Exploring Semantic Concepts for Complex Event Analysis in Unconstrained Video Clips

Xiaojun Chang

Centre for Quantum Computation & Intelligent Systems

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

July, 2016

I would like to dedicate this thesis to my loving parents and wife.

Acknowledgements

This thesis cannot be done without many important people in my work and life. My supervisor, Dr. Yi Yang, is the one who led me into the realm of machine learning and multimedia analysis. Under his supervision, I have learned a lot of theoretic knowledge and research skills. His passion for exploring new ideas and rigorous attention to detail affected me profoundly. Dr. Feiping Nie is my associate supervisor. I am always impressed by the way he thinks about every research problem. His self-motivation, hard-working spirit and rigorous research attitude have set an excellent example for me. My friend, Dr. Yao-Liang Yu, is a master of maths and statistics. I always admire how knowledgeable he is on maths and statistics. He has taught me a lot of mathematical theory and skills that are very useful in my work. Dr. Alexander G. Hauptmann was my supervisor when I visited Carnegie Mellon University. I have appreciated his guidance immensely.

I would also like to thank many mates at UTS and CMU. Zhongwen Xu, Yan Yan, Pingbo Pan and Linchao Zhu have helped me so much when I was in the group. Their sincerity and kind heart impressed me to give warm care to every other person. I will show my gratitude to Shoou-I Yu, Zhigang Ma and Xuanchong Li. Their acompany makes my research and life much easier.

Lastly, I would like to thank my wife and my parents for their continuous support for my academic pursuit. I wish them health and happiness.

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.
I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Xiaojun Chang
13/07/2016

Abstract

Modern consumer electronics (*e.g*. smart phones) have made video acquisition convenient for the general public. Consequently, the number of videos freely available on the Internet has been exploding, thanks also to the appearance of large video hosting websites (*e.g*. Youtube). Recognizing complex events from these unconstrained videos has been receiving increasing interest in the multimedia and computer vision field. Compared with visual concepts such as actions, scenes and objects, event detection is more challenging in the following aspects. Firstly, an event is a higher level semantic abstraction of video sequences than a concept and consists of multiple concepts. Secondly, a concept can be detected in a shorter video sequence or even in a single frame but an event is usually contained in a longer video clip. Thirdly, different video sequences of a particular event may have dramatic variations.

The most commonly used technique for complex event detection is to aggregate low-level visual features and then feed them to sophisticated statistical classification machines. However, these methodologies fail to provide any interpretation of the abundant semantic information contained in a complex video event, which impedes efficient high-level event analysis, especially when the training exemplars are scarce in real-world applications. A recent trend in this direction is to employ some high-level semantic representation, which can be advantageous for subsequent event analysis tasks. These approaches lead to improved generalization capability and allow zero-shot learning (*i.e*. recognizing new events that are never seen in the training phase). In addition, they provide a meaningful way to aggregate low-level features, and yield more interpretable results, hence may facilitate other video analysis tasks such as retrieval on top of many low-level features, and have roots in object and action recognition.

Although some promising results have been achieved, current event analysis systems still have some inherent limitations. 1) They fail to consider the fact that only a few shots in a long video are relevant to the event of interest while others are irrelevant or even misleading. 2) They are not capable of leveraging the mutual benefits of Multimedia Event Detection (MED) and Multimedia Event Recounting (MER), especially

when the number of training exemplars is small. 3) They did not consider the differences of the classifier's prediction capability on individual testing videos. 4) The unreliability of the semantic concept detectors, due to lack of labeled training videos, has been largely unaddressed. To solve these challenges, in this thesis, we aim to develop a series of statistical learning methods to explore semantic concepts for complex event analysis in unconstrained video clips. Our works are summarized as follows:

In Chapter 2, we propose a novel semantic pooling approach for challenging tasks on long untrimmed Internet videos, especially when only a few shots/segments are relevant to the event of interest while many other shots are irrelevant or event misleading. The commonly adopted pooling strategies aggregate the shots indifferently in one way or another, resulting in a great loss of information. Instead, we first define a novel notion of semantic saliency that assess the relevance of each shot with the event of interest. We then prioritize the shots according to their saliency scores since shots that are semantically more salient are expected to contribute more to the final event analysis. Next, we propose a new isotonic regularizer that is able to exploit the constructed semantic ordering information. The resulting nearly-isotonic SVM classifier exhibits higher discriminative power in event detection and recognition tasks. Computationally, we develop an efficient implementation using the proximal gradient algorithm, and we prove new and closed-form proximal steps.

In Chapter 3, we develop a joint event detection and evidence recounting framework with limited supervision, which is able to leverage the mutual benefits of MED and MER. Different from most existing systems that perform MER as a post-processing step on top of the MED results, the proposed framework simultaneously detects high-level events and localizes the indicative concepts of the events. Our premise is that a good recounting algorithm should not only explain the detection result, but should also be able to assist detection in the first place. Coupled in a joint optimization framework, recounting improves detection by pruning irrelevant noisy concepts while detection directs recounting to the most discriminative evidences. To better utilize the powerful and interpretable semantic video representation, we segment each video into several shots and exploit the rich temporal structures at shot level. The consequent computational challenge is carefully addressed through a significant improvement of the current ADMM algorithm, which, after eliminating all inner loops and equipping novel closed-form solutions for all intermediate steps, enables us to efficiently process extremely large

video corpora.

In Chapter 4, we propose an **E**vent-**D**riven **C**oncept **W**eighting framework to automatically detect events without the use of visual training exemplars. In principle, zero-shot learning makes it possible to train an event detection model based on the assumption that events (*e.g.* birthday party) can be described by multiple mid-level semantic concepts (*e.g.* "blowing candle", "birthday cake"). Towards this goal, we first pre-train a bundle of concept classifiers using data from other sources, which are applied on all test videos to obtain multiple prediction score vectors. Existing methods generally combine the predictions of the concept classifiers with fixed weights, and ignore the fact that each concept classifier may perform better or worse for different subset of videos. To address this issue, we propose to learn the optimal weights of the concept classifiers for each testing video by exploring a set of online available videos which have free-form text descriptions of their content. To be specific, our method is built upon the local smoothness property, which assumes that visually similar videos have comparable labels within a local region of the same space.

In Chapter 5, we develop a novel approach to estimate the reliability of the concept classifiers without labeled training videos. The **EDCW** framework proposed in Chapter 4, as well as most existing works on semantic event search, ignore the fact that not all concept classifiers are equally reliable, especially when they are trained from other source domains. For example, "face" in video frames can now be reasonably accurately detected, but in contrast, the action "brush teeth" remains hard to recognize in short video clips. Consequently, a relevant concept can be of limited use or even misuse if its classifier is highly unreliable. Therefore, when combining concept scores, we propose to take their relevance, predictive power, and reliability all into account. This is achieved through a novel extension of the spectral meta-learner, which provided a principled way to estimate classifier accuracies using purely unlabeled data.

Contents

List of Figures

Publications

This thesis consists of the following publications:

- Chapter 2:
  - **X. Chang**, Y. Yang, E. P. Xing and Y. Yu: "Complex Event Detection using Semantic Saliency and Nearly-Isotonic SVM", International Conference on Machine Learning (ICML), 2015. *ERA Rank A**
- Chapter 3:
  - **X. Chang**, Y. Yu, Y. Yang, and A. Hauptmann: "Searching Persuasively: Joint Event Detection and Evidence Recounting with Limited Supervision", ACM International Conference on Multimedia (ACM MM), 2015. *ERA Rank A**
- Chapter 4:
  - **X. Chang**, Y. Yang, G. Long, C. Zhang and A. Hauptmann: "Dynamic Concept Composition for Zero-Example Event Detection", AAAI Conference on Artificial Intelligence (AAAI), 2016. *ERA Rank A**
  - **X. Chang**, Y. Yang, A. Hauptmann, E. P. Xing and Y. Yu: "Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection", International Joint Conferences on Artificial Intelligence (IJCAI), 2015. *ERA Rank A**
- Chapter 5:
  - **X. Chang**, Y. Yu, Y. Yang, and E. P. Xing: "They Are Not Equally Reliable: Semantic Event Search using Differentiated Concept Classifiers", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. *ERA Rank A*

The following are the papers published during the course of the Ph.D but not included in this thesis:

- **X. Chang** and Y. Yang: "Semi-Supervised Feature Analysis by Mining Correlations among Multiple Tasks", IEEE Transactions on Neural Networks and Learning Systems, 2016. *ERA Rank A**

- **X. Chang**, Z. Ma, Y. Yang, Z. Zheng and A. Hauptmann: "Bi-Level Semantic Representation Analysis for Multimedia Event Detection", IEEE Transactions on Cybernetics, 2016. *ERA Rank A*
- **X. Chang**, F. Nie, Y. Yang, C. Zhang and H. Huang: "Convex Sparse PCA for Unsupervised Feature Analysis", ACM Transactions on Knowledge Discovery from Data, 2016. *ERA Rank B*
- M. Luo, F. Nie, **X. Chang**, Y. Yang, A. Hauptmann and Q. Zheng: "Avoiding Optimal Mean Robust PCA/2DPCA with Non-greedy L1-norm Maximization", International Joint Conferences on Artificial Intelligence (IJCAI), 2016. *ERA Rank A**
- **X. Chang**, F. Nie, Y. Yang and X. Zhou: "A Convex Formulation for Spectral Shrunk Clustering", AAAI Conference on Artificial Intelligence (AAAI), 2015. *ERA Rank A**
- **X. Chang**, F. Nie, Y. Yang, X. Zhou and C. Zhang: "Compound Rank-k Projections for Bilinear Analysis", IEEE Transactions on Neural Networks and Learning Systems, 2015. *ERA Rank A**
- Y. Yang, Z. Ma, F. Nie, **X. Chang** and A. Hauptmann: "Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization", International Journal of Computer Vision, 2015. *ERA Rank A**
- S. Wang, F. Nie, **X. Chang**, L. Yao, X. Li and Q. Zheng: "Unsupervised Feature Analysis with Class Margin Optimization", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), 2015. *ERA Rank A*
- **X. Chang**, F. Nie, Y. Yang and H. Huang: "A Convex Formulation for Semi-Supervised Multi-Label Feature Selection", AAAI Conference on Artificial Intelligence (AAAI), 2014. *ERA Rank A**

Chapter 1

Introduction

Multimedia content analysis has been progressing from simple concepts Qian et al. [2014]; Shao et al. [2014] to complex events Ni et al. [2013]. The growing research effort is of great interest to improve user experience in multimedia search, particularly video search. This is generally realized by bridging the semantic gap between the low-level features and the high-level descriptions Habibian et al. [2013]; Hauptmann et al. [2007].

Pioneering work of video analysis focused primarily on surveillance videos, sports videos and *etc*. Dhar et al. [2011]; Hwang et al. [2011]; Wang and Forsyth [2009]. Video content detection is one of the fundamental tasks of video analysis. Its goal is to detect the occurrence of a category of interest from a set of infinite classes. As it is not reasonable to know all the infinite classes during the training phrase, building up a discriminative classifying hyperplane is difficult, which in turn makes detection a challenging task. Detection is readily applied to many real-world applications. For example, when a user queries "birthday party" on Youtube, the back-end system detects this from its tremendous video archives and provides the user with retrieval results. Within the last decades, video content detection has evolved from detecting simple concepts (*e.g*., objects, scenes or human actions) to more meaningful complex events such as "birthday party". A complex event is a higher level semantic abstraction of video sequence than a concept and consists of multiple concepts. For example, a "birthday party" event can be described by multiple objects (*e.g*., birthday cake), scene (*e.g*., in a park), actions (*e.g*., blowing candle) and acoustic concepts (*e.g*., music). Detection of these complex events is formidably difficult as they contain many objects, various actions, multiple scenes and have significant intra-class variations. In addition, they usually take place in much longer video clips where many frames can be uninformative so it is unlikely to infer an event with a single frame or a few frames.

Existing MED systems aim to gain good detection performance with different number of positive examples. Shirahama *et al*. have presented a Hidden Conditional Random Field based algorithm using 100 positive training examples for MED

Shirahama et al. [2014]. Ma *et al*. have proposed to adapt knowledge from auxiliary domain for MED when only 10 positive training exemplars are provided Ma et al. [2014]. Chang *et al*. have proposed an event-driven concept weighting algorithm for zero-shot MED Chang et al. [2016].

On the other hand, an MED system is desired to be intelligent with human-understandable reasoning that provides a list of concepts relevant to the target event. With such a motivation, a few researchers have worked on using semantic concept classifiers for MED. These findings demonstrate that semantic representation is comparable to hand-crafted low-level features in performance while has the advantage of identifying event-related concepts for reasoning. Technically, semantic representation for event video is obtained by training a variety of concept detectors related to objects, scenes and actions, *etc*. and then applying these detectors to the videos Habibian et al. [2013]. For current MED task, the concept detectors are usually trained using multiple image/video archives. These archives may have disparate structures leading to semantic representations with different capabilities for event detection.

In this chapter, we thoroughly review related literatures, clearly elaborate our motivations and contributions, and finally present the organization of the entire thesis.

1.1 Related Work

Reflecting the challenge of managing video content, the National Institute of Standards and Technology (NIST) hosts an annual competition on a variety of retrieval tasks, of which the multimedia event detection (MED) task has received increasing attention. According to the number of positive training exemplars, MED can be grouped into 100Ex (complex event detection), 10Ex (few-exemplar event detection) and 0Ex (zero-exemplar event detection). This section will review the related literatures accordingly.

1.1.1 Complex Event Detection

In 2010, TRECVID launched the MED challenge aiming to encourage new technologies for detecting more complicated multimedia events, *e.g*., making a sandwich. Ever since, researchers have been making effort to improve MED from different perspectives. Jiang *et al*. have proposed to use high-level features atop low-level ones by encoding high-level features into graphs and diffusing the scores obtained from low-level features on the established graph for MED Jiang et al. [2012]. The final prediction is derived from multiple graphs, each of which corresponds to a high-level feature. Ye *et al*. have presented a new video representation built upon joint audio-visual bi-modal words Ye et al. [2012a]. The proposed bi-modal words are obtained by exploring the relation across the quantized words extracted from both

the visual and audio modalities in video. A few pooling strategies are subsequently harnessed to formulate the bi-modal into the final video representation. Ma *et al*. have proposed to treat the negative videos unequally by assigning fine-grained labels to them in their classification model, motivated by that many negative videos may resemble the positive videos in different degrees Ma et al. [2013b]. A joint model is developed to optimize the fine-grained labels with the video representations for a more robust classifier. Younessian *et al*. have utilized multi-modal knowledge consisting of Automatic Speech Recognition (ASR) transcripts, acoustic concept indexing and visual semantic indexing for MED Younessian et al. [2012]. Liu *et al*. have studied how to use the high-level face information to assist in MED considering that people are usually the central subjects in videos Liu et al. [2014]. Yu *et al*. have proposed a way to learn abstract semantic "regions" during the feature pooling, thus being able to remove the "junk" information in video for better detection performance Yu et al. [2013]. Ma *et al*. have proposed an approach that exploits the external concept-based videos and target event videos simultaneously to learn an intermediate representation coupled with the event classifier Ma et al. [2013a]. Gkalelis *et al*. have combined a novel nonlinear discriminant analysis method and a linear Support Vector Machine to improve the computational efficiency for MED Gkalelis and Mezaris [2014]. In line with the popularity of deep Convolutional Neural Networks (CNN) features in computer vision community, several researchers also apply this technique to complex event detection. For instance, Xu *et al*. propose to effectively leverage deep CNNs to advance event detection, where only frame level static descriptors can be extracted by the existing CNN toolkits Xu et al. [2015]. The authors propose using a set of latent concept descriptors as the frame descriptor, which enriches visual information while keeping it computationally affordable. Gan *et al*. similarly develop a deep CNN representation that can be used for both event detection and evidence recounting Gan et al. [2015].

On top of various features, a bunch of statistical machine learning algorithms have been proposed for event detection Chang et al. [2015b]; Yu et al. [2014c], among which SVM is the most popular learning tool. Tang *et al*. introduce a hierarchical method for combining features based on AND/OR graph structure, where nodes in the graph represent combinations of different sets of features Tang et al. [2013]. Their method automatically learns the structure of the AND/OR graph using score-based structure learning, and we introduce an inference procedure that is able to efficiently compute structure scores. Vahdat *et al*. present a compositional model for video event detection that leverages a multiple kernel learning algorithm that incorporates structured latent variables Vahdat et al. [2013]. The kernelized latent variable framework allows the model to select and match test video segments with those that are extracted from the pool of training of videos. In Ma et al. [2013c], Ma *et al*. propose to leverage attributes at video level, *i.e*., the semantic labels of external videos are used as attributes. Building upon a correlation vector which correlates the attributes and the complex

event, the authors incorporate video attributes latently as extra informative cues into the event detector learned from complex event videos. Xu *et al*. propose utilizing related exemplars for complex event detection using multiple features Xu et al. [2014]. This is the first multiple feature fusion method which can deal with the related exemplars. Their method adaptively utilizes the related exemplars by cross-feature learning. They use ordinal labels to represent the multiple relevance levels of the related videos.

1.1.2 Few-Exemplar Event Detection

Given that precisely labeled positive exemplars are difficult to get in a real system, a few researchers have investigated how to gain reasonable performance with few positive examples. The presupposition is that it is non-trivial to collect positive exemplar videos which convey the precise semantics of a particular event and excludes any irrelevant information Yang et al. [2013]. In contrast, it is much easier to collect a video exemplar which is related to a specific event of interest, but does not necessarily contain all the essential elements of the event. The main challenge is how to utilize the related exemplars for complex event detection. It is not reasonable to assign hard labels to these related exemplars. The related exemplars can be of any other event, *e.g*., both "thesis proposal" and "people date" are related to "marriage proposal". None of the existing systems in the TRECVID competition has ever used the related exemplars. They usually use these related exemplars as negative exemplars.

A common strategy is to utilize auxiliary resources to compensate the scarce information from the few positive examples. Ma *et al*. propose using knowledge adaption from another source for MED even if the features of the source and the target are partially different, but overlapping Ma et al. [2014]. The authors exert a sophisticated method to alleviate the negative impact of the irrelevance and noise to optimize the event detector. VideoStory, proposed in Habibian et al. [2014b], combines correlated term labels if the combination improves the video classifier prediction. It prevents the combination of correlated terms which are visually dissimilar by optimizing a joint-objective balancing descriptiveness and predictability. Some researchers have also advocated using related videos to improve MED with 10 positive training examples. Related videos are those closely but not fully associated with the event of interest. Yang *et al*. have presented an algorithm that automatically evaluates how positive the related exemplars are for the detection of an event and uses them on an exemplar-specific basis for learning the event classification model Yang et al. [2013]. A weighted margin SVM formulation is developed by Tzelepis *et al*. to effectively incorporate related class observations for learning the classifier Tzelepis et al. [2013]. Notwithstanding, much effort is still needed in this realm since it is more consistent with the practical requirement regarding labeled training data.

1.1.3 Zero-Exemplar Event Detection

The goal of zero-exemplar event detection is to detect an event of interest, only based on a given event definition, without using any visual training examples. The event definition is usually given in the form of an event name and its description. This zero-exemplar event detection goes beyond conventional zero-shot image recognition of objects and scenes which relies on a training set of related classes and a pre-specified class-to-attribute mappings Farhadi et al. [2009]; Lampert et al. [2009]. In recent years, much research attention has been paid to this challenging zero-shot event detection because of its high practical value. Reflecting this challenge, the NIST has launched the corresponding TRECVID benchmark task NIST [2013, 2014]. Most existing zero-exemplar event detection systems first represent videos and the event queries using a semantic representation, following by ranking all the searching videos based on the cosine similarity with the event query.

Attributes and term-attributes have been widely exploited to represent the videos Chen et al. [2014]; Cui et al. [2014]; Wu et al. [2014]; Ye et al. [2015]. To step further, researchers have proposed many extensions. For example, Habibian *et al*. propose to train classifiers for combination of concepts composed by Boolean logic operators Habibian et al. [2014a]. The authors present an algorithm that automatically discovers composite concepts from existing video-level concept annotations and demonstrate that by combining concepts into composite concepts, more accurate classifiers for the concept vocabulary can be trained. In Mazloom et al. [2015], the authors propose learning a set of relevant frames as the concept prototypes from web video examples without the need for frame-level annotations, and using them for representing an event video. Jiang *et al*. propose adjusting the attribute scores using an ontology structure Jiang et al. [2015]. Beyond these attribute-based approaches, researchers have also considered automatic speech recognition (ASR) and optical character recognition (OCR) to enrich the semantic video representations Dalton et al. [2013]; Wu et al. [2014].

1.1.4 Semantic Representation

As early as in 2005, Hauptmann has suggested that proper usage of semantic concept is beneficial for improving the performance of video analysis Hauptmann [2005]. Complex event detection, as a higher-level of video analysis, could also benefit from semantic concepts, which has been demonstrated by recent works.

