the events under the specified high-level categories. This way, the hierarchical structure of EventNet ontology is helpful in relieving the limitations of short text based semantic matching, and helps improve concept-matching accuracy. Table 4.2 lists some difficult events from TRECVID MED and their top matched events with and without leveraging the EventNet structure. As can be seen, our method can achieve more

relevant matching results than using phrase-based semantic similarity alone. After we obtain the top matched events, we can further choose concepts based on their semantic similarity to the query event. Quantitative evaluations between the matching methods can be found in Section 4.8.4.

4.8 Experiments

In this section, we evaluate the effectiveness of the EventNet concept library in concept-based event detection. We first introduce the dataset and experiment setup, and then report the performance of different methods in the context of various event detection tasks, including zero-shot event retrieval and semantic recounting. After this, we study the efforts of leveraging the EventNet structure for matching concepts in zero-shot event retrieval. Finally, we will treat the $95K$ videos over 500 events in EventNet as a video event benchmark and report the baseline performance of using the CNN model in event detection.

4.8.1 Dataset and Experiment Setup

Dataset. We use two benchmark video event datasets as the test sets of our experiments to verify the effectiveness of the EventNet concept library. (1) *TRECVID 2013 MED* dataset [MED, 2010]. It contains $32,744$ videos that span over 20 event classes and the distracting background, whose names are "E1: *birthday party*", "E2: *changing a vehicle tire*", "E3: *flash mob gathering*", "E4: *getting a vehicle unstuck*", "E5: *grooming an animal*", "E6: *making a sandwich*", "E7: *parade*", "E8: *parkour*", "E9: *repairing an appliance*", "E10: *working on a sewing project*", "E11: *attempting a bike trick*", "E12: *cleaning an appliance*", "E13: *dog show*", "E14: *giving directions to a location*", "E15: *marriage proposal*", "E16: *renovating a home*", "E17: *rock climbing*", "E18: *town hall meeting*", "E19: *winning a race without a vehicle*", and "E20: *working on a metal crafts project*". We follow the original partition of this dataset in TRECVID MED evaluation, which partitions the dataset into a training set with $7,787$ videos and a test set with $24,957$ videos. (2) *Columbia Consumer Video (CCV)* dataset [Y.-G. Jiang and Loui, 2011]. It contains $9,317$ videos that span over 20 classes, which are "E1: *basketball*", "E2: *baseball*", "E3: *soccer*", "E4: *ice skating*", "E5: *skiing*", "E6: *swimming*", "E7: *biking*", "E8: *cat*", "E9: *dog*", "E10: *bird*", "E11: *graduation*", "E12: *birthday*", "E13: *wedding reception*", "E14: *wedding ceremony*", "E15: *wedding dance*",

"E16: *music performance*", "E17: *non-music performance*", "E18: *parade*", "E19: *beach*", and "E20: *playground*". The dataset is further divided into $4,659$ training videos and $4,658$ test videos. Because we focus on zero-shot event detection, we do not use the training videos in these two datasets, but only test the performance on the test set. For supervised visual recognition, features from deep learning models, e.g., the last few layers of deep learning models learned over ImageNet $1K$ or $20K$) can be directly used to detect events [M. Jain and Snoek, 2014]. However, the focus of this paper is on the semantic description power of the event-specific concepts, especially in recounting the semantic concepts in event detection, and finding relevant concepts for retrieving events not been seen before (zero-shot retrieval).

Feature Extraction. On the two evaluation event video datasets, we extract the same features we did on EventNet videos. In particular, we sample one frame every 2 seconds from a video, and extract the $4,096$-dimensional deep learning features from the CNN model trained on EventNet video frames. Then we run SVM-based concept models over each frame, and aggregate the score vectors in a video as the semantic concept feature of the video.

Comparison Methods and Evaluation Metric. We compare different concept based video representations produced by the following methods. (1) **Classemes** [L. Torresani and Fitzgibbon, 2010]. It is a $2,659$-dimensional concept representation whose concepts are defined based on LSCOM concept ontology. We directly extract Classemes on each frame and then average them across the video as video-level concept representation. (2) Flickr Concept Representation (**FCR**) [Y. Cui and Chang, 2014]. For each event, the concepts are automatically discovered from the tags of Flickr images in the search results of event query and organized based on WikiHow ontology. The concept detection models are based on the binary multiple kernel linear SVM classifiers trained with the Flickr images associated with each concept. Five types of low-level features are adopted to represent Flickr images and event video frames. (3) ImageNet-$1K$ CNN Concept Representation (**ICR-1K**). In this method, we directly apply the network architecture in [A. Krizhevsky and Hinton, 2012] to train a CNN model over 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest that covers $1,000$ different classes [J. Deng and Fei-Fei, 2009]. After model training, we apply the CNN model on the frames from both TRECVID MED and CCV datasets. Concept scores of the individual frames in a video are averaged to form the concept scores of the video. We treat the $1,000$ output scores as the concept based video representation from ImageNet-$1K$. (4)

ImageNet-20K CNN Concept Representation (**ICR-20K**). We apply the same network architecture as ICR-1K to train a CNN model using over 20 million images that span over $20,574$ classes from the latest release of ImageNet [J. Deng and Fei-Fei, 2009]. We treat the $20,574$ concept scores output from the CNN model as the concept representation. Notably, ICR-1K and ICR-20K represent the most successful visual recognition achievements in the computer vision area, which can be applied to justify the superiority of our EventNet concept library over the state-of-the-art. (5) Our proposed EventNet-CNN Concept Representation (**ECR**), where we use our EventNet concept library to generate concept based video representations. (6) Some state-of-the-art results reported in the literature. Regarding the evaluation metric, we adopt AP, which approximates the area under precision/recall curve, to measure the performance on each event in our evaluation datasets. Finally, we calculate mAP over all event classes as the overall evaluation metric.

4.8.2 Task I: Zero-Shot Event Retrieval

Our first experiment evaluates the performance of zero-shot event retrieval, where we do not use any training videos, but completely depend on the concept scores on test videos. To this end, we use each event name in the two video datasets as a query to match the two most relevant events, and choose the 15 most relevant EventNet concepts based on semantic similarity, and then average the scores of these 15 concepts as the zero-shot event detection score of the video, through which a video ranking list can be generated. Notably, the two most relevant events mentioned above are automatically selected based on the semantic similarity matching method described in Section 4.7. For Classemes and FCR, we follow the setting in [Y. Cui and Chang, 2014] to choose 100 relevant concepts based on semantic similarity using ConceptNet and the concept matching method described in [Y. Cui and Chang, 2014]. For ICR-1K and ICR-20K, we choose 15 concepts using the same concept matching method.

Figure 4.6 shows the performance of different methods on two datasets, respectively. From the results, we obtain the following observations: (1) event specific concept representations, including FCR and ECR outperform the event independent concept representation Classemes. This is because the former not only discovers semantically relevant concepts of the event, but also leverages the contextual information about the event in the training samples of each concept. In contrast, the latter only borrows concepts that are not specifically designed for events, and the training images

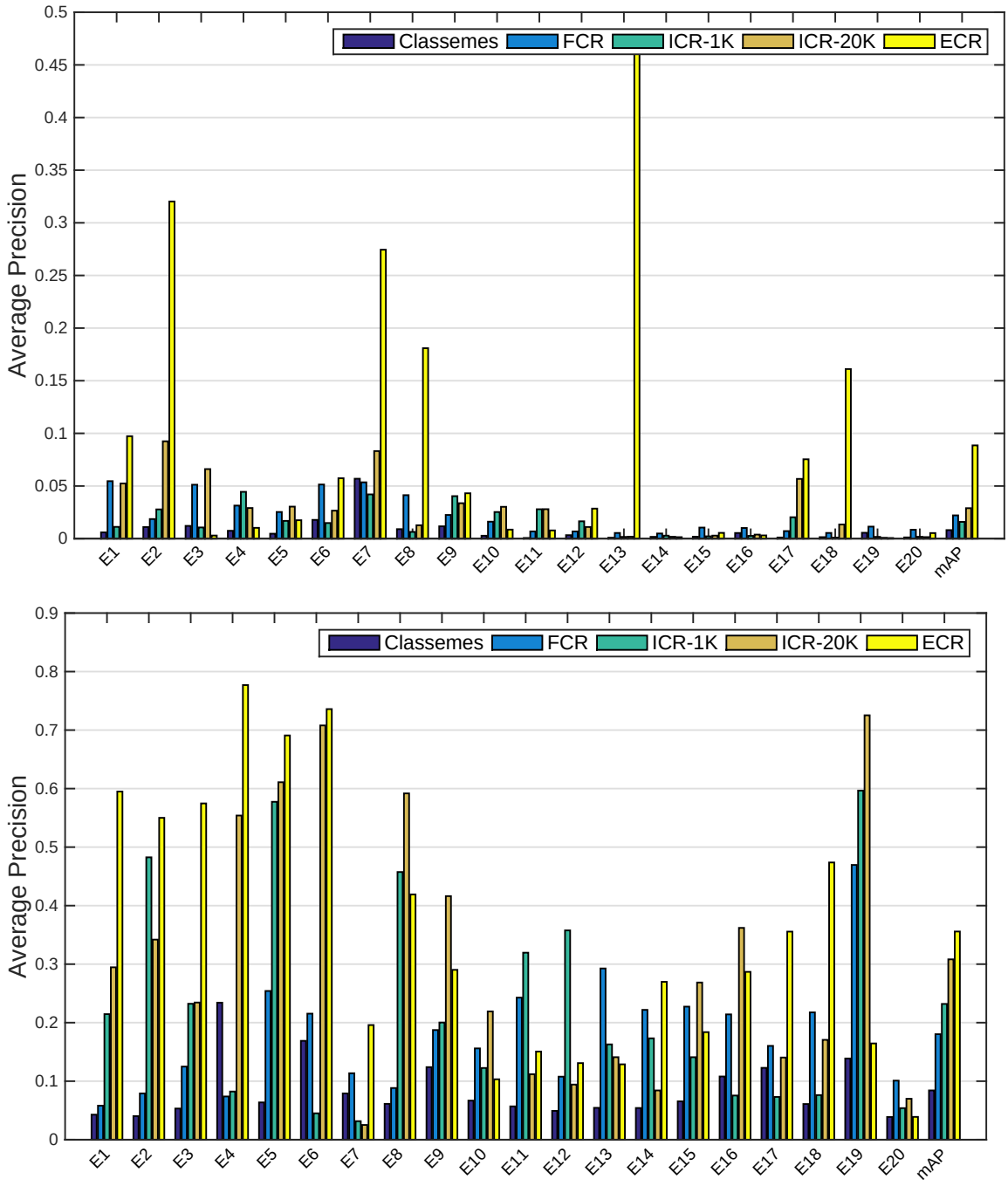


Figure 4.6: Performance comparisons on zero-shot event retrieval task (left: MED; right: CCV). This figure is best viewed in color.

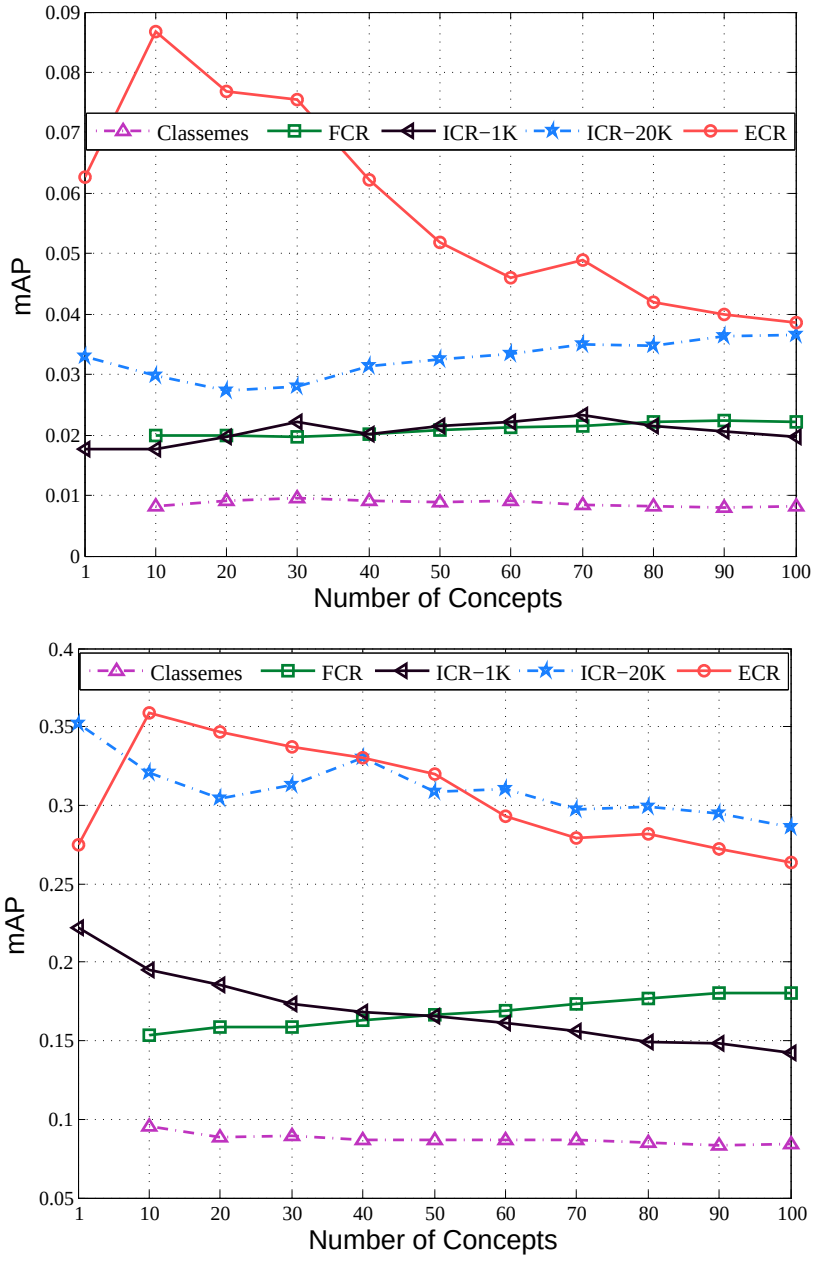


Figure 4.7: Zero-shot event retrieval performance with different number of concepts (left: MED; right: CCV). The results of Classemes and FCR are from literature, in which the results when concept number is 1 are not reported.

for concept classifiers do not contain the event-related contextual information. (2) Concept representations trained with deep CNN features, including ICR-20K and ECR, produce much higher performance than the concept representations learned from low-level features including Classemes and FCR for most of the events. This is reasonable because the CNN model can extract learning based features that have been shown to achieve strong performance. (3) Although all are trained with deep learning features, ECR generated by our proposed EventNet concept library performs significantly better than ICR-1K and ICR-20K, which are generated by deep learning models trained on ImageNet images. The reason is that concepts in EventNet are more relevant to events than the concepts in ImageNet which are mostly objects independent of events. From this result, we can see that our EventNet concepts even outperformed the concepts from the state-of-the-art visual recognition system, and it is believed to be a powerful concept library for the task of zero-shot event retrieval.

Notably, our ECR achieves significant performance gains over the best baseline ICR-20K, where the mAP on TRECVID MED increases from 2.89% to 8.86% with 207% relative improvement. Similarly, the mAP on CCV increases from 30.82% to 35.58% (15.4% relative improvement). Moreover, our ECR achieves the best performance on most event categories on each dataset. For instance, on the event "E02: *changing a vehicle tire*" from the TRECVID MED dataset, our method outperforms the best baseline ICR-20K by 246% relative improvement. On the TRECVID MED dataset, the reason for the large improvement on "E13: *dog show*" is that the matched events contain exactly the same event "dog show" as the event query. The performance on E10 and E11 is not so good because the automatic event matching method matched them to wrong events. When we use the EventNet structure to correct the matching errors as described in Section 4.8.4, we achieve higher performance on these events.

In Figure 4.7, we show the impact on zero-shot event retrieval performance when the number of concepts changes using the concept matching method described in Section 4.7, i.e., we first find the matched events, and then select the top-ranked concepts that belong to these events. We select the number of events until the desired number of concepts is reached. On TRECVID MED, we can see consistent and significant performance gains for our proposed ECR method over others. However, on the CCV dataset, ICR-20K achieves similar or even better performance when several concepts are adopted. We conjecture that this occurs because the CCV dataset contains a number

Figure 4.8: Event video recounting results: each row shows evenly subsampled frames of a video and the top 5 concepts detected in the video.

of object categories, such as "E8: cat" and "E9: dog", which might be better described by the visual objects contained in the ImageNet dataset. Alternatively, all the events in TRECVID MED are highly complicated, and they might be better described by EventNet. It is worth mentioning that mAP first increases and then decreases as we choose more concepts from EventNet. This is because our concept matching method always ranks the most relevant concepts on top of the concept list. Therefore, involving many less relevant concepts ranked at lower positions (after the 10th position in this experiment) in the concept list might decrease performance. In Table 4.3, we compare our results with the state-of-the-art results reported on the TRECVID MED 2013 test set with the same experiment setting. We can see that our ECR method outperforms these results by a large margin.

Method	mAP (%)
Selective concept [A. Habibian and Snoek, 2014; M. Mazloom and Snoek, 2013]	4.39
Bi-concept [A. Habibian and Snoek, 2014; M. Rastegari and Farhadi, 2013]	3.45
Composite concept [A. Habibian and Snoek, 2014]	5.97
Weak concept [S. Wu and Natarajan, 2014]	3.48
Annotated concept [J. Liu and Friedland, 2013]	6.50
Our EventNet concept	**8.86**

Table 4.3: Comparisons between our ECR with other state-of-the-art concept based video representation methods built on visual content. All results are obtained in the task of zero-shot event retrieval on TRECVID MED 2013 test set.

4.8.3 Task II: Semantic Recounting in Videos

Given a video, semantic recounting aims to annotate the video with the semantic concepts detected in the video. Because we have the concept-based representation generated for the videos using the concept classifiers described earlier, we can directly use it to produce recounting. In particular, we rank the $4,490$ event-specific concept scores on a given video in descending order, and then choose the top-ranked ones as the most salient concepts that occur in the video. Figure 4.8 shows the recounting results for some sample videos from the TRECVID MED and CCV datasets. As can be seen, the concepts generated by our method precisely reveal the semantics presented in the videos.

It is worth noting that the EventNet ontology also provides great benefits for developing a real-time semantic recounting system that requires high efficiency and accuracy. Compared with other concept libraries that use generic concepts, EventNet allows selected execution of a small set of concepts specific to an event. Given a video to be recounted, it first predicts the most relevant events, and then applies only those concepts that are specific to these events. This unique two-step approach can greatly improve the efficiency and accuracy of multimedia event recounting because only a small number of event-specific concept classifiers need to be started after event detection.

4.8.4 Task III: Effects of EventNet Structure for Concept Matching

As discussed in Section 4.7, because of the limitations of text based similarity matching, the matching result of an event query might not be relevant. In this case, the EventNet structure can help users find relevant events and their associated concepts from the EventNet concept library. Here we first perform quantitative empirical studies to verify this. In particular, for each event query, we manually specify two suitable categories from the top 19 nodes of the EventNet tree structure, and then match events under these categories based on semantic similarity. We compare the results obtained by matching all events in EventNet (i.e., without leveraging the EventNet structure) with the results obtained by the method we described above (i.e., with leveraging the EventNet structure). For each query, we apply each method to select 15 concepts from the EventNet library, and then use them to perform zero-shot event retrieval.

Method (mAP %)	MED	CCV
Without Leveraging EventNet Structure	8.86	35.58
With Leveraging EventNet Structure	**8.99**	**36.07**

Table 4.4: Comparison of zero-shot event retrieval using the concepts matched without leveraging EventNet structure (top row) and with leveraging EventNet structure (bottom row).

Table 4.4 shows the performance comparison between the two methods. From the results, we can see that event retrieval performance can be improved if we apply the concepts matched with the help of EventNet structure, which proves the usefulness of EventNet structure for the task of concept matching.