Essentially, using semantic concepts for complex event detection is to learn semantic representation of event videos via a large corpus of concepts covering object, scene and action. A multitude of approaches have been proposed for such purpose and they all seem promising for better performance gain. Yang *et al*. propose to discover data-driven concepts mapped to a low dimensional space using deep belief

nets (DBNs), after which a compact and robust sparse representation is learned to jointly model the concepts from multiple modalities (audio, scene and motion) Yang and Shah [2012]. Mazloom *et al*. propose to use concept detectors to learn a semantic signature representation to capture the variance in semantic appearance of events Mazloom et al. [2013b]. Merler *et al*. have introduced semantic model vectors that are learned with many semantic classifiers, each being an ensemble of SVM models trained from thousands of labeled web images, as an intermediate level semantic representation for complex event detection Merler et al. [2012b]. Jiang *et al*. propose to encode semantics into graphs, diffuse the scores on the established graph and fuse them with those obtained from low-level features to get the final prediction Jiang et al. [2012]. Ma *et al*. propose to leverage semantics at video level and build the classifier learning upon a correlation vector which correlates the semantics and the complex event Ma et al. [2013c]. Habibian *et al*. study how to create an effective vocabulary for complex event detection by considering the number, type, specificity and quality of the concept detectors Habibian et al. [2013]. Ye *et al*. build a large-scale event-specific concept library that is used a mid-level representation of complex events in videos Ye et al. [2015]. Wu *et al*. leverage multimedia archives with text descriptions to learn a large bank of concepts, in addition to using several off-the-shelf concept detectors, speech, and video text for representing videos and apply their approach to zero-shot complex event detection Wu et al. [2014]. Can *et al*. propose to exploit the dependencies of object-based concepts using a Markov Random Field based model, which enables the modeling of likelihood of these concepts in videos Can and Manmatha [2014].

1.1.5 Event Recounting

In event recounting, we are interested in knowing why a certain detection/recognition decision, *e.g*. this video contains the "horse riding competition" event, is made. Usually, a recounting algorithm is expected to return some evidences (*e.g*. sample frames) to support the decision. Event recounting is very useful since it helps locate the video segment that contains the event of interest. In order to recount a multimedia event *semantically* and *comprehensibly*, it is useful to characterize an event as a juxtaposition of various semantic concepts, such as actions, scenes and objects, which are more descriptive and meaningful. Thus, unlike event detection or recognition, mid-level concept representations of the video contents, thanks to their interpretability, are usually preferred to low-level features.

Most existing event recounting approaches focus on the temporal domain. In Liu et al. [2013c]; Sun et al. [2014], the authors apply object and action detectors or low-level visual features to localize temporal key evidences, and a video-level classifier is trained to rank the key frames or shots. These approaches built on the assumption that classifiers that can distinguish positive and negative exemplars can also be used to

distinguish the informative shots. However, they treat the shots or key frames equally, which may be dominated by the ubiquitous but non-informative ones. To overcome this limitation, Lai et al. [2014] formulated the recounting problem as multiple instance learning, which aims to learn a instance-level event detection and recounting model by selecting the informative shots or key frames during the training process. Tan et al. [2011a] proposed a rule-based recounting approach to collect the evidence, involving human knowledge heavily. In Gupta et al. [2009a], a generative And-OR graph model is used to represent the causal relationship between action concepts only. None of these works explicitly exploited saliency information.

1.2 Motivations and Contributions

The main motivation of this thesis is to explore semantic concepts for complex event analysis in unconstrained videos, from sufficient to scarce positive exemplars. Specifically, four approaches have been introduced in this thesis, which are **S**emantic **P**ooling for complex event analysis (**SP**), **J**oint **E**vent **D**etection and evidence **R**ecounting with limited supervision (**JEDaR**), **E**vent-**D**riven **C**oncept **W**eighting for semantic video search (**EDCW**) and semantic event search using **D**ifferentiated **C**oncept **C**lassifiers (**DCC**).

In **SP**, we propose a new semantic pooling approach for challenging event analysis tasks (*e.g.* event detection, event recognition and event recounting) on long untrimmed Internet videos, especially when only a few shots/segments are relevant to the event of interest while many other shots are irrelevant or even misleading. The commonly adopted pooling strategies aggregate the shots indifferently in one way or another, resulting in a great loss of information. Differently, we first define a novel notion of semantic saliency that assesses the relevance of each shot with the event of interest. We then prioritize the shots according to their saliency scores since shots that are semantically more salient are expected to contribute more to the final event analysis. Prioritizing objects according to saliency Koch and Ullman [1987] is ubiquitous in visual tasks such as segmentation Rahtu et al. [2010] and video summarization Lee et al. [2012]. Next, to overcome the small sample size issue due to limited training data, we propose to train an "informed" classifier that puts larger weights on more relevant shots. Leveraging on this ordering bias, we are able to significantly enhance the discriminative power of the statistical classifier. To achieve this goal, we propose a new isotonic regularizer that is able to exploit the constructed semantic ordering information. The isotonic regularizer encourages the classifier to put more weights on more salient shots. Although our isotonic regularizer is not convex, the popular proximal gradient algorithm can still be applied, hence enjoying the strong convergence guarantee recently established in Bolte et al. [2014a]. The key component, namely the proximal map of the isotonic regularizer, despite

being nonconvex, is solved globally and exactly in linear time through a sequence of reductions. The final algorithm, which we call nearly-isotonic SVM (NI-SVM), is very efficient and runs quickly on large real video datasets. For comparison, we also propose an alternative convex variant, although its performance is found to be inferior. We validate the proposed approach through extensive experiments conducted on three real-world unconstrained video datasets (CCV, MEDTest 2013 and MEDTest 2014), and achieve state-of-the-art performances measured by the mean average precision.

In **JEDaR**, we propose a joint framework to simultaneously classify high-level events and locate semantic evidences for each complex event. Different from most existing systems which perform MER as a post-processing step on top of the MED results, the proposed approach is capable of leveraging the mutual benefits of the two tasks. After extracting a semantic, albeit noisy, video representation, we introduce a recounting model based on recent sparse regularizers Chen et al. [2001a]; Rudin et al. [1992]; Yuan and Lin [2006b] that can localize key evidences both on concept-wise and temporal-wise, and a detection model based on the infinite push support vector machine (SVM) Rudin [2009] that greatly enhances the discriminative power. In a nutshell, our recounting model assists detection by filtering out noisy irrelevant information while simultaneously our detection model guides recounting by directing it to the most discriminative evidences. We further segment each video into multiple shots to exploit the rich temporal information, although this in turn creates a severe computational challenge, especially on large-scale datasets. Therefore, we propose a significantly improved alternating direction method of multiplier (ADMM) algorithm. In contrast to existing works Lai et al. [2014]; Rakotomamonjy [2012], we prove novel closed-form solutions for all intermediate updates that allow us to remove all inner loops, resulting in tremendously improved efficiency. The per-step complexity of our algorithm scales only linearly with the problem size. We test the proposed algorithm on the recent TRECVID MEDTest 2013 NIST [2013] and MEDTest 2014 datasets NIST [2014], and achieve very promising results for both detection and recounting.

In **EDCW**, we focus on automatically detecting events in unconstrained videos without the use of any visual training exemplars. Following previous work on zero-shot learning Lampert et al. [2009]; Palatucci et al. [2009], we regard an event as a composition of multiple mid-level semantic concepts. These semantic concept classifiers are shared among events and can be trained using other resources. We then learn a skip-gram model Mikolov et al. [2013b] to assess the semantic correlation of the event description and the pre-trained vocabulary of concepts, based on which we automatically select the most relevant concepts for each event of interest. This step is carried out without any visual training data. This concept bundle view of an event aligns with the cognitive science literature, where humans are found to conceive objects as bundles of attributes Roach and Lloyd [1978]. The concept prediction scores on the testing videos are combined to obtain a final ranking of the presence of the event of interest. However, most existing zero-shot event detection systems aggregate the

prediction scores of the concept classifiers with fixed weights. This clearly assumes all the predictions of a concept classifier share the same weight and fails to consider the differences in the classifier's prediction capability on individual testing videos. A concept classifier does in fact have different prediction capability on different testing videos, and some videos are correctly predicted while others are not. Therefore, instead of using a fixed weight for each concept classifier, a promising alternative is to estimate the specific weight for each testing video to alleviate the individual prediction errors from imperfect concept classifiers and achieve robust detection. The main building blocks of the proposed approach for zero-example event detection can be described as follows. We first rank the semantic concepts for each event of interest using the skip-gram model, based on which the relevant concept classifiers are selected. The aggregation process can be cast as an information propagation procedure which propagates the weights learned on individual online available videos to the individual unlabeled testing videos, which ensures that visually similar videos have similar aggregated scores and offers the ability to infer weights for the testing videos. We then use ℓ_∞ norm infinite push constraint to minimize the number of positive videos ranked below the highest-scored negative videos, which ensures that positive videos mostly have higher aggregated scores than negative videos. In this way, we learn the optimal weights for each testing video and push positive videos to rank above negative videos as possible. To validate the effectiveness of the proposed approach, we conduct extensive experiments on the latest TRECVID MEDTest 2014 NIST [2014], MEDTest 2013 NIST [2013] and CCV Jiang et al. [2011] datasets. The experimental results confirm the superiority of the proposed approach.

Most of existing works on semantic event search Habibian et al. [2013]; Jiang et al. [2014b]; Rastegari et al. [2013]; Wu et al. [2014] ignore the fact that *not all concept classifiers are equally reliable*, especially when they are pre-trained from other source domains. For example, "face" in video frames can now be reasonably accurately detected, but in contrast, the action "brush teeth" remains hard to recognize in short video clips. Consequently, a relevant concept can be of limited use or even misuse if its classifier is highly unreliable. Therefore, in **DCC**, when combining concept scores, we propose to take their relevance, predictive power, and reliability all into account. The significant challenge here is that we do not supply the learning algorithm with any labeled training data. To account for the unreliability of concept classifiers, we propose to use the warped spectral meta-learner in Parisi et al. [2014] to estimate the concept accuracies and combine them in a principled and purely unsupervised manner. We provide an efficient implementation based on the recent generalized conditional gradient (GCG), which is the key to conduct event search in large-scale video datasets. With essential modifications, we take the positive semidefinite constraint into account. Although the original objective function is nonconvex, we can combine the global GCG algorithm with a local fast solver for the nonconvex problem. The intuition is that both the original objective function and the smooth unconstrained problem

share the same set of global minimizers, and by combining them we gain both global optimality and local fast convergence, especially because the latter nonconvex problem has no constraint at all. Extensive experiments have been conducted on three real video datasets (MEDTest 2014 NIST [2014], MEDTest 2013 NIST [2013] and CCV Jiang et al. [2011] datasets), and achieve state-of-the-art performances.

In summary, the main contributions of this thesis lie in the following four aspects:

- We propose a new semantic pooling approach for challenging event analysis tasks on long untrimmed Internet videos, especially when only a few shots/segments are relevant to the event of interest while many other shots are irrelevant or even misleading.
- We design a joint framework that simultaneously detects high-level events and localizes the indicative concepts of the events. The proposed framework is able to leverage the mutual benefits of event detection and event recounting tasks.
- We present a novel event-driven concept weighting approach for zero-example event detection to learn the optimal weights of related concept classifiers for each testing video by exploring a set of online available videos which have free-form text descriptions of their content.
- To account for the unreliability of concept classifiers, we propose to use the warped spectral meta-learner to estimate the concept accuracies and combine them in a principled and purely unsupervised manner.

1.3 Thesis Structure

The remaining parts of this thesis are organized as follows:

- Chapter 2 introduces the **SP** approach for complex event detection, event recognition and event recounting.
- Chapter 3 introduces the **JEDaR** approach for joint event detection and recounting with limited supervision.
- Chapter 4 introduces the **EDCW** approach for semantic event search without using any visual training data.
- Chapter 5 introduces the **DCC** approach for semantic event search, taking the unreliability of the concept classifiers into consideration.
- Chapter 6 concludes this thesis and also presents possible future works.

Chapter 2

Semantic Pooling for Complex Event Analysis

The key to many event analysis tasks is a compact and discriminative representation of the video contents. Deep learning approaches, *e.g*. convolutional neural networks (CNNs), have become increasingly popular in this regard. The standard way Aly et al. [2013]; Yu et al. [2014b] is to extract local descriptors using CNNs on each frame of a video clip and then aggregate video-wise, through either average-pooling or max-pooling or even more complicated pooling strategies, *e.g*. Cao et al. [2012]; Li et al. [2013]; van de Sande et al. [2010]. While effective in reducing size, pooling may result in the loss of structural or temporal information. On the other hand, retaining all frame features may not be desirable either, for computational or statistical reasons, especially in light of the limited number of training examples.

Instead, in this chapter we consider an intermediate strategy. We first split each video into multiple shots, and for each shot we randomly sample one key frame whose extracted features will be used to represent the entire shot. Instead of conducting pooling on the shot-level, we prioritize the shots according to their "relevance" to the event of interest. Next, to overcome the small sample size issue due to limited training data, we propose to train an "informed" classifier that puts larger weights on more relevant shots. Leveraging on this ordering bias, we are able to significantly enhance the discriminative power of the statistical classifier.

More precisely, in §2.1 we propose a new prioritizing procedure based on the notion of *semantic saliency*. Prioritizing objects according to saliency Koch and Ullman [1985] is ubiquitous in visual tasks such as segmentation Rahtu et al. [2010] and video summarization Lee et al. [2012]. However, instead of borrowing an existing saliency algorithm, we prefer a more "supervised" notion that is closely related to our event analysis tasks. To this end, we first train 1,534 concept detectors using online available datasets (*e.g*., Trecvid SIN dataset, Google sports Karpathy et al. [2014], UCF101 dataset Soomro et al. [2012] and YFCC dataset Thomee et al. [2015]), resulting in

Figure 2.1: Two Internet video examples, where the same event *Rock Climbing* happened in very different time frames. The number in each frame indicates its saliency score, which describes how this keyframe is relevant to the specified event. We use this saliency information to prioritize the video shot representations.

a probability vector for each shot that indicates the relative presence of individual concepts. Then, using the skip-gram model Mikolov et al. [2013b] in natural language processing, we pre-learn a relevance vector that measures the *a priori* relevance of each concept name with the textual description (provided in most video event datasets) of the event of interest. Lastly, by taking a weighted combination of the probability vector (likelihood) and the relevance vector (prior), we obtain the proposed semantic saliency of each shot. Rearranging the shots according to their saliency scores yields the desired prioritization.

The prioritized shot-level representations of each video can then be used to perform event analysis tasks. For event detection, we feed the prioritized representations into a linear large margin classifier such as support vector machines (SVM). Intuitively, shots with higher semantic saliency scores are expected to be more relevant to the event of interest, hence providing more discriminative information. To incorporate this carefully constructed side information, we propose, in §2.2, a new isotonic regularizer that encourages the classifier to put more weights on more salient shots. Our isotonic regularizer is not convex, but as we show in §2.3, the popular proximal gradient algorithm can still be applied, hence enjoying the strong convergence guarantee recently established in Bolte et al. [2014b]. The key component, namely the proximal

map of the isotonic regularizer, despite being nonconvex, is solved globally and exactly in linear time through a sequence of reductions. The final algorithm, which we call nearly-isotonic SVM (NI-SVM), is very efficient and runs quickly on large real video datasets. For comparison, we also propose an alternative convex variant, although its performance is found to be inferior.

In §2.4 we extend NI-SVM to the event recognition task by combining the multiclass support vector machines with our isotonic regularizer. After properly smoothing the multiclass hinge loss we can again apply the proximal gradient algorithm, whose per-step complexity scales only linearly with the problem size. In §2.5 we show how to use our NI-SVM to perform event recounting.

In §2.6 we validate the proposed approach through extensive experiments conducted on three real-world unconstrained video datasets (CCV, MED13, MED14), and achieve state-of-the-art performances measured by the mean average precision.

2.1 Prioritization using Semantic Saliency

As we mentioned before, a good feature extraction module is vital for event analysis. Thus we first describe our feature extraction method. Since not all video shots are equally relevant to the event of interest, we develop in this section a new prioritization procedure to reorder them. Then in §2.2 and §2.4 we propose the nearly-isotonic SVM classifier to exploit this ordering information. The overall system is illustrated in Figure 2.2 and we discuss it block by block in the sequel.

2.1.1 Feature extraction

To extract representative features from videos, we first segment each video into m shots $[\mathbf{v}_1, \ldots, \mathbf{v}_m]$ using the color histogram difference as the indication of the shot boundary. Other segmentation or change-point detection algorithms may also be used. For simplicity, we randomly sample one key frame from each shot and extract the frame level CNN descriptors using the architecture of Simonyan and Zisserman [2015]. The key insight in Simonyan and Zisserman [2015] is that by using smaller convolution filters (3×3) and very deep architecture (16-19 layers), significant improvement on the ImageNet Challenge 2014 can be achieved. Due to its excellent performance on images, we therefore choose to apply the same architecture to our video datasets after sampling key frames. With some abuse of notation, the extracted CNN features (from fc6, the first fully-connected layer) of all m shots are still written collectively as $[\mathbf{v}_1, \ldots, \mathbf{v}_m] \in \mathbb{R}^{d \times m}$. In our experiments, we set m as the smallest number of keyframes for all videos. For example, in the Trecvid MED14 dataset, the shortest video has $m = 535$ keyframes. We do not explicitly model temporal information in this work, although conceivably it could further aid our system.

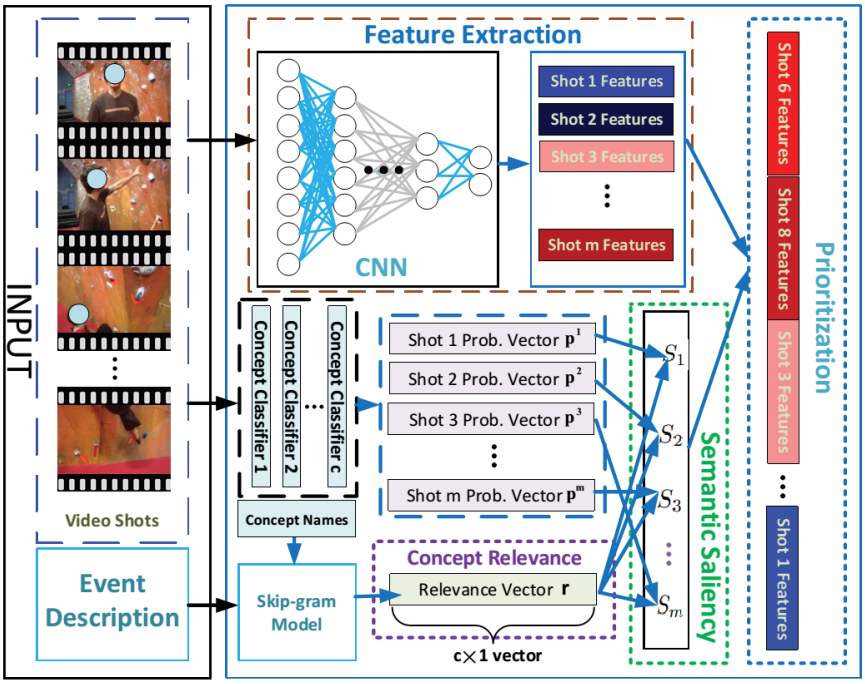


Figure 2.2: Each input video is divided into multiple shots, and each event has a short textual description. CNN is used to extract features (§2.1.1). Semantic concept names and skip-gram model are used to derive a probability vector (§2.1.2) and a relevance vector (§2.1.3), which are combined to yield the new semantic saliency and used for prioritizing shots in the classifier training (§2.1.4).

2.1.2 Concept detectors

We collect a concept vocabulary of $c = 1,534$ concepts from online available datasets (*e.g.*, Trecvid SIN dataset, Google sports Karpathy et al. [2014], UCF101 dataset Soomro et al. [2012] and YFCC dataset Thomee et al. [2015]), each accompanied with an entity description (*e.g.*, *rope climbing, skiing, fencing, diving, playing piano, horse race*). These concepts can be used to aid event analysis. For example, we would expect concepts such as *horse race* and *horse riding* to be relevant to the event "horse competition." Thus we train a detector/classifier for each concept in the vocabulary. All c concept classifiers/detectors will be applied to each video shot $\mathbf{v}_j$, resulting in a c-dimensional probability vector $\mathbf{p} \in \mathbb{R}_+^c$, with the entry p_k standing for the (relative) probability of having the k-th concept appear in the shot $\mathbf{v}_j$. Finally, we will combine the probability vector $\mathbf{p}$ and the concept relevance (defined next) to yield the semantic saliency scores.

2.1.3 Concept relevance

Events come with short textual descriptions. For example, the event *dog show* in the Trecvid MED14 NIST [2014] is defined as "a competitive exhibition of dogs." We exploit this textual information by learning a semantic relevance score between the event description and the individual concept names (note that stop words are removed using NLTK Bird [2006]). More precisely, a skip-gram model Mikolov et al. [2013a] can be trained using the English Wikipedia dump (`http://dumps.wikimedia.org/enwiki/`). The skip-gram model learns a D-dimensional vector space representation of words by fitting the joint probability of the co-occurrence of surrounding contexts on large unstructured text data, and places semantically similar words near each other in the embedding vector space. Thus, it is able to capture a large number of precise syntactic and semantic word relationships. To learn the vector representation of short phrases consisting of multiple words, we aggregate the word embeddings using Fisher vectors Sánchez et al. [2013] and follow mostly Clinchant and Perronnin [2013] except our embedding vectors are from the skip-gram model. In the Fisher vector, each phrase (*i.e.*, a set of words) is described as the gradient of the log-likelihood of these observations on an underlying probabilistic model, and we use vlfeat Vedaldi and Fulkerson [2010] to generate the Fisher vector for each phrase. After normalizing the length of the respective vector representations, we compute the cosine distance between the event description and each concept name, resulting in a relevance vector $\mathbf{r} \in \mathbb{R}^c$, where r_k measures the *a priori* relevance of the k-th concept to the event of interest. Note that the concept relevance vector $\mathbf{r}$ is event dependent, and we repeat the procedure for each event in our training data.