4.8.5 Task IV: Multi-Class Event Classification

The 95,321 videos over 500 event categories in EventNet can also be seen as a benchmark video dataset to study large-scale event detection. To facilitate direct comparison, we provide standard data partitions and some baseline results over these partitions. It is worth noting that one important purpose of designing the EventNet video dataset is to use it as a testbed for large scale event detection models, such as deep convolutional Neutral Network. Therefore, in the following, we summarize a baseline implementation using the state-of-the-art CNN models, as was done in [J. Deng and

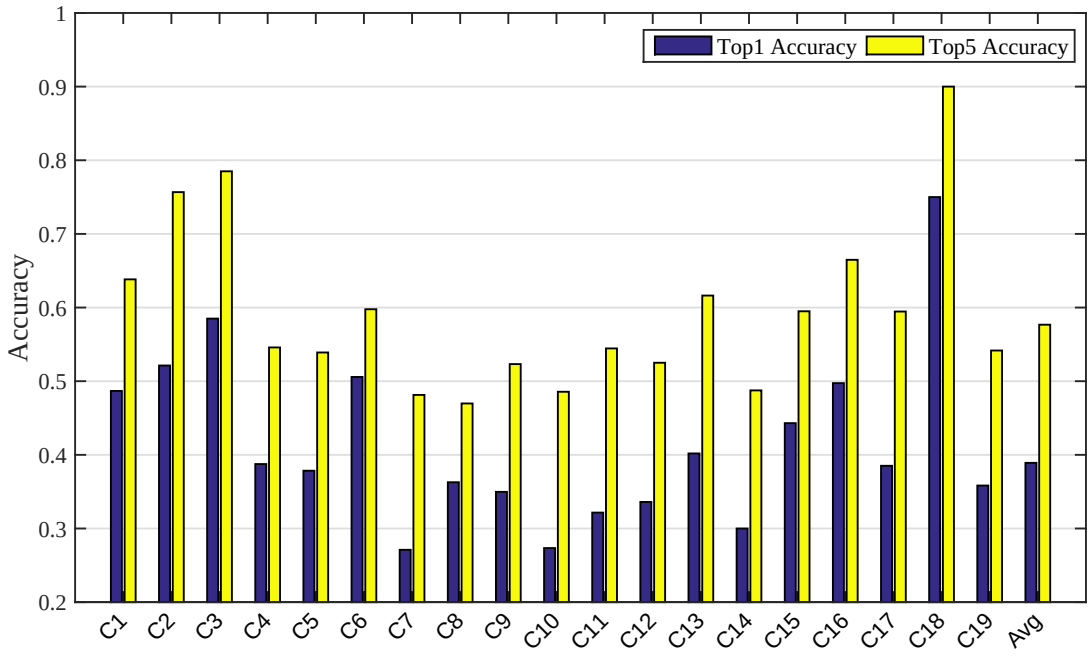


Figure 4.9: Top-1 and top-5 event classification accuracies over 19 high-level event categories of EventNet structure, in which the average top-1 and top-5 accuracy are 38.91% and 57.67%.

Fei-Fei, 2009].

Data Division. Recall that each of the 500 events in EventNet has approximately 190 videos. In our experiment, we divide the videos and adopt 70% of the videos as the training set, 10% as validation set, and 20% as the test set. In all, there are approximately $70K$ (2.8 million frames), $10K$ (0.4 million frames), and $20K$ (0.8 million frames) training, validation, and test videos, respectively.

Deep Learning Model. We adopt the same network architecture and learning setting of the CNN model described in Section 4.6.1 as our multi-class event classification model. In the training process, for each event, we treat the frames sampled from the training videos of an event as positive training samples and feed them into the CNN model for model training. Seven days are required to finish $450K$ iterations of training. In the test stage, to produce predictions for a test video, we take the average of the frame-level probabilities over sampled frames in a video and use it as the video-level prediction result.

Evaluation Metric. Regarding evaluation metric, we adopt the most popular top-1 and top-5 accuracy commonly used in large scale visual recognition, where the top-1 (top-5) accuracy is a

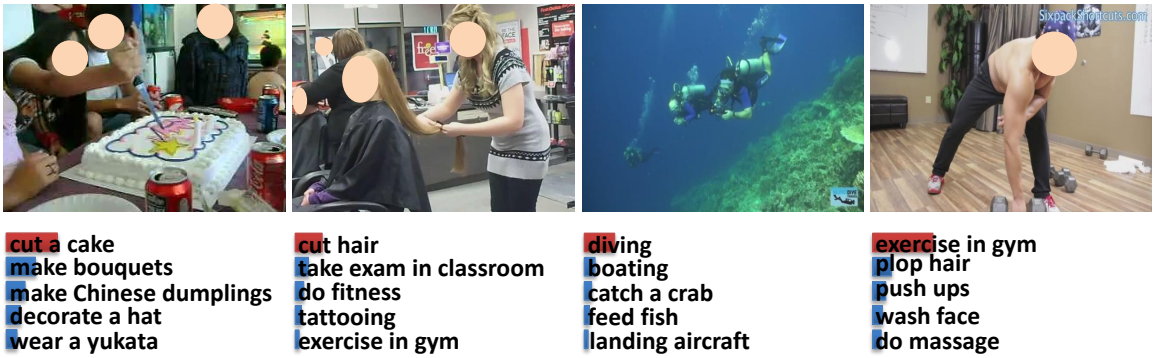


Figure 4.10: Event detection results of some sample videos. The 5 events with the highest detection scores are shown in the descending order. The bar length indicates the score of each event. Event with the red bar is the ground truth.

fraction of the test videos for which the correct label is among the top-1 (5) labels predicted to be most probable by the model.

Results. We report the multi-class classification performance by 19 high-level categories of events in the top layer of the EventNet ontology. To achieve this, we collect all events under each of the 19 high-level categories in EventNet (e.g., 68 events under "home and garden"), calculate the accuracy of each event and then calculate their mean value over the events within this high-level category. As seen in Figure 4.9, most high-level categories show impressive classification performance. To illustrate the results, we choose four event video frames and show their top-5 prediction results in Figure 4.10.

4.9 Summary and Discussion

We introduced EventNet, a large scale structured event-driven concept library, for representing complex events in video. The library contains 500 events mined from WikiHow and 4,490 event-specific concepts discovered from YouTube video tags, for which large margin classifiers are trained with deep learning features over 95,321 YouTube videos. The events and concepts are further organized into a tree structure based on the WikiHow ontology. Extensive experiments on two benchmark event datasets showed major performance improvement of the proposed concept library over zero-

shot event retrieval task. We also showed that the tree structure of EventNet helps match the event queries to semantically relevant concepts. For future work, we will continue to expand EventNet by continuously discovering more events from WikiHow, YouTube, and other knowledge resources. We will also pursue tree structured event modeling that incorporates the hierarchical relationship of events in EventNet.

Chapter 5

Large Scale Video Event and Concept Ontology Applications

5.1 Introduction

In this chapter, we present several applications using our large-scale video event and concept ontology. In particular, the novel functions of our EventNet system include: interactive browser, semantic search, and live tagging of user-uploaded videos. In each of the modules, we emphasize the unique ontological structure embedded in EventNet and utilize it to achieve a novel experience. For example, the event browser leverages the hierarchical event structure discovered from the crowdsource forum WikiHow to facilitate intuitive exploration of events, the search engine focuses on retrieval of hierarchical paths that contain events of interest rather than events as independent entities, and finally the live detection module applies the event models and associated concept models to explain why a specific event is detected in an uploaded video. To the best of our knowledge, this is the first interactive system that allows users to explore high level events and associated concepts in videos in a systematic structured manner.

5.2 Application I: Event Ontology Browser

Our system supports users to browse the EventNet tree ontology in an interactive and intuitive manner. When a user clicks a non-leaf category node, the child category nodes are expanded along

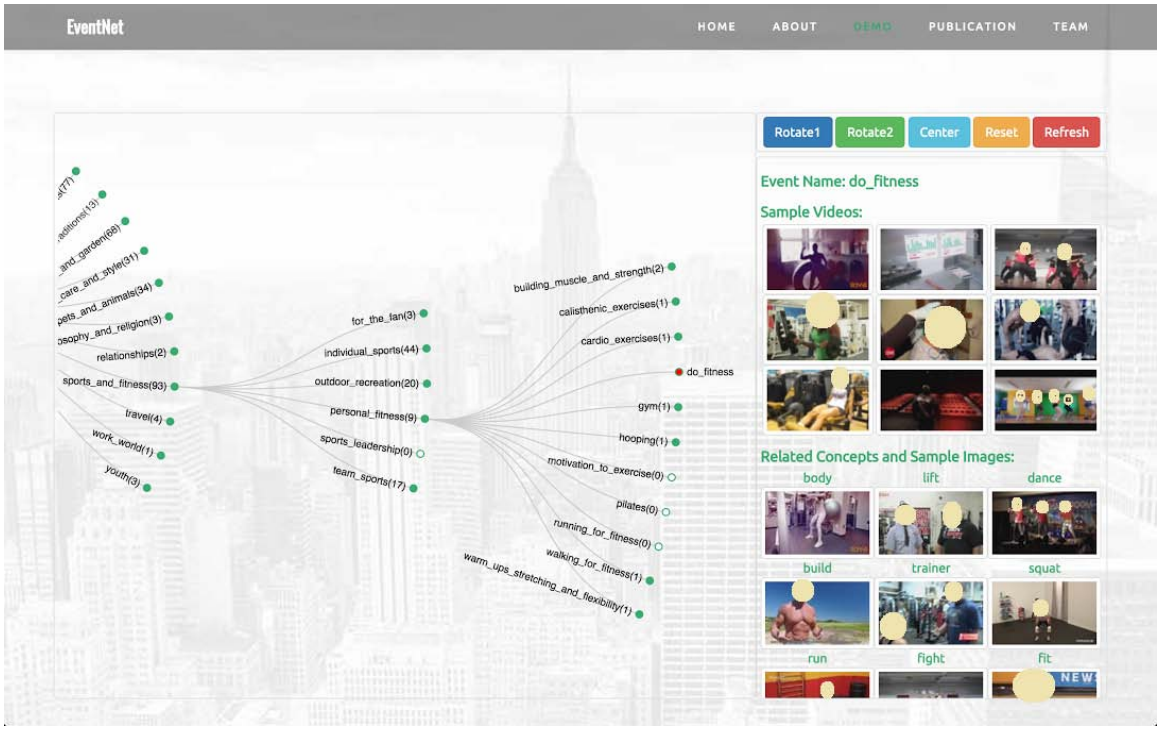


Figure 5.1: Visualization interface for event ontology browsing. Example videos and related concepts of the selected event are shown.

with any event attached to this category (the event node is filled in red, whereas the category node is in green). When the user clicks an event, the exemplary videos for this event is shown with a dynamic GIF animation of the keyframes extracted from a sample video. Concepts specific to the event are also shown with representative keyframes of the concept. We specifically adopt the expandable, rotatable tree as the visualization tool (as shown in Figure 5.1) because it maintains a nice balance between the depth and breadth of the scope when the user navigates through layers and siblings in the tree.

5.3 Application II: Semantic Search of Events in the Ontology

We adopt a unique search interface that is different from the conventional ones by allowing users to find hierarchical paths that match user interest, instead of treating events as independent units. This design is important for fully leveraging the ontology structure information in EventNet. For each event in EventNet, we generate its text representation by combining all words of the category

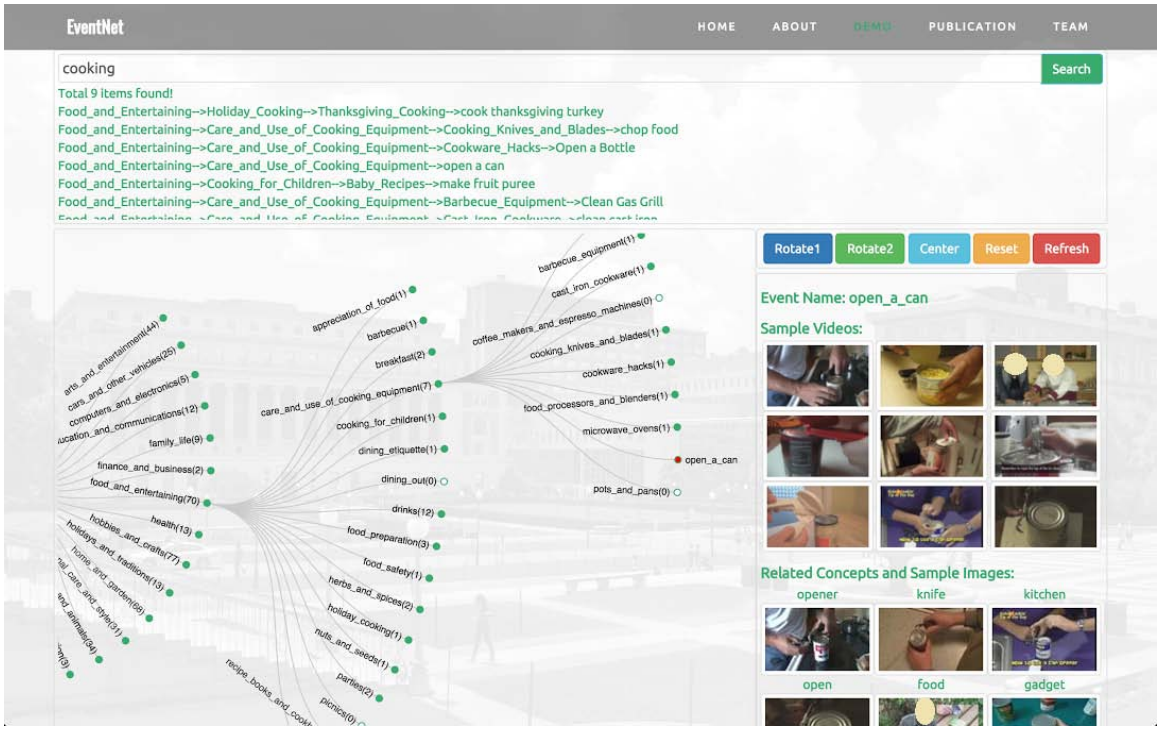


Figure 5.2: Interface for searching events embedded in the EventNet ontology.

names from the root node to the current category that contains the event, plus the name of the event. Such texts are used to set up search indexes in Java Lucene [Luc, 2015]. When the user searches for keywords, the system returns all the paths in the index that contain the query keywords. If the query contains more than one word, the path with the more matched keywords is ranked higher in the search result. After the search, the users can click each returned event, and our system dynamically expands the corresponding path of this event, and visualizes it using the tree browser described in the previous section. This not only helps users quickly find target events, but also helps suggest additional events to the user by showing events that could exist in the sibling categories in the EventNet hierarchy. Figure 5.2 shows the interface of the search function.

5.4 Application III: Automatic Video Tagging

EventNet includes an upload function that allows users to upload any video and use pre-trained detection models to predict the events and concepts present in the video. For each uploaded video, EventNet extracts one frame every 10 seconds. Each frame is then resized to 256 by 256 pixels and

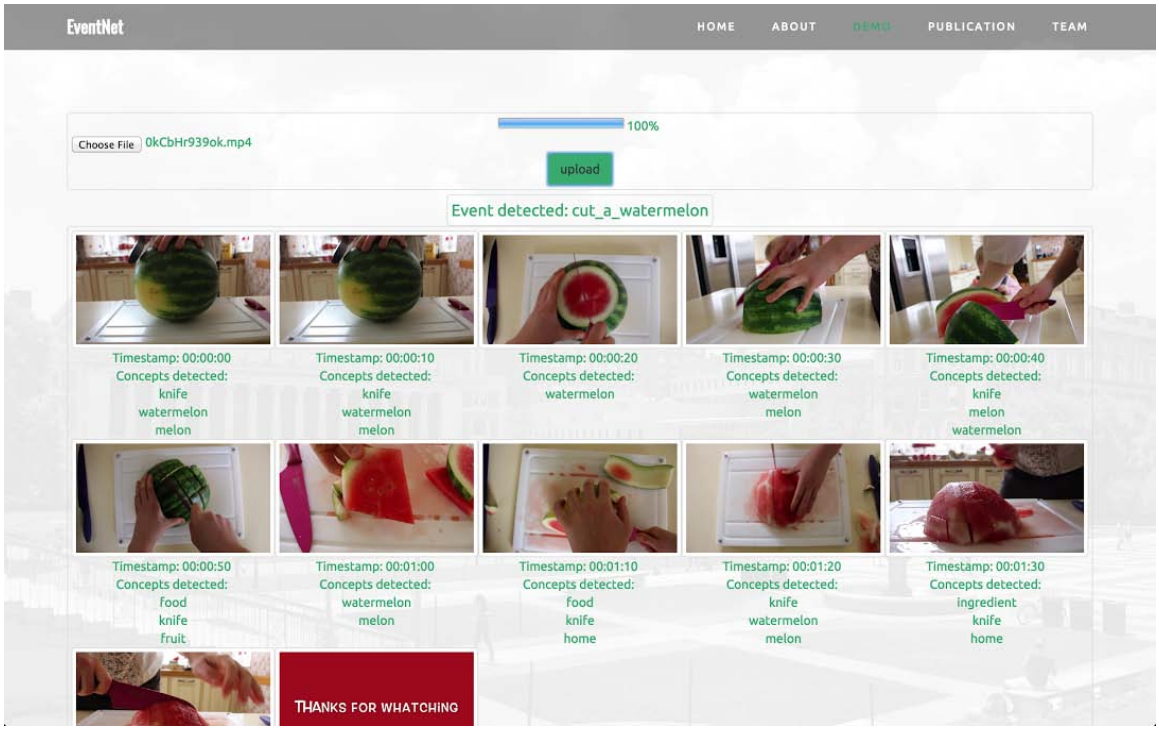


Figure 5.3: Interface of automatic tagging of user uploaded videos.

fed to the deep learning model described earlier. We average the 500-dimensional detection scores across all extracted frames, and use the average score vector as the event detection scores of the video. To present the final detection result, we only show the top event with highest score as the event prediction of the video. For concept detection, we use the feature in the second last layer of the deep learning model computed over each frame, and then apply the binary SVM classifiers to compute the concept scores on each frame. We show the top-ranked predicted concepts under each sampled frame of the uploaded video. Figure 5.3 shows the tagging results of the event and concept for an uploaded video, indicating the high accuracy of the tagging result. It is worth mentioning that our tagging system is very fast and satisfies real-time requirements. For example, when we upload a 10 MB video, the tagging system can generate tagging results in 5 seconds on a single regular workstation, demonstrating the high efficiency of the system.

5.5 Summary and Discussion

In this chapter, we demonstrated novel applications on EventNet, the largest event ontology existing today (to the best of our knowledge) with a hierarchical structure extracted from the popular crowdsource forum WikiHow. The system provides efficient event browsing and search interfaces, and supports live video tagging with high accuracy. It also provides a flexible framework for future scaling up by allowing users to add new event nodes to the ontology structure.

In the future, we will continue to expand EventNet with more daily events. In order to precisely detect the event specific concepts, we will focus on the research of event and concept spatial and temporal localizations.

Part III

Event Detection with Multi-Modality Representations and Multi-Source Fusion

Detecting complex events in videos is intrinsically a multi-modality and multi-source fusion problem. On one hand, because joint cross-modality patterns (e.g., audio-visual pattern) often exist in videos, we tend to propose a novel joint multi-modality representation to discover the intrinsic correlations among modalities that help detect video event. On the other hand, because combining features from multiple sources often produces performance gains reported in literatures [Y.-G. Jiang and Chang, 2010; Bach *et al.*, 2004], we further propose a robust late fusion method for video event detection with effective multiple source fusion.

We first propose a new multi-modality representation, called *bi-modal* words, to explore representative joint audio–visual patterns. In particular, we build a bipartite graph to model the relationship across the quantized words extracted from the visual and audio modalities. Partitioning over the bipartite graph is then applied to produce the bi-modal words that reveal the joint patterns across modalities. Different pooling strategies are then employed to re-quantize the visual and audio words into the bi-modal words and form bi-modal BoW representations. Because it is difficult to predict the suitable number of bi-modal words, we generate bi-modal words at different levels (i.e., codebooks with different sizes), and use Multiple Kernel Learning(MKL) to combine the resulting multiple representations during event classifier learning. Experimental results on three popular datasets show that the proposed method achieves statistically significant performance gains over methods using individual visual and audio feature alone and existing popular multi-modal fusion methods.We also find that average pooling is particularly suitable for bi-modal representation, and using multiple kernel learning (MKL) to combine multi-modal representations at various granularities is helpful.

Next, we propose a rank minimization method to fuse the predicted confidence scores from multiple sources, each of which is obtained based on a certain type of feature. In particular, we convert each confidence score vector obtained from one model into a pairwise relationship matrix, where each entry characterizes the comparative relationship of scores of two test samples. Our hypothesis is that the relative score relationships are consistent among component models up to certain sparse deviations, despite the large variations that can exist in the absolute values of the raw scores. Then we formulate the score fusion problem as seeking a shared rank-2 pairwise relationship matrix based on the original score matrix from individual models that can be decomposed into the common rank-2 matrix and sparse deviation errors. A robust score vector is then extracted to fit

the recovered low-rank score relationship matrix. We formulate the problem as a nuclear norm and ℓ_1 norm optimization objective function, and employ the Augmented Lagrange Multiplier (ALM) method for the optimization. Our method is isotonic (i.e., scale invariant) to the numeric scales of the scores that originate from different models. We experimentally show that the proposed method achieves significant performance gains on video event detection.

Chapter 6

Discovering Joint Audio-Visual Representation for Video Event Detection

6.1 Introduction

Automatically detecting complex events in diverse Internet videos is a topic that is receiving increasing research attention in computer vision and multimedia. Currently, a large portion of Internet videos is captured by amateur consumers without professional post-editing. This makes the task of event recognition extremely challenging, because such videos contain large variations in lighting, viewpoint, camera motion, etc. Fig. 6.1 shows sample frames from six videos that contain the event "Feeding an animal". In addition to the variations mentioned above, the "high-level" nature of the event categories (e.g., different types of animals in this event) sets a big challenge in event recognition.

Fortunately, besides the visual frames shown in Fig. 6.1, videos also contain audio information that provides an additional useful clue for event detection. In other words, the events captured in the videos are multimodal and videos of the same event typically show consistent audio–visual patterns. For example, an "explosion" event is best manifested by the transient burst of sound along with visible smoke and flame after the incident. Other examples include strong temporal synchroniza-

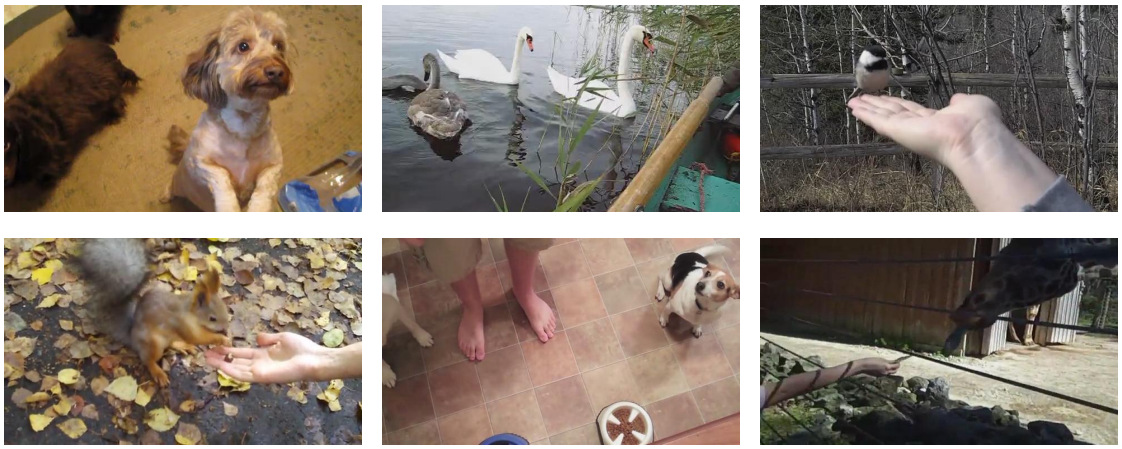


Figure 6.1: Example video frames of event "feeding an animal" defined in TRECVID Multimedia Event Detection Task 2011. As can be seen, event detection in such unconstrained videos is a highly challenging task since the content is extremely diverse.

tion (e.g., horse running with audible footsteps) or loose association (e.g., people feeding an animal while talking about the feeding action). Therefore, we believe that successful event detection solutions should effectively harness both audio and visual modalities.

Most existing works fused multimodal features in a superficial fashion, such as early fusion that concatenates feature vectors before classification, or late fusion that combines prediction scores after classification. To better characterize the relationship between audio–visual modalities in videos, we propose an audio–visual bi-modal BoW representation. First, we apply the typical BoW representation to build audio and visual BoW representations, where the codebooks are generated using standard k-means clustering separately. Subsequently, a bipartite graph is constructed to capture joint co-occurrence statistics between the quantized audio and visual words. A bi-modal codebook is then generated by spectral clustering, which partitions the graph into a set of visual/audio word groups, and each group is treated as a joint bi-modal word. Finally, the original individual feature in each modality (audio or visual) is re-quantized based on the bi-modal codewords, using popular feature pooling methods. In addition, as given that it is difficult (if not impossible) to predict a suitable number of bi-modal words, we generate bi-modal codebooks of different sizes and employ MKL to combine their respective representations for event model learning. The flowchart for our approach is illustrated in Fig. 6.2.

The main contributions are summarized as follows:

- We propose an audio–visual bi-modal BoW representation, that effectively explores the underlying structure of the joint audio–visual feature space of complex unconstrained videos. Our representation is very easy to implement because only the classical bipartite graph partition technique is used to generate the bi-modal words. Compared with the original audio or visual BoW representations, the joint bi-modal BoW not only outperforms simple early/late fusion, but also greatly reduces the dimensionality of the final video representation.

- Other than fixing the number of codewords as most existing works on visual/audio word-based representations do, we propose generating bi-modal codewords at different granularities (multiple codebooks of different sizes) and adopt MKL [Jhuo and Lee, 2010; A. Kembhavi and Davis, 2009; A. Vedaldi and Zisserman, 2009] to incorporate multiple bi-modal BoW representations for event detection, which further improves the detection accuracy.

The rest of the chapter is organized as follows. We first review related works in Sect. 6.2. Section 6.3 discusses typical representations of audio and visual features. Section 6.4 introduces our proposed audio–visual bi-modal BoW representation. Extensive experimental evaluations on three popular datasets will be given in Sect. 6.5. Finally, we conclude this work in Sect. 6.6.

6.2 Related Works

Fusing complementary audio and visual information is important in video content analysis, and has been attempted in many prior works. For example, Jiang *et al.* [Y.-G. Jiang and Chang, 2010] adopted average late fusion, which uses the average prediction scores of multiple independently trained classifiers. On the contrary, the work in [L. Bao, 2011] averaged the kernel matrices obtained from audio and visual features before classification, and it is known as the early fusion method. Unlike these superficial fusion methods, our bi-modal BoW representation characterizes the joint patterns across the two modalities, which can uncover their underlying relationships rather than simple combination.

There are also several interesting works on joint audio–visual analysis, especially for object tracking and detection. For instance, Beal *et al*. [M. Beal and Attias, 2003] developed a joint

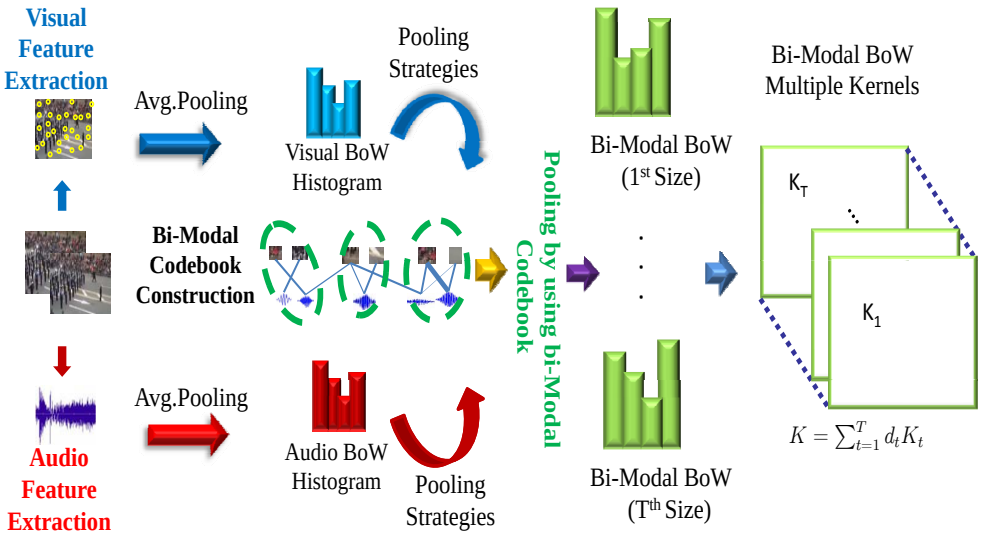


Figure 6.2: The framework of our proposed joint bi-modal word representation. We first extract audio and visual features from the videos and then quantize them into audio and visual BoW histograms respectively. After that, a bipartite graph is constructed to model the relations across the quantized words extracted from both modalities, in which each node denotes a visual or an audio word and edges between two nodes encode their correlations. By partitioning the bipartite graph into a number of clusters, we obtain several bi-modal words that reveal the joint audio-visual patterns. With the bi-modal words, the audio and visual features in the original BoW representations are re-quantized into a bi-modal BoW representation. Finally, bi-modal codebooks of various sizes are combined in a multiple kernel learning framework for event model learning.

probability model of audio and visual cues for object tracking. Cristani *et al.* [M. Cristani and Murino, 2007] attempted to synchronize foreground objects and audio sounds in the task of object detection. One limitation of these methods is that they only considered videos in a fully controlled environment, which is much easier than the unconstrained videos managed in this work.

More recently, Jiang *et al*. [W. Jiang and Loui, 2009] proposed Short-Time Audio–Visual Atom as the joint audio–visual feature for video concept classification. First, visual regions are tracked within short-term video slices to generate visual atoms, and audio energy onsets are located to generate audio atoms. Then the regional visual features extracted from the visual atoms and the spectrogram features extracted from the audio atoms are concatenated to form an audio–visual atom feature representation. Finally, a discriminative joint audio–visual codebook is constructed

on the audio–visual atoms using multiple instance learning, and finally the codebook-based BoW features are generated for semantic concept detection. As an extension of this work, in [Jiang and Loui, 2011], the authors further proposed Audio–Visual Grouplets by exploring temporal audio–visual interactions, where an audio–visual grouplet is defined as a set of audio and visual codewords grouped based on their strong temporal correlations in videos. In particular, the authors conducted foreground/background separation in both audio and visual channels, and then formed four types of audio–visual grouplets by exploring the mixed-and-matched temporal audio–visual correlations, which provide discriminative audio–visual patterns for classifying semantic concepts. Despite the close relatedness with our work, the above two methods require performing object or region tracking, which is extremely difficult and computationally expensive, particularly for unconstrained Internet videos. Several other works demonstrated the success of utilizing audio and visual information for recognition [Y.-G. Jiang and Shah, 2013; G. Potamianos and Matthews, 2004; J.-C. Wang and Wang, 2012], but they are restricted to videos that contain emotional music or talking faces. On the contrary, our method is proposed for more general situations and avoids using expensive and unreliable region segmentation and tracking.

Methodologically, our work uses the bipartite graph partitioning technique [Dhillon, 2001] to obtain the bi-modal codebooks. Bipartite graph partitioning has been widely adopted in many applications. For example, Liu *et al*. [J. Liu and Savarese, 2011] used a bipartite graph to model the co-occurrence of two related views based on visual vocabularies, and the graph partitioning algorithm was applied to find visual word co-clusters. The generated co-clusters not only transfer knowledge across different views, but also allow cross-view action recognition. To model the co-occurrence relations between words from different domains, Pan *et al*. [Dhillon, 2001] adopted a bipartite graph and spectral clustering to discover cross-domain word clusters. This way, the clusters can reduce the gap between different domains, and achieve good performance in cross-domain sentiment classification. In contrast to these applications that focus on cross-domain/view learning, we propose using a bipartite graph to discover the correlations between audio and visual words. Another algorithm used in our approach is MKL [A. Rakotomamonjy and Grandvalet, 2009; A. Vedaldi and Zisserman, 2009], which has been frequently adopted in many computer vision and multimedia tasks.

6.3 Unimodal Feature Representations

Before introducing the bi-modal BoW representation, let us briefly describe the popular unimodal BoW feature representations, that are the basis of our approach. Typical audio/visual BoW representation involves three steps: first, a set of descriptors (visual/audio) is extracted from a video corpus. Then the descriptors are used to generate visual/audio codebooks using k-means clustering. Each cluster describes a common pattern of the descriptors, and it is usually referred to as a codeword. With the codebook, feature pooling is performed to aggregate all the descriptors in each video[1] to form a single fixed dimensional feature vector.

We describe the visual/audio descriptors applied in this work as follows:

- **Static Sparse SIFT Appearance Feature**. The effectiveness of SIFT descriptors [Lowe, 2004] has been proven in numerous object and scene recognition tasks. It is therefore adopted to characterize the static visual information in video frames. Following the work of [Y.-G. Jiang and Chang, 2010], we adopt two versions of sparse keypoint detectors: Difference of Gaussians [Lowe, 2004] and Hessian Affine [Mikolajczyk and Schmid, 2004], to find local keypoints in the frames. Each keypoint descriptor is described by a 128-dimensional SIFT vector. To reduce computational costs, we sample one frame every 2 seconds. Finally, the SIFT features within a frame are further quantized using a SIFT codebook and form a $5,000$-dimensional BoW histogram.

- **Motion-based STIP Feature**. Motion information is always an important clue for video content recognition. For this, we adopt the commonly used Spatial–Temporal Interest Points (STIP) [L. Laptev and Rozenfeld, 2008]. STIP extracts space–time local volumes with significant variations in both space and time. We apply Laptev's algorithm [Laptev and Lindeberg, 2003] to locate the volumes and compute the corresponding descriptors. In particular, a local volume is described by the concatenation of Histogram Of Gradients (HOG) and Histogram of Optical Flow (HOF). This leads to a 144-dimensional vector for each volume, which is then quantized with a codebook to produce a $5,000$- dimensional BoW histogram.

[1] Normally, event detection is performed at the video level, i.e., to detect whether a video contains an event of interest. Therefore, we represent each video by a feature vector.

- **Acoustic MFCC Feature**. In addition to the aforementioned visual features, audio information provides another important clue for video event detection [J.-C. Wang and Wang, 2012]. To utilize this, we adopt the popular MFCC [Pols, 1966a] and compute a 60-dimensional MFCC feature for every temporal window of 32ms. The features are densely computed with nearby windows that have 50% overlap. Finally, the MFCC features are quantized into a $4,000$ -dimensional BoW histogram, in the same way as we quantize the visual features.

With these unimodal features, for each video clip, we have a $10,000$-dimensional visual BoW representation by concatenating the BoW histograms generated from SIFT and STIP ($5,000 + 5,000$), and a $4,000$ dimensional audio BoW representation. These are used to compute the bi-modal representation discussed in the next section.

6.4 Joint Audio-Visual Bi-Modal Words

We now introduce the audio–visual bi-modal representation in detail. We first introduce the construction of the bipartite graph based on the audio BoW and visual BoW representation, and the way of generating the bi-modal codewords. Then we describe three pooling strategies used for re-quantizing the original visual/audio BoW into the joint audio–visual bi-modal BoW representation. Finally, we discuss integrating the bi-modal BoW representations generated at different granularities using MKL.

6.4.1 Audio-Visual Bipartite Graph Construction

Let $\mathcal{D} = \{d_i\}_{i=1}^{n}$ be a training collection with n videos. Denote the audio BoW feature of video d_i as $\mathbf{h}_i^a$ and its visual BoW feature as $\mathbf{h}_i^v$, i.e., $d_i = \{\mathbf{h}_i^a, \mathbf{h}_i^v\}$, where $\mathbf{h}_i^a$ is $4,000$-dimensional and $\mathbf{h}_i^v$ is $10,000$-dimensional. These features are ℓ_1 normalized such that the sum of its entries equals to 1. In addition, we use $\mathcal{W}^a = \{w_1^a, \ldots, w_{m_a}^a\}$ and $\mathcal{W}^v = \{w_1^v, \ldots, w_{m_v}^v\}$ to denote the sets of audio and visual words respectively, where $w_i^a \in \mathcal{W}^a$ represents an audio word and $w_i^v \in \mathcal{W}^v$ indicates a visual word, and m_a and m_v denote the number of audio and visual words, respectively. The total number of audio and visual words is $m = m_a + m_v$.

We further define an undirected graph $G = (V, E)$ between the audio and visual words, where V and E denote the set of vertices and edges, respectively. Let V be a finite set of vertices $V =$

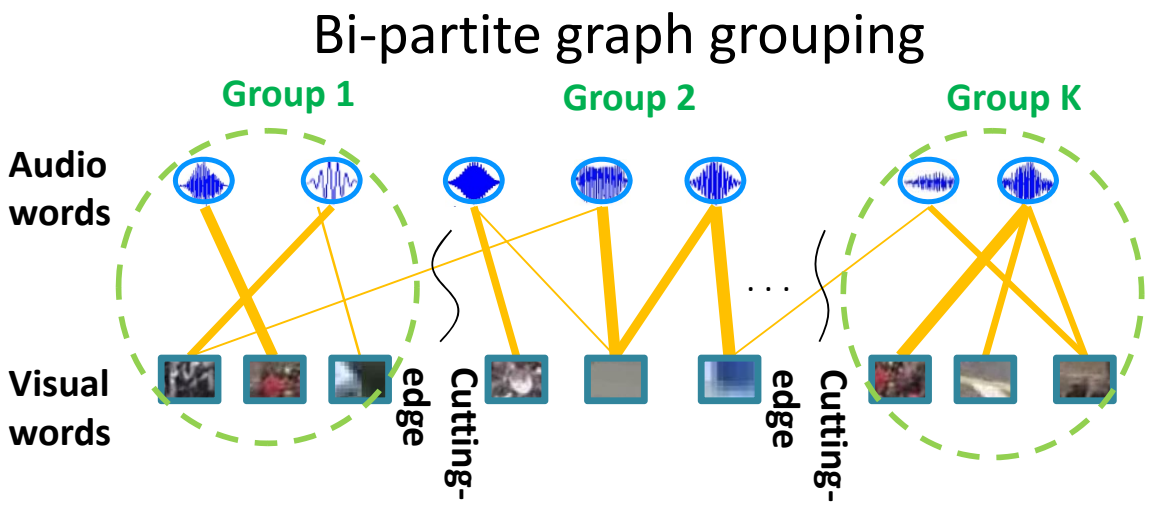


Figure 6.3: An illustration of the bipartite graph constructed between audio and visual words, where the upper vertices denote the audio words and the lower vertices denote the visual words. Each edge connects one audio word and one visual word, which is weighted by the correlation measure calculated based on Eq. (6.1). In this figure, the thickness of the edge reflects the value of the weight.

$V^a \bigcup V^v$, where each vertex in V^a corresponds to an audio word in $\mathcal{W}^a$ and each vertex in V^v corresponds to a visual word in $\mathcal{W}^v$. An edge in E connects two vertices in V^a and V^v, and there is no intra-set edge that connects the two vertices in V^a or V^v, respectively. This graph $G = (V, E)$, where $V = V^a \bigcup V^v$, is commonly called a *bipartite* graph. To measure the correlation between an audio $w_k^a \in \mathcal{W}^a$ and visual $w_l^v \in \mathcal{W}^v$ word, we assign a non-negative weight s_{kl} to any edge $e_{kl} \in E$, which is defined as follows,

$$s_{kl} = \frac{\sum_{i=1}^{n} \mathbf{h}_i^a(k)\mathbf{h}_i^v(l)}{\sum_{i=1}^{n} \mathbf{h}_i^a(k) \sum_{i=1}^{n} \mathbf{h}_i^v(l)}, \tag{6.1}$$

where $\mathbf{h}_i^a(k)$ denotes the entry of $\mathbf{h}_i^a$ that corresponds to the kth audio word w_k^a and $\mathbf{h}_i^v(l)$ denotes the entry of $\mathbf{h}_i^v$ that corresponds to the lth visual word w_l^v.

In Eq. (6.1), the numerator measures the summation of the joint probability of the audio w_k^a and visual w_l^v words, where the summation is calculated over the entire video collection. This value essentially reveals the correlation of the audio and visual words. On the other hand, the denominator acts as a normalization term, that penalizes the audio and/or visual words that appear frequently in the video collection. It is also worth noting that the choice of the correlation measure in Eq. (6.1)

is flexible. We can also estimate the weight s_{kl} by applying other methods, such as Pointwise Mutual Information (PMI) [J. Liu and Savarese, 2011]. Figure 6.3 gives a conceptual illustration of a bipartite graph constructed from the joint statistics of the audio and visual words.

6.4.2 Discovering Bi-Modal Words

We adopt the standard bipartite graph partitioning method to discover audio–visual bi-modal words. Following [Dhillon, 2001], we begin with a bipartitioning method over the bipartite graph, and then extend it into the multipartitioning scenario.

Recall that we have a bipartite graph $G = (V, E)$ between the audio and visual words. Given a partitioning of the vertex set V into two subsets V_1 and V_2, the cut between them can be defined as the sum of all edge weights that connect the vertices from the two subsets,

$$\mathrm{cut}(V_1, V_2) = \sum_{k \in V_1, l \in V_2} s_{kl}. \tag{6.2}$$

The bipartite partition problem over the bipartite graph is to find the vertex subsets V_1^* and V_2^* such that $\mathrm{cut}(V_1^*, V_2^*) = \min_{V_1, V_2} \mathrm{cut}(V_1, V_2)$. To this end, we define the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{m \times m}$ associated with the bipartite graph G as,

$$L_{kl} = \begin{cases} \sum_l s_{kl}, & k = l, \\ -s_{kl}, & k \neq l \text{ and } e_{kl} \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

Given a bipartitioning of V into V_1 and V_2, we further define a partition vector $\mathbf{p} \in \mathbb{R}^m$ that characterizes this division, where the ith entry describes the partitioning state of $i \in V$,

$$p_i = \begin{cases} +1, & i \in V_1, \\ -1, & i \in V_2. \end{cases} \tag{6.4}$$

With the above definitions, it can be proven that the graph cut can be equally written in the following form,

$$\mathrm{cut}(V_1, V_2) = \frac{1}{4}\mathbf{p}^\top \mathbf{L}\mathbf{p} = \frac{1}{4} \sum_{(i,j) \in E} s_{ij}(p_i - p_j)^2. \tag{6.5}$$

However, it can be easily seen from Eq. (6.5) that the cut is minimized by a trivial solution when all p_i's are either $+1$ or -1. To avoid this problem, a new objective function is used to achieve not

only the minimized cut, but also a balanced partition. Formally, the objective function is defined as follows,

$$Q(V_1, V_2) = \frac{\text{cut}(V_1, V_2)}{\text{weight}(V_1)} + \frac{\text{cut}(V_1, V_2)}{\text{weight}(V_2)}, \tag{6.6}$$

where $\text{weight}(V_i) = \sum_{k,l \in V_i} s_{kl}$, $i = 1, 2$. Then it can be proven that the eigenvector that corresponds to the second smallest eigenvalue of the generalized eigenvalue problem $\mathbf{Lz} = \lambda \mathbf{Dz}$ (where $\mathbf{D}$ is a diagonal matrix with $D(k,k) = \sum_l s_{kl}$) provides a real relaxed solution of the discrete optimization problem in Eq. (6.6) [Lutkepohl, 1997]. To obtain the eigenvector that corresponds to the second smallest eigenvalue, [Dhillon, 2001] proposed a computationally efficient solution through Singular Value Decomposition (SVD). In particular, for the given bipartite graph G, we have

$$\mathbf{L} = \begin{pmatrix} \mathbf{D}_1 & -\mathbf{S} \\ -\mathbf{S}^\top & \mathbf{D}_2 \end{pmatrix}, \text{and } \mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_2 \end{pmatrix}, \tag{6.7}$$

where $\mathbf{S} = [s_{kl}]$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices such that $D_1(k,k) = \sum_l s_{kl}$ and $D_2(l,l) = \sum_k s_{kl}$. Let the normalized matrix $\hat{\mathbf{S}} = \mathbf{D}_1^{-1/2} \mathbf{S} \mathbf{D}_2^{-1/2}$, it can be proven that the eigenvector that corresponds to the second smallest eigenvalue of $\mathbf{L}$ can be expressed in terms of the left and right singular vectors that corresponds to the second largest singular value of $\hat{\mathbf{S}}$ as follows,

$$\mathbf{z}_2 = \begin{bmatrix} \mathbf{D}_1^{-1/2} \mathbf{u}_2 \\ \mathbf{D}_2^{-1/2} \mathbf{v}_2 \end{bmatrix}, \tag{6.8}$$

where $\mathbf{z}_2$ is the eigenvector that corresponds to the second smallest eigenvalue of $\mathbf{L}$, $\mathbf{u}_2$ and $\mathbf{v}_2$ are, respectively, the left and right singular vectors that corresponds to the second largest singular value of $\hat{\mathbf{S}}$.