2.1.4 Semantic saliency

Lastly, we define the semantic saliency score of each video shot as a weighted combination of the concept probability vector $\mathbf{p}$ (the likelihood, different for each video shot, see §2.1.2) and the concept relevance vector $\mathbf{r}$ (the prior, same for all shots, see §2.1.3):

$$s := \mathbf{p}^\top \mathbf{r} = \sum_{k=1}^{c} p_k r_k. \tag{2.1}$$

Repeating this for each shot $\mathbf{v}_j, j = 1, \ldots, m$, of a video generates its saliency vector $\mathbf{s} = [s_1, \ldots, s_m]$. Note that this saliency vector $\mathbf{s}$ is event dependent, and we derive it separately for each event in our training data. Intuitively, the saliency score s_j evaluates the importance of the j-th shot to the event of interest. The most salient shots are those most likely to contain the specified event, hence they should carry more weight in the

final classifier boundary. Thus we prioritize the shots by reordering them such that

$$s_1 \geq s_2 \geq \cdots \geq s_m, \tag{2.2}$$

i.e., the shots are ranked in a descending order of saliency. Importantly, note that different videos are likely reordered differently. After prioritization, it is desirable to train an event classifier that takes this valuable ordering information into account, which motivates the isotonic regularizer that we propose in the next section. Note that all our results can be extended to a partial ordering, *i.e.*, allowing some shots to be incomparable (when their scores are very close, for instance).

The definition of our semantic saliency essentially follows the zero-shot learning framework of Lampert et al. [2009]. It is convenient because it is fully automatic. A potentially superior approach is to extract relevant keyframes from the positive training exemplars and use these to define saliency. However, the downside of this alternative is that it requires some human intervention/labeling.

2.2 Nearly-Isotonic Support Vector Machines for event detection

As described above, we represent each video $V^i, i = 1, \ldots, n$, as a matrix $[\mathbf{v}_1^i, \ldots, \mathbf{v}_m^i]$, where $\mathbf{v}_j^i \in \mathbb{R}^d$ are the extracted CNN features from the j-th shot. In this section we consider event detection, that is, decide whether or not a test video belongs to an event of interest. As the usual practice, we model the event detection as a binary classification problem, and we reorder the shot-level CNN features according to their semantic saliency scores as defined in §2.1.4. The resulting feature representation is still denoted as $V^i = [\mathbf{v}_1^i, \ldots, \mathbf{v}_m^i]$ for notational simplicity.

To perform event detection, we then employ the large margin *binary* support vector machines (SVM):

$$\min_{W \in \mathbb{R}^{d \times m}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle V^i, W \rangle) + \lambda \cdot \Omega(W), \tag{2.3}$$

where $\lambda \geq 0$ is the regularization constant, and the loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ measures the discrepancy between the true label $y_i \in \{1, -1\}$ and the prediction $\langle V^i, W \rangle := \sum_{j,k} V_{jk}^i W_{jk}$. For instance, we can use

- the least squares loss: $\ell(y, t) = \frac{1}{2}(y - t)^2$;
- the hinge loss: $\ell(y, t) = (1 - yt)_+$, where as usual $(t)_+ := \max\{t, 0\}$ is the positive part;
- the squared hinge loss: $\ell(y, t) = \frac{1}{2}(1 - yt)_+^2$;

- the logistic loss: $\ell(y,t) = \log(1+\exp(-yt))$.

Note that the hinge loss is not differentiable, however, both the squared hinge loss and the logistic loss can be used instead as a smooth approximation. To detect a test video V, we use the usual sign rule:

$$\hat{y} = \operatorname{sign}(\langle V, W\rangle). \tag{2.4}$$

The regularizer Ω in (2.3) is introduced to induce some desirable structures on the classifier weight matrix W, and will play a major role in the sequel. In vanilla SVM, $\Omega(W) = \|W\|_{\mathsf{F}}^2$ (the squared Frobenius norm), which penalizes large weight matrices to avoid overfitting. Another useful regularizer is $\Omega(\mathbf{w}) = \|W\|_1$ (the ℓ_1-norm, sum of absolute values), which encourages sparsity hence is effective for feature selection. However, neither of the above-mentioned regularizers is able to exploit the order information that we carefully constructed in §2.1. In fact, both of them are invariant to column reorderings. Instead, we propose below a new isotonic regularizer that respects the prioritization we performed on the shots using their saliency scores.

2.2.1 The isotonic regularizer

Let us assume momentarily that $d = 1$, *i.e.*, there is only a single feature. This assumption, although unrealistic, simplifies our presentation and will be removed later. As mentioned, we want to learn a weight vector that respects the saliency order in our shot-level features, since more relevant shots are expected to contribute more to the final detection boundary. This motivates us to consider the following isotonic regularizer:

$$\|\mathbf{w}\|_\imath := \sum_{j=2}^{m} (|w_j| - |w_{j-1}|)_+. \tag{2.5}$$

To see the rationale behind, let us use the absolute value $|w_j|$ of the weight vector to indicate the contribution of the j-th shot to the final decision rule $\operatorname{sign}(\sum_j v_j w_j)$. Since the shots are arranged in decreasing order of relevance, we would expect roughly $|w_1| \geq |w_2| \geq \cdots \geq |w_m|$, *i.e.*, the weights (in magnitude) align well with the saliency order we constructed in §2.1.4. If this is the case, the regularizer $\|\mathbf{w}\|_\imath$ would be 0, *i.e.* incurring no penalty. On the other hand, we pay a linear cost for violating any of the saliency orders, *i.e.*, if instead $|w_j| > |w_{j-1}|$ for some j, we suffer a cost equal to the difference $|w_j| - |w_{j-1}|$. Clearly, the more we deviate from a pair of saliency order, the more we are penalized. Equipping $\Omega(\mathbf{w}) = \|\mathbf{w}\|_\imath$ in (2.3) we obtain a new classification method which we call the nearly-isotonic SVM (NI-SVM):

$$\min_{W} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \langle V^i, W\rangle) + \lambda \cdot \|W\|_\imath. \tag{2.6}$$

Exploiting order information in statistical estimation has a long history, see the wonderful book Barlow et al. [1972] for early applications. Similar regularizers to (2.5) have also appeared recently. For instance, Tibshirani et al. [2011] dropped the absolute values in (2.5) and considered:

$$\|\mathbf{w}\|_+ := \sum_{j=2}^{m} (w_j - w_{j-1})_+, \tag{2.7}$$

while Yang et al. [2012] replaced the positive part in (2.5) with the absolute value:

$$\|\mathbf{w}\|_a := \sum_{j=2}^{m} \Big| |w_j| - |w_{j-1}| \Big|. \tag{2.8}$$

The well-known total variation (semi)norm Rudin et al. [1992]:

$$\|\mathbf{w}\|_{\mathsf{tv}} := \sum_{j=2}^{m} |w_j - w_{j-1}| \tag{2.9}$$

is also widely used in image denoising problems. These alternative regularizers bear a clear similarity to ours, however, we believe our formulation (2.5) is more appropriate for our setting (see §2.6.2 below for empirical verification): The weight vector $\mathbf{w}$ has signed entries, and the order we aim to force is one-directional. Indeed, the alternative regularizer (2.8) will always incur a cost except when $|w_j| = |w_{j-1}|$, a condition that is too stringent to be useful in our setting. Similarly, for two negative entries $0 > w_j > w_{j-1}$, the alternative regularizer (2.7) would incur an unnecessary penalty $w_j - w_{j-1} > 0$. The same problem also occurs for the total variation norm (2.9). Note that all four regularizers (2.5), (2.7), (2.8), and (2.9) are nondifferentiable while (2.5) and (2.8) are also nonconvex. Nevertheless, we can still design an efficient algorithm for solving NI-SVM (with regularizer (2.5)). Before that, however, let us mention how to extend to multiple features ($d > 1$).

2.2.2 Extending to multiple features

When $d > 1$, each video representation V^i is a matrix in $\mathbb{R}^{d\times m}$, hence accordingly the linear classifier we learn is indexed by the weight matrix $W \in \mathbb{R}^{d\times m}$. Inspecting the NI-SVM formulation (2.6), we note first that the loss term extends immediately: the standard inner product $\langle V^i, W\rangle$ in $\mathbb{R}^{d\times m}$ extends straightforwardly for any d. For the isotonic regularizer, we need to summarize all d importance measures (each contributed by a feature). There are multiple ways to achieve this, and we consider

two particularly convenient ones here:

$$\|W\|_{\imath,1} := \sum_{i=1}^{d} \|W_{i,:}\|_\imath = \sum_{i=1}^{d} \sum_{j=2}^{m} (|W_{i,j}| - |W_{i,j-1}|)_+, \tag{2.10}$$

$$\|W\|_{\imath,2} := \sum_{j=2}^{m} (\|W_{:,j}\|_2 - \|W_{:,j-1}\|_2)_+, \tag{2.11}$$

where $W_{i,:}$ (resp. $W_{:,j}$) is the i-th row (resp. j-th column) of the matrix W. The first regularizer (2.10) simply sums the vector isotonic regularizer along each feature dimension, while the second regularizer (2.11) first aggregates the shot-level importance by computing the Euclidean norm of the d weights and then applies the vector isotonic regularizer on top. When $d = 1$, both (2.10) and (2.11) reduce to the vector isotonic regularizer (2.5), but we expect them to behave differently when $d > 1$. The corresponding NI-SVM formulation (2.6) with the matrix regularizers (2.10) and (2.11) will be called respectively NI-SVM$_1$ and NI-SVM$_2$.

2.2.3 A convex alternative

The matrix isotonic regularizers (2.10) and (2.11) are not convex, making the corresponding NI-SVM$_1$ and NI-SVM$_2$ formulations also nonconvex. In this section we propose a simple *convex* alternative, mainly as a comparison baseline against the above nonconvex formulations.

To this end, we add a nonnegative constraint on the classifier weight matrix W:

$$\min_{W \geq 0} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle V^i, W \rangle) + \lambda \cdot \|W\|_\imath, \tag{2.12}$$

where $\|W\|_\imath$ can be either $\|W\|_{\imath,1}$ (NI-SVM$_{1+}$) or $\|W\|_{\imath,2}$ (NI-SVM$_{2+}$). Note that the convexity in (2.12) is gained by placing a restriction on the classifier, which in turn may jeopardize its prediction performance (verified in our experiments). On the other hand, the nonnegative constraint encourages a sparse weight matrix W, in a spirit similar to nonnegative matrix factorization Lee and Seung [1999], since our video representation V^i is nonnegative as well. This may in turn be beneficial in interpretation tasks.

2.3 Solving NI-SVM by the proximal gradient

The matrix isotonic regularizers (2.10) and (2.11) are both nonsmooth and nonconvex, making the optimization of the NI-SVM formulation (2.6) a very challenging task. Fortunately, the proximal gradient algorithm (see *e.g.* Fukushima and Mine [1981])

has been recently extended in Bolte et al. [2014b] to handle semialgebraic functions that need not be convex or smooth. Thus, below we first briefly recall the proximal gradient algorithm, and then we show how to efficiently implement its key component (the proximal map) for our NI-SVM formulation.

2.3.1 The proximal gradient

We first recall that a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is semialgebraic if its graph $\{(\mathbf{w}, f(\mathbf{w})) : \mathbf{w} \in \operatorname{dom} f\}$ is a semialgebraic set, *i.e.*, a finite union of finite intersections of the sets $\{\mathbf{z} \in \mathbb{R}^{d+1} : p_1(\mathbf{z}) < 0, p_2(\mathbf{z}) = 0\}$, where p_1, p_2 are polynomial functions. By definition of course polynomials are semialgebraic. Furthermore, many functions used in practice are semialgebraic, including all functions we use in this work[1]. Restricting to semialgebraic functions f and g, we can now consider the general composite minimization problem:

$$\min_{\mathbf{w}} \; f(\mathbf{w}) + g(\mathbf{w}). \tag{2.13}$$

For our NI-SVM formulation (2.6), the loss term corresponds to the function f and the matrix isotonic regularizer corresponds to the function g. Since we do not assume convexity, we will be satisfied with convergence to a critical point[2].

Lemma 2.1 (Bolte et al. [2014b], Proposition 3)**.** *Let f and g be semialgebraic functions, and f is differentiable whose gradient ∇f is L-Lipschitz continuous:*

$$\forall \mathbf{x}, \forall \mathbf{y}, \; \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2. \tag{2.14}$$

Then, for any step size $\eta \in (0, 1/L)$, the following iteration converges to a critical point of the minimization problem (2.13)*, provided that the iterates $\{\mathbf{w}\}$ are bounded:*

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla f(\mathbf{w}), \tag{2.15}$$

$$\mathbf{w} \leftarrow \mathsf{P}_g^{\eta}(\mathbf{w}) := \left\{ \operatorname*{argmin}_{\mathbf{z}} \tfrac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2 + g(\mathbf{z}) \right\}. \tag{2.16}$$

The above iterations (2.15)-(2.16), which we call proximal gradient (PG), are very simple: it amounts to a usual gradient step w.r.t. f first, and then a proximal step w.r.t. g using the *proximal map* P_g^{η}. The proximal map is a natural generalization of the

[1] Except for the logistic loss. However, all results still hold since all we really need is the function to be "definable"—a very technical term that we refrain from giving details, but see Bolte et al. [2014b] and the references therein.

[2] For nonsmooth functions, we define the critical points to be the set $\{\mathbf{w} : 0 \in \partial(f + g)(\mathbf{w})\}$, where ∂f is the subdifferential of f, a strict generalization of the usual gradient, see Rockafellar and Wets [1998] for details.

Euclidean projection operator onto a closed set, and it is well-defined if the function g does not decrease faster than a quadratic function. For instance, when g is the 1-norm, then $\mathsf{P}^{\eta}_{\|\cdot\|_1}(\mathbf{w}) = \operatorname{sign}(\mathbf{w}) \cdot (|\mathbf{w}| - \eta)_+$ is the well-known soft-shrinkage operator. The boundedness assumption in Lemma 2.1 is not restrictive: it is satisfied as long as the sublevel sets $\{\mathbf{w} : f(\mathbf{w}) + g(\mathbf{w}) \leq \alpha\}$ are bounded, which is usually met.

Since evaluating the gradient ∇f is straightforward (for the nondifferentiable hinge loss, we will see how to approximate it using the logistic loss in §2.4), the efficiency of PG (2.15)-(2.16) hinges on our capability of computing the proximal map (2.16) quickly, which itself is a minimization problem. This is a great advantage of the PG algorithm: it encapsulates the nonconvexity and nonsmoothness of the function g entirely into the proximal map (2.16). If the function g is "simple" enough so that its proximal map can be solved in closed-form, then we bypass the nonconvex and nonsmooth issue completely. This is indeed the case for our matrix isotonic regularizers (2.10) and (2.11), as we demonstrate next.

2.3.2 Proximal map for the isotonic regularizer

In this subsection we show that the proximal map for both matrix isotonic regularizers (2.10) and (2.11) can be computed exactly in linear time. This is achieved through a sequence of reductions.

Reducing to the vector case:

We first reduce the proximal maps for the matrix isotonic regularizers (2.10) and (2.11):

$$\mathsf{P}^{\eta}_{\|\cdot\|_{\imath,1}}(W) := \operatorname*{argmin}_{Z} \tfrac{1}{2\eta}\|W - Z\|_{\mathsf{F}}^2 + \|Z\|_{\imath,1} \tag{2.17}$$

$$\mathsf{P}^{\eta}_{\|\cdot\|_{\imath,2}}(W) := \operatorname*{argmin}_{Z} \tfrac{1}{2\eta}\|W - Z\|_{\mathsf{F}}^2 + \|Z\|_{\imath,2} \tag{2.18}$$

to their vector cousin:

$$\mathsf{P}^{\eta}_{\|\cdot\|_{\imath}}(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \tfrac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2 + \|\mathbf{z}\|_{\imath}, \tag{2.19}$$

where $\mathbf{w}, \mathbf{z} \in \mathbb{R}^m$ and $\|\cdot\|_{\imath}$ is defined in (2.5).

For (2.17) this reduction is obvious as $\|W\|_{\imath,1}$ is separable in rows of the matrix W, so we need only apply (3.16) to each row of W independently. For (2.18) its objective function expands as follows:

$$\frac{1}{2\eta}\sum_{j=1}^{m}\|W_{:,j} - Z_{:,j}\|_2^2 + \sum_{j=2}^{m}(\|Z_{:,j}\|_2 - \|Z_{:,j-1}\|_2)_+. \tag{2.20}$$

Now consider the decomposition $Z = \Theta\Lambda$, where each column of Θ has unit Euclidean norm and Λ is diagonal with z_i in the i-th diagonal. Clearly, the regularizer $\|Z\|_{\imath,2}$ only

depends on Λ, and for fixed Λ, the first term in (2.20) is minimized precisely when $\Theta_{:,j} = \frac{W_{:,j}}{\|W_{:,j}\|_2}$ for all j. Plugging it back we can solve the diagonal matrix $\Lambda = \mathrm{diag}(\mathbf{z})$ by:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \quad \frac{1}{2\eta}\sum_{j=1}^{m}(\|W_{:,j}\|_2 - z_j)^2 + \|\mathbf{z}\|_\iota, \tag{2.21}$$

which clearly is in the form of the vector problem (3.16).

Therefore, we need only focus on the vector proximal map (3.16). Note that the isotonic regularizer $\|\mathbf{w}\|_\iota$ is not convex, thus its proximal map in (3.16) is not a convex problem. Nevertheless, we will show how to solve it exactly and *globally* in linear time.

Reducing to the convex case:

Crucially, we observe that the vector isotonic regularizer $\|\mathbf{z}\|_\iota$ is invariant to the sign changes of any component z_i, but the quadratic term $\frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2$ is minimized when the signs of $\mathbf{w}$ and $\mathbf{z}$ match. Thus, at any minimizer of (3.16) we must have $\mathrm{sign}(w_i) = \mathrm{sign}(z_i)$ for all i, further reducing the vector problem (3.16) to:

$$\mathsf{P}^{\eta}_{\kappa+\|\cdot\|_\iota}(|\mathbf{w}|) := \operatorname*{argmin}_{\mathbf{z}} \tfrac{1}{2\eta}\Big\|\mathbf{z}-|\mathbf{w}|\Big\|_2^2 + \kappa(\mathbf{z}) + \|\mathbf{z}\|_\iota \tag{2.22}$$

$$= \operatorname*{argmin}_{\mathbf{z}\geq 0} \tfrac{1}{2\eta}\Big\|\mathbf{z}-|\mathbf{w}|\Big\|_2^2 + \|\mathbf{z}\|_\iota, \tag{2.23}$$

where $|\mathbf{w}|$ is the component-wise absolute value of $\mathbf{w}$, and

$$\kappa(\mathbf{z}) = \begin{cases} 0, & \text{if } \mathbf{z}\geq 0 \\ \infty, & \text{otherwise} \end{cases}. \tag{2.24}$$

If we can solve (2.22), now a convex problem thanks to the nonnegative constraint, then we can immediately recover

$$\mathsf{P}^{\eta}_{\|\cdot\|_\iota}(\mathbf{w}) = \mathsf{P}^{\eta}_{\kappa+\|\cdot\|_\iota}(|\mathbf{w}|)\cdot\mathrm{sign}(\mathbf{w}). \tag{2.25}$$

(The multiplication on the right-hand side is component-wise.)

Reducing to the total variation norm:

Two elementary observations turn out to be key in solving (2.22) efficiently: (a). Under the nonnegative constraint $\mathbf{z}\geq 0$, we have

$$2\|\mathbf{z}\|_\iota = \|\mathbf{z}\|_{\mathsf{tv}} + z_m - z_1, \tag{2.26}$$

which follows from applying the identity $2(t)_+ = t + |t|$ to each term $(|z_j| - |z_{j-1}|)_+$. (b). The function κ in (2.24), *i.e.*, the nonnegative constraint, is invariant to permutations of the coordinates.