Finally, we need to use $\mathbf{z}_2$ to find the approximated optimal bipartitioning by assigning each $\mathbf{z}_2(i)$ to the clusters $\mathcal{C}_j$ $(j = 1, 2)$ such that the following sum-of-squares criterion is minimized,

$$\sum_{j=1}^{2} \sum_{\mathbf{z}_2(i) \in \mathcal{C}_j} (\mathbf{z}_2(i) - m_j)^2, \tag{6.9}$$

where m_j is the cluster center of $\mathcal{C}_j$ $(j = 1, 2)$.

The above objective function can be practically minimized by directly applying the k-means clustering method on the 1-dimensional entries of $\mathbf{z}_2$. The bipartitioning method can be easily extended to a general case of finding K audio-visual clusters [Dhillon, 2001]. Suppose we have $l =$

$\lceil \log_2 K \rceil$ singular vectors $\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_{l+1}$, and $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_{l+1}$, then we can form the following matrix with l columns,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{D}_1^{-1/2}\mathbf{U} \\ \mathbf{D}_2^{-1/2}\mathbf{V} \end{bmatrix}, \tag{6.10}$$

where $\mathbf{U} = [\mathbf{u}_2, \ldots, \mathbf{u}_{l+1}]$ and $\mathbf{V} = [\mathbf{v}_2, \ldots, \mathbf{v}_{l+1}]$. Based on the obtained matrix $\mathbf{Z}$, we further run k-means method on it to obtain K clusters of audio–visual words, which can be represented as follows,

$$\mathcal{B} = \{B_1, \ldots, B_K\}, \tag{6.11}$$

where each B_i consists of the audio word subset $\mathcal{W}_i^a$ and the visual word subset $\mathcal{W}_i^v$ falls in the same bi-modal cluster. Note that either $\mathcal{W}_i^a$ or $\mathcal{W}_i^v$ can be empty, indicating that only one modality forms a consistent pattern within the bi-modal word B_i (e.g., visual words that corresponds to the background scene).

The above graph partition method needs to compute eigenvectors of the Laplacian matrix, and thus has a computational complexity of $\mathcal{O}(m^3)$ in general, where m is the total number of audio and visual words. We implement the method using MATLAB with a Six-Core Intel Xeon Processor X5660 (2.8 GHz) and 32 GB memory. A total of 32 minutes is required to group $14,000$ audio and visual words into $2,000$ bi-modal words in the experiment on the CCV dataset (cf. Sect. 6.5.1).

6.4.3 Bi-Modal BoW Generation

After generating the bi-modal codewords, we need to map the original visual and audio descriptors to the new codebook. The main purpose here is to fuse the original two visual and audio representations into one joint representation to be used for event classification. For this, we adopt three different quantization strategies. Given a video $d_i = (\mathbf{h}_i^a, \mathbf{h}_i^v)$, the audio–visual bi-modal BoW representations generated by average pooling, max pooling, and hybrid pooling are described as follows.

Average Pooling Average pooling treats the audio and visual words as equally important. Formally, the bi-modal BoW generation strategy is described as follows,

$$\mathbf{h}_i^{\text{avg}}(k) = \frac{\sum_{w_p^a \in \mathcal{W}_k^a, w_q^v \in \mathcal{W}_k^v} (\mathbf{h}_i^a(p) + \mathbf{h}_i^v(q))}{|\mathcal{W}_k^a| + |\mathcal{W}_k^v|}, \tag{6.12}$$

Algorithm 1 Audio-Visual Bi-Modal BoW Representation Generation Procedure

1: **Input:** Training video collection $\mathcal{D} = \{d_i\}$ where each d_i is represented as a multi-modality representation $d = \{\mathbf{h}_i^a, \mathbf{h}_i^v\}$; Size of the audio-visual bi-modal codebook K.
2: Create the correlation matrix $\mathbf{S}$ between the audio and visual words by calculating the co-occurrence probability over $\mathcal{D}$ by Eq. (6.1).
3: Calculate matrix $\mathbf{D}_1$, $\mathbf{D}_2$ and $\hat{\mathbf{S}}$ respectively.
4: Apply SVD on $\hat{\mathbf{S}}$ and select $l = \lceil \log_2 K \rceil$ of its left and right singular vectors $\mathbf{U} = [\mathbf{u}_2, \ldots, \mathbf{u}_{l+1}]$ and $\mathbf{V} = [\mathbf{v}_2, \ldots, \mathbf{v}_{l+1}]$.
5: Calculate $\mathbf{Z} = (\mathbf{D}_1^{-1/2}\mathbf{U}, \mathbf{D}_2^{-1/2}\mathbf{V})^\top$.
6: Apply k-means clustering algorithm on $\mathbf{Z}$ to obtain K clusters, which form the audio-visual words $\mathcal{B} = \{B_1, \ldots, B_K\}$.
7: Apply a suitable pooling strategy to re-quantize each video into the audio-visual bi-modal BoW representation.
8: **Output:** Audio-visual BoW representation.

where w_p^a means the pth audio word, w_q^v represents the qth visual word, and $\mathbf{h}_i^{\text{avg}}(k)$ denotes the entry in the bi-modal BoW $\mathbf{h}^{avg}$ that corresponds to a given audio–visual bi-modal word $B_k = (\mathcal{W}_k^a, \mathcal{W}_k^v)$. $|\mathcal{W}_k^a|$ and $|\mathcal{W}_k^v|$ denote the cardinalities of $\mathcal{W}_k^a$ and $\mathcal{W}_k^v$ respectively. As we can see in Eq.(6.12), the measure of the entry in the bi-modal representation is the average value of the entries of the audio and visual words in the original BoW representations. We call such bi-modal BoW generation strategy *average pooling* because of its relatedness with regard to the pooling strategy in sparse coding [Y.-L. Boureau and Lecun, 2010].

Max Pooling Max pooling selects the maximum summation in the original audio or visual words as the quantization value of the given audio–visual bi-modal word, formally defined as follows,

$$\mathbf{h}_i^{\max}(k) = \max\big(\sum_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p), \sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)\big). \tag{6.13}$$

Hybrid Pooling We also propose a hybrid pooling strategy that integrates average and max pooling. Intuitively, the visual features from the visual scene in the video tend to persist over a certain interval when the camera does not move too fast. Therefore, we use average pooling to aggregate information in the interval. Max pooling is employed for the audio information because

audio features tend to be transient in time. Formally, the hybrid pooling strategy can be defined as follows,

$$\mathbf{h}_i^{\text{hyb}}(k) = \frac{1}{2}\big(\max_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p) + \frac{\sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)}{|\mathcal{W}_k^v|}\big), \tag{6.14}$$

where the average pooling aggregates the two entries of the audio and visual words obtained from max and average pooling, respectively.

Algorithm 1 provides the detailed flow of the generation procedure of the bi-modal BoW representation.

6.4.4 Combining Multiple Joint Bi-Modal Representations

As is true for any BoW-based representation, it is extremely difficult (if not impossible) to identify a suitable number of codewords. Existing works mostly set a fixed number (a few thousand) codewords, that have been empirically observed to work well in practice. Because our bi-modal words are generated on top of the audio and visual words, the problem becomes more complicated given that there is no (even empirical) evidence of a suitable word number for the joint codebook. Using a small bi-modal codebook can result in ambiguous audio–visual patterns within a bi-modal word. On the other hand, the joint audio–visual patterns can be separated immensely if the codebook size is too large.

To alleviate the effect of codebook size, we propose generating the bi-modal BoW representation at different granularities, i.e., with different codebook sizes. The representations are then combined through the well-known MKL framework. In particular, suppose we have the joint bi-modal BoW representations generated from T bipartite graph partitioning with different resolutions (i.e., cluster number), and denote the kernel matrix that corresponds to the histogram generated at the tth resolution as $K_t(h, h')$. MKL seeks an optimal combination $K(h, h') = \sum_{t=1}^{T} d_t K_t(h, h')$ with the constraints $d_t \geq 0,\ \forall t$ and $\sum_{t=1}^{T} d_t = 1$. By using this $K(h, h')$ for event classification, performance can usually be boosted compared with using a single kernel. Many MKL frameworks [Jhuo and Lee, 2010; A. Kembhavi and Davis, 2009; A. Vedaldi and Zisserman, 2009] have been proposed and demonstrated for visual classification. We adopt the widely used simpleMKL framework [A. Rakotomamonjy and Grandvalet, 2009] because of its sound performance and efficiency. In this MKL framework, each kernel K_t, is associated with a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_t$, and the decision function is in the form $f(h) + b = \sum_t f_t(h) + b$ where each f_t is

associated with $\mathcal{H}_t$. The objective of simpleMKL is as follows:

$$\begin{aligned} \min_{f_t,b,\xi,d} \quad & \frac{1}{2}\sum_t \frac{1}{d}\|f_t\|^2_{\mathcal{H}_t} + C\sum_i \xi_i \\ s.t. \quad & y_i \sum_i f_t(x_i) + y_i b \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \\ & \sum_t d_t = 1, \quad d_t >= 0, \quad \forall t, \end{aligned} \tag{6.15}$$

To solve the above objective, we use the simpleMKL solver [A. Rakotomamonjy and Grandvalet, 2009].

6.5 Experiments

In this section, we evaluate our proposed audio–visual bi-modal representation for video event detection using three datasets: TRECVID MED 2011 dataset, a large dataset that consists of both TRECVID MED 2010 and TRECVID MED 2011, and CCV dataset.

6.5.1 Datasets

TRECVID MED 2011 Dataset. TRECVID MED [Nis, 2011] is a challenging task for the detection of complicated high-level events in unconstrained Internet videos. Our first dataset is the MED 2011 development set, which includes five events "Attempting a board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony", and "Working on a woodworking project". This dataset consists of $10,804$ videos from $17,566$ minutes of web videos, and it is partitioned into training ($8,783$ videos) and test ($2,021$ videos) sets. The training set contains approximately 100 positive videos for each event (most videos in the training set are background videos that do not contain any of the five events). Within each class, there exist complicated content variations, thus making the task extremely challenging.

TRECVID MED 2010+2011 Dataset. We also consider the earlier MED 2010 dataset [Nis, 2010]. Because the MED 2010 dataset is too small (fewer than 2,000 videos), we combine TRECVID MED 2010 with MED 2011 [Nis, 2010; Nis, 2011] to form a larger and more challenging event detection dataset. MED 2010 has three events "Assembling a shelter", "Batting a run in", and "Making a

cake", each with 50 positive videos for training and 50 for testing. This combined dataset consists of $14,272$ videos that fall into 8 event categories, and it is partitioned into training ($10,527$ videos) and test ($3,745$ videos) sets. Note that the combination also provides another opportunity for re-examining the performance of the five MED 2011 events when more background videos are added (i.e., the MED 2010 videos).

CCV Dataset. This dataset [Y.-G. Jiang and Loui, 2011] contains $9,317$ YouTube videos annotated over 20 semantic categories, where $4,659$ videos are used for training and the remaining $4,658$ are used for testing. Most of the 20 categories are events, with a few categories that belong to objects or scenes. To facilitate benchmark comparison, we report the performance of all 20 categories.

6.5.2 Experiment Setup

As discussed earlier, we adopt the SIFT BoW (5,000 dimensions) and STIP BoW (5,000 dimensions) representations as the visual features while using the MFCC BoW (4,000 dimensions) as the audio representation. One-vs-all SVM is used to train a classifier for each evaluated event. To obatain the optimal SVM trade-off parameter for each method, we partition the training set into 10 subsets and then perform 10-fold cross validation. Moreover, we adopt the χ^2 kernel because of its outstanding performance in many BoW-based applications, which is calculated as $k(x,y) = exp(-\frac{d_{\chi^2(x,y)}}{\sigma})$, where σ follows the previous work [G. Ye and Chang, 2012], $d_{\chi^2(x,y)}$ is defined as $d_{\chi^2(x,y)} = \sum_{i=1} \frac{(x(i)-y(i))^2}{x(i)+y(i)}$, and σ is by default set as the mean value of all pairwise distances in the training set.

For performance evaluation, we follow previous works [Y.-G. Jiang and Loui, 2011; Natarajan, 2011] and use AP, which approximates the area under a precision–recall curve. We calculate AP for each event, and then use mAP across all event categories in each dataset as the final evaluation metric.

In the following experiments, we systematically evaluate the performance of the following methods:

1. Single Feature (SF). We only report the best performance achieved by one of the three audio/visual features.

2. Early Fusion (EF). We concatenate the three types of BoW features into a long vector with

$14,000$ dimensions.

3. Late Fusion (LF). We use each feature to train an independent classifier and then average the output scores of the three classifiers as the final fusion score for event detection.

4. Average Pooling based Bi-Modal BoW (BMBoW-AP), where the average pooling is employed to generate the bi-modal BoW.

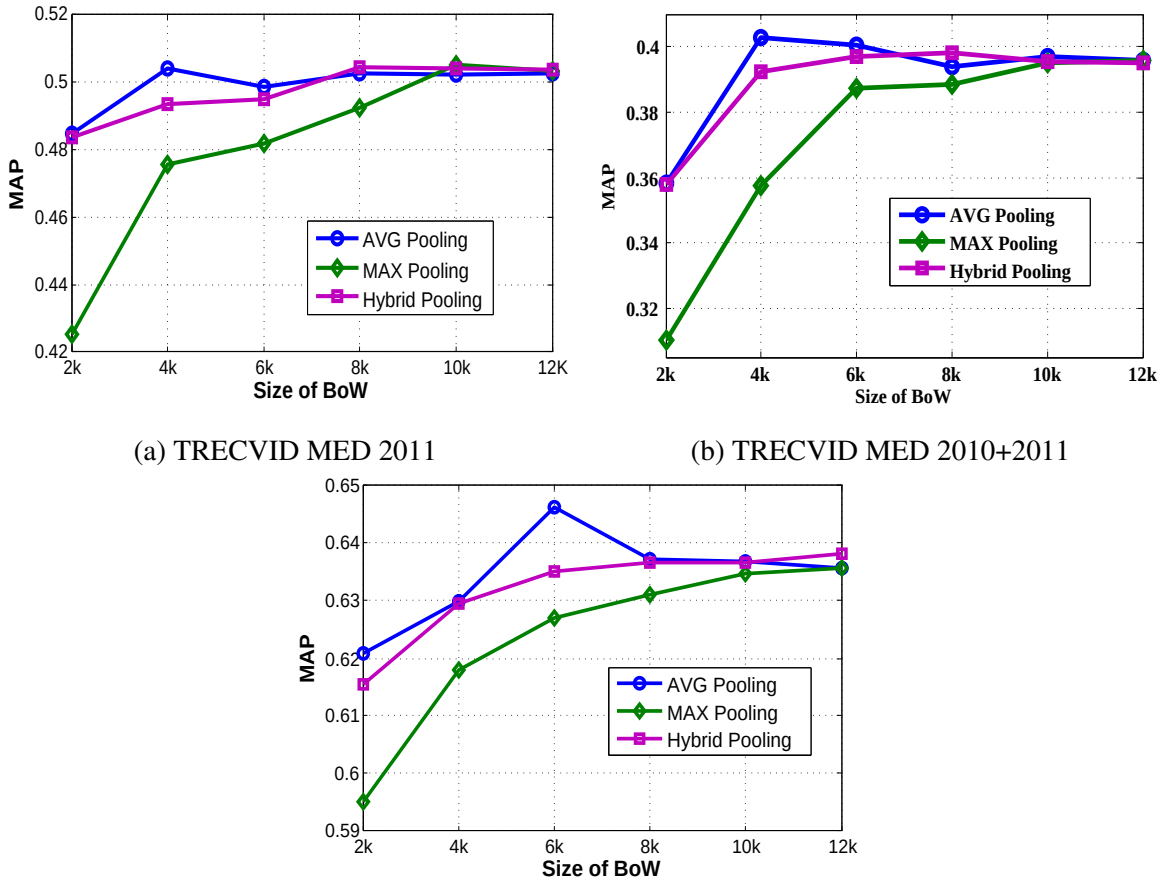
5. Max Pooling based Bi-Modal BoW (BMBoW-MP), where we use max pooling to generate the bi-modal BoW.

6. Hybrid Pooling based Bi-Modal BoW (BMBoW-HP), which applies the hybrid pooling to generate the bi-modal BoW.

7. MKL based Bi-Modal BoW (BMBoW-MKL), which uses MKL to combine multiple bi-modal BoW representations. We use all the codebook sizes as shown in Figure 6.4.

6.5.3 Effect of Codebook Size and Pooling Strategies

We first experimentally evaluate the performance of different codebook sizes and pooling strategies. Because the sizes of audio and visual modalities are 4,000 and 10,000, respectively, we expect each bipartite partitioning to provide a high correlation between audio and visual modalities. Therefore, we increase the size of the bi-modal codebook from 2,000 to 12,000 and discuss the mAP performance with different pooling strategies in Fig. 6.4. We can see that average pooling tends to show better stability than max and hybrid pooling when the codebook size varies, which demonstrates that average pooling is more suitable for the bi-modal BoW quantization. This could be because average pooling captures joint audio–visual patterns, whereas hybrid/max pooling incurs significant information loss caused by considering only the maximum response of audio/visual information. For codebook size, 4,000 seems a good number for MED datasets, but for CCV, large codebooks with 6,000–10,000 bi-modal words seem to be more effective. Note that for such a large bi-modal codebook, there are codewords that contain only word from the audio or visual channel. It makes sense to have such words because, although we would like to discover the correlations between the two modalities, not all words are correlated. Therefore it is good to leave some audio/visual



(a) TRECVID MED 2011

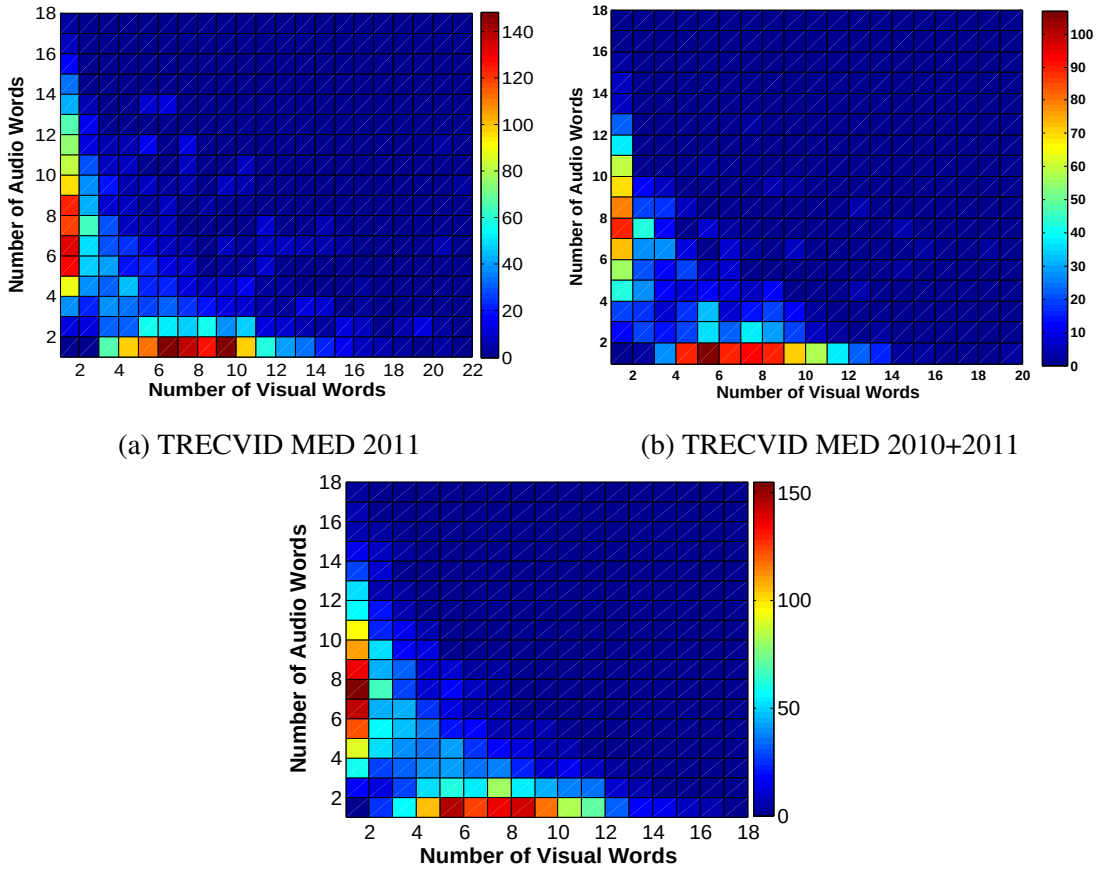
(b) TRECVID MED 2010+2011

(c) CCV Dataset

Figure 6.4: Performance with different bi-modal codebook size and pooling strategies.

words independent in the bi-modal representation. This inconsistent observation also confirms the usefulness of aggregating multiple bi-modal codebooks, which is evaluated later.