Denote $h(\mathbf{z}) = z_m - z_1$, and recall from (2.22) that we need to solve the proximal map of the function

$$\kappa(\mathbf{z}) + \|\mathbf{z}\|_\imath = \kappa(\mathbf{z}) + \tfrac{1}{2}(\|\mathbf{z}\|_{\mathsf{tv}} + h(\mathbf{z})). \tag{2.27}$$

Then we have the following decomposition rule:

Theorem 2.2. *Denote $\mathbf{e}_i$ the i-th canonical basis in $\mathbb{R}^m$. For any $\eta \geq 0$ and for all $\mathbf{w} \in \mathbb{R}^m$,*

$$\mathsf{P}^\eta_{\kappa+\|\cdot\|_\imath}(\mathbf{w}) = \mathsf{P}^\eta_\kappa\Big(\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}\big(\mathsf{P}^{\eta/2}_h(\mathbf{w})\big)\Big), \text{ where} \tag{2.28}$$

$$\mathsf{P}^\eta_\kappa(\mathbf{w}) = (\mathbf{w})_+, \tag{2.29}$$

$$\mathsf{P}^{\eta/2}_h(\mathbf{w}) = \mathbf{w} + \tfrac{\eta}{2}(\mathbf{e}_1 - \mathbf{e}_m). \tag{2.30}$$

Theorem 2.2 is a special case of Theorem 2.3 below. The key insight here is that to compute the proximal map of the left-hand side $\kappa(\mathbf{z}) + \|\mathbf{z}\|_\imath$ in (2.27), we need only apply the proximal maps of each of the three terms on the right-hand side of (2.27). We have not specified the proximal map of the total variation norm $\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}$, however, it has a well-known linear time exact algorithm, see *e.g.* Davies and Kovac [2001].

2.3.3 Adding more regularizers

In some applications it may be desirable to add other regularizers. For instance, the squared 2-norm can be used to avoid overfitting and the 1-norm may be needed for feature selection. Pleasantly, we can easily incorporate these additional regularizers, without complicating the algorithm at all, thanks to the following result:

Theorem 2.3. *With the same setup as in Theorem 2.2, we have for any $\eta, \gamma \geq 0$ and for all $\mathbf{w} \in \mathbb{R}^m$:*

$$\mathsf{P}^\eta_{\kappa+\|\cdot\|_\imath+\gamma\|\cdot\|_2^2}(\mathbf{w}) = \mathsf{P}^\eta_{\gamma\|\cdot\|_2^2}\Big[\mathsf{P}^\eta_\kappa\Big(\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}\big(\mathsf{P}^{\eta/2}_h(\mathbf{w})\big)\Big)\Big] \tag{2.31}$$

$$\mathsf{P}^\eta_{\kappa+\|\cdot\|_\imath+\gamma\|\cdot\|_1}(\mathbf{w}) = \mathsf{P}^\eta_{\gamma\|\cdot\|_1}\Big[\mathsf{P}^\eta_\kappa\Big(\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}\big(\mathsf{P}^{\eta/2}_h(\mathbf{w})\big)\Big)\Big]. \tag{2.32}$$

Proof. We repeatedly apply the results in Yu [2013] to the identity (2.27).

Consider (2.31) first. Since the function h is linear, Theorem 3 of Yu [2013] implies that

$$\mathsf{P}^\eta_{\kappa+\|\cdot\|_\imath+\gamma\|\cdot\|_2^2}(\mathbf{w}) = \mathsf{P}^\eta_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}+\frac{1}{2}h+\gamma\|\cdot\|_2^2}(\mathbf{w}) \tag{2.33}$$

$$= \mathsf{P}^\eta_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}+\gamma\|\cdot\|_2^2}\big(\mathsf{P}^\eta_{\frac{1}{2}h}(\mathbf{w})\big) \tag{2.34}$$

$$= \mathsf{P}^\eta_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}+\gamma\|\cdot\|_2^2}\big(\mathsf{P}^{\eta/2}_h(\mathbf{w})\big), \tag{2.35}$$

where it is easy to verify that

$$\mathsf{P}^{\eta}_{\frac{1}{2}h}(\mathbf{w}) = \mathsf{P}^{\eta/2}_{h}(\mathbf{w}) = \mathbf{w} + \tfrac{\eta}{2}(\mathbf{e}_1 - \mathbf{e}_m). \tag{2.36}$$

Next, since both κ and the total variation norm $\|\cdot\|_{\mathsf{tv}}$ are positive homogeneous, applying Theorem 4 of Yu [2013] we obtain

$$\mathsf{P}^{\eta}_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}+\gamma\|\cdot\|_2^2}(\mathbf{w}) = \mathsf{P}^{\eta}_{\gamma\|\cdot\|_2^2}\left(\mathsf{P}^{\eta}_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}}(\mathbf{w})\right). \tag{2.37}$$

Lastly, since κ is permutation invariant, Corollary 4 of Yu [2013] gives

$$\mathsf{P}^{\eta}_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}}(\mathbf{w}) = \mathsf{P}^{\eta}_{\kappa}\Big(\mathsf{P}^{\eta}_{\frac{1}{2}\|\cdot\|_{\mathsf{tv}}}(\mathbf{w})\Big) = \mathsf{P}^{\eta}_{\kappa}\left(\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}(\mathbf{w})\right). \tag{2.38}$$

Assembling the above results we obtain (2.31).

For (2.32), we use the same reasoning as above to get

$$\begin{aligned}
\mathsf{P}^{\eta}_{\kappa+\|\cdot\|_\iota+\gamma\|\cdot\|_1}(\mathbf{w}) &= \mathsf{P}^{\eta}_{\kappa+\frac{1}{2}\|\cdot\|_{\mathsf{tv}}+\frac{1}{2}h+\gamma\|\cdot\|_1}(\mathbf{w}) && (2.39)\\
&= \mathsf{P}^{\eta}_{\kappa+\gamma\|\cdot\|_1}\Big[\mathsf{P}^{\eta}_{\frac{1}{2}\|\cdot\|_{\mathsf{tv}}}\Big(\mathsf{P}^{\eta}_{\frac{1}{2}h}(\mathbf{w})\Big)\Big] && (2.40)\\
&= \mathsf{P}^{\eta}_{\kappa+\gamma\|\cdot\|_1}\Big[\mathsf{P}^{\eta/2}_{\|\cdot\|_{\mathsf{tv}}}\Big(\mathsf{P}^{\eta/2}_{h}(\mathbf{w})\Big)\Big]. && (2.41)
\end{aligned}$$

Since both the function κ and $\gamma\|\cdot\|_1$ are separable in coordinates, we can reduce their proximal maps into the one dimensional case. It is then easily verified that

$$\mathsf{P}^{\eta}_{\kappa+\gamma\|\cdot\|_1}(\mathbf{w}) = \mathsf{P}^{\eta}_{\kappa}\Big(\mathsf{P}^{\eta}_{\gamma\|\cdot\|_1}(\mathbf{w})\Big) = \mathsf{P}^{\eta}_{\gamma\|\cdot\|_1}\Big(\mathsf{P}^{\eta}_{\kappa}(\mathbf{w})\Big). \tag{2.42}$$

Assembling the results again we obtain (2.32). □

Of course, if we take $\gamma = 0$, then $\mathsf{P}^{\eta}_{\gamma\|\cdot\|_2^2}(\mathbf{w}) = \mathsf{P}^{\eta}_{\gamma\|\cdot\|_1}(\mathbf{w}) = \mathbf{w}$, allowing us to recover Theorem 2.2 as a special case.

Putting things together:

We summarize the above reductions and steps in Algorithm 1. Despite the nonconvexity and nonsmoothness, Algorithm 1 computes the proximal maps of the matrix isotonic regularizers (2.10) and (2.11) *globally* and *exactly* in linear time. It is clear that each iteration of the proximal gradient costs $O(dmn)$ in time complexity while it costs $O(dm)$ in space complexity.

Conveniently, the proximal gradient Algorithm 1 we developed above for NI-SVM can be easily recycled for the convex alternative in §2.2.3, with only a single slight change: We do not backup or restore the sign (*e.g.* omitting lines 14 and 19 in Algorithm 1).

Algorithm 1: Proximal Gradient for NI-SVM

Input: $W \in \mathbb{R}^{d\times m}$, regularization λ, γ, step size η.
repeat
  $W \leftarrow W - \frac{\eta}{n}\sum_i \ell'(y_i, \langle V^i, W\rangle)V^i$; `// grad`
  $W \leftarrow \begin{cases} \texttt{prox\_row}\,(W, \eta, \lambda, \gamma)\,; & \texttt{// for (2.10)} \\ \texttt{prox\_col}\,(W, \eta, \lambda, \gamma)\,; & \texttt{// for (2.11)} \end{cases}$
until *convergence*;
Procedure `prox_row` $(W, \eta, \lambda, \gamma)$
  for *j= 1, . . . , d* **do**
    $W_{j,:} \leftarrow \texttt{prox\_vec}\,(W_{j,:}, \eta, \lambda, \gamma)$
Procedure `prox_col` $(W, \eta, \lambda, \gamma)$
  $\mathbf{w} \leftarrow (\|W_{:,1}\|_2, \ldots, \|W_{:,m}\|_2)$
  $\mathbf{w} \leftarrow \texttt{prox\_vec}\,(\mathbf{w}, \eta, \lambda, \gamma)$
  $W \leftarrow W \cdot \operatorname{diag}\left(\frac{w_1}{\|W_{:,1}\|_2}, \ldots, \frac{w_m}{\|W_{:,m}\|_2}\right)$
Procedure `prox_vec` $(\mathbf{w}, \eta, \lambda, \gamma)$
  $\mathbf{s} \leftarrow \operatorname{sign}(\mathbf{w}), \mathbf{w} \leftarrow |\mathbf{w}|$; `// omitted for (2.12)`
  $\mathbf{w} \leftarrow \mathbf{w} + \frac{\lambda\eta}{2}(\mathbf{e}_1 - \mathbf{e}_m)$
  $\mathbf{w} \leftarrow \mathsf{P}^{\eta\lambda/2}_{\|\cdot\|_{\mathsf{tv}}}(\mathbf{w})$
  $\mathbf{w} \leftarrow (\mathbf{w})_+$
  $\mathbf{w} \leftarrow \begin{cases} (\mathbf{w} - \gamma\eta)_+ & \texttt{// for (2.31)} \\ \frac{1}{1+2\gamma\eta}\mathbf{w} & \texttt{// for (2.32)} \end{cases}$
  $\mathbf{w} \leftarrow \mathbf{s} \cdot \mathbf{w}$; `// omitted for (2.12)`

2.3.4 Comparison against a smooth "ℓ_2-ish" regularizer

To better appreciate the decomposition rule in Theorem 2.2, let us consider an "ℓ_2-ish" alternative regularizer

$$\Phi(\mathbf{w}) := \sum_{j=2}^{m} (|w_j| - |w_{j-1}|)_+^2, \tag{2.43}$$

which penalizes more heavily if the violation of the order is large (and vice versa). Using a similar sign match argument, the proximal map P^{η}_{Φ} reduces to (for $\eta = 1/2$ say)

$$\min_{\mathbf{w}\geq 0} \ \|\mathbf{w} - \mathbf{z}\|_2^2 + \phi(\mathbf{w}), \ \text{where} \ \ \phi(\mathbf{w}) := \sum_{j=2}^{m} (w_j - w_{j-1})^2.$$

Surprisingly, although the above minimization problem appears even simpler than (2.23) (in the sense of the smoothness of the objective), we are not aware of a linear time exact algorithm for it. In particular, the naive decomposition rule $\mathsf{P}_\kappa^\eta(\mathsf{P}_\phi^\eta)$, for κ defined in (2.24), does **not** hold. This partly explains why our specific choice of the isotonic regularizer in (2.5) is convenient.

2.4 Multiclass NI-SVM for event recognition

In this section we consider the event recognition problem, that is, to decide which of the k events does a test video $V = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ belong to. Like previous work we model event recognition as a multiclass classification problem, *i.e.* the label $y \in \{1, \ldots, k\}$. In the following we extend our NI-SVM formulation (2.6) to this multiclass setting by following the work of Crammer and Singer [2001].

For each event e and video V^i, following §2.1 we compute its saliency score vector $\mathbf{s}^{i,e}$, which then induces a permutation matrix $P^{i,e} \in \mathbb{R}^{d\times m}$ such that $V^{i,e} = V^i P^{i,e}$, *i.e.* we reorder the video representation V^i so that its saliency vector is ordered decreasingly. For each event e we train a classifier represented as $W^e \in \mathbb{R}^{d\times m}$, and we define the multiclass loss as

$$\ell_i = \ell_i(W^1, \ldots, W^k) \tag{2.44}$$

$$= \max_{e=1,\ldots,k} \langle V^{i,e}, W^e\rangle - \langle V^{i,y_i}, W^{y_i}\rangle + 1 - \mathbb{1}_{e=y_i}, \tag{2.45}$$

where $\mathbb{1}_{e=y_i} = \begin{cases} 1, & \text{if } e = y_i \\ 0, & \text{otherwise} \end{cases}$ is the indicator function. Clearly the multiclass loss ℓ_i couples the k classifiers due to the max operator in (2.45): it is zero if the true prediction $\langle V^{i,y_i}, W^{y_i}\rangle$ is larger than any other prediction $\langle V^{i,e}, W^e\rangle, e \neq y_i$ by at least a margin of size 1, otherwise we pay a linear cost w.r.t. the most confusing wrong label.

Now we are ready to present the multiclass NI-SVM formulation:

$$\min_{W^1,\ldots,W^k} \underbrace{\frac{1}{n}\sum_{i=1}^n \ell_i(W^1, \ldots, W^k)}_{f} + \underbrace{\sum_{e=1}^k \lambda\|W^e\|_{\iota,p} + \gamma\|W^e\|_p^p}_{g}. \tag{2.46}$$

Depending on whether $p = 1$ or $p = 2$, we will denote (2.46) as NI-SVM$_1^{\mathrm{m}}$ or NI-SVM$_2^{\mathrm{m}}$. By adding the nonnegative constraint $W^e \geq 0$, we again have a convex alternative, which will be denoted respectively as NI-SVM$_{1+}^{\mathrm{m}}$ and NI-SVM$_{2+}^{\mathrm{m}}$. To recognize a test video V, we resort to the max-prediction rule:

$$\hat{y} = \operatorname*{argmax}_{e=1,\ldots,k} \langle V^e, W^e\rangle, \tag{2.47}$$

where ties are broken arbitrarily. Of course for $k = 2$, the multiclass formulation (2.46) reduces to the binary formulation (2.6) (with the hinge loss).

We can again optimize the multiclass formulation (2.46) using the proximal gradient, except one small problem: the multiclass hinge loss (2.45) is not differentiable. Nevertheless, there is a well-known smoothing technique to get around this issue, using the following inequality:

$$\mu \log \sum_{e=1}^{k} \exp(a_e/\mu) - \mu \log k \leq \max\{a_1, \ldots, a_k\} \tag{2.48}$$

$$\leq \mu \log \sum_{e=1}^{k} \exp(a_e/\mu). \tag{2.49}$$

Therefore, as long as the smoothing parameter μ is small, we can adequately approximate the max function using the log-sum-exp function, which is clearly differentiable (with Lipschitz continuous gradient). Applying this technique to the multiclass loss (2.45) we get rid of the nonsmoothness of the loss, and make the proximal gradient applicable again. Of course, the binary hinge loss can be dealt with analogously, although it is often easier to use simply the squared binary hinge loss.

As to the proximal map of the regularizer g in (2.46), we need only apply the steps in Section 2.3.2 independently to each of the k classifier weights W^e, thanks to separability. The entire procedure is very similar to Algorithm 1 hence we do not reproduce the pseudo-code here. It suffices to note that each iteration costs $O(dmnk)$ in time and $O(dmk)$ in space.

2.5 Event Recounting using NI-SVM

Note that event recounting aims at providing comprehensible evidences to justify a detection/recognition decision. Here we present a simple scoring approach on top of NI-SVM to perform event recounting. The idea is that the NI-SVM classifier is designed to assign larger weights to more semantically salient shots, thus it is natural to use these weights to compute a recounting score for each shot. More specifically, for each test video $V = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ that is declared to contain the event of interest, we compute the recounting score for its j-th shot as follows:

$$\mathcal{RS}_j = \langle \mathbf{v}_j, \mathbf{w}_j \rangle, \quad j = 1, \ldots, m, \tag{2.50}$$

where recall that $W = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$ is the classifier weight returned by (the binary) NI-SVM. Then, we rank the shots using the scores $\mathcal{RS}_j$ above and return the top ones

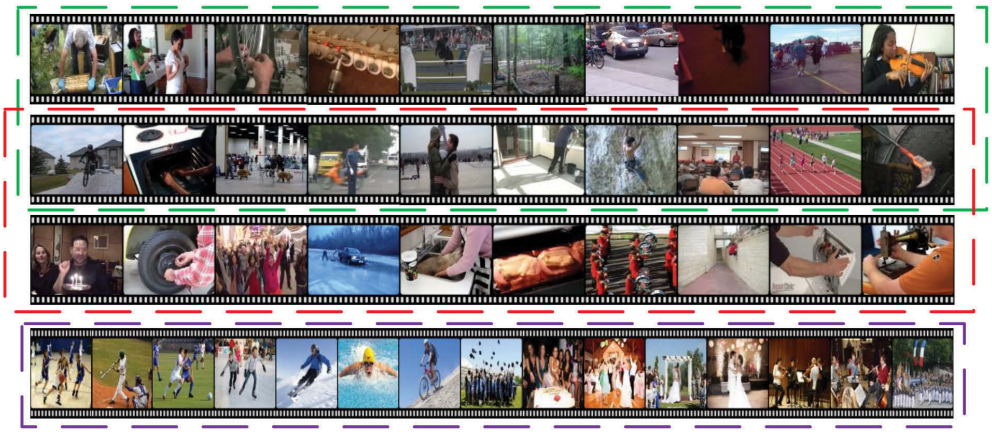


Figure 2.3: Example videos from the MED14 (green), MED13 (red), and CCV_{sub} (purple) datasets . Note that MED14 and MED13 have 10 events in common.

as evidence. Note that our decision rule for event detection is

$$\hat{y} = \text{sign}(\langle V, W \rangle) = \text{sign}\left(\sum_{j=1}^{m} \mathcal{RS}_j \right), \tag{2.51}$$

hence a shot with a large recounting score $\mathcal{RS}_j$ is likely to contribute a lot to or even determine the decision.

A similar scoring approach can be used to recount event recognition results. Interestingly, in this multiclass scenario, if a test video is declared to belong to event e, then shots with large *negative* scores $\langle \mathbf{v}_j, \mathbf{w}_j^{e'} \rangle$ for all $e' \neq e$ can also be considered as evidences to support event e.

2.6 Experiments

In this section we carry out extensive experiments to validate the proposed approach, on three event analysis tasks: event detection, event recounting, and event recognition.

2.6.1 Experimental Setup

Datasets:

We test on the following three real video datasets, of which some exemplar videos are provided in Figure 2.3.

1. MED14: The Trecvid MEDTest 2014 dataset NIST [2014] contains approximately 100 positive training exemplars per event, and all events share $\sim$ 5000

negative training exemplars. The test set has about 23,000 videos. There are in total 20 events, with descriptions in NIST [2014]. To our best knowledge, this is the largest (35,914 videos in total) public dataset for event analysis.

2. MED13 NIST [2013]: Similar as MED14. Note that 10 of its 20 events overlap with those of MED14.

3. CCV_{sub}: The official Columbia Consumer Video dataset Jiang et al. [2011] contains 9,317 videos in total with 20 semantic categories, including scenes like "beach", objects like "cat", and events like "baseball" and "parade". For our purpose we only use the 15 event categories. For each event we use its own training data as positive and all other training data as negative, totaling 4,659 training videos and 4,658 testing videos.

Concept classifier vocabulary:

We pre-train 1,534 concept classifiers using Trecvid SIN dataset (346 classes), Google sports (478 classes) Karpathy et al. [2014], UCF101 dataset (101 classes) Soomro et al. [2012] and YFCC dataset (609 classes) Thomee et al. [2015]. We first extract improved dense trajectory features with the code provided in Wang and Schmid [2013] and encode with the fisher vector representation Perronnin et al. [2010]. Then, on top of the extracted low-level features the cascade SVM Graf et al. [2004] was trained for each semantic concept. We further extract the improved dense trajectory features of all the shots for each video in the same fashion and apply the concept detectors to each shot.

Parameter tuning:

As mentioned in §2.1.1 we use the CNN architecture in Simonyan and Zisserman [2015] to extract 4096 features on one keyframe per video shot. The regularization constants λ and γ are selected using cross-validation from the range $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$.

2.6.2 Event Detection

In this section, we evaluate the performance of the proposed NI-SVM for complex event detection. According to the NIST standard NIST [2013, 2014], we detect each event separately.

Evaluation metric:

According to the NIST standard NIST [2013, 2014], we evaluate the event detection performance by the mean Average Precision (mAP). Average precision is a single-valued metric approximating the area under the precision-recall curve, and is widely used in information retrieval tasks. Denote R as the number of true relevant videos in a test dataset. At any index j, let R_j be the number of relevant videos in the

top j list. Then, AP is defined as

$$\text{AP} = \frac{1}{R} \sum_{j=1}^{n} \frac{R_j}{j} \times I_j, \tag{2.52}$$

where $I_j = 1$ if the j-th video is relevant (positive) and 0 otherwise. When all relevant videos are ranked on top of the irrelevant ones, AP achieves its largest value 1.0. Thus, a larger AP usually indicates a better performance.