We also show the density of audio and visual words within each bi-modal word in Fig. 6.5. Here, each grid in the map denotes the frequency of bi-modal words composed of certain numbers of audio (vertical coordinate) and visual (horizontal coordinate) words. The portion of words in the entire bi-modal codebook that contain both visual and audio information is estimated, and it is found to be approximately 47% for the TRECVID MED 2011 dataset, 39% on the combined TRECVID MED 2010 and 2011 dataset, and 36% for the CCV dataset, respectively. This confirms the significant effect of the audio–visual correlations in the joint bi-modal representation. Therefore,



(a) TRECVID MED 2011

(b) TRECVID MED 2010+2011

(c) CCV Dataset

Figure 6.5: The density of audio and visual words in the bi-modal words.

the bi-modal word is also an important component of a large event detection system that achieves the best performance [Natarajan, 2011] in TRECVID MED 2011. We observe that several bi-modal words contain more visual than audio words, or the opposite of having more audio than visual words (i.e., the bins close to the x or y axis in Fig. 6.5). This could be because some large visual or audio patterns are only correlated to a small clue in the other modality. For instance, a birthday scene with many visual characteristics might be only highly correlated to cheering sounds.

6.5.4 Performance Comparison on TRECVID MED 2011 Dataset

Let us now evaluate the seven methods listed in Sect.6.5.2. Figure 6.7 shows the results on the MED 2011 dataset. We fix the size of the bi-modal codebook to be 4,000 with the exception of the

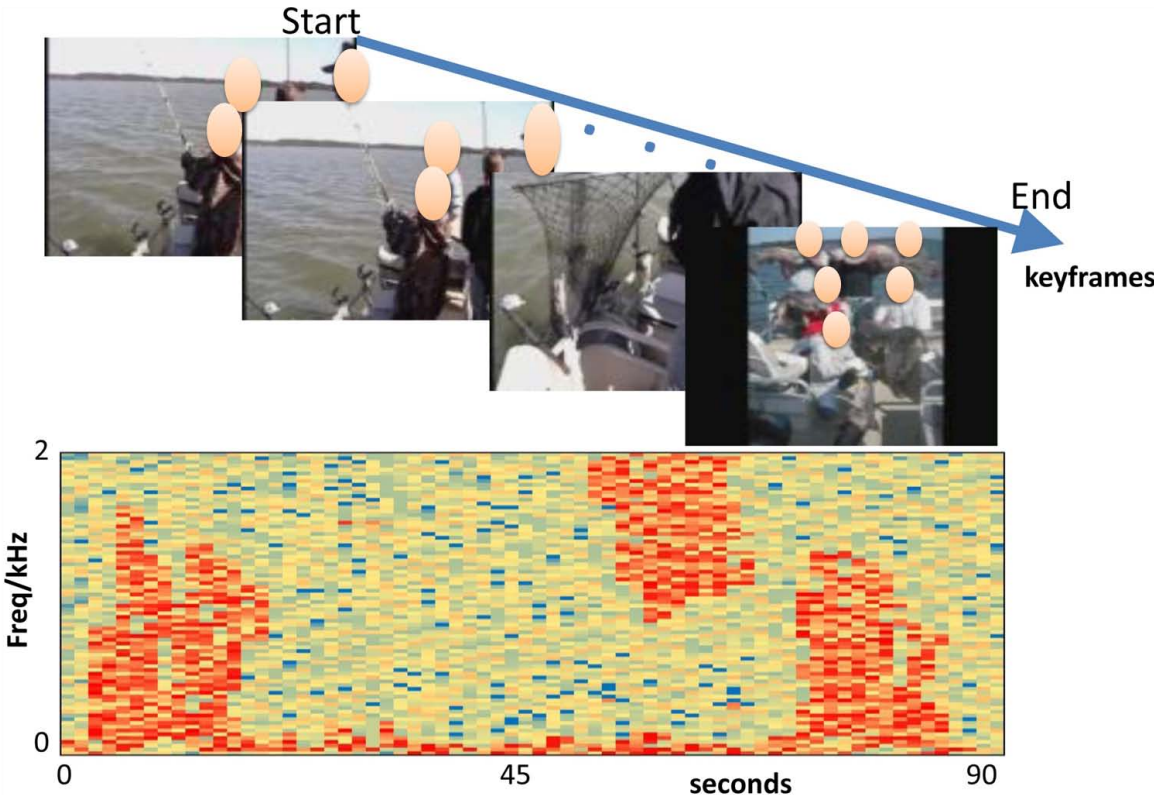


Figure 6.6: An example of audio-visual correlations in the event "Landing a fish" from the TRECVID MED 2011 dataset. We see that there are clear audio patterns correlating with the beginning and the end (fish successfully landed) of the event.

BMBoW-MKL method, which combines multiple codebook sizes as shown in Fig. 6.4. In addition, we adopt average pooling in BMBoW-MKL, because—as is shown—it outperforms max pooling and hybrid pooling. Based on the results, we obtain the following findings:

1. Our proposed bi-modal word representation outperforms all other baseline methods in terms of mAP, which proves the effectiveness of this approach. In particular, it outperforms the most popularly used early fusion and late fusion methods by a large margin. This is because that the bi-modal words not only capture the correlation between audio and visual information, but also aggregate their mutual dependence.

2. As an important but quite obvious conclusion, the bi-modal word representation performs significantly better than all the single features, which verifies the merits of considering multi-

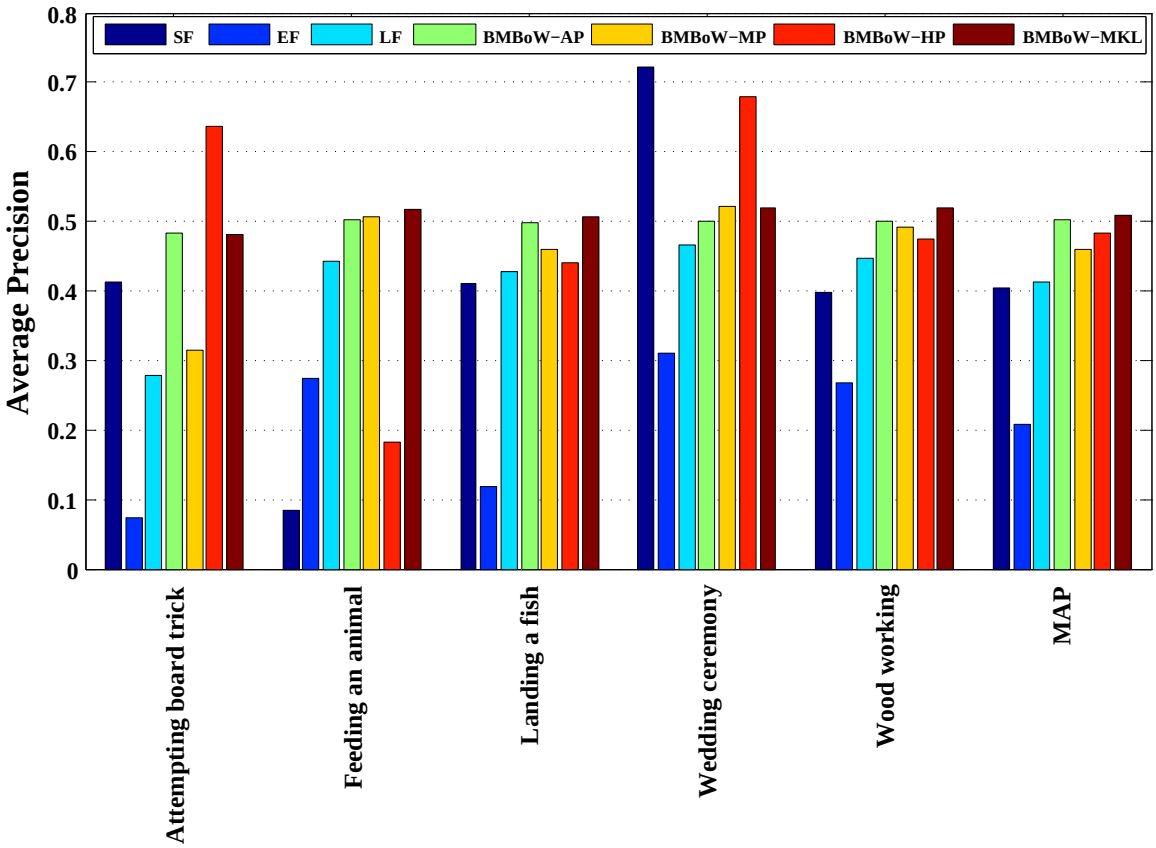


Figure 6.7: Per-event performance on TRECVID MED 2011 dataset. This figure is best viewed in color.

modality in the task of video event detection.

3. As indicated in the introduction, visual and audio information of the same event category often present consistent joint patterns. This not only holds for events with intuitive audio–visual correlations, such as "Batting a run in", but it is also true for events that contain fewer audio clues. Fig. 6.6 shows an example. In the event "Landing a fish", although the soundtrack is mostly quite silent, at the start and after the fish is successfully landed, there are some clear audio patterns. Our method can capture such local correlations, which is the main reason that it performs better than the simple fusion strategies.

4. BMBoW-AP tends to provide better results that BMBoW-MP, which could be because the former captures joint audio–visual patterns, whereas the latter incurs significant information

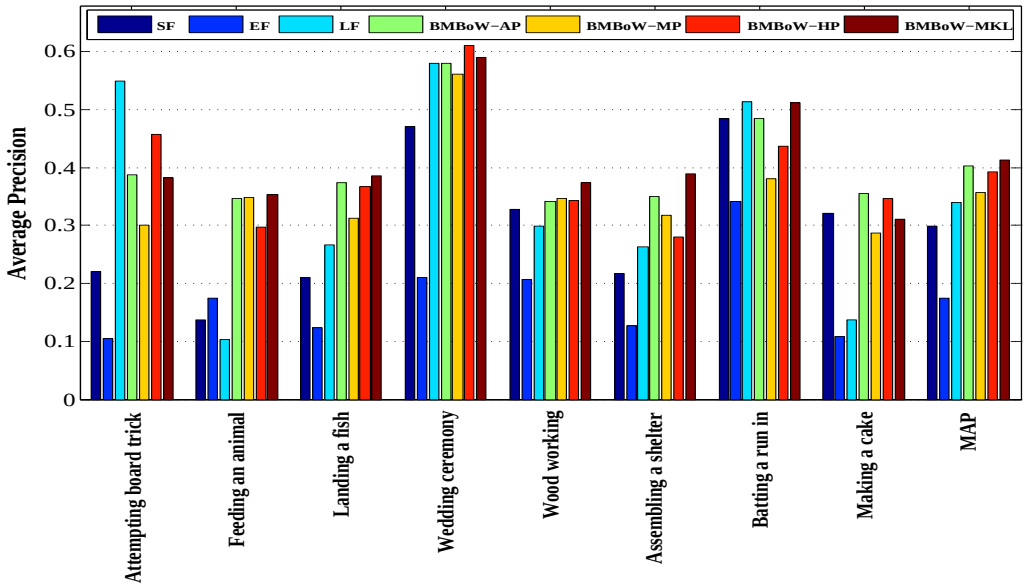


Figure 6.8: Per-event performance comparison on TRECVID MED 2010+2011 dataset, which includes eight events. This figure is best viewed in color.

loss caused by selecting only the maximum contribution between two modalities.

5. BMBoW-HP outperforms BMBoW-MP, because it utilizes perhaps the more suitable pooling strategies for different modalities (i.e., max pooling for the transient audio signal and average pooling for the persistent visual signal). For some events, BMBoW-HP even achieves better results than BMBoW-AP, indicating that selecting the maximum response of audio signals could help reveal the semantic clue of the videos. However, in general, BMBoW-AP is the best among the three pooling strategies.

6. BMBoW-MKL shows better results than all the methods based on a single bi-modal codebook, confirming the fact that using multiple codebooks is helpful.

6.5.5 Performance Comparison on TRECVID MED 2010+2011 Dataset

Figure 6.8 shows the per-event performance for all the methods on this combined dataset. From the results, we can see that the MKL-based bi-modal representation, i.e., BMBoW-MKL, achieves the best performance. In particular, it outperforms BMBoW-AP, BMBoW-MP, and BMBoW-HP by 0.96%, 5.53%, and 1.96% respectively in terms of mAP. Among the three pooling methods,

average pooling offers the best result. In addition, comparing the results of the five 2011 events on this combined dataset with those of MED 2011, we also observe that the performances of early and late fusion are not as stable as that of the bi-modal representations when more background videos are added. For example, late fusion performs quite badly for the "Attempting board trick" event on MED 2011, but it is very good on the combined dataset.

6.5.6 Performance Discussion on CCV Dataset

Figure 6.9 further shows the per-category performance comparison on the CCV dataset, where the bi-modal codebook size is set at $6,000$, with the exception of the BMBoW-MKL method. Again, the results show that BMBoW-MKL achieves the best performance in terms of mAP. It outperforms BMBoW-AP, BMBoW-MP, and BMBoW-HP by 1.16%, 2.26%, and 7.36%, respectively. Moreover, BMBoW-MKL also achieves the best performance on most event categories. For instance, on event "graduation", it outperforms the best baseline method SF by 15.05%. In addition, compared with the best baseline EF, our method achieves the highest relative performance gain on categories "birds " and "Wedding ceremony". This could be because these two categories contain more significant audio–visual correlations than the other categories. For example, the appearance of birds is often accompanied with a singing sound. Meanwhile, people's actions in a wedding ceremony are always accompanied by background music. In general, we expect high impact of the proposed bi-modal features on other events that share strong audio–visual correlations, such as the ones mentioned above.

6.5.7 Statistical Significance Testing

We also measure the statistical significance between the best baseline and BMBoW-MKL on the three datasets. We use a popular measure for statistical significance testing, the p-value, which is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true [Pte, 2015]. We can reject the null hypothesis when the p-value is less than the significance level, which is often set at 0.05. When the null hypothesis is rejected, the result is said to be statistically significant. To obtain the p-value, we sample 50% of the test set from each dataset and repeat the experiment 1000 times. For each round, we compute the paired mAP differences $D_i = MAP_{\mathrm{BMBoW-MKL}}(i) - MAP_{\mathrm{Baseline}}(i)$, where $i = 1, 2, \ldots, 1000$. Then we

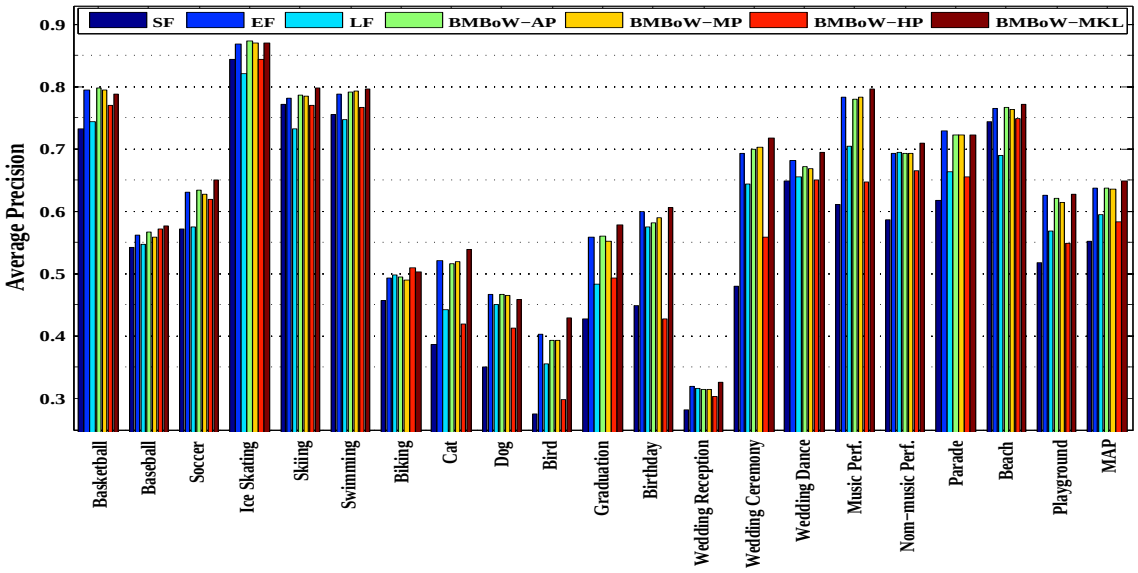


Figure 6.9: Per-category performance comparison on CCV dataset. This figure is best viewed in color.

make the assumption that the null hypothesis is $D_i < 0, i = 1, 2, \ldots, 1000$, based on which, the p-value can be defined as the percentage of D_i that is below 0. We find that the p-values obtained on the MED 2011, MED 2010+2011, and CCV datasets are 0.015, 0.018, and 0.022, respectively, which are well below 0.05 and show that the null hypothesis can be rejected. Therefore, we can conclude that our method has achieved statistically significant improvements over the best baseline on the three datasets.

6.6 Summary

In this chapter, we introduced a bi-modal representation to better explore the power of multi-modality representation in video event detection. The proposed method uses a bipartite graph to model the relationship between visual and audio words and partitions the graph to generate audio–visual bi-modal words. Several popular pooling methods were evaluated to generate BoW representation using bi-modal words, and average pooling was found to be the best performer. Extensive experiments on three datasets consistently showed that the proposed bi-modal representation significantly outperforms early and late fusion, which are currently the most widely used multimodal fusion methods. In addition, because there is no single codebook size that is suitable in all cases,

we proposed using multiple bi-modal codebooks and MKL to combine BoW representations based on different codebooks. The results showed that using MKL and multiple bi-modal codebooks is always helpful. With these findings, we conclude that many state-of-the-art video event detection systems might have overlooked the importance of joint audio–visual modeling. We would also like to underline that—although some promising results from the perspective of bi-modal words' were shown in this chapter—advanced joint audio-visual representations continues to be a topic that deserves more in-depth studies in the future. It is also interesting and important to construct a larger dataset for evaluating these new representations.

Chapter 7

Robust Multi-Source Fusion with Rank Minimization

7.1 Introduction

Multiple features are often considered in video event detection because a single feature cannot provide sufficient information. Systems that combine multiple features have also been proven to improve the classification performance in various visual classification tasks [Bach *et al.*, 2004; Gehler and Nowozin, 2009; Y.-G. Jiang and Chang, 2010].

There are two popular strategies for fusing features: early fusion and late fusion. Early fusion, also known as feature level fusion, has been widely used in the computer vision and multimedia communities [Bach *et al.*, 2004; Gehler and Nowozin, 2009; Y.-G. Jiang and Chang, 2010]. One characteristic method is to represent the features as multiple kernel matrices, and then combine them in the kernel space. One of the most successful feature fusion methods is MKL [Bach *et al.*, 2004], which learns a linear or non-linear kernel combination and the associated classifier simultaneously. However, MKL might not produce better performance in real-world applications. In [Gehler and Nowozin, 2009], the authors prove that even simple feature combination strategies that are much faster than MKL, can achieve highly comparable results with MKL.

The other strategy is late fusion. It aims at combining the confidence scores of the models constructed from different features, where each confidence score measures the possibility of classifying a test sample into the positive class by one specific model. Compared with early fusion, late fusion

is easier to implement and often shows to be effective in practice. However, one problem with this strategy comes from the possible heterogeneity among the confidence scores provided by different models. In practice, such heterogeneity results from the variation of the discriminative capability of each model in a certain feature space, thus producing incomparable confidence scores at different numeric scales. This makes the direct combination of confidence scores from different models inappropriate, posing a great challenge to the late fusion task.

Existing solutions to this problem typically assume that the confidence scores of the individual models are the posterior probabilities of the samples belonging to the positive class. Because this assumption is not generally true, a normalization step is required to normalize the scores to a common scale such that the combination can be performed [Jain *et al.*, 2005]. However, the main issues with these existing methods are two-fold. First, the choice of normalization schemes is data-dependent and requires extensive efforts in empirical validation [Jain *et al.*, 2005]. Second, they blindly combine all confidence scores, including considerable noise caused by the incorrect predictions made by the models, which could deteriorate fusion performance.

In this chapter, we propose a robust late fusion method, that not only achieves isotonicity (i.e., scale invariance) among the numeric scores of different models, but also recovers a robust prediction score for the individual test sample via removing the prediction error. Given a confidence score vector $\mathbf{s} = [s_1, s_2, \ldots, s_m]$ of a model, where each s_i denotes the score of the ith test sample, and m is the sample number. We first convert $\mathbf{s}$ into a pairwise relationship matrix T such that $T_{jk} = 1$ if $s_j > s_k$, $T_{jk} = -1$ if $s_j < s_k$, $T_{jk} = 0$ if $s_j = s_k$. The matrix T is a skew-symmetric matrix that encodes the comparative relationship of every two test samples under the given model. We apply the above conversion on the score vector of each model, and obtain multiple relationship matrices. This way, the real-valued confidence scores are converted into the integer-valued isotonic pairwise relationships, that address the scale variance problem. Moreover, although the ideal score fusion vector $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_m]$ is unknown, suppose we have a real-valued matrix $\hat{T}$ where $\hat{T}_{jk} = \hat{s}_j - \hat{s}_k$, we can find a rank-2 factorization of $\hat{T}$ such that $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$. By doing so, we can recover the unknown score fusion vector.

Based on the above assumptions, our late fusion method attempts to find a rank-2 relationship matrix from the multiple pairwise relationship matrices. In particular, it infers a common low-rank pairwise relationship matrix by novel joint decompositions of the original pairwise relationship

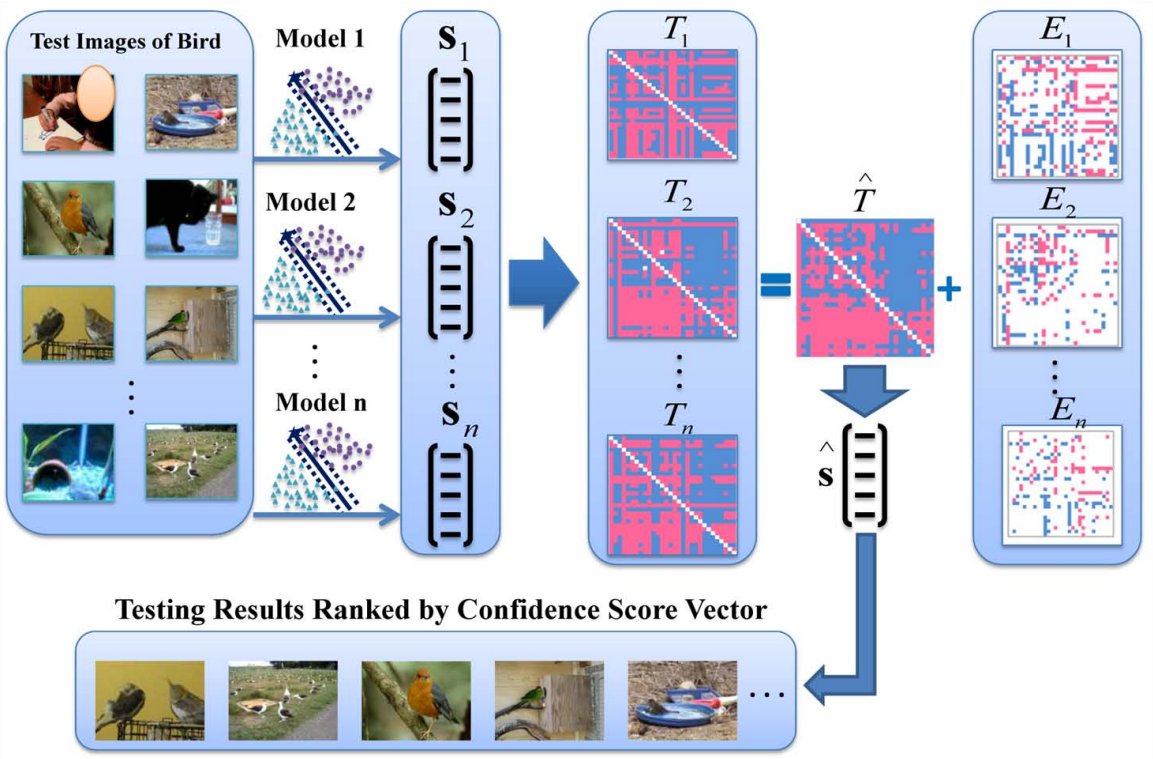


Figure 7.1: An illustration of our proposed method. Given n confidence score vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ obtained from n models, we convert each $\mathbf{s}_i$ into a comparative relationship matrix T_i that encodes the pairwise comparative relation of scores of every two testing images under the ith model. Then we seek a shared rank-2 matrix $\hat{T}$, through which each original matrix T_i can be reconstructed by an additive sparse residue matrix E_i. Finally, we recover from the matrix $\hat{T}$ a confidence score vector $\hat{\mathbf{s}}$ that can more precisely perform the final prediction.

matrices into combinations of the shared rank-2 and sparse matrices. We hypothesize that such common rank-2 matrix can robustly recover the true comparative relationships among the test samples. The joint decomposition process is valuable because each pairwise comparative relationship in the original matrix might be incorrect, yet the joint relationships from multiple matrices might be complementary and could be used to collectively refine the results. Moreover, the individual sparse residue essentially contains the prediction errors for each pair of test samples made by one model.

The fusion procedure is formulated as a constrained nuclear norm and ℓ_1 norm minimization problem, that is convex and can be solved efficiently with the ALM [Lin *et al.*, 2009] method. In addition, we also develop a Graph Laplacian regularized robust late fusion method that incorporates the information from different types of low-level features, which further enhances the performance. Figure 7.1 illustrates the framework of our proposed method. Extensive experiments confirm the effectiveness of the proposed method, achieving a relative performance gain over the state-of-the-art visual tasks. We also show that the proposed multi-source fusion method provides a robust fusion scheme for complex video event detection.

7.2 Related Work

Combining multiple diverse and complementary features is a recent trend in visual classification. A popular feature combination strategy in computer vision is MKL [Bach *et al.*, 2004], which learns an optimized kernel combination and the associated classifier simultaneously. Varma *et al*. [Varma and Ray, 2007] used MKL to combine multiple features and achieved good results on image classification. A recent work in [Gehler and Nowozin, 2009] fully investigated the performance of MKL and proved that MKL might not be more effective than the average kernel combination. Unlike this line of research, we focus on late fusion that works by combining the confidence scores of the models obtained from different features.

There are numerous score late fusion methods in the literature. For example, Jain *et al*. [Jain *et al.*, 2005] transformed the confidence scores of multiple models into a normalized domain, and then combined the scores through a linear weighted combination. In [Nandakumar *et al.*, 2008], the authors used the Gaussian mixture model to estimate the distributions of the confidence scores, and then fused the scores based on the likelihood ratio test. The discriminative model fu-

sion method [Smith *et al.*, 2003] treated the confidence scores from multiple models as a feature vector and then constructed a classifier for different classes. Terrades *et al*. [Terrades *et al.*, 2009] formulated the late fusion as a supervised linear combination problem that minimized the misclassification rates under the ℓ_1 constraint on the combination weights. In contrast, we focus on a novel late fusion method that not only achieves isotonicity but also removes the prediction errors made by individual models.

Methodologically, our work is motivated by recent advances in low rank matrix recovery [Gleich and Lim, 2011; Wright *et al.*, 2009]. One representative is Robust PCA introduced in [Wright *et al.*, 2009], which decomposed a corrupted matrix into low-rank and sparse components. On the contrary, our work attempts to discover a shared low rank matrix from the joint decomposition of multiple matrices into combinations of the shared low rank and sparse components. In [Gleich and Lim, 2011], the authors used a rank minimization method to complete the missing values of the user-item matrix, and then used these values to extract the rank for each item. This is essentially different from our work, which considers multiple complete score matrices for the purpose of robust late fusion.

7.3 Robust Late Fusion with Rank Minimization

In this section, we introduce our Robust Late Fusion (RLF) method. We first explain how to construct the relationship matrix, and then describe the problem formulation.

7.3.1 Pairwise Relationship Matrix Construction

Given the confidence score vector of a model $\mathbf{s} = [s_1, s_2, \ldots, s_m]$, where each s_i denotes the confidence score of the ith test sample and m is the number of test samples, we can construct a $m \times m$ pairwise comparative relationship matrix T where the (j, k)th entry is defined as

$$T_{jk} = \text{sign}(s_j - s_k), \tag{7.1}$$

Obviously, the obtained matrix T encodes the comparative relationship of every two test samples under the given model. In particular, $T_{jk} = 1$ denotes that the jth test sample is more confident to be classified as positive than the kth test sample, whereas $T_{jk} = -1$ denotes the opposite comparative

relationship. Meanwhile, when $T_{jk} = 0$, we believe that the jth and kth samples have the same confidence to be positive.

Compared with confidence scores, the pairwise comparative relationship matrix is a relative measurement that quantizes the real-valued scores into three integers. By converting the absolute values of the raw scores into the pairwise comparative relationships, we naturally arrive at an isotonic data representation that can be used as the input for our late fusion method.

Here, we also consider the reverse problem: given a relative score relationship matrix T, how do we reconstruct the original ranks or scores? If T is consistent, namely all the transitive relationships are satisfied (if $s_i > s_j$ and $s_j > s_k$, then $s_i > s_k$), a compatible rank list can be derived easily. If T is continuous valued (as is the case of the recovered matrix $\hat{T}$ described in the next section), we assume that there exist compatible score vectors $\hat{\mathbf{s}}$ that can be used to explain the relations encoded in $\hat{T}$, i.e., $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$. This formulation naturally leads to a nice property $\mathrm{rank}(\hat{T}) = 2$, which provides a strong rationale to justify using the low-rank optimization method in discovering a common robust $\hat{T}$ when fusing scores from multiple models.

7.3.2 Problem Formulation

Suppose we have a pairwise comparative relationship matrix T constructed from the confidence score vector produced by a model. The entries in T summarize the prediction ability of the given model, where some entries correctly characterize the comparative relationships of the test samples, whereas other entries are incorrect because of the wrong prediction made by the model. Intuitively, the correct entries in T are consistent among the test sample pairs, and hence tend to form a global structure. Moreover, the incorrect entries in T often appear irregularly within the matrix, which can be seen as the sparse errors.

To capture the underlying structure information of the correct entries while removing the error entries that degrade prediction performance, we consider a matrix decomposition problem as follows:

$$\begin{aligned} &\min_{\hat{T},E} \|E\|_1, \\ &\text{s.t.} \quad T = \hat{T} + E,\ \hat{T} = -\hat{T}^\top,\ \mathrm{rank}(\hat{T}) = 2, \end{aligned} \tag{7.2}$$

where $\mathrm{rank}(\hat{T})$ denotes the rank of matrix $\hat{T}$ and $\|E\|_1$ is the ℓ_1 norm of a matrix. By minimizing the objective function, we actually decompose the original matrix T into a rank-2 component $\hat{T}$ and

a sparse component E, which not only recovers the true rank relationships among the test samples, but also removes the incorrect predictions as noises. Finally, the skew-symmetric constraint $\hat{T} = -\hat{T}^\top$ enforces the decomposed $\hat{T}$ to continue to be a pairwise comparative matrix.

The above optimization problem is difficult to solve because of the discrete nature of the rank function. Instead, we consider a tractable convex optimization that provides a good surrogate for the problem:

$$\begin{aligned} &\min_{\hat{T},E} \|\hat{T}\|_* + \lambda\|E\|_1, \\ &\text{s.t.} \quad T = \hat{T} + E,\ \hat{T} = -\hat{T}^\top, \end{aligned} \tag{7.3}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, i.e., the sum of the singular values of the matrix, and λ is a positive tradeoff parameter. Because our implementation for nuclear norm minimization is based on Singular Value Thresholding (SVT), we can continue to truncate the singular values until the rank-2 constraint is satisfied (See section 7.4). Therefore, we can still obtain an exact rank-2 $\hat{T}$ based on the above objective function.

Until now, our formulation considers only one pairwise comparative relationship matrix, and hence cannot be used for the fusion purpose. Suppose we have a set of n pairwise comparative relationship matrices $T_1, \ldots, T_n$, where each T_i is constructed from the score vector $\mathbf{s}_i$ of the ith model. Our robust late fusion is formulated as follows:

$$\begin{aligned} &\min_{\hat{T},E_i} \|\hat{T}\|_* + \lambda \sum_{i=1}^{n} \|E_i\|_1, \\ &\text{s.t. } T_i = \hat{T} + E_i,\ i = 1, \ldots, n, \\ &\quad \hat{T} = -\hat{T}^\top. \end{aligned} \tag{7.4}$$

Compared with the single matrix decomposition in Eq. (7.3), the above objective function attempts to find a shared low-rank pairwise comparative relationship matrix through the joint decompositions of multiple pairwise matrices into pairs of low-rank and sparse matrices. As a result, the $\hat{T}$ matrix recovers the true consistent comparative relationships across multiple relationship matrices. Moreover, each E_i encodes the prediction errors made by one specific model. With the proposed framework, we can robustly recover the comparative relationships among test samples.

7.4 Optimization and Score Recovery

Low-rank matrix recovery is well studied in the literature [Cai *et al.*, 2008; Wright *et al.*, 2009]. However, our optimization problem differs from these existing methods in that we have a skew-symmetric constraint. Fortunately, the following theorem shows that if SVT is used as the solver for rank minimization, this additional constraint can be neglected [Gleich and Lim, 2011].

Theorem 1. *Given a set of n skew-symmetric matrices T_i, the solution for the problem in Eq. (7.4) from the SVT solver (shown in Algorithm 2) is a skew-symmetric matrix $\hat{T}$ if the spectrums between the dominant singular values are separated.*

The theorem can be proven based on the SVD property of a skew-symmetric matrix, which can be found in the supplementary material. Therefore, we can directly employ the existing SVT-based rank minimization methods to solve our problem. It is well known that ALM uses SVT for rank minimization, and shows excellent performance in terms of both speed and accuracy. Therefore, we choose the ALM method for the optimization. We first convert Eq. (7.4) into the following equivalent problem:

$$\begin{aligned}\min_{\hat{T},E_i} \|\hat{T}\|_* + \lambda \sum_{i=1}^{n} \|E_i\|_1 + \sum_{i=1}^{n} \langle Y_i, T_i - \hat{T} - E_i \rangle \\ + \frac{\mu}{2} \sum_{i=1}^{n} \|T_i - \hat{T} - E_i\|_F^2,\end{aligned} \tag{7.5}$$

where Y_i are Lagrange multipliers for the constraints $T_i = \hat{T} + E_i$, $\mu > 0$ is a penalty parameter and $\langle \cdot, \cdot \rangle$ denotes the inner-product operator. Then the optimization problem can be solved by the inexact ALM algorithm as shown in Algorithm 2. Step 4 is solved via the SVT operator [Cai *et al.*, 2008], whereas step 5 is solved via the solution in [Hale *et al.*, 2008]. Note that after the singular value truncating in step 4, even the number of singular values is truncated (see the proof of Theorem 1), and thus the rank of $\hat{T}$ is reduced. During the iterations, we repeat the above truncating operation until the rank-2 constraint in step 8 is satisfied. (i.e., only two non-zero singular values are retained after the progressive truncating). This way, we obtain a rank-2 skew-symmetric matrix.

We implement Algorithm 2 on the 64-bit MATLAB platform of an Intel XeonX5660 workstation with 2.8 GHz CPU and 8 GB memory, and observe that the iterative optimization converges

Algorithm 2 Solving Problem of Eq. (7.4) by Inexact ALM

1: **Input:** Comparative relationship matrix T_i, $i = 1, 2, \ldots, n$, parameter λ, number of samples m.

2: **Initialize:** $\hat{T} = 0$, $E_i = 0$, $Y_i = 0$, $i = 1, \ldots, n$, $\mu = 10^{-6}$, $max_\mu = 10^{10}$, $\rho = 1.1$, $\varepsilon = 10^{-8}$.

3: **repeat**

4: Fix the other term and update $\hat{T}$ by
$(U, \Lambda, V) = SVD(\frac{1}{n\mu}\sum_{i=1}^{n} Y_i + \frac{1}{n}\sum_{i=1}^{n} T_i - \frac{1}{n}\sum_{i=1}^{n} E_i))$, $\hat{T} = U\mathcal{S}_{\frac{1}{\mu}}[\Lambda]V^\top$, where $\mathcal{S}$ is a shrinkage operator for singular value truncating defined as:

$$\mathcal{S}_\varepsilon[x] = \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

5: Fix the other term and update E_i by $E_i = \mathcal{S}_{\frac{\lambda}{\mu}}[T_i + \frac{Y_i}{\mu} - \hat{T}]$.

6: Update the multipliers $Y_i = Y_i + \mu(T_i - \hat{T} - E_i)$.

7: Update the parameter μ by $\mu = \min(\rho\mu, max_\mu)$.

8: **until** $\max_i \|T_i - \hat{T} - E_i\|_\infty < \varepsilon$ and $\mathrm{rank}(\hat{T}) = 2$.

9: **Output:** $\hat{T}$.

fast. For example, in the Oxford Flower 17 classification experiment (see section 7.6.1), one iteration between step 4 and step 7 in Algorithm 2 can be finished within 0.8 seconds. Furthermore, because each optimization sub-problem in Algorithm 2 monotonically decreases the objective function, the algorithm converges.

After obtaining the optimized matrix $\hat{T}$, we want to recover an m-dimensional confidence score vector $\hat{\mathbf{s}}$ that can better estimate the prediction results. Based on our rank-2 assumption mentioned before, we expect that $\hat{T}$ is generated from $\hat{\mathbf{s}}$ as $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$. The authors in [Jiang *et al.*, 2010] proved that $(1/m)\hat{T}\mathbf{e}$ can provide the best least-square approximation of $\hat{\mathbf{s}}$ which can be formally described as follows:

$$(1/m)\hat{T}\mathbf{e} = \arg\min_{\hat{\mathbf{s}}} \|\hat{T}^\top - (\hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top)\|_F^2. \tag{7.6}$$

Therefore, we can treat $(1/m)\hat{T}\mathbf{e}$ as the recovered $\hat{\mathbf{s}}$ after late fusion. Note that vector $\hat{\mathbf{s}}$ is no longer the original confidence score vector generated by the model, but instead is the true consistent confident patterns across different models.

7.5 Extension with Graph Laplacian

Thus far, the proposed late fusion relies only on the confidence scores of multiple models without utilizing any low-level feature information. In this section, we show that our RLF method can be easily extended to incorporate the information of multiple low-level features, which further improves fusion performance.

Suppose we have n types of low-level features associated with m test samples. For the ith feature type, $i \in \{1, 2, ...n\}$, the graph Laplacian regularizer $\Psi^i(\hat{T})$ can be defined as follows [Chung, 1997]:

$$\Psi^i(\hat{T}) = \frac{1}{2}\sum_{j,k=1}^{m} P^i_{jk}\|\hat{\mathbf{t}}_j - \hat{\mathbf{t}}_k\|_2^2 = tr(\hat{T}^\top L^i \hat{T}), \tag{7.7}$$

where $P^i = (Q^i)^{-\frac{1}{2}} W^i (Q^i)^{-\frac{1}{2}}$ is a normalized weight matrix of W^i. W^i denotes the pairwise similarity between the test samples calculated based on the ith feature. Q^i is a diagonal matrix whose (l,l)-entry is the sum of the lth row of W^i. $L^i = I - P^i$ is the graph Laplacian matrix with I denoting an identity matrix. $\hat{\mathbf{t}}_j$ and $\hat{\mathbf{t}}_k$ denote the jth and kth rows of the low-rank matrix $\hat{T}$, each of which actually measures the pairwise comparative relationships of the given test sample with regard to the other test samples.

The intuition behind the graph regularizer is that highly similar test samples in the feature space should have similar comparative relationships with regard to the other test samples (and hence similar prediction scores). Such a regularizer is helpful for robust learning, and allows our model to not only inherit the discriminative capability from each model, but also utilize the complementary information of multiple features.

In this work, we choose the nearest neighbor graph for the multi-feature graph regularizer. Given m test samples represented as the ith feature type $\{x_1^i, x_2^i, ...x_m^i\}$. For each test sample x_j^i, we find its K nearest neighbors and place an edge between x_j^i and its neighbors. The entry W_{jk}^i in the weight matrix W^i associated with the graph is defined as

$$W_{jk}^i = \begin{cases} \exp(-\frac{d_{\chi^2}(x_j^i, x_k^i)}{\sigma}), & \text{if } j \in \mathcal{N}_K(k) \text{ or } k \in \mathcal{N}_K(j), \\ 0, & \text{otherwise}, \end{cases} \tag{7.8}$$

where $\mathcal{N}_K(j)$ denotes the index set for the K-nearest neighbors of sample x_j^i (we set $K = 6$ in this work), $d_{\chi^2}(x_j^i, x_k^i)$ is the χ^2 distance between two samples, and σ is the radius parameter of the Gaussian function, which is set at the mean value of all pairwise χ^2 distances between the samples.