Competitors:

For our NI-SVM formulations we consider both the least squares loss $\ell(y,t) = \frac{1}{2}(t-y)^2$ and the squared[1] hinge loss $\ell(y,t) = (1-yt)_+^2$. We use the subscript $_1$ and $_2$ respectively to distinguish the matrix isotonic regularizers (2.10) and (2.11). A further subscript $_+$ is used to signal the convex alternative in §2.2.3. More precisely, we compare the following algorithms:

- LS_A: least squares loss with average-pooling on the video shots. Note that pooling is performed on the selected m keyframes, for fairness and efficiency.
- LS_M: least squares loss with max-pooling.
- LS_T: least squares loss without pooling, but the shots are prioritized according to their saliency scores.
- NI-LS_1: least squares loss with isotonic regularizer (2.10).
- NI-LS_2: least squares loss with isotonic regularizer (2.11).
- NI-LS_{1+}: nonnegative convex version of NI-LS_1.
- NI-LS_{2+}: nonnegative convex version of NI-LS_2.

Similarly, for the squared hinge loss, we replace "LS" throughout with "SVM". As suggested in §2.3.3, additional ℓ_2^2 and ℓ_1 regularizers can be incorporated. We also compare against some state-of-the-art alternatives in §2.6.2.

Results against standard baselines:

We report the experimental results in Table 2.1 and Table 2.2 (least squares loss) and Table 2.3 and Table 2.4 (squared hinge loss), where full details on the MED14 dataset are documented. The average performances on MED13 and CCV_sub are also recorded at the bottom of Table 2.1, Table 2.2, Table 2.3 and Table 2.4, with full details deferred to the supplement.

We make a few observations from Table 2.1, Table 2.2, Table 2.3 and Table 2.4:

1) Average-pooling outperforms max-pooling on average and in most events.

2) LS_T and SVM_T perform significantly better than their pooling counterparts. This confirms that pooling, if naively done, can be detrimental. However, LS_T and SVM_T do not directly benefit from prioritizing the shots: their classifier weights ignore the ordering information.

[1]The convergence guarantee in Lemma 2.1 requires the loss to be smooth, hence excludes the usual hinge loss.

Table 2.1: Mean average precisions (mAP) with least squares loss ℓ_2^2 regularization on MED14 (full details), MED13 (summary), and CCV_{sub} (summary). A larger mAP indicates better performance.

		ℓ_2^2 regularized						
Dataset	ID	LS_{A}	LS_{M}	LS_{T}	NI-LS_1	NI-LS_{1+}	NI-LS_2	NI-LS_{2+}
MED14	21	0.149	0.114	0.205	0.223	0.218	0.222	0.213
	22	0.106	0.094	0.126	0.157	0.136	0.144	0.143
	23	0.735	0.714	0.814	0.853	0.808	0.831	0.789
	24	0.027	0.025	0.031	0.022	0.058	0.042	0.043
	25	0.009	0.009	0.011	0.011	0.015	0.038	0.021
	26	0.077	0.063	0.104	0.110	0.111	0.094	0.099
	27	0.142	0.148	0.135	0.205	0.143	0.188	0.195
	28	0.349	0.308	0.378	0.410	0.399	0.384	0.389
	29	0.233	0.151	0.314	0.385	0.348	0.359	0.297
	30	0.099	0.115	0.126	0.099	0.143	0.126	0.141
	31	0.761	0.760	0.738	0.779	0.746	0.819	0.812
	32	0.207	0.113	0.218	0.231	0.225	0.299	0.264
	33	0.511	0.499	0.536	0.571	0.542	0.632	0.509
	34	0.366	0.343	0.402	0.431	0.428	0.515	0.428
	35	0.423	0.329	0.418	0.463	0.435	0.503	0.413
	36	0.122	0.127	0.138	0.135	0.142	0.172	0.176
	37	0.347	0.279	0.326	0.412	0.331	0.443	0.325
	38	0.035	0.031	0.029	0.043	0.025	0.045	0.071
	39	0.417	0.341	0.388	0.524	0.411	0.521	0.453
	40	0.075	0.072	0.132	0.123	0.128	0.195	0.098
	mAP	0.259	0.232	0.273	0.309	0.289	0.329	0.294
MED13	mAP	0.298	0.281	0.356	0.369	0.351	0.383	0.360
CCV_{sub}	mAP	0.727	0.705	0.741	0.767	0.739	0.773	0.757

3) NI-LS, with either matrix isotonic regularizers, further outperforms LS_{T}, demonstrating that properly exploiting the ordering information can significantly improve the performance. Moreover, the matrix isotonic regularizer (2.11) (subscript $_2$) generally performs better than the matrix isotonic regularizer (2.10) (subscript $_1$).

4) The squared hinge loss on average performs better than the least squares loss, unanimously across all methods.

5) Additional ℓ_2^2-norm regularization (left panel) generally outperforms additional ℓ_1-norm regularization (right panel). We hypothesize that it is because the CNN features we use are very discriminative hence sparsity does not help here.

6) The convex variants (with subscript $_+$) have poorer performance than the nonconvex counterparts (but still competitive against average-pooling), possibly because the nonnegative constraint is too restrictive. Empirically (results not shown), we also found that the nonconvex variants are quite robust against initializations (random or using the convex variant), likely because we are able to solve the proximal maps exactly.

Table 2.2: Mean average precisions (mAP) with least squares loss (ℓ_1 regularization) on MED14 (full details), MED13 (summary), and CCV_{sub} (summary). A larger mAP indicates better performance.

		ℓ_1 regularized						
Dataset	ID	LS_{A}	LS_{M}	LS_{T}	NI-LS_1	NI-LS_{1+}	NI-LS_2	NI-LS_{2+}
MED14	21	0.132	0.105	0.186	0.218	0.226	0.163	0.157
	22	0.097	0.084	0.103	0.099	0.112	0.121	0.084
	23	0.722	0.698	0.732	0.821	0.743	0.713	0.714
	24	0.026	0.019	0.035	0.047	0.041	0.053	0.039
	25	0.008	0.010	0.008	0.018	0.009	0.009	0.007
	26	0.066	0.062	0.078	0.085	0.092	0.087	0.081
	27	0.133	0.117	0.154	0.182	0.178	0.158	0.139
	28	0.334	0.304	0.331	0.382	0.326	0.351	0.334
	29	0.218	0.200	0.255	0.296	0.269	0.288	0.256
	30	0.091	0.085	0.092	0.099	0.089	0.089	0.086
	31	0.744	0.728	0.725	0.778	0.735	0.778	0.759
	32	0.194	0.178	0.199	0.197	0.208	0.185	0.147
	33	0.502	0.493	0.487	0.532	0.495	0.513	0.452
	34	0.352	0.331	0.398	0.388	0.396	0.411	0.375
	35	0.418	0.369	0.312	0.364	0.318	0.404	0.365
	36	0.111	0.103	0.098	0.082	0.099	0.113	0.109
	37	0.332	0.318	0.338	0.361	0.342	0.312	0.284
	38	0.042	0.036	0.041	0.061	0.043	0.048	0.033
	39	0.425	0.409	0.427	0.465	0.422	0.446	0.413
	40	0.068	0.061	0.072	0.094	0.099	0.098	0.093
	mAP	0.251	0.236	0.254	0.278	0.262	0.267	0.246
MED13	mAP	0.306	0.295	0.337	0.342	0.341	0.332	0.308
CCV_{sub}	mAP	0.718	0.703	0.735	0.738	0.733	0.716	0.743

Results against state-of-the-art alternatives

We further compare the performance of the proposed approaches against some state-of-the-art alternatives. The results are summarized in Table 2.5, from which we observe that the proposed approach, with either the least squares loss or the squared hinge loss, achieves the best accuracy on all three datasets. For instance, the proposed approach (with squared hinge loss) outperforms the second best approach by a margin as large as 4%, which is 0.344 *vs*. 0.329, on the most recent and challenging MED14 dataset. These experimental results indicate that prioritizing the video shots using semantic saliency and exploiting the ordering information using nearly-isotonic SVM can largely improve the performance of complex event detection.

Results against different isotonic regularizers:

In this section, we compare different isotonic regularizers that have appeared in the literature:

- $\|\mathbf{w}\|_\imath := \sum_{j=2}^{m}(|w_j| - |w_{j-1}|)_+$; proposed in this work.
- $\|\mathbf{w}\|_+ := \sum_{j=2}^{m}(w_j - w_{j-1})_+$; Tibshirani et al. [2011].
- $\|\mathbf{w}\|_a := \sum_{j=2}^{m}\Big||w_j| - |w_{j-1}|\Big|$; Yang et al. [2012].

Table 2.3: Mean average precisions (mAP) with squared hinge loss (ℓ_2^2 regularization) on MED14 (full details), MED13 (summary), and CCV_{sub} (summary). A larger mAP indicates better performance.

		ℓ_2^2 regularized						
Dataset	ID	SVM_{A}	SVM_{M}	SVM_{T}	NI-SVM_{1}	NI-SVM_{1+}	NI-SVM_{2}	NI-SVM_{2+}
MED14	21	0.153	0.131	0.215	0.241	0.232	0.238	0.226
	22	0.117	0.105	0.148	0.163	0.151	0.156	0.155
	23	0.752	0.726	0.833	0.864	0.821	0.848	0.805
	24	0.037	0.035	0.052	0.031	0.067	0.051	0.053
	25	0.012	0.015	0.018	0.021	0.024	0.051	0.034
	26	0.086	0.078	0.127	0.123	0.135	0.108	0.109
	27	0.154	0.161	0.145	0.217	0.158	0.208	0.206
	28	0.356	0.317	0.396	0.421	0.409	0.395	0.397
	29	0.247	0.163	0.362	0.398	0.353	0.369	0.312
	30	0.109	0.126	0.143	0.112	0.157	0.138	0.157
	31	0.775	0.781	0.772	0.792	0.763	0.843	0.836
	32	0.221	0.125	0.235	0.222	0.241	0.312	0.277
	33	0.523	0.514	0.538	0.584	0.559	0.645	0.521
	34	0.381	0.358	0.426	0.448	0.441	0.528	0.461
	35	0.441	0.337	0.438	0.475	0.452	0.521	0.438
	36	0.147	0.141	0.158	0.154	0.161	0.192	0.198
	37	0.359	0.292	0.374	0.431	0.354	0.462	0.343
	38	0.048	0.045	0.047	0.052	0.036	0.061	0.084
	39	0.431	0.354	0.456	0.543	0.432	0.542	0.473
	40	0.097	0.093	0.132	0.141	0.143	0.209	0.114
	mAP	0.272	0.245	0.301	0.322	0.304	0.344	0.310
MED13	mAP	0.311	0.292	0.360	0.381	0.364	0.392	0.374
CCV_{sub}	mAP	0.738	0.716	0.753	0.779	0.750	0.783	0.768

- $\|\mathbf{w}\|_{\text{tv}} := \sum_{j=2}^{m} |w_j - w_{j-1}|$; this is the well-known total variational norm.

All of the above isotonic regularizers are extended to the matrix setting as illustrated in §5.1.6. Figure 2.4 gives full details of the results on MED14, where we use the least squares loss, an additional ℓ_2^2 regularizer, and the matrix extension (2.11) of the isotonic regularizers. Similar results on MED13 and CCV_{sub} can be found respectively in Figure 2.5 and Figure 2.6 in the supplement. As expected, our isotonic regularizer $\|\cdot\|_{\imath}$ achieves the best overall performance.

Sensitivity against tuning parameters λ and γ:

Next, we conduct experiments to assess the sensitivity of NI-SVM w.r.t. the regularization parameters λ and γ. To be more specific, we first fix $\gamma = 1$, which is the median of its allowed range of values, and we record the mAP by varying λ in Figure 2.7, from which we observe that the performance is relatively robust against the parameter λ. Generally speaking, the best performance is obtained when λ is in the range of $\{10^{-3}, 10^{-2}, 10^{-1}\}$.

Then, we fix λ at the median value 1 and test the sensitivity against the parameter γ. The mAP with varying γ is shown in Figure 2.8, from which we see that the performance degrades when γ is overly large. The best performance is obtained when γ is in the range of $\{10^{-2}, 10^{-1}, 10^{0}\}$.

Table 2.4: Mean average precisions (mAP) with squared hinge loss (ℓ_1 regularization) on MED14 (full details), MED13 (summary), and CCV_{sub} (summary). A larger mAP indicates better performance.

		ℓ_1 regularized						
Dataset	ID	SVM_{A}	SVM_{M}	SVM_{T}	NI-SVM_{1}	NI-SVM_{1+}	NI-SVM_{2}	NI-SVM_{2+}
MED14	21	0.147	0.118	0.247	0.225	0.241	0.182	0.174
	22	0.109	0.096	0.129	0.112	0.137	0.143	0.108
	23	0.741	0.712	0.753	0.835	0.759	0.735	0.728
	24	0.041	0.037	0.066	0.059	0.071	0.069	0.053
	25	0.014	0.017	0.028	0.023	0.026	0.027	0.024
	26	0.081	0.078	0.095	0.095	0.112	0.103	0.098
	27	0.145	0.131	0.173	0.194	0.189	0.171	0.153
	28	0.348	0.321	0.352	0.398	0.345	0.364	0.348
	29	0.231	0.217	0.275	0.312	0.281	0.302	0.271
	30	0.109	0.098	0.114	0.107	0.096	0.097	0.096
	31	0.761	0.743	0.775	0.791	0.751	0.792	0.772
	32	0.207	0.191	0.183	0.208	0.214	0.198	0.154
	33	0.516	0.506	0.524	0.547	0.512	0.527	0.468
	34	0.364	0.348	0.385	0.396	0.407	0.428	0.393
	35	0.431	0.384	0.387	0.378	0.336	0.417	0.383
	36	0.128	0.116	0.108	0.102	0.109	0.123	0.114
	37	0.348	0.323	0.351	0.374	0.368	0.347	0.298
	38	0.059	0.048	0.049	0.078	0.056	0.059	0.048
	39	0.436	0.421	0.442	0.473	0.437	0.458	0.428
	40	0.082	0.075	0.088	0.108	0.104	0.103	0.109
	mAP	0.265	0.249	0.276	0.291	0.278	0.282	0.261
MED13	mAP	0.318	0.309	0.342	0.354	0.382	0.344	0.320
CCV_{sub}	mAP	0.731	0.716	0.745	0.751	0.745	0.729	0.755

Table 2.5: Comparing against state-of-the-art alternatives, with additional ℓ_2^2 regularization and matrix isotonic regularizer (2.11).

Dataset	Tang et al. [2012]	Lai et al. [2014]	Li et al. [2013]	Karpathy et al. [2014]	NI-LS_2	NI-SVM_2
MED14	0.275	0.296	0.288	0.304	*0.329*	**0.344**
MED13	0.346	0.362	0.353	0.371	*0.383*	**0.392**
CCV_{sub}	0.734	0.747	0.741	0.758	*0.773*	**0.783**

Sensitivity against random initializations:

The formulation of NI-SVM is nonconvex, hence in theory it could have multiple local optima. In practice we observed that the proximal gradient in Algorithm 1 always converged to a reasonable solution. To test this point, we repeatedly run Algorithm 1 20 times, each with a different initialization. We also tried to initialize Algorithm 1 with the (globally) optimal solution of the convex variant in §2.2.3. The results w.r.t. AP on all three datasets are recorded in Figure 2.9. It is clear that the convex variants (with subscript $_+$) have stable performance w.r.t. different initializations, thanks to convexity. The nonconvex variants exhibit small variations, and if we initialize it by the solution of the convex variant, we get slightly worse but stable performance. Overall, Algorithm 1 is able to converge to a reasonable solution rather quickly.

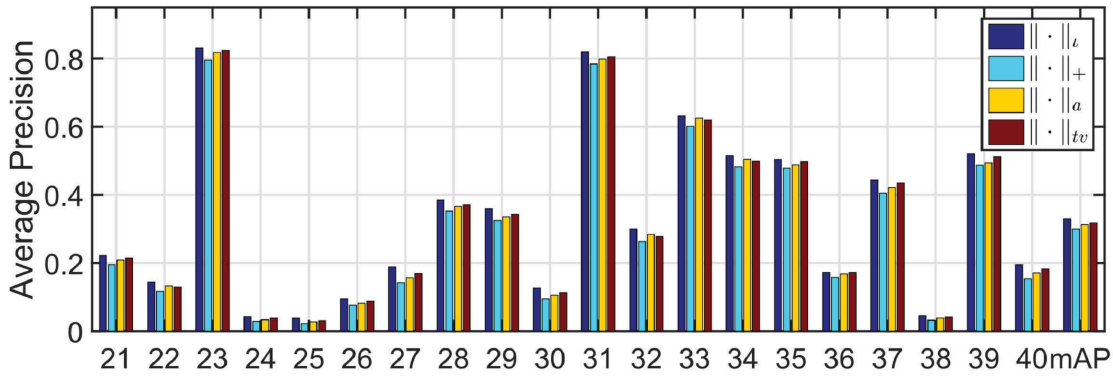


Figure 2.4: Comparing different isotonic regularizers on the MED14 dataset. The x-axis indicates the event ID.

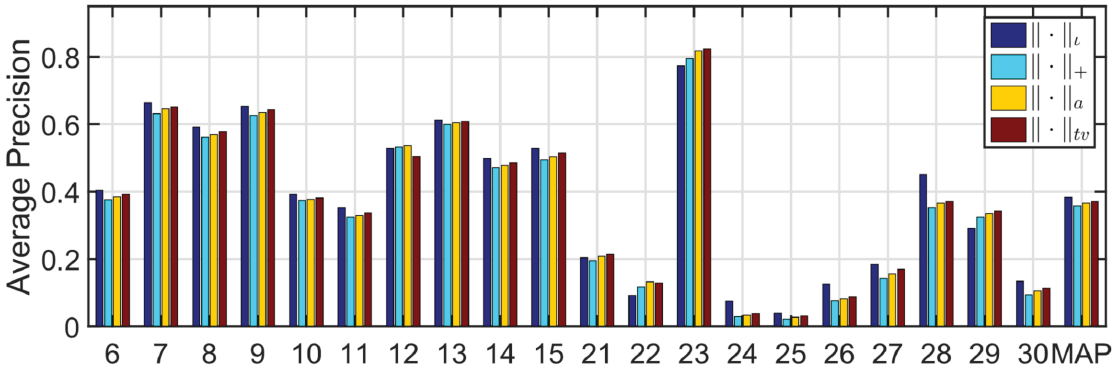


Figure 2.5: Comparing different isotonic regularizers on the MED13 dataset. The x-axis indicates the event ID.

Table 2.6: Average video length, shot length, and accuracy of the proposed recounting method.

	MED13	MED14
Average Video Length	164.3 seconds	188.4 seconds
Average Shot Length	6.1 seconds	8.3 seconds
Average Accuracy	91.4 %	85.7 %

2.6.3 Event recounting

We conduct experiments on event recounting in this section, using the scoring method detailed in §2.5. We only consider the event detection setting for succinctness.

Evaluation metric:

Evaluating the performance of event recounting algorithms is challenging for the

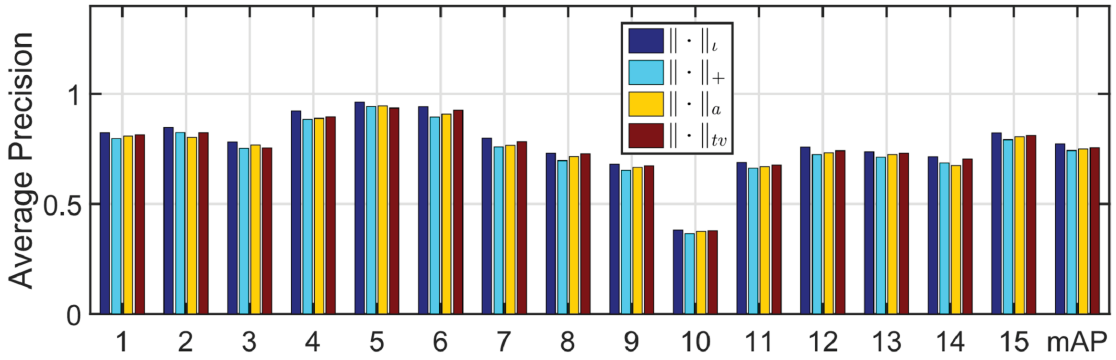
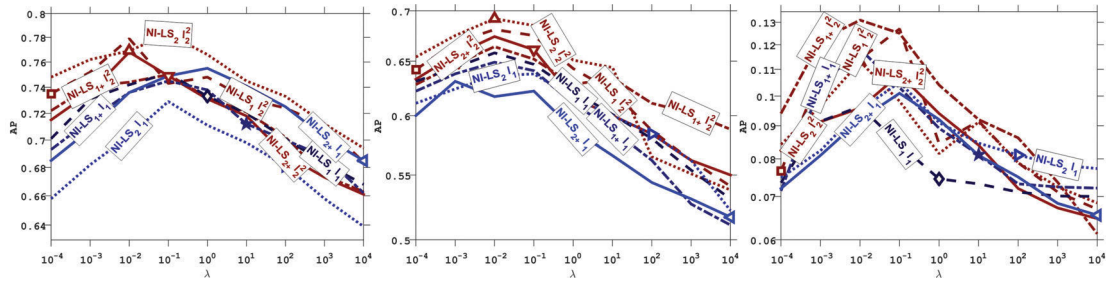


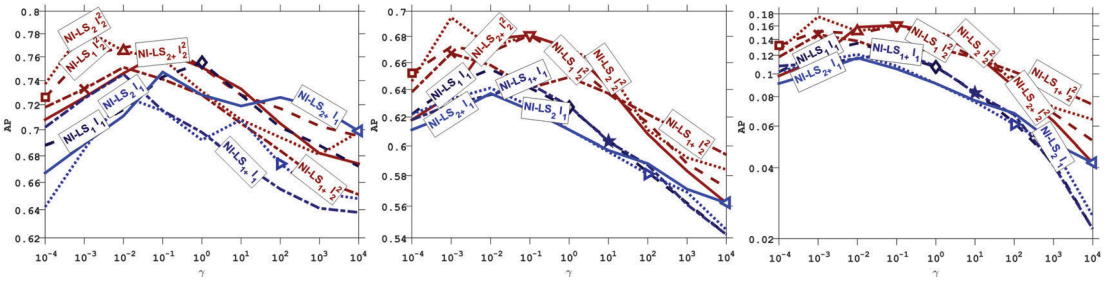
Figure 2.6: Comparing different isotonic regularizers on the MED13 dataset. The x-axis indicates the event ID.