Based on the above definition, we arrive at the following objective function with a multi-feature graph Laplacian regularizer (λ and γ are two positive tradeoff parameters):

$$\begin{aligned} &\min_{\hat{T}, E_i} \|\hat{T}\|_* + \lambda \sum_{i=1}^{n} \|E_i\|_1 + \gamma \sum_{i=1}^{n} \Psi^i(\hat{T}), \\ &s.t.\ T_i = \hat{T} + E_i,\ i = 1, \ldots, n, \\ &\quad \hat{T} = -\hat{T}^\top, \end{aligned} \tag{7.9}$$

Because the multi-feature graph Laplacian regularizer is a differentiable function of $\hat{T}$, the above objective can be easily solved by the ALM method. This can be realized by replacing the updating of $\hat{T}$ in step 4 of Algorithm 2 with the following updating rule.

$$\begin{aligned} &(U, \Lambda, V) \\ &= SVD\big((nI + \frac{2\gamma}{\mu} \sum_{i=1}^{n} L^i)^{-1} (\frac{1}{\mu} \sum_{i=1}^{n} U_i + \sum_{i=1}^{n} T_i - \sum_{i=1}^{n} E_i)\big), \\ &\hat{T} = U \mathcal{S}_{\frac{1}{\mu}}[\Lambda] V^\top,\ \ \hat{T} = (\hat{T} - \hat{T}^\top)/2, \end{aligned} \tag{7.10}$$

where I is an identity matrix. Because the input matrix for SVD is no longer skew-symmetric, in order to ensure the skew-symmetric constraint, we use $\hat{T} = (\hat{T} - \hat{T}^\top)/2$ to project $\hat{T}$ into a skew-

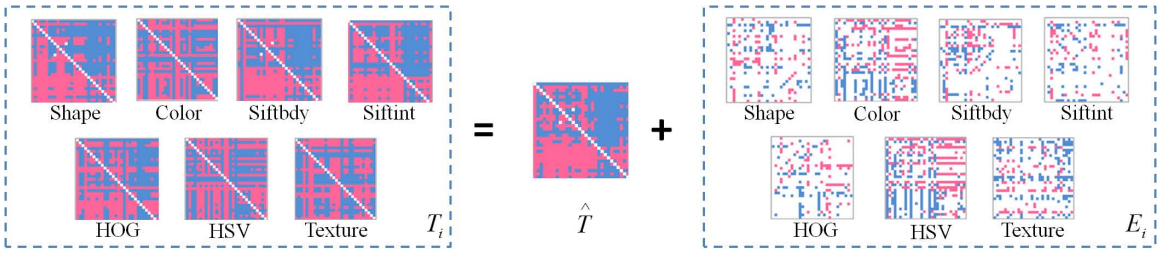


Figure 7.2: Visualization of the low rank and sparse matrices obtained by our RLF method from seven different confidence score vectors of Oxford Flower 17 dataset, each of which is generated by training a binary classifier based on one feature. To ease visualization, we sample a 30×30 sub-matrix from each 340×340 matrix. Blue cells denote the values above 0, purple cells denote the values below 0, and white cells denote 0 values. The obtained matrix $\hat{T}$ is skew-symmetric. This figure is best viewed in color.

symmetric matrix [Gleich and Lim, 2011]. After obtaining the optimized $\hat{T}$, we can recover a score vector $\hat{\mathbf{s}}$ by Eq. (7.6) which can be used for the final prediction.

7.6 Experiment

In this section, we first evaluate our proposed method on a general visual classification task, e.g., Oxford Flower 17 classification task, to prove that the proposed method is a general robust multi-source fusion framework for visual classification. Then, we further show the promising results achieved by the proposed method over video event detection tasks, where 8% relative performance gain over the state-of-the-art is achieved. The following early and late fusion methods are compared in our experiments: (1) Kernel Average. This method is in fact an early fusion method, that averages multiple kernel matrices into a single kernel matrix for model learning. (2) MKL. We use Simple MKL [A. Rakotomamonjy and Grandvalet, 2009] to train the SVM classifier and determine the optimal weight for each kernel matrix simultaneously. (3) Average Late Fusion. After obtaining the normalized confidence score from each model, we average them as the fusion score for classification. (4) Our proposed Robust Late Fusion (RLF) method. (5) Our proposed Graph-regularized Robust Late Fusion (GRLF) method.

Without loss of generality, we use the one-vs-all SVM as the model for generating confidence

scores. Because the one-vs-all SVM is a binary classifier that works on unbalanced numbers of positive and negative training samples, we employ the AP that is popularly applied in the binary visual classification task as the evaluation metric. Then we calculate the mAP across all categories of the dataset as the final evaluation metric.

We use cross validation to determine the appropriate parameter values for each method. In particular, we vary the values of the regularization parameters λ and γ in our method on the grid for $\{10^{-3}, 10^{-2}, ..., 10^{3}\}$, and then choose the best values based on validation performance. With regard to the parameter setting for MKL, we follow the parameter setting strategies suggested in [Gehler and Nowozin, 2009]. For the SVM classifier, we apply χ^2 kernel as the kernel matrix for each method, which is calculated as $\exp\left(-\frac{1}{\sigma}d_{\chi^2}(x,y)\right)$, where σ is set as the mean value of all pairwise distances on the training set. The tradeoff parameter C of SVM is selected from $\{10^{-1}, 10^{0}, ..., 10^{3}\}$ through cross validation.

7.6.1 Experiment on Oxford Flower 17

In this section, we present the results for the Oxford Flower 17 dataset [Nilsback and Zisserman, 2006]. This dataset contains flower images of 17 categories with 80 samples per category. The dataset has three predefined separations with 680 (17×40), 340 (17×20), and 340 (17×20) training, test, and validation images, respectively. The author of [Nilsback and Zisserman, 2008] provides the pre-computed distance matrices for the three separations. We directly apply these matrices in our experiment. The matrices are computed from seven different types of features including color, shape, texture, HOG, clustered HSV values, SIFT feature [Lowe, 2004] on the foreground internal region (SIFTint), and SIFT feature, on the foreground boundary (SIFTbdy). The details of the features can be found in [Nilsback and Zisserman, 2008]. For each method, the best parameter is selected via cross validation on the validation set.

Table 7.1 lists the performance of different methods in comparison, where we also list the best individual features (SIFTint). From the results, we can see that: (1) all fusion methods generate better results than SIFTint, which clearly verifies the advantages of multi-model fusion. (2) Our proposed RLF method clearly outperforms the other baseline methods, because it seeks a robust scale-invariant low-rank fusion matrix from the outputs of multiple classifiers; (3) Our proposed GRLF method outperforms the RLF method, thus demonstrating that involving multiple features

further improves performance. In Figure 7.2, we visualize the low-rank and sparse matrices obtained by applying our method on one category of the Oxford Flower 17 dataset. As can be seen, our proposed method tends to find a shared structure while removing the noise information as sparse matrices. Note that the obtained matrix $\hat{T}$ is skew-symmetric, which well verifies the conclusion in theorem 1, i.e., when the input matrices are skew-symmetric, even without the skew-symmetric constraint, our algorithm naturally produces a skew-symmetric matrix.

Method	MAP
SIFTint	0.749 ± 0.013
Kernel Average	0.860 ± 0.017
MKL	0.863 ± 0.021
Average Late Fusion	0.869 ± 0.021
Our RLF Method	$\mathbf{0.898 \pm 0.019}$
Our GRLF Method	$\mathbf{0.917 \pm 0.017}$

Table 7.1: MAP comparison on Oxford Flower 17 dataset.

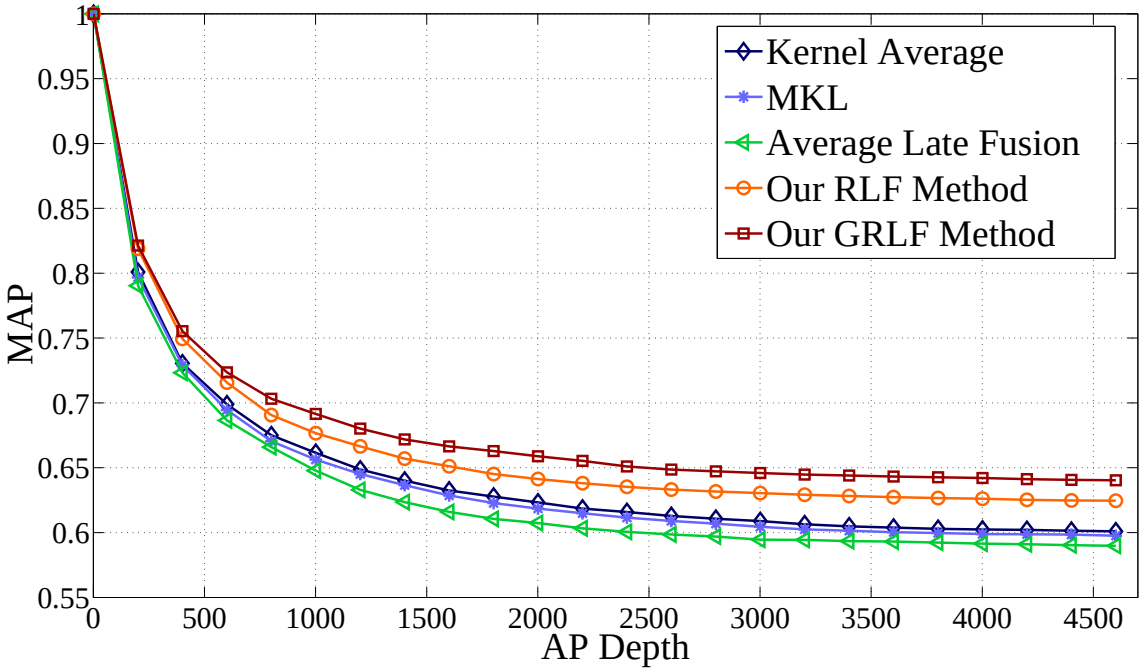


Figure 7.3: MAP comparison at variant depths on CCV dataset.

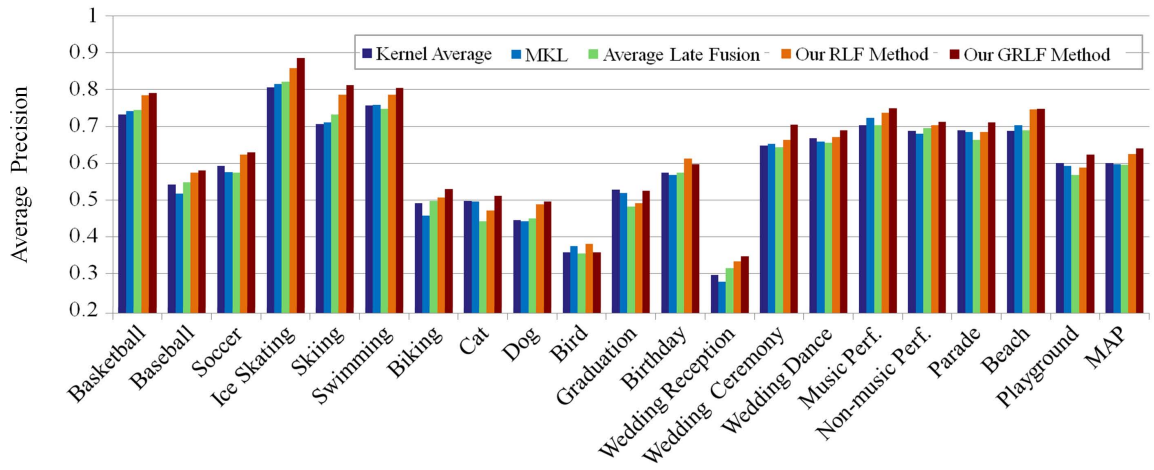


Figure 7.4: AP comparison of different methods on CCV dataset. This figure is best viewed in color.

7.6.2 Experiment on CCV

For the second dataset of the experiments, we use the large scale CCV [Y.-G. Jiang and Loui, 2011]. This dataset contains $9,317$ web videos over 20 semantic categories, where $4,659$ are used for training and the remaining $4,658$ videos are used for testing. In our experiment, we use the three types of the features provided by the dataset [Y.-G. Jiang and Loui, 2011], which includes $5,000$-dimensional SIFT, $5,000$-dimensional spatial-temporal interest points (STIP) [Laptev and Lindeberg, 2003], and $4,000$-dimensional MFCC [Pols, 1966b] BoW features.

To obtain the optimal parameter for each method, we partition the training set into three subsets, and then perform three-fold cross-validation. Figure 7.3 shows the mAP performance at different returned depths (the number of top ranking test samples to be included in the result evaluation). From the results, we can see that our method achieves significant and consistent mAP improvement over the other baseline methods at variant returned depths. Figure 7.4 shows the per-category AP performance comparisons of all methods. As shown, the performances of all the baseline methods are quite similar to each other, which is consistent with the results in section 7.6.1. The proposed GRLF method shows the best performance on most events. In particular, in terms of mAP it outperforms the Kernel Average, MKL and Average Late Fusion methods by 7.2%, 6.6% and 7.6%, relatively. Here, the Average Late Fusion result is directly quoted from [Y.-G. Jiang and Loui, 2011], which clearly demonstrates that our method is superior over the state-of-the-art method in the literature.

7.6.3 Experiment on TRECVID MED 2011

TRECVID MED is a challenging task for the detection of complicated high-level events. We test our proposed method on the TRECVID MED 2011 development dataset [web, 2011], that includes five events "Attempting board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony", and "Wood working". The training and test sets consist of $8,783$ and $2,021$ video shots respectively. For low-level features, we extract $5,000$-dimensional SIFT, $5,000$-dimensional STIP, and $4,000$-dimensional MFCC BoW features. Again, one-versus-all SVM with χ^2 kernel is used to train the model. Three-fold cross-validation on the training set is used for parameter tuning.

Figure 7.5 shows the per-event performance for all the methods in comparison. From the results, we obtain the following observations: (1) our proposed RLF method produces better results than all baseline methods in terms of mAP. (2) The GRLF method further outperforms the RLF method and achieves better performance on four out of five events, which well verifies the advantages of bringing the low-level features into the late fusion task. (3) mAP for our proposed GRLF method is 0.509, which is relatively 10.4% higher than the best baseline performance (Average Late Fusion method with mAP: 0.461). This confirms the superiority of our method. Figure 7.6 shows the mAP at different returned depths for all methods.

7.6.4 Discussion

Consistency of the recovered matrix. Given a real-valued rank-2 skew-symmetric matrix $\hat{T}$, the score vector $\hat{\mathbf{s}}$ can be recovered from $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$. Based on the analysis in [Jiang *et al.*, 2010], even if we have inconsistent entries in $\hat{T}$, optimization results of Eq. (7.6) can still provide the best approximation of $\hat{\mathbf{s}}$, thus overcoming any remaining inconsistency issues. This has also been verified by our experiment results, where there is no inconsistency in the final score vectors recovered from the rank-2 matrices obtained by our method over the three datasets.

Tradeoff between low-rankness and sparsity. Notably, our method can achieve a good tradeoff between low-rankness and sparsity. If there are many classification errors associated with the ith model, the decomposed additive term E_i is dense with many non-zero entries. This can be illustrated in Figure 7.2, where the denser the matrix E_i, the worse is the performance obtained by the corresponding component model. For example, the classification performance of the HSV feature is the worst among the seven features, and thus its additive noise matrix is the densest. This further

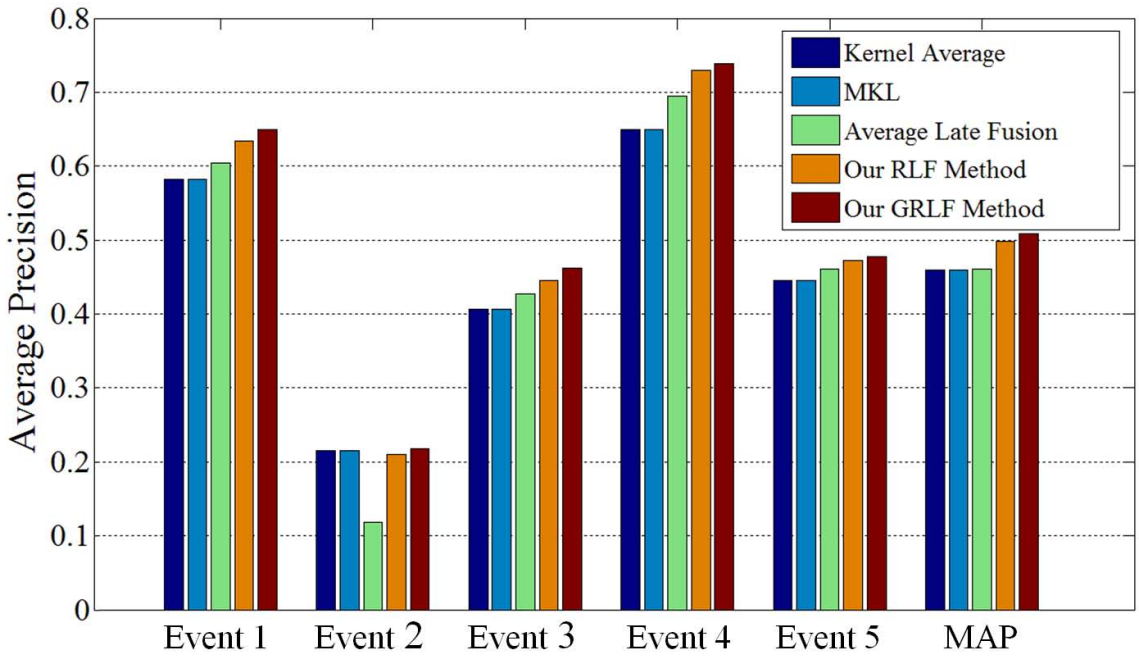


Figure 7.5: AP comparison on TRECVID MED 2011 development dataset. The five events from left to right are "Attempting board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony", and "Wood working". This figure is best viewed in color.

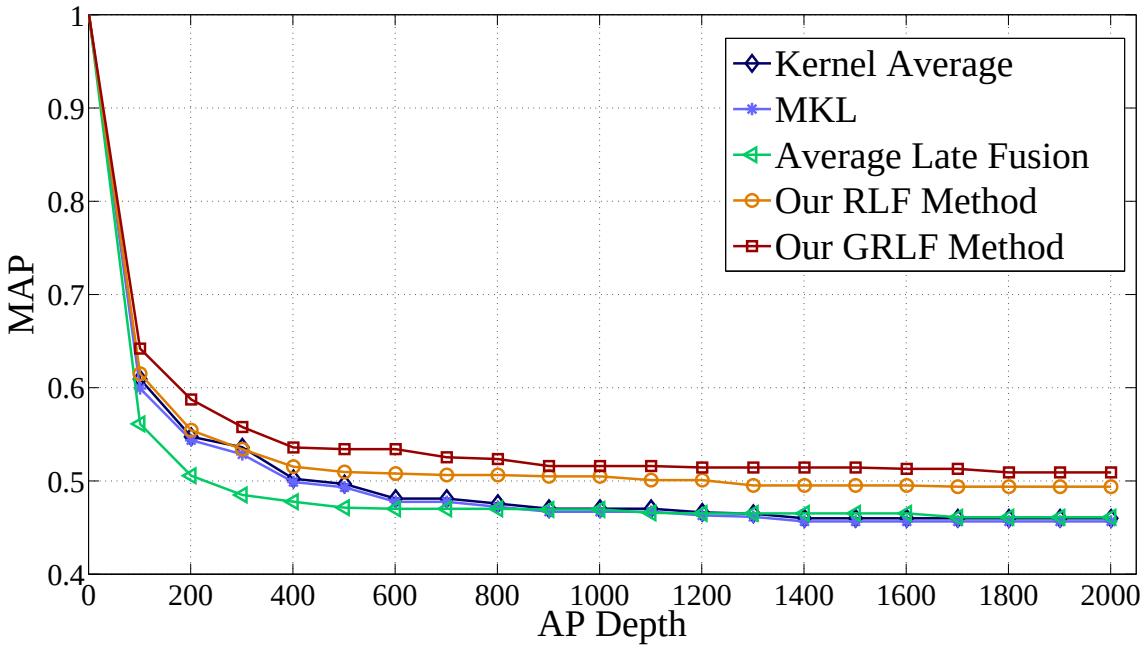


Figure 7.6: MAP comparison of different methods at variant depths on TRECVID MED 2011 development dataset.

verifies the advantage of our method to obtain balanced tradeoff between low-rankness of the score relationships and the sparsity of the score errors.

Out-of-sample extension. We can adopt a simple nearest-neighbor method to manage the out-of-sample problem for our robust late fusion model. When a new test sample $\mathbf{x}_{m+1}$ represented with n feature types $\{\mathbf{x}_{m+1}^1, \ldots, \mathbf{x}_{m+1}^n\}$ comes, we can find its nearest neighbors $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ where each $\mathbf{x}^i$ is the nearest neighbor of $\mathbf{x}_{m+1}^i$ in terms of the ith feature type. Then the fusion score can be obtained by $\hat{s}(\mathbf{x}_{m+1}) = \sum_{i=1}^{n} \frac{W(\mathbf{t}_{m+1}^i, \mathbf{x}^i)}{\sum_{i=1}^{n} W(\mathbf{t}_{m+1}^i, \mathbf{x}^i)} \hat{s}(\mathbf{x}^i)$, where $W(\mathbf{t}_{m+1}^i, \mathbf{x}^i)$ denotes the feature similarity based on the ith feature type, $\hat{s}(\mathbf{x}^i)$ is the fusion score of sample $\mathbf{x}^i$.

7.7 Summary

We introduced a robust rank minimization method for multi-source fusion. We first convert each confidence score vector of a model into a pairwise comparative relationship matrix, so that the confidence scores of different models can be manipulated in an isotonic manner. Then the late

fusion is formulated as a matrix decomposition problem where a shared matrix is inferred from the joint decomposition of multiple pairwise relationship matrices into pairs of low-rank and sparse components. Extensive experiments on various visual classification tasks showed that our method outperforms the state-of-the-art early and late fusion methods. In the future, we will investigate the fusion of more complex models to consider multi-class or multi-label problems in computer vision and multimedia applications.

Part IV

Conclusions

Chapter 8

Conclusions

8.1 Contribution Summarization

This thesis was dedicated to developing robust and efficient solutions for large-scale video event detection systems. We focused on two techniques: large-scale video and concept ontology construction and large-scale video event detection with multi-modality representations and multi-source fusion. The first part focused on developing automatic methodologies for large-scale event and concept ontology discovery and design so that video events can be represented by the detection of mid-level concept features. The second part focused on pursuing cross-modality cross-source correlation discovery so that video event can be predicted by a multi-modal representation based model and robust fusion across different sources.

The main contributions of the thesis are as follows:

1. Large-scale event and concept ontology construction: we proposed an automatic framework for discovering event-driven concepts. By leveraging the external knowledge bases, we built the largest video event ontology (to the best of our knowledge), *EventNet*, that includes 500 complex events and discovered $4,490$ event-specific concepts. Dramatic performance gains were achieved especially for unseen novel event detections with EventNet ontology. Based on the proposed EventNet framework, we constructed the first interactive system (to the best of our knowledge) that allows users to explore high level events and associated concepts in videos in a systematic structured manner. Several useful applications for the EventNet system were shown, e.g., interactive browser, semantic search, live tagging of user-uploaded videos, etc.

2. Event detection with multi-modality representations and multi-source fusion: in order to discover joint patterns among multiple modalities, we proposed multi-modality representation, called bi-modal word, that discovers the joint audio-visual patterns in videos by the bipartite graph partitioning. We experimentally showed that the bi-modal representations provide promising results over state-of-the-art features, and thus they can be used for robust multi-modality feature representations. In order to incorporate the heterogeneous complimentary information from multiple sources, we further proposed a robust rank minimization based fusion method for fusing the confidence scores of multiple models. Extensive experiment results showed that the proposed robust late fusion method can not only be considered a robust fusion method for video detection task, but it can also be considered a general fusion framework for various visual classification tasks.

8.2 Open Issues and Future Direction

Although the proposed methods achieved promising results, there are many open issues remaining. Here, we list a few topics for future research.

1. EventNet expansion: although we proposed an automatic methodology to discover event-driven concepts, the recent construction of EventNet ontology still depends heavily on subjective evaluation with manual selections and clarifications of the chosen events. In order to continuously expand EventNet ontology in the future, we need to further explore a systematic method for discovering events either of WikiHow articles, or with help from crowdsourcing in the public domain. Once the novel events and event-driven concepts are discovered, we need to find the most appropriate node to be attached to the EventNet ontology in an incremental manner.

2. Event and concept spatial and temporal localizations: the recent EventNet system can only detect events and concepts at the video level. Some advanced methodologies applied in the recent THUMOS challenge [THU, 2015] provide some potential solutions for localizing event and concept in certain regions of specific video frames. In the future, we will further explore the spatial temporal event and concept localization problem so that more precise and accurate detections are accomplished.

3. Cross-modality representation from multiple modalities: for the current bi-modal word representation, only joint audio-visual patterns from two modalities are analyzed. In the future, we

will explore cross-modality patterns over more diverse modalities, e.g., text, static image, motion, audio, speech, music, etc.

4. Adaptive robust late fusion: the proposed robust late fusion with rank minimization is promising when the test data arrives in as a batch mode. When the test data arrives as a sequential mode, the optimization problem needs to be incrementally making the current fusion method infeasible. In the future, we will further explore an adaptive robust refinement method so that the algorithm can be applied efficiently in sequential mode.

Part V

Bibliography

Bibliography

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

[A. Kembhavi and Davis, 2009] R. Miezianko S. McCloskey A. Kembhavi, B. Siddiquie and Larry S. Davis. Incremental multiple kernel learning for object recognition. In *ICCV*, 2009.

[A. Krizhevsky and Hinton, 2012] I. Sutskever A. Krizhevsky and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[A. Rakotomamonjy and Grandvalet, 2009] S. Canu A. Rakotomamonjy, F.R. Bach and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 2009.

[A. Ritter and Clark, 2012] O. Etzioni A. Ritter, Mausam and S. Clark. Open domain event extraction from twitter. In *KDD*, 2012.

[A. Tamrakar and Sawhney, 2012] Q. Yu J. Liu O. Javed-A. Divakaran H. Cheng A. Tamrakar, S. Ali and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

[A. Vedaldi and Zisserman, 2009] M. Varma A. Vedaldi, V. Gulshan and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[Ayache and Quénot, 2008] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *ECIR*, 2008.

[Bach *et al.*, 2004] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[Baptiste Mazin and Gousseau, 2012] Julie Delon Baptiste Mazin and Yann Gousseau. Combining color and geometry for local image matching. In *ICPR*, 2012.

[Bird, 2006] S. Bird. Nltk: The natural language toolkit. In *ACL*, 2006.

[C. Schuldt and Caputo, 2004] I. Laptev C. Schuldt and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[Cai *et al.*, 2008] Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *UCLA CAM Report*, 2008.

[Chang and Lin, 2011] C.-C. Chang and C.-J. Lin. Libsvm : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[Chung, 1997] Fan Chung. Spectral graph theory. *Regional Conference Series in Mathematics*, 1997.

[CI, 2007] Connolly CI. Learning to recognize complex actions using conditional random fields. In *International Conference on Advances in Visual Computing*, 2007.

[D. Field, 1987] et al D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 1987.

[Dhillon, 2001] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, 2001.

[Florian, 2014] Florian. Bbn viser trecvid 2014 multimedia event detection and multimedia event recounting systems. In *NIST TRECVID Workshop*, 2014.

[G. Potamianos and Matthews, 2004] J. Luettin G. Potamianos, C. Neti and I. Matthews. Audio-visual automatic speech recognition: an overview. *Issues in visual and audio-visual speech processing*, 2004.

[G. Ye and Chang, 2012] D. Liu Y.G. Jiang D.-T. Lee G. Ye, I.-H. Jhuo and S.-F. Chang. Joint audio-visual bi-modal codewords for video event detection. In *ICMR*, 2012.

[Gehler and Nowozin, 2009] Peter Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[Gleich and Lim, 2011] David F. Gleich and Lek-Heng Lim. Rank aggregation via nuclear norm minimization. In *KDD*, 2011.

[Guangnan Ye and Chang, 2012] I-Hong Jhuo Guangnan Ye, Dong Liu and Shih-Fu Chang. Robust late fusion with rank minimization. In *CVPR*, 2012.

[Guangnan Ye and Chang, 2015] Hongliang Xu Dong Liu Guangnan Ye, Yitong Li and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *MM*, 2015.

[H. Wang and Liu, 2011] C. Schmid H. Wang, A. Klaser and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[H. Xu and Chang, 2015] Y. Li D. Liu H. Xu, G. Ye and S.-F. Chang. Large video event ontology browsing, search and tagging (eventnet demo). In *MM*, 2015.

[Hale *et al.*, 2008] Elaine T. Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for l1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 2008.

[I. Laptev and Rozenfeld, 2008] C. Schmid I. Laptev, M. Marszalek and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[Izadinia and Shah, 2012] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.

[J.-C. Wang and Wang, 2012] I.-H. Jhuo Y.-Y. Lin J.-C. Wang, Y.-H. Yang and H.-M. Wang. The acousticvisual emotion guassians model for automatic generation of music video. In *ACM MM*, 2012.

[J. Chen and Chang, 2014] G. Ye D. Liu J. Chen, Y. Cui and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014.

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

[J. Goldberger and Salakhutdinov, 2004] G. Hinton J. Goldberger, S. Roweis and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004.

[J. Liu and Friedland, 2013] O. Javed Q. Yu I. Chakraborty W. Zhang A. Divakaran H. Sawhney J. Allan R. Manmatha J. Foley M. Shah A. Dehghan M. Witbrock J. Curtis J. Liu, H. Cheng and G. Friedland. Sri-sarnoff aurora system at trecvid 2013 multimedia event detection and recounting. *NIST TRECVID Workshop*, 2013.

[J. Liu and Savarese, 2011] B. Kuipers J. Liu, M. Shah and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

[J. Liu and Sawhney, 2012] O. Javed S. Ali A. Tamrakar A. Divakaran H. Cheng J. Liu, Q. Yu and H. Sawhney. Video event recognition using concept attributes. In *WACV*, 2012.

[J. Revaud and Jégou, 2013] C. Schmid J. Revaud, M. Douze and H. Jégou. Event retrieval in large video collections with circulant temporal encoding. In *CVPR*, 2013.

[J. van Gemert and Geusebroek, 2010] A. Smeulders J. van Gemert, C. Veenman and J-M Geusebroek. Visual word ambiguity. *TPAMI*, 2010.

[Jain *et al.*, 2005] Anil K. Jain, Karthik Nandakumar, and Arun Ross. Score normalizationin multimodal biometric systems. *Pattern Recognition*, 2005.

[Jhuo and Lee, 2010] I.H. Jhuo and D.-T. Lee. Boosting-based multiple kernel learning for image re-ranking. In *ACM MM*, 2010.

[Jia, 2013] `http://caffe.berkeleyvision.org`, 2013.

[Jiang and Loui, 2011] W. Jiang and A. Loui. Audio-visual grouplet: Temporal audio-visual interactions for general video concept classification. In *ACM MM*, 2011.

[Jiang *et al.*, 2010] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 2010.

[K. Saenko and Darrell, 2010] M. Fritz K. Saenko, B. Kulis and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[K. Soomro and Shah, 2012] A. Zamir K. Soomro and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. *CRCV-TR*, 2012.

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

[K. Tang and Koller, 2012] L. Fei-Fei K. Tang and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[K. Van De Sande and Snoek, 2010] T. Gevers K. Van De Sande and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 2010.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[Kuehne H and T, 2011] Garrote E-Poggio T Kuehne H, Jhuang H and Serre T. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[L. Bao, 2011] et al L. Bao. Informedia @ trecvid 2011. In *TRECVID Workshop*, 2011.

[L. Duan and Luo, 2010] I. W. Tsang L. Duan, D. Xu and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[L. Han and Weese, 2013] T. Finin J. Mayfield L. Han, A. Kashyap and Jonathan Weese. Umbc ebiquity-core: Semantic textual similarity systems. In *ACL*, 2013.

[L.-J. Li and Xing, 2010] L. Fei-Fei L.-J. Li, H. Su and E. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[L. Laptev and Rozenfeld, 2008] C. Schmid L. Laptev, M. Marszaek and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[L. Torresani and Fitzgibbon, 2010] M. Szummer L. Torresani and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[Laptev and Lindeberg, 2003] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[Lin *et al.*, 2009] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of a corrupted low-rank matrix. *UIUC Technical Report*, 2009.

[Liu and Shah, 2008] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[Liu and Singh, 2004] H. Liu and P. Singh. Conceptnet: A pratical commonsence reasoning toolkit. *BT Technology Journal*, 2004.

[Liu J and M, 2009] Luo J Liu J and Shah M. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[Lowe, 2004] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[Luc, 2015] `https://lucene.apache.org/core/`, 2015.

[Lutkepohl, 1997] H. Lutkepohl. Handbook of matrices. *Chichester: Wiley*, 1997.

[M. Beal and Attias, 2003] N. Jojic M. Beal and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 2003.

[M. Blank and Basri, 2005] E. Shechtman M. Irani M. Blank, L. Gorelick and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[M. Cristani and Murino, 2007] M. Bicego M. Cristani and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transcations on Multimedia*, 2007.

[M. Jain and Snoek, 2014] J. Gemert M. Jain and C. Snoek. University of amsterdam at thumos challenge 2014. In *Thumos Challenge*, 2014.

[M. Mazloom and Snoek, 2013] K. Sande M. Mazloom, E. Gavves and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[M. Merler and Natsev, 2012] L. Xie G. Hua M. Merler, B. Huang and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE TMM*, 2012.

[M. Rastegari and Farhadi, 2013] D. Parikh M. Rastegari, A. Diba and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013.

[MED, 2010] `http://www.nist.gov/itl/iad/mig/med.cfm`, 2010.

[Mikolajczyk and Schmid, 2004] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 2004.

[Miller, 1995] G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 1995.

[N. Ikizler-Cinbis and Sclaroff, 2009] R. Cinbis N. Ikizler-Cinbis and S. Sclaroff. Learning actions from the web. In *ICCV*, 2009.

[Nandakumar *et al.*, 2008] Karthik Nandakumar, Yi Chen, Sarat C. Dass, and Anil K. Jain. Likelihood ratio-based biometric score fusion. *TPAMI*, 2008.

[Natarajan P, 2008] Nevatia R Online Natarajan P. Online, real-time tracking and recognition of human actions. In *IEEE workshop on motion and video computing*, 2008.

[Natarajan, 2011] P. Natarajan. Bbn viser trecvid 2011 multimedia event detection system. In *NIST TRECVID Workshop*, 2011.

[Nilsback and Zisserman, 2006] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *ICCV*, 2006.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[Nis, 2010] `http://www.nist.gov/itl/iad/mig/med10.cfm/`, 2010.

[Nis, 2011] `http://www.nist.gov/itl/iad/mig/med11.cfm/`, 2011.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[P. Natarajan and Zhuang, 2012] S. Vitaladevuni P. Natarajan, S. Wu and X. Zhuang. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[Patterson and Hays, 2012] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[Pols, 1966a] L. Pols. Spectral analysis and identification of dutch vowels in monosyllabic words. *Doctoral dissertion, Free University, Amsterdam*, 1966.

[Pols, 1966b] L. C. W. Pols. Spectral analysis and identification of dutch vowels in monosyllabic words. *Doctoral dissertion, Free University, Amsterdam, The Netherlands*, 1966.

[Pte, 2015] `http://en.wikipedia.org/wiki/P-value`, 2015.

[Rodriguez MD and M, 2008] Ahmed J Rodriguez MD and Shah M. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[S. Lazebnik and Ponce, 2006] C. Schmid S. Lazebnik and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

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

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

[Smith *et al.*, 2003] John R. Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *ICME*, 2003.

[T. Berg and Shih, 2010] A. Berg T. Berg and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[T. Ojala and Maenpaa, 2002] M. Pietikainen T. Ojala and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002.

[Terrades *et al.*, 2009] Oriol Ramos Terrades, Ernest Valveny, and Salvatore Tabbone. Optimal classifier fusion in a non-bayesian probabilistic framework. *TPAMI*, 2009.

[THU, 2015] `http://www.thumos.info/home.html`, 2015.

[ucf, 2015] `server.cs.ucf.edu/~ision/data/UCF50.rar`, 2015.

[Varma and Ray, 2007] Manik Varma and Debajyoti Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[W. Jiang and Loui, 2009] S.-F. Chang D. Ellis W. Jiang, C. Cotton and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *ACM MM*, 2009.

[web, 2011] `http://www.nist.gov/itl/iad/mig/med11.cfm/`, 2011.

[Weinland D and E, 2006] Ronfard R Weinland D and Boyer E. Free viewpoint action recognition using motion history volumes. *Computer Vision Image Underst*, 2006.

[Wik, 2015] `http://www.wikihow.com/Main-Page`, 2015.

[Wright *et al.*, 2009] John Wright, Yigang Peng, Yi Ma, Arvind Ganesh, and Shankar Rao. Robust principal component analysis:exact recovery of corrupted low-rank matrices by convex optimization. In *NIPS*, 2009.

[Wu and Palmer, 1994] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, 1994.

[Y. Cui and Chang, 2014] J. Chen Y. Cui, D. Liu and S.-F. Chang. Building a large concept bank for representing events in video. *arXiv:1403.7591*, 2014.

[Y.-G. Jiang and Chang, 2010] G. Ye S. Bhattacharya D. Ellis M. Shah Y.-G. Jiang, X. Zeng and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

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

[Y.-G. Jiang and Loui, 2011] S.-F. Chang D. Ellis Y.-G. Jiang, G. Ye and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.

[Y.-G. Jiang and Shah, 2013] S.-F. Chang Y.-G. Jiang, S. Bhattacharya and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, 2013.

[Y.-L. Boureau and Lecun, 2010] J. Ponce Y.-L. Boureau and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, 2010.

[Yanai and Barnard, 2005] K. Yanai and K. Barnard. Image region entropy: A measure of "visualness" of web images associated with one concept. In *ACM MM*, 2005.

[Yang and Shah, 2012] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *ECCV*, 2012.

[Z. Ma and Hauptmann, 2013a] Y. Cai-N. Sebe Z. Ma, Y. Yang and A. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM MM*, 2013.

[Z. Ma and Hauptmann, 2013b] Z. Xu-S. Yan N. Sebe Z. Ma, Y. Yang and A. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.

Part VI

Appendices

Appendix A

Low Rank Theorem

Theorem 2. *Given a set of n skew-symmetric matrices T_i, Algorithm 1 produces a skew-symmetry matrix $\hat{T}$ if the spectrums between the dominant singular values are separated.*

Proof. To prove the above theorem, we first introduce some properties of skew-symmetric matrices. Assuming T is a skew-symmetric matrix with the property $T = -T^\top$, then the eigenvalues of T are pure-imaginary and come in complex-conjugate pairs while the rank of T is even. Then the following lemma characterizes the Singular Value Decomposition (SVD) of a skew-symmetric matrix. □

Lemma 1. *If T is an $n \times n$ skew-symmetric matrix ($T = -T^\top$) with eigenvalues $i\lambda_1$, $-i\lambda_1$,$i\lambda_2$, $-i\lambda_2$,...,$i\lambda_j$,$-i\lambda_j$ where $\lambda_p > 0$, $j = \lfloor (n/2) \rfloor$, and i denotes the imaginary unit. Then the SVD of T is given by $T = UDV^\top$ where*

$$D = \begin{pmatrix} \lambda_1 & & & & & & & \\ & \lambda_1 & & & & & & \\ & & \lambda_2 & & & & & \\ & & & \lambda_2 & & & & \\ & & & & . & & & \\ & & & & & . & & \\ & & & & & & . & \\ & & & & & & & \lambda_j & \\ & & & & & & & & \lambda_j \end{pmatrix}$$

Proof. Real-valued matrix T can be decomposed as a real-valued orthogonal matrix X and a real-valued block-upper-triangular matrix Z, with 2-by-2 blocks along the diagonal by Murnaghan-Wintner form:

$$T = XZX^{\top}. \tag{A.1}$$

Since T is skew-symmetric, the decomposed component Z is also skew-symmetric. Then Z has a block-diagonal form:

$$Z = \begin{pmatrix} 0 & \lambda_1 & & & & & & \\ -\lambda_1 & 0 & & & & & & \\ & & 0 & \lambda_2 & & & & \\ & & -\lambda_2 & 0 & & & & \\ & & & & . & & & \\ & & & & & . & & \\ & & & & & & . & \\ & & & & & & 0 & \lambda_j \\ & & & & & & -\lambda_j & 0 \end{pmatrix}$$

Furthermore, the SVD of the matrix $\left(\begin{smallmatrix} 0 & \lambda_1 \\ -\lambda_1 & 0 \end{smallmatrix}\right)$ is given by

$$\begin{pmatrix} 0 & \lambda_1 \\ -\lambda_1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_1 \end{pmatrix} \times \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Assume matrix A and matrix B are defined as follows:

$$A = \begin{pmatrix} 0 & 1 & & & & & & \\ 1 & 0 & & & & & & \\ & & 0 & 1 & & & & \\ & & 1 & 0 & & & & \\ & & & & . & & & \\ & & & & & . & & \\ & & & & & & . & \\ & & & & & & 0 & 1 \\ & & & & & & 1 & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} -1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ & & -1 & 0 & & & & \\ & & 0 & 1 & & & & \\ & & & & . & & & \\ & & & & & . & & \\ & & & & & & . & \\ & & & & & & -1 & 0 \\ & & & & & & 0 & 1 \end{pmatrix}$$

Then, the real-valued matrix T has the following decomposition form: $T = XADBX^\top$. We construct U and V such that $U = XA$ and $V = BX^\top$, which are real and orthogonal. We thus complete the theorem which constructs the SVD of T.

□

Next, we use the following lemma to illustrate that the best rank-k approximation to a skew-symmetric matrix T is also skew-symmetric.

Lemma 2. *Assuming T is an $n \times n$ skew-symmetric matrix ($T = -T^\top$) with eigenvalues $i\lambda_1$, $-i\lambda_1$,$i\lambda_2$, $-i\lambda_2$,...,$i\lambda_j$,$-i\lambda_j$,...,$i\lambda_{n/2}$,$-i\lambda_{n/2}$ where $\lambda_p > 0$, the top-k magnitude of the singular value pairs are given by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_j, k = 2j$, then the best rank-$k$ approximation of T in an orthogonally invariant norm is also skew-symmetric.*

Proof. Because the best rank-k approximation of T in an orthogonally invariant norm is given by the k largest singular values and vectors, there is a gap in the spectrum between the kth and $(k+1)$th singular value. Then in the SVD form from Lemma 1, if we truncate the singular values into the k largest ones, it also produces a skew-symmetric matrix.

Finally, we use the above lemma to prove that given a set of n skew-symmetric observation matrices T_i, our ALM-based algorithm produces a skew-symmetry matrix $\hat{T}$ if the target rank is even and the dominant singular values are separated.

Clearly, from Algorithm 1, $\hat{T}^{(0)}$ is skew-symmetric. In each iteration, we compute SVD of a skew-symmetric matrix and truncate the singular values below the threshold, then $\hat{T}^{(i+1)}$, which is

the approximation of the skew-symmetric matrix $\hat{T}^{(i)}$ is also skew-symmetric. Finally, the algorithm converges to a skew-symmetric matrix $\hat{T}$. □