(a) CCV (b) MED13 (c) MED14

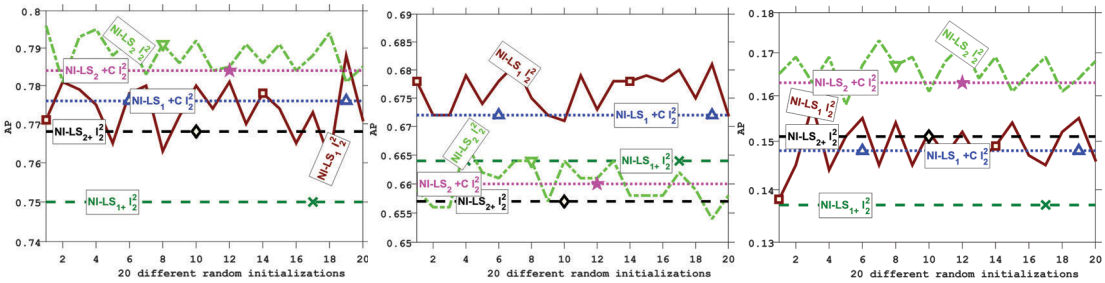
Figure 2.7: Performance sensitivity w.r.t. λ on the CCV_{sub}, MED13, and MED14 datasets.

following reasons: (1) there is no ground truth information provided; (2) there are relatively few previous works that can be compared against. Instead, we compare to a natural baseline as follows. We first train a video-level event classifier, and then apply it to the shots to rank them accordingly. Lastly, the top ranked shots are returned as evidence. Following the NIST pipeline NIST [2013, 2014], we use Amazon Mechanical Turk to invite 10 volunteers for the evaluation purpose. Before evaluation, the volunteers are asked to read the event description in text, and watch five positive videos in the training set. Then, our evaluation system randomly chooses 10 positive videos from the test set, and presents the top evidence shots generated by the baseline and our proposed method. The judges are asked to decide if these evidence shots are relevant (true/false), and which method gives more informative evidence shots (better/similar/worse). To make fair comparison, the judges are not aware which shot is generated by which method during evaluation. Based on the judges' responses, we consider two metrics: 1) average accuracy, which is the percentage of relevant evidence shots; 2) relative performance, which counts judges' preferences of the baseline or the



(a) CCV (b) MED13 (c) MED14

Figure 2.8: Performance sensitivity w.r.t. γ on the CCV_{sub}, MED13, and MED14 datasets.



(a) CCV (b) MED13 (c) MED14

Figure 2.9: We repeat running Algorithm 1 twenty times, each with a different random initialization (except methods with "C" which are initialized using the solution of the convex variant). Here an additional ℓ_2^2 regularizer is used and y-axis measures the mAP.

proposed approach.

Results:

We summarize the average length of the test videos, the average length of evidence shots returned by our proposed approach, and the average accuracy derived from the judge's responses in Table 2.6. The results are quite promising: the proposed approach achieves 91.4% and 85.7% average accuracy by returning only 3.8% and 4.5% evidence shots in the original videos, respectively. This clearly demonstrates that the classifier weights of NI-SVM are reasonable in capturing the relative importance of the individual shots, in a way that is comprehensible to humans. The results also seem to indicate that MED14 is more challenging than MED13.

The judges' preferences between the proposed method and the baseline are averaged and recorded in Table 2.7. It is clear that the proposed method is subjectively better for most events on both datasets.

Table 2.7: Event recounting results on MED13 and MED14. For each event, we randomly pick 10 positive test videos and ask 10 judges to rate *better, similar, or worse* between the proposed method and the baseline. The results among judges are averaged.

	MED13				MED14		
ID	Better	Worse	Similar	ID	Better	Worse	Similar
E006	7	1	2	E021	7	3	0
E007	9	0	1	E022	3	4	3
E008	6	1	3	E023	6	2	2
E009	7	2	1	E024	6	1	3
E010	8	2	0	E025	5	3	2
E011	6	3	1	E026	7	1	2
E012	7	1	2	E027	6	3	1
E013	4	6	0	E028	7	1	2
E014	6	2	2	E029	5	2	3
E015	4	1	5	E030	4	1	5
E021	7	2	1	E031	8	0	2
E022	3	4	3	E032	7	1	2
E023	5	3	2	E033	8	1	1
E024	6	1	3	E034	5	0	5
E025	6	3	1	E035	4	1	5
E026	7	2	1	E036	6	1	3
E027	4	4	2	E037	5	3	2
E028	6	1	3	E038	4	5	1
E029	5	3	2	E039	4	0	6
E030	4	2	4	E040	7	1	2
Total	117	44	39	Total	114	33	52

Table 2.8: Mean F_1 score (mF_1) with ℓ_2^2 regularization on MED14, MED13, and $\mathrm{CCV}_{\mathrm{sub}}$ datasets. A larger mF_1 indicates better performance.

		ℓ_2^2 regularized						
Dataset		$\mathrm{SVM}^{\mathrm{m}}_{\mathrm{A}}$	$\mathrm{SVM}^{\mathrm{m}}_{\mathrm{M}}$	$\mathrm{SVM}^{\mathrm{m}}_{\mathrm{T}}$	$\text{NI-SVM}^{\mathrm{m}}_{1}$	$\text{NI-SVM}^{\mathrm{m}}_{1+}$	$\text{NI-SVM}^{\mathrm{m}}_{2}$	$\text{NI-SVM}^{\mathrm{m}}_{2+}$
MED14	mF_1	0.384	0.346	0.397	0.413	0.403	0.447	0.422
MED13	mF_1	0.485	0.452	0.494	0.526	0.503	0.535	0.521
$\mathrm{CCV}_{\mathrm{sub}}$	mF_1	0.816	0.787	0.824	0.848	0.826	0.861	0.850

2.6.4 Event recognition

In this section we evaluate the multiclass $\text{NI-SVM}^{\mathrm{m}}$ proposed in §2.4 for event recognition.

Evaluation metric:

A widely adopted evaluation metric in the multiclass setting is the F_1 score, which is simply the harmonic mean of the recall (r) and precision (p):

$$\mathsf{F}_1 = \frac{2rp}{r+p}, \tag{2.53}$$

where recall (r) is the fraction of relevant videos retrieved by the system and precision (p) is the fraction of retrieved videos that are relevant. The F_1 scores for different events are averaged.

Table 2.9: Mean F_1 score (mF_1) with ℓ_1 regularization on MED14, MED13, and CCV_{sub} datasets. A larger mF_1 indicates better performance.

		ℓ_1 regularized						
Dataset		$\text{SVM}^{\text{m}}_{\text{A}}$	$\text{SVM}^{\text{m}}_{\text{M}}$	$\text{SVM}^{\text{m}}_{\text{T}}$	$\text{NI-SVM}^{\text{m}}_{1}$	$\text{NI-SVM}^{\text{m}}_{1+}$	$\text{NI-SVM}^{\text{m}}_{2}$	$\text{NI-SVM}^{\text{m}}_{2+}$
MED14	mF_1	0.371	0.358	0.379	0.395	0.387	0.398	0.363
MED13	mF_1	0.468	0.457	0.473	0.501	0.516	0.482	0.458
CCV_{sub}	mF_1	0.798	0.784	0.816	0.823	0.821	0.814	0.826

Competitors:

We compare the following multiclass algorithms:

- $\text{SVM}^{\text{m}}_{\text{A}}$: the multiclass SVM Crammer and Singer [2001] with average-pooling on the video shots.
- $\text{SVM}^{\text{m}}_{\text{M}}$: the multiclass SVM Crammer and Singer [2001] with max-pooling.
- $\text{SVM}^{\text{m}}_{\text{T}}$: the multiclass SVM Crammer and Singer [2001] without pooling, but the shots are prioritized according to their saliency scores.
- $\text{NI-SVM}^{\text{m}}_{1}$: the proposed method with isotonic regularizer (2.10).
- $\text{NI-SVM}^{\text{m}}_{2}$: the proposed method with isotonic regularizer (2.11).
- $\text{NI-SVM}^{\text{m}}_{1+}$: nonnegative convex version of $\text{NI-SVM}^{\text{m}}_{1}$.
- $\text{NI-SVM}^{\text{m}}_{2+}$: nonnegative convex version of $\text{NI-SVM}^{\text{m}}_{2}$.

We can again add additional ℓ_1 or ℓ_2^2 regularizer, without any computational cost.

Note that in the multiclass setting, we only use training videos belonging to some event while recall that in event detection, for each event we use a lot more training videos as negatives, especially those that do not belong to any event. Thus, the results here cannot be directly compared to the ones in §2.6.2.

Results:

The experimental results are tabulated in Tables 2.8 and 2.9, from which we make the following observations: (1) Similar as in event detection, average-pooling consistently performs better than max-pooling on all three datasets; (2) $\text{SVM}^{\text{m}}_{\text{T}}$ further outperforms average-pooling, verifying that pooling can also be detrimental for event recognition, if naively done; (3) The proposed multiclass NI-SVM^{m} variants achieve the best performance on all three datasets, confirming again the benefits of exploiting the semantic ordering information. (4) Additional ℓ_2^2-norm regularization generally outperforms additional ℓ_1-norm regularization, although we found the latter usually leads to much sparser solution (hence may result in significantly reduced test time).

2.7 Summary of This Chapter

Based on the observation that not all video shots are equally relevant to an event of interest, we propose to prioritize the video shots using a novel notion of semantic saliency. Through a suitable isotonic regularizer we design the "informed" nearly-isotonic SVM classifier (nisvm) that is able to exploit the carefully constructed ordering information. An efficient proximal gradient implementation, with new and closed-form proximal steps, is developed. We further extend nisvm to the multi-class setting to perform event recognition. Extensive experiments on three real video datasets are conducted to validate the proposed algorithms on video analysis tasks such as event detection, recognition, and recounting. In the future, we plan to incorporate temporal and spatial information in a more refined notion of saliency. We also plan to explore nisvm in other applications.

Chapter 3

Searching Persuasively: Joint Event Detection and Evidence Recounting with Limited Supervision

In this chapter, we propose a joint framework, illustrated in Figure 3.1, to simultaneously classify high-level events and locate semantic evidences for each complex event. After extracting a semantic, albeit noisy, video representation, we introduce a recounting model based on recent sparse regularizers Chen et al. [2001b]; Rudin et al. [1992]; Yuan and Lin [2006a] that can localize key evidences both concept-wise and temporal-wise, and a detection model based on the infinite push support vector machine (SVM) Rudin [2009] that greatly enhances the discriminative power. In a nutshell, our recounting model assists detection by filtering out noisy irrelevant information while simultaneously our detection model guides recounting by directing it to the most discriminative evidences. We further segment each video into multiple shots to exploit the rich temporal information, although this in turn creates a severe computational challenge, especially on large-scale datasets. Therefore, our second contribution is a significantly improved alternating direction method of multiplier (ADMM) algorithm. In contrast to existing works *e.g*. Lai et al. [2014]; Rakotomamonjy [2012], we prove novel closed-form solutions for all intermediate updates that allow us to remove all inner loops, resulting in tremendously improved efficiency. The per-step complexity of our algorithm scales only *linearly* with the problem size. We test our algorithm on the recent TRECVID MEDTest 2013 dataset NIST [2013] and MEDTest 2014 dataset NIST [2014], and achieve very promising results for both detection and recounting.

We summarize our contributions as follows:

- Unlike most previous works that address detection and recounting separately, we integrate event detection and recounting into a joint optimization framework, allowing us to exploit the mutual benefits of the two tasks, particularly when training

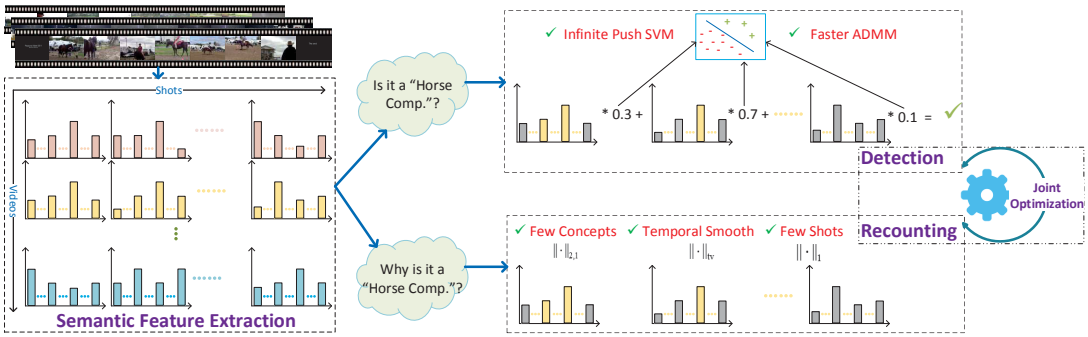


Figure 3.1: The proposed framework simultaneously conducts event detection and evidence recounting. Illustrated on the particular *Horse Competition* event. We first segment each video into multiple shots upon which we extract *semantic* features. Then we iterate between the detection model and the recounting model. We employ the infinite push SVM Rudin [2009] for detection and develop a fast ADMM algorithm for it. Sparse regularizers are used to localize indicative concepts for recounting.

data is scarce.

- We propose a very efficient ADMM algorithm that is key to perform recounting and detection jointly on large scale shot-level video representations.
- We conduct extensive experiments on the recent large-scale MEDTest 13 and MEDTest 14 datasets, and obtain very promising results for both detection and recounting.

Notations: We use $\|\cdot\|_{\mathsf{F}}, \|\cdot\|_1, \|\cdot\|_2$ for the Frobenius norm, 1-norm, and 2-norm, respectively. For matrix A, we use the notation $A_{i,j}, A_{i,:}, A_{:,j}$ for its ij-th entry, i-th row, and j-th column, respectively. We split our m training videos into p positive exemplars $\mathfrak{P}$ and n negative exemplars $\mathfrak{N}$, where of course $m = p + n$. Frequently we will use A_+ to denote the positive part, *i.e.* $\max\{A, 0\}$, understood componentwise for vectors or matrices. The standard inner product $\langle A, B\rangle = \sum_{ij} A_{i,j}B_{i,j}$ is used throughout.

3.1 Joint Detection and Recounting

We start by introducing our semantic representation of the video ensemble, which unfortunately is usually noisy and not directly applicable for the recounting task. Then, we motivate our joint training protocol by separately detailing the recounting model and the detection model. The two models are combined into a joint framework to allow *simultaneous* detection and recounting.

3.1.1 Semantic Concept Representation

Videos, in their raw format, are represented by pixel values of each frame, which are not robust against small variations. Thus, in many video analysis tasks it is a common practice to first extract low level features, such as the improved dense trajectories Wang and Schmid [2013]. To reduce the high dimensionality of low level features and increase their discriminative power, various pooling procedures can be applied on top Ikizler-Cinbis and Sclaroff [2010]; Tamrakar et al. [2012]. However, low level features usually do not have semantic meanings, thus are not suitable for interpretation purposes, such as the recounting task we consider in this work. In recent years, semantic representations based on *concepts, attributes,* and *actions* have been popular in video event detection, recognition and recounting Habibian et al. [2013]; Liu et al. [2013b]; Mazloom et al. [2013b]; Merler et al. [2012a]; Sun and Nevatia [2014]; Tan et al. [2011b]. Usually, these semantic representations are trained on top of low level features.

In this work we represent each video by its confidence scores on c pre-defined concept classes $\mathcal{C} = \{C_1, C_2, \ldots, C_c\}$. The construction of these concepts is detailed in Section 5.2 below. In order to identify the key *temporal* evidences, we further split each video into s shots/clips. Here, for simplicity, we assume each video has the same number of shots, but the algorithm can be easily extended to the heterogeneous setting. Each shot (in the i-th video) is encoded by a confidence score vector $\mathbf{v}_t \in [0,1]^c, t = 1, \ldots, s$, where for each $k = 1, \ldots, c$, $(\mathbf{v}_t)_k$ indicates the confidence of the k-th concept being present in the t-th shot. Thus, the i-th video is encoded as the matrix $V^i = [\mathbf{v}_1, \ldots, \mathbf{v}_s] \in [0,1]^{c \times s}$, and we use $V = [V^1, \ldots, V^m] \in [0,1]^{c \times sm}$ to represent the entire video ensemble. Here m is the total number of videos. Building on this semantic representation, our goal is to decide whether or not each video V^i belongs to a certain event, and also identify the key evidences to support our classification results, i.e., based on what concepts appeared in which shots our detection algorithm made its judgment.

3.1.2 The Joint Training Protocol

Motivations: Our joint training protocol for simultaneous event detection and recounting is motivated by the following observations: 1). The concept detectors we use are far from being perfect, partly because of the relatively small number of positive exemplars and partly because of the cross-domain training. Therefore, there is a considerable amount of noise in our semantic representation, hence it is only reasonable to treat each semantic representation V^i as a noisy perturbation of some ground truth R^i. 2). Usually, only a few concepts are relevant for detecting a certain event Bhattacharya et al. [2014]. For instance, in detecting the birthday party event, we would expect concepts like "birthday cake" or "blowing candle" to be highly

discriminative while others would be less useful or even misleading (worsened by the imperfections of our semantic representation). Thus it makes sense to find a clean and discriminative representation R^i that has many zero rows, *i.e*. containing only few relevant concepts. Moreover, since each relevant concept likely appears only in very few shots and in a temporally smooth way, we expect each row of R^i to contain only few *consecutive* nonzero entries. 3). Event recounting is different from video summarization (*e.g*. Das et al. [2013]; Gupta et al. [2009b]), in the sense that not all concepts appearing in our video are equally useful. In particular, event recounting is detection oriented: We are only interested in identifying the few concepts, along with their temporal positions, that are highly discriminative for our detection module. In other words, event recounting acts more like a recurrent supervised learning task: It aims at both explaining and serving the detection module. Ideally, we would like our "clean" representation R^i to be highly discriminative.

Our recounting model and detection model take the above observations into account. Specifically, our recounting model "denoises" the semantic representation so that it only contains few concepts appearing in few shots in a temporally smooth manner. This denoised representation localizes key evidences and is fed into the infinite push support vector machine (SVM) Agarwal [2011]; Rakotomamonjy [2012]; Rudin [2009] to gain more discriminative power. Unlike previous recounting works Tsai et al. [2014], we conduct event detection and event recounting *simultaneously* in a joint framework.

Recounting Model: Our recounting model finds the latent "clean" representation (*i.e*., ground truth) $R = [R^1, \ldots, R^m]$ by solving the following denoising problem:

$$\min_R \tfrac{1}{2}\|V - R\|_{\mathsf{F}}^2 + \Omega(R), \tag{3.1}$$

where the regularizer Ω encodes our observation that only few concepts are relevant and they appear in few shots in a temporally smooth manner. Specifically, we use

$$\Omega(R) = \alpha\|R\|_{2,1} + \beta\|R\|_1 + \gamma\sum_{i=1}^{m}\|R^i\|_{\mathsf{tv}}, \tag{3.2}$$

where the group norm Yuan and Lin [2006a]

$$\|R\|_{2,1} = \sum_{k=1}^{c}\|R_{k,:}\|_2 \tag{3.3}$$

encourages many rows of R to be zeroed out (*i.e*. few concepts are relevant); the total variation semi-norm Rudin et al. [1992]

$$\|R^i\|_{\mathsf{tv}} = \sum_{k=1}^{c}\sum_{t=2}^{s}|R^i_{k,t} - R^i_{k,t-1}| \tag{3.4}$$

encourages piecewise constant entries (i.e. temporally smooth) in each row of R^i; and finally the 1-norm $\|R\|_1$ encourages sparsity Chen et al. [2001b] (i.e., relevant concepts appear only in few shots). By inspecting the support patterns (*i.e.* nonzero entries) of the "clean" representation R we can localize the key evidences hence perform recounting. However, so far the recounting model (3.1) is purely unsupervised hence may not be helpful for event detection. To enhance its discriminative power, we will couple it with our supervised detection model below. Note that we detect each event separately (as required in the NIST standard NIST [2013, 2014]). Therefore, each event will have its own clean representation R, and may choose different indicative concepts. To our best knowledge, the formulation of our recounting model is new.

Detection Model: Our detection model follows the usual approach. It finds a discriminative linear classifier, parameterized by an appropriate matrix W, to distinguish the positive and negative "clean" representations R. From now on we split the m training videos into positive exemplars $\mathfrak{P}$ and negative exemplars $\mathfrak{N}$, with size respectively p and n. For each negative exemplar $j \in \mathfrak{N}$, the quantity $\sum_{i\in\mathfrak{P}} \mathbb{1}(\langle W, R^i\rangle < \langle W, R^j\rangle)$ counts the number of positive exemplars i that are ranked below j by the linear classifier W. These ranking errors with respect to each negative exemplar are combined to yield a loss that we aim to minimize. For computational tractability we upper bound the discrete 0-1 loss $\mathbb{1}(\delta < 0)$ by the *convex* hinge loss $(1-\delta)_+$, where as usual $(\delta)_+ := \max\{\delta, 0\}$ is the positive part. Since we usually pay more attention, if not exclusively, to the *top* of the rank list, we focus on minimizing the maximum ranking error among all negative exemplars $j \in \mathfrak{N}$:

$$\ell(W; R) := \max_{j\in\mathfrak{N}} \frac{1}{p} \sum_{i\in\mathfrak{P}} (1 - \langle W, R^i - R^j\rangle)_+. \tag{3.5}$$

Adding an appropriate regularizer Φ to control the model complexity, we obtain the infinite push support vector machine Agarwal [2011]; Rakotomamonjy [2012]; Rudin [2009]:

$$\min_W \ \ell(W; R) + \Phi(W). \tag{3.6}$$

Usual choices for Φ include the (squared) 2-norm $\Phi(W) = \lambda\|W\|_{\mathsf{F}}^2$ Agarwal [2011] and the sparse 1-norm $\Phi(W) = \lambda\|W\|_1$ Rakotomamonjy [2012]. Since our "clean" representation R is already sought to be sparse, we will mostly use the 2-norm in our experiments, due to its better performance. If desired, loss functions other than the worst ranking error ℓ in Equation (3.5) can also be used.

Joint Framework: So far the recounting model and the detection model are separate hence not helping each other. To integrate them, we propose the following joint optimization framework:

$$\min_{W,R} \ \ell(W; R) + \tfrac{1}{2}\|V - R\|_{\mathsf{F}}^2 + \Phi(W) + \Omega(R), \tag{3.7}$$

where ℓ and Ω are given respectively in (3.5) and (3.2) above. By coupling the two models jointly in the ranking loss ℓ, we expect to exploit the mutual benefits of the two tasks. Indeed, the detection model works on the clean representation R that is supplied by the recounting model, avoiding any noisy or irrelevant concepts. Conversely, the detection model also directs the recounting model to find discriminative evidences that are tailored for detection. More conveniently, the joint optimization problem (3.7) is bi-convex, meaning that fixing either one of W and R results in a convex problem that is immune to local minima. Thus we will use a coordinate descent algorithm to learn W and R, one at a time and iteratively. However, standard convex optimization toolboxes cannot be naively applied here for two reasons: 1). The training ensemble has a large size $c \times s \times m$, putting a stringent time complexity on the numerical algorithm (preferably linear-time); 2). The regularizer Ω is a highly non-smooth function, leading to extremely slow convergence if not properly handled. One obvious remedy for the first issue is to apply a pooling procedure to aggregate the s shot-level information, but it leads to a significant loss of temporal information. Instead, we circumvent the computational difficulties by developing a significantly improved alternating direction method of multipliers (ADMM) algorithm in the next section.

3.2 The Optimization Scheme

In this section we develop a much improved linear time ADMM algorithm for our joint model (3.7). First, we rewrite the ranking loss ℓ by introducing an auxiliary variable A for decoupling:

$$\min_{W,R,A} \quad g(A) + \tfrac{1}{2}\|V - R\|_{\mathsf{F}}^2 + \Phi(W) + \Omega(R), \tag{3.8}$$

$$\text{s.t.} \quad \forall i \in \mathfrak{P}, j \in \mathfrak{N},\ A_{i,j} = 1 - \langle W, R^i - R^j \rangle, \tag{3.9}$$

where $g(A) = \frac{1}{p} \max_{j \in \mathfrak{N}} \sum_{i \in \mathfrak{P}} (A_{i,j})_+$. Next we introduce the Lagrangian multiplier matrix Γ and a quadratic penalty term to eliminate the linear constraint (3.9):

$$\begin{aligned} \min_{W,R,A} \ & \langle \Gamma, A + \mathcal{R}(W) - E \rangle + \tfrac{1}{2}\|A + \mathcal{R}(W) - E\|_{\mathsf{F}}^2 \\ & + \tfrac{1}{2}\|V - R\|_{\mathsf{F}}^2 + \Phi(W) + \Omega(R) + g(A), \end{aligned} \tag{3.10}$$

where E is the all 1's matrix and $\mathcal{R}(W) = \mathcal{W}(R)$ is the matrix whose ij-th entry is $\langle W, R^i - R^j \rangle$. We use $\mathcal{R}(W)$ to emphasize the linearity in W when R is fixed and use $\mathcal{W}(R)$ to highlight the linearity in R when W is fixed.

The optimization variables in (3.10) are now uncoupled, and the usual ADMM algorithm iteratively solves them one at a time. At a high level, we switch iteratively between the detection model (parameterized by W, A) and the recounting model

(parameterized by R), until they collaboratively reach a consensus. Interestingly, the subproblems w.r.t. W, A and w.r.t. R are completely analogous, thus largely simplifying the subsequent developments. We discuss each subproblem in more details below, and address some severe computational challenges there.

3.2.1 Optimizing W while fixing R

In this step we fix R as a constant and solve for W and A. The updates from the vanilla ADMM algorithm are as follows:

$$W \leftarrow \arg\min_{W} \tfrac{1}{2}\|A + \mathcal{R}(W) - E + \Gamma\|_{\mathsf{F}}^2 + \Phi(W) \tag{3.11}$$

$$A \leftarrow \arg\min_{A} \tfrac{1}{2}\|A + \mathcal{R}(W) - E + \Gamma\|_{\mathsf{F}}^2 + g(A) \tag{3.12}$$

$$\Gamma \leftarrow A + \mathcal{R}(W) - E + \Gamma. \tag{3.13}$$

Unfortunately, executing the above steps is not easy (except for the Lagrangian multiplier Γ). Previous work Rakotomamonjy [2012] proposed two dedicated iterative subroutines for the W-step (3.11) and the A-step (3.12), respectively. However, the whole algorithm requires three nested loops, significantly slowing down the convergence. Moreover, the numerical errors in the inner loops may also accumulate and affect the outer loop. Essentially the same algorithm (hence same drawback) was adopted in Lai et al. [2014]. Here we eliminate all inner loops based on two novel ideas.

Firstly, the difficulty in the W-step (3.11) is mostly due to the (non-diagonal) quadratic term induced by the linear map $\mathcal{R}(W)$ (recall that R is fixed). Fortunately, we can bypass this difficulty by linearizing the quadratic term at the current iterate W, which is known as the inexact Uzawa step in the ADMM literature He and Yuan [2012]. Namely, we replace (3.11) with two steps:

$$\tilde{W} \leftarrow W - \tfrac{1}{\mu}\mathcal{R}^\top[A + \mathcal{R}(W) - E + \Gamma] \tag{3.14}$$

$$W \leftarrow \mathsf{P}_{\Phi}^{\eta}(\tilde{W}), \tag{3.15}$$

where the first step is a simple gradient update with step size $1/\mu$ while the second step is a proximal update with respect to the regularizer Φ. The latter, usually referred to as the proximal map Yu [2013], is defined for any function f with parameter μ as

$$\mathsf{P}_{f}^{\eta}(\mathbf{x}) = \arg\min_{\mathbf{z}} \tfrac{\mu}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + f(\mathbf{z}). \tag{3.16}$$

It is a strict and natural generalization of the usual Euclidean projection operator (where f is the indicator function of some constraint set). For "simple" functions,

the proximal map admits a closed-form solution, for instance,

$$\mathsf{P}^{\eta}_{\Phi}(W) = \begin{cases} \frac{\mu}{\lambda+\mu} W, & \text{if } \Phi = \frac{\lambda}{2}\|\cdot\|^2_{\mathsf{F}} \\ \operatorname{sign}(W) * (|W| - \frac{\lambda}{\mu})_+, & \text{if } \Phi = \lambda\|\cdot\|_1 \end{cases}, \tag{3.17}$$

where the algebraic operations are componentwise. Thus, the W-step (3.14) and (3.15) can now be performed in linear time, without the need of any nested loop at all.

Secondly, using the proximal notation in (3.16) we note that the A-step in (3.12) is simply the proximal update $\mathsf{P}^1_g(-\mathcal{R}(W) + E - \Gamma)$. Surprisingly, we prove here that this proximal update in fact admits a closed-form solution, which, to our best knowledge, has not been derived previously. Our result is based on the following new theorem, where we recall that $\mathbf{z}_+$ denotes the positive part, i.e., $(\mathbf{z}_+)_i = \max\{z_i, 0\}$.

Theorem 3.1. *If the function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ satisfies*

$$\forall i,\ \forall \mathbf{z},\ \ f(\mathbf{z}) \geq f(z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_d), \tag{3.18}$$

then for all $\mu > 0$,

$$\mathbf{w} - \mathbf{w}_+ + \mathbf{u} \in \arg\min_{\mathbf{z}} \tfrac{\mu}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 + f(\mathbf{z}_+) \tag{3.19}$$

$$\textit{if and only if}\quad \mathbf{u} \in \arg\min_{\mathbf{z}} \tfrac{\mu}{2}\|\mathbf{w}_+ - \mathbf{z}\|_2^2 + f(\mathbf{z}). \tag{3.20}$$

Note that Theorem 3.1 does not even require the function f to be convex: The condition (3.18) basically says that on each coordinate i, f attains its minimum at 0. Many familiar functions, for instance p-norms for all $p \geq 0$, satisfy this condition. The significance of the established equivalence in Theorem 3.1 is that it allows us to swap the positive part (a nonlinear operation) from an *optimization variable* $\mathbf{z}$ in (3.19) to a *fixed input variable* $\mathbf{w}$ in (3.20). This simple swapping can lead to enormous computational savings, especially when f is "simple". Indeed, for the A-step in (3.12), $g = f(A_+)$, where $f(A) = \frac{1}{p}\|A\|_{\infty,1}$ is the max norm of the 1-norms of each column of A. Since the proximal map $\mathsf{P}^{\eta}_{\|\cdot\|_{\infty,1}}$ has been derived in (nearly) closed-form in *e.g.* Quattoni et al. [2009], applying Theorem 3.1 we immediately have a closed-form for the A-step (3.12):

$$A \leftarrow \mathsf{P}^1_g(\tilde{A}) = \tilde{A} - \tilde{A}_+ + \mathsf{P}^p_{\|\cdot\|_{\infty,1}}(\tilde{A}_+), \tag{3.21}$$

where $\tilde{A} = -\mathcal{R}(W) + E - \Gamma$. The computational complexity is linear after sorting Quattoni et al. [2009], without any nested loop at all. In §3.3.1 below we empirically verify that our closed-form solution (3.21) significantly improves the nested loops in Lai et al. [2014]; Rakotomamonjy [2012]. We expect Theorem 3.1 to be useful for other applications involving the infinite push ranking loss.

To summarize, for fixed R we can perform the update w.r.t. W and A in linear time in a single step. The same idea can be recycled to update R, which we discuss next.

3.2.2 Optimizing R while fixing W

This step is analogous to the previous step where W is optimized. For brevity we only mention the few differences. Again we linearize the quadratic term induced by $\mathcal{W}(R) = \mathcal{R}(W)$ at the current iterate R, and we perform the following updates (for some step size $\mu > 1$):

$$\tilde{R} \leftarrow R - \tfrac{1}{\mu}\mathcal{W}^\top[A + \mathcal{W}(R) - E + \Gamma] + \tfrac{1}{\mu}(V - R) \tag{3.22}$$

$$R \leftarrow \mathsf{P}^{\eta}_{\Omega}(\tilde{R}) \tag{3.23}$$

$$A \leftarrow \arg\min_{A} \tfrac{1}{2}\|A + \mathcal{W}(R) - E + \Gamma\|_{\mathsf{F}}^2 + g(A) \tag{3.24}$$

$$\Gamma \leftarrow A + \mathcal{W}(R) - E + \Gamma, \tag{3.25}$$

where the first step is a simple gradient update; the third A-step is solved using (3.21) as before; the fourth Γ-step is also straightforward (standard matrix operations). The second proximal update w.r.t. the regularizer Ω in our recounting model is more involved, and has not been considered in previous work. Fortunately, we prove that it can still be easily computed by decomposing the individual regularizers as follows:

Theorem 3.2. $\mathsf{P}^{\eta}_{\Omega}(R) = \mathsf{P}^{\eta}_{\alpha\|\cdot\|_{2,1}}\Big(\mathsf{P}^{\eta}_{\beta\|\cdot\|_1}\big(\mathsf{P}^{\eta}_{\gamma\|\cdot\|_{\mathsf{tv}}}(R)\big)\Big)$.

Very crucially, all three proximal maps on the right-hand side of Theorem 3.2 have known exact linear time algorithms. By composing them we immediately obtain the proximal map for the sum regularizer Ω, completing the R-step in (3.23).

To summarize, for fixed W we can again update w.r.t. R in linear time in a single step, without any nested loop at all.

3.2.3 Combining the W and R steps

For fixed R, iterating (3.14), (3.15), (3.21), (3.13) by at most $O(\frac{1}{\epsilon})$ steps yields an ϵ-optimal solution for W He and Yuan [2012]. Similarly, for fixed W, iterating (3.22), (3.23), (3.24), (3.25) by at most $O(\frac{1}{\epsilon})$ steps yields an ϵ-optimal solution for R. The two procedures can be alternated until convergence, however, we notice that they share the same A-step and Γ-step. Therefore it seems reasonable to combine the two procedures into one. Another way of thinking about the combination is that since we are alternating the two procedures it would be wasteful to wait until each procedure completely converges. Instead, we simply run each procedure for a *single* iteration and then switch to another immediately. As we observed in the experiments, this "eager" switching strategy works very well.

We summarize the combined and improved ADMM algorithm for solving our joint training protocol (3.7) in Algorithm 2, and we wish to point out that Algorithm 2 is very general and accommodates various regularizers Φ (*e.g.* those in (3.17)).

Algorithm 2: Faster ADMM for solving JEDaR (3.7)

Input: $V, \alpha, \beta, \gamma, p, \Phi$. Initialize: $W, R, A, \Gamma, \mu > 1$.
repeat
  $W \leftarrow W - \frac{1}{\mu}\mathcal{R}^\top[A + \mathcal{R}(W) - E + \Gamma]$;
  $W \leftarrow \mathsf{P}^{\eta}_{\Phi}(W)$; `// Equation (3.17)`
  $A \leftarrow -\mathcal{R}(W) + E - \Gamma$;
  $A \leftarrow A - A_+ + \mathsf{P}^{p}_{\|\cdot\|_{\infty,1}}(A_+)$;
  $\Gamma \leftarrow A + \mathcal{R}(W) - E + \Gamma$;
  $R \leftarrow R - \frac{1}{\mu}\mathcal{W}^\top[A + \mathcal{W}(R) - E + \Gamma] + \frac{1}{\mu}(V - R)$;
  $R \leftarrow \mathsf{P}^{\eta}_{\alpha\|\cdot\|_{2,1}}\Big(\mathsf{P}^{\eta}_{\beta\|\cdot\|_1}\big(\mathsf{P}^{\eta}_{\gamma\|\cdot\|_{\text{tv}}}(R)\big)\Big)$;
until *convergence*;

Computationally, Algorithm 2 is very appealing since each iteration incurs only a cost linear in the size of the training data. To appreciate the efficiency of our algorithm, we compared it with the full algorithm in Rakotomamonjy [2012] (whose model requires $\alpha = \beta = \gamma = 0$ and $\Phi = \lambda\|\cdot\|_1$). In general 400x speedups were achieved even on moderate problem sizes (*e.g.* $m = 800, cs = 100$). This tremendous speedup is the key for us to train our joint framework on large real multimedia datasets.

3.3 Experiments

In this section we conduct thorough experimental evaluations of the proposed **J**oint **E**vent **D**etection **a**nd **R**ecounting framework, abbreviated as **JEDaR**.

Datasets: We test on two real-world datasets: the TRECVID MEDTest 2013 NIST [2013] and the TRECVID MEDTest 2014 NIST [2014]. Both are collected by the NIST for the TRECVID competition. Each dataset consists of 20 complex events with 10 events in common. Specifically, the MEDTest 2013 dataset has events E006 to E015 and E021 to E030, while the MEDTest 2014 contains events E021 to E040. These events include *changing a vehicle tire*, *grooming an animal*, *etc*. Please refer to NIST [2013, 2014] for the complete list of event names.

Setup: For all our experiments we strictly follow the ***10Ex evaluation procedure*** outlined by the NIST TRECVID event kit. According to the rules, we ***detect each event separately***, totaling 20 individual detection tasks for each dataset. For each event, we have ***10*** positive videos from the event kit training data, along with approximately 5,000 negative videos from the background training data. The testing data has about 23,000 videos. We report both event detection and recounting results for each event.

Competitors: We compare our method JEDaR against current state-of-the-art alternatives. Support Vector Machine (SVM) and Ridge Regression (RR) are the

most widely used classifiers in the TRECVID MED 2014 competition among the top ranked teams and recent research reports. Therefore, the two algorithms are used as the baseline. We also compare to more recent state-of-the-art alternative methods, including dynamic pooling with segment-pairs (DPSP) Li et al. [2013], VD-HMM Tang et al. [2012], and ELM Sun and Nevatia [2014]. All comparisons follow the same rules of the official TRECVID MEDTest 2013 and MEDTest 2014 data splits.

Concept detectors: 3,135 concept detectors are pre-trained using the TRECVID SIN dataset (346 categories), google sports (478 categories) Jiang et al. [2014a]; Karpathy et al. [2014], ucf101 dataset (101 categories) Jiang et al. [2014a]; Soomro et al. [2012], YFCC dataset (609 categories) Jiang et al. [2014a]; Thomee et al. [2015] and DIY dataset (1,601 categories) Jiang et al. [2014a]; Yu et al. [2014a]. The improved dense trajectory features are extracted with the code provided in Wang and Schmid [2013] and encoded them with the fisher vector representation Perronnin et al. [2010]. Then, on top of the extracted low-level features the cascade SVM Graf et al. [2004] was trained for each concept detector. We further split each video in the TRECVID MEDTest 2013 and 2014 datasets into s shots, using the color histogram difference as an indication of the boundary. We applied the pre-trained concept detectors to each shot and obtained a 3135-dimensional representation. The average accuracy of the concept detectors is relatively low due to the small number of positive training examples, therefore justifying our recounting model (3.1) which tries to *denoise* the noisy concept representations.

Parameter Tuning: Our Algorithm 2 has a few parameters and we tune them as follows. We use the recounting model (3.1) alone, without coupling with the detection model (3.2), to first tune the regularization constant α so that roughly $5 \sim 10$ rows of the semantic representation R are nonzero. Then we further tune β so that each row contains roughly 10 nonzero entries. These two parameters are fixed in our subsequent experiments. Next we cross-validate the parameter λ from the range $\{0.01, 0.1, 1, 10, 100\}$, and similarly for the parameter γ from the same range. We have conducted a sensitivity analysis and found the results to be relatively robust once the parameters are in a reasonably large region. The parameters for other competing algorithms are set according to their respective implementations.

3.3.1 Efficiency of our closed-form solution

We first demonstrate the efficiency of our ADMM Algorithm 2, where the key is Theorem 3.1 that provides a closed-form solution in (3.21) for the A-step in (3.12). Previous work Lai et al. [2014]; Rakotomamonjy [2012] dedicated a nested loop iterative subroutine for this step. We randomly generated the input (fixed) square matrix W with varying sizes, and compared the objective values in the A-step (3.12) (the smaller the better). In this section we set $\alpha = \beta = \gamma = 0$ since this is the setting that Lai et al. [2014]; Rakotomamonjy [2012] can handle, although our algorithm

Table 3.1: Experimental comparisons for 10Ex event detection on TRECVID MEDTest 2013. Mean average precision (mAP) is used as the evaluation metric. Results are presented in percentages. Larger mAP indicates better performance. From the results we observe that the proposed algorithm outperforms the state-of-the-art competitors, indicating the superiority of the proposed JEDaR.

	MEDTest 2013					
Event ID	SVM	RR	DPSP	VDHMM	ELM	JEDaR
E006	26.54	27.68	31.29	35.74	38.25	44.66
E007	38.75	39.85	44.48	50.86	54.39	57.63
E008	47.29	48.78	54.82	61.62	65.17	64.22
E009	37.76	36.82	38.75	41.88	45.54	47.94
E010	17.12	17.47	20.64	23.56	26.38	28.49
E011	8.88	9.53	11.28	13.34	14.29	16.77
E012	25.34	25.57	30.66	34.82	38.14	39.38
E013	56.82	57.24	62.53	67.44	66.58	68.25
E014	37.76	38.11	42.48	47.26	47.41	52.33
E015	17.68	18.72	22.69	27.85	31.82	35.34
E021	11.96	12.58	11.24	15.57	14.75	19.53
E022	2.76	2.87	2.72	4.84	7.54	8.77
E023	27.47	27.72	30.75	38.32	41.29	46.26
E024	3.17	2.98	3.12	2.86	3.16	3.98
E025	0.81	0.92	0.87	1.05	1.12	1.27
E026	4.16	4.48	4.85	5.42	5.88	6.23
E027	10.58	10.29	12.56	11.25	13.96	15.62
E028	16.91	18.64	20.82	22.43	25.25	27.41
E029	11.78	12.47	14.28	15.12	17.84	19.63
E030	10.96	11.26	13.65	14.88	15.92	15.26
mean	20.73	21.19	23.72	26.81	28.73	30.95

efficiently extends to any α, β, γ, thanks to the new Theorem 3.2. We used the default setting in Rakotomamonjy [2012]: The maximum number of outer and inner iterations are 200 and 50, respectively. Figure 3.2 confirms the huge computational saving we enjoy thanks to Theorem 3.1. As can be seen from Figure 3.2, the running time of Rakotomamonjy [2012] increased sharply with the input size, but still failed to converge to the optimum. In comparison, our closed-form solution (3.21) only took a negligible time and returned a much smaller (in fact optimal) objective value. For the full ADMM algorithm, the difference is even larger, since we chose to linearize the W-step while Lai et al. [2014]; Rakotomamonjy [2012] used another nested loop. This tremendous speedup is the key to train our joint framework on the large TRECVID MEDTest datasets, especially when we segment each video into multiple shots to exploit temporal information and to perform recounting (see below).

3.3.2 Event Detection Result

As the common practice, we evaluate the event detection performance using the (mean) average precision. The results of our method JEDaR and the competitors (quoted from the respective papers) on both the MEDTest 2013 and MEDTest 2014 datasets are

Table 3.2: Experimental comparisons for 10Ex event detection on TRECVID MEDTest 2014 datasets. Mean average precision (mAP) is used as the evaluation metric. Results are presented in percentages. Larger mAP indicates better performance. From the results we observe that the proposed algorithm outperforms the state-of-the-art competitors, indicating the superiority of the proposed JEDaR.

	MEDTest 2014					
Event ID	SVM	RR	DPSP	VDHMM	ELM	JEDaR
E021	11.96	12.58	11.24	15.57	14.75	19.53
E022	2.76	2.87	2.72	4.84	7.54	8.77
E023	27.47	27.72	30.75	38.32	41.29	46.26
E024	3.17	2.98	3.12	2.86	3.16	3.98
E025	0.81	0.92	0.87	1.05	1.12	1.27
E026	4.16	4.48	4.85	5.42	5.88	6.23
E027	10.58	10.29	12.56	11.25	13.96	15.62
E028	16.91	18.64	20.82	22.43	25.25	27.41
E029	11.78	12.47	14.28	15.12	17.84	19.63
E030	10.96	11.26	13.65	14.88	15.92	15.26
E031	55.98	56.74	61.82	64.24	67.85	69.41
E032	22.89	23.57	24.64	26.48	27.43	28.28
E033	38.74	39.85	41.47	44.28	46.54	46.27
E034	25.88	26.59	27.83	28.74	29.78	31.63
E035	34.85	35.28	38.86	40.63	42.86	45.32
E036	13.28	13.75	15.28	16.45	16.52	19.27
E037	33.15	33.74	35.87	36.12	47.81	49.25
E038	0.92	0.85	0.78	0.91	1.14	1.43
E039	26.94	27.36	31.43	23.22	28.72	30.58
E040	13.25	13.67	16.54	21.41	15.83	18.62
mean	18.32	18.78	20.47	21.71	23.56	25.21

recorded in Table 3.1 and Table 3.2. It is clear that the proposed method JEDaR performs the best on both datasets. Comparing against the SVM baseline, we see that JEDaR significantly improves the detection performance for nearly all events, with mAP of 30.95% vs SVM's 20.73% on the MEDTest 2013 dataset and mAP of 25.21% vs 18.32% on the MEDTest 2014 dataset. We attribute the improvement to our joint training framework that integrates recounting with detection. Thanks to our recounting model, many irrelevant noisy concepts are pruned away, which in turn largely improves the detection performance.

Next we discuss the state-of-the-art alternatives that we were able to compare against. From Table 3.1 and Table 3.2, we have the following observations: 1). RR performs similar to SVM for most events, achieving for instance mAP of 18.78% vs 18.32% on the MEDTest 2014 dataset. This is in accordance with past experiences of several research groups in the TRECVID MED competition. 2). DPSP Li et al. [2013] improves RR (and SVM), with mAP 20.47% vs 18.78% on the MEDTest 2014 dataset. This is probably because DPSP identifies the segments that are most informative for detecting a given event and also dynamically determines the pooling operator most suited for each sequence. 3). VD-HMM Tang et al. [2012] further improves the performance by discovering and assigning sequences of states that

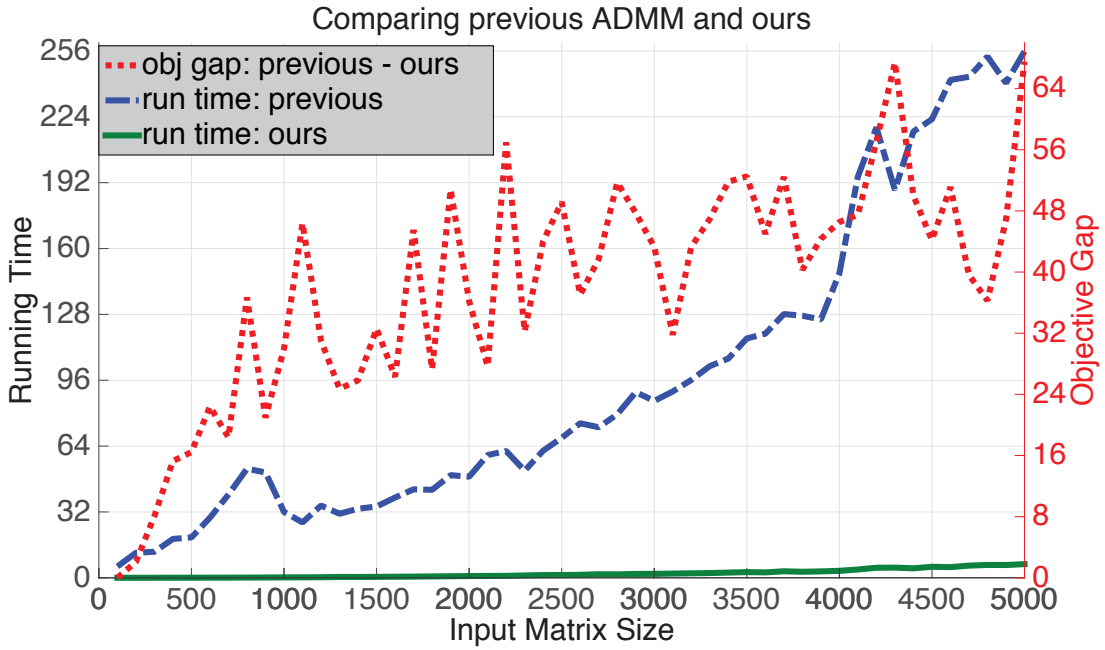


Figure 3.2: Comparison between our closed-form solution (3.21) and the previous nested loop iterative subroutine in Rakotomamonjy [2012]. Even after spending significantly more time (blue dashed vs. solid green), the latter still has not converged yet (red dot).

are most discriminative for a given event. Note that none of the above methods considered detection and recounting in a joint framework. 4). Similar to our method, ELM Sun and Nevatia [2014] also considers detection and recounting jointly, and it achieves the second best performance, *e.g.*, with mAP 23.56% on the MEDTest 2014. However, it explicitly models the evidence locations which involve a large number of latent variables, making learning less efficient especially when training data is scarce. Overall, the advantage of DPSP, VD-HMM and ELM over RR and SVM suggests that finding important segments generally leads to better performance. 5). Lastly, the proposed method JEDaR achieves the best performance on both datasets. This confirms the benefits of our joint training framework, in which recounting improves detection by pruning irrelevant noisy concepts while detection directs recounting to the most discriminative evidences.

3.3.3 Event Recounting Result

In the TRECVID Multimedia Event Recounting (MER) task, a video description is defined as a video shot with a starting frame, an ending frame and a textual description.

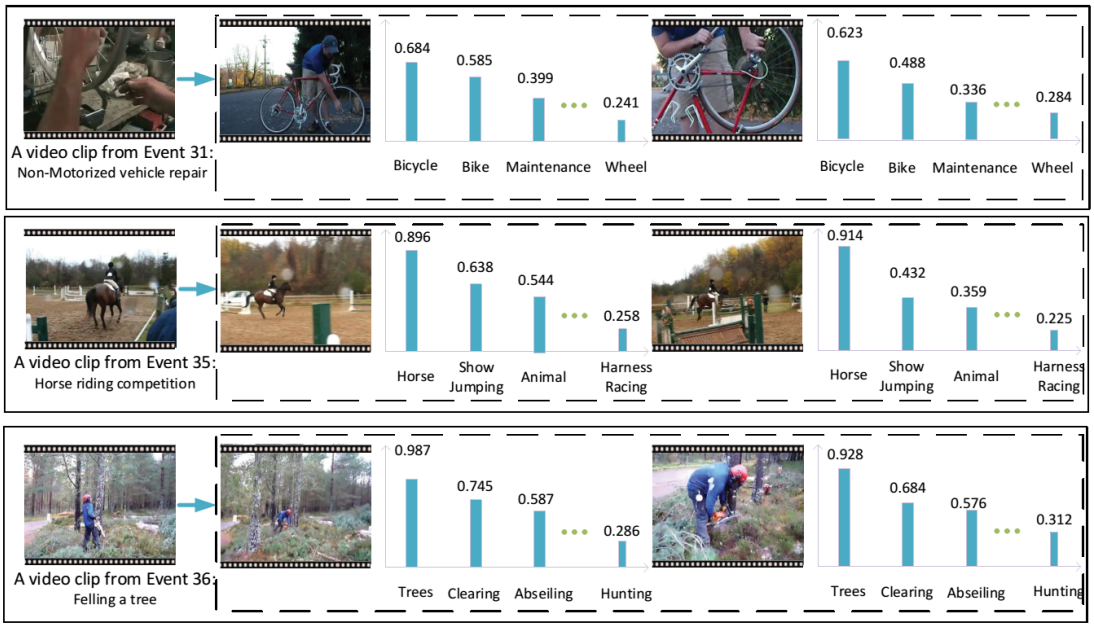


Figure 3.3: Example recounting results generated by the proposed method on the TRECVID MEDTest 2014 dataset. The video events are *non-motorized vehicle repair*, *horse riding competition* and *felling a tree*. The first two relevant shots are displayed.

We use the proposed method to obtain such results as follows: We average the SVM weight W along the shot dimension and pick few top concepts for each event. These concepts are highly discriminative. Then, for each shot of the test video, we examine its top concept scores and decide if it is useful for recounting. The evaluation of event recounting is not easy, since a). there is no ground truth information, and b). few previous works can be compared with. Here we compare to the natural baseline which conducts detection and recounting in separate stages. Following the NIST evaluation pipeline NIST [2013, 2014], we invited 10 volunteers to serve as judges. Before evaluation, we showed each judge the event category descriptions in text, in addition to five positive videos in the training set. Then, we randomly picked 10 positive videos from the test set, and presented informative shots generated by the baseline and the proposed method. The judges were asked to determine which shots are more discriminative. To make a fair comparison, we do not tell the judges which shot is generated by which method during evaluation. Two evaluation metrics are employed: 1). average accuracy, which measures the percentage of correctly labeled shots; and 2). relative performance, which counts judges' preferences of the baseline or the proposed method.

Table 3.3 summarizes the average length of the test videos, the average length of shots generated by the proposed method, and the average accuracy derived from

Table 3.3: Average video length, average shot length generated by the proposed method, and average accuracy derived from the judges' labels.

	MEDTest 2013	MEDTest 2014
Average Video Length	164.3 seconds	188.4 seconds
Average Shot Length	6.4 seconds	8.6 seconds
Average Accuracy	89.7 %	83.2 %

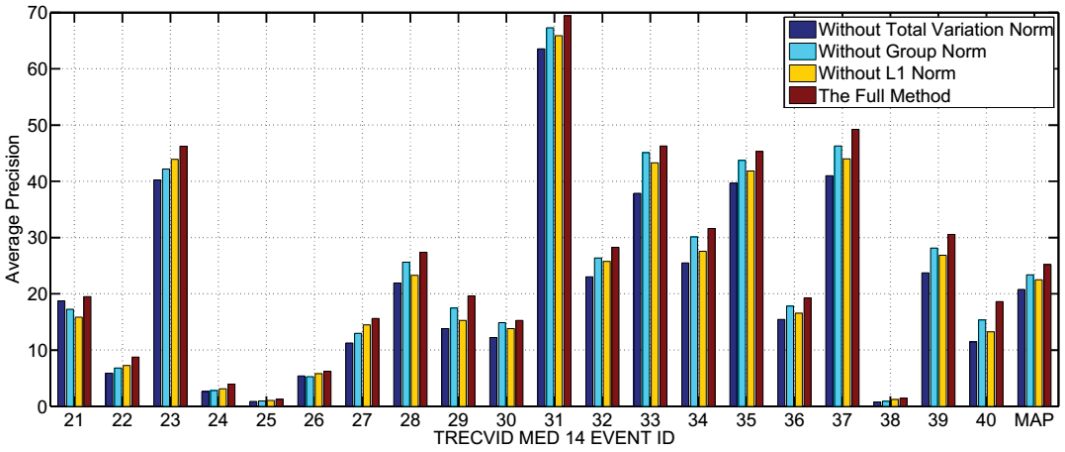


Figure 3.4: The average precisions of different sparse regularizers Ω on the TRECVID MEDTest 2014 dataset. Comparisons are made among our method by a) dropping group norm ($\alpha = 0$); b) dropping the ℓ_1 norm ($\beta = 0$); c) dropping total variation norm ($\gamma = 0$); and d) the full method.

the judges' labels. We observe that the proposed method achieves 89.7% and 83.2% average accuracy by selecting only 3.8% and 4.5% of shots in the original videos, respectively. This clearly demonstrates that our method is capable of localizing reasonable evidences that are accountable to the (positive) detection outcome and are comprehensible to humans. The results seem to indicate that the MEDTest 2014 dataset is more challenging than the MEDTest 2013 dataset.

The judges' preferences between the proposed method and the baseline are averaged and recorded in Table 3.4. It is clear that the proposed method is subjectively better for most of the events on both datasets. To verify, we show some of the recounting results in Figure 3.3 for the MEDTest 2014 dataset. These results again confirm that the proposed method successfully localizes the key evidences that also match human's intuition, for instance, as one would expect, "horse", "show jumping" and "animal" are all very indicative evidences for the *horse riding competition* event.

Table 3.4: Event recounting results on the TRECVID MEDTest 2013 and 2014 datasets. For each event, we randomly pick 10 positive test videos and ask 10 judges to rate *better, similar, or worse* between the proposed method and the baseline. The results among judges are aggregated via averaging.

	MEDTest 2013				MEDTest 2014		
ID	Better	Worse	Similar	ID	Better	Worse	Similar
E006	6	2	2	E021	8	1	1
E007	8	1	1	E022	2	5	3
E008	7	2	1	E023	4	4	2
E009	5	2	3	E024	7	0	3
E010	9	1	0	E025	5	2	3
E011	7	1	2	E026	6	2	2
E012	5	2	3	E027	7	2	1
E013	3	6	1	E028	5	3	2
E014	7	1	2	E029	4	2	4
E015	3	1	6	E030	3	1	6
E021	8	1	1	E031	8	0	2
E022	2	5	3	E032	6	2	2
E023	4	4	2	E033	6	1	3
E024	7	0	3	E034	5	0	5
E025	5	2	3	E035	3	1	6
E026	6	2	2	E036	4	1	5
E027	7	2	1	E037	6	2	2
E028	5	3	2	E038	3	6	1
E029	4	2	4	E039	3	0	7
E030	3	1	6	E040	7	1	2
Total	111	41	48	Total	102	36	62

3.3.4 Sensitivity Analysis

We conduct some sensitivity analysis in this section, to draw further insights of the proposed joint training framework.

Effects of Ω in recounting model: We first study the influence of different norms in the sparse regularizer Ω on MEDTest 2014 dataset. Recall that $\Omega(R) = \alpha\|R\|_{2,1} + \beta\|R\|_1 + \gamma\sum_i \|R^i\|_{\mathsf{tv}}$ is used in our recounting model to enforce different structured sparse information about the evidences. We drop one of the three terms in Ω and record its influence on the detection accuracy. In details, we compare: a). without the group norm, *e.g.* $\alpha = 0$; b). without the ℓ_1 norm, *e.g.* $\beta = 0$; and c). without the total variation norm, *e.g.* $\gamma = 0$. The results are summarized in Figure 3.4. We see that the full method (*i.e.*, none of α, β, γ is set 0) consistently performs the best, indicating the usefulness of all three sparse regularizers. Dropping the total variation norm (*i.e.* $\gamma = 0$) deteriorates the performance the most, indicating the utter importance of temporal smoothness. Dropping the group norm (*i.e.* $\beta = 0$) affects the performance the least (on average). This is because the ℓ_1 norm also has the effect of zeroing out irrelevant concepts. The comparison against the separate training model, *i.e.* $\alpha = \beta = \gamma = 0$, is performed below. Overall the benefits of incorporating these sparse regularizers in our joint detection and recounting framework are significant.

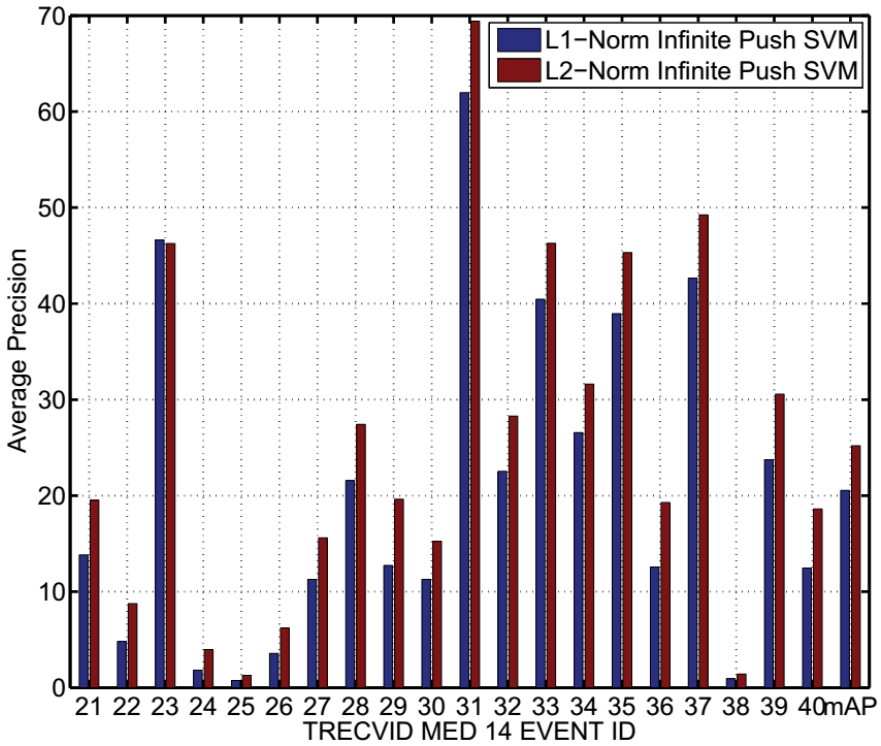


Figure 3.5: Performance comparison for the infinite-push SVM with $\Phi = \ell_2^2$ and $\Phi = \ell_1$ on the TRECVID MEDTest 2014 dataset. Results are presented in percentages.

Effects of Φ in detection model:

As discussed in Section 3.1.2, the proposed algorithm directly applies to different regularizers Φ in the infinite push SVM detection model (3.6), including the usual (squared) ℓ_2-norm and the sparse ℓ_1-norm. We conduct experiments to compare their performances on MEDTest 2014 dataset. The results are summarized in Figure 3.5. We observe that $\Phi = \ell_2^2$ outperforms $\Phi = \ell_1$ for all events, leading to a significantly higher mAP. This is because the "clean" representation R is already sought to be sparse in our recounting model, thus imposing further sparsity in the detection model often hurts the detection performance. On the other hand, the learned SVM weight W for $\Phi = \ell_1$ is much sparser than $\Phi = \ell_2^2$, and may be advantageous when testing time is a bigger concern.

Joint vs. Separate:

In our last experiment we validate the effectiveness of the proposed joint detection and recounting framework. By setting $\alpha = \beta = \gamma = 0$, we obtain a separate model that first performs event detection then followed by recounting. The results, summarized in Table 3.5, clearly demonstrate that the proposed joint training method consistently outperforms the separate training method on all events, usually by a large

Table 3.5: Performance comparison of the separate training method (*i.e.* event detection only with $\alpha = \beta = \gamma = 0$) and our joint training method. Mean average precision (mAP) is used as the evaluation metric. Results are presented in percentages. A larger mAP indicates better performance.

	MEDTest 2013			MEDTest 2014	
ID	Separate	Joint	ID	Separate	Joint
E006	38.25	44.66	E021	12.68	19.53
E007	49.26	57.63	E022	5.86	8.77
E008	57.45	64.22	E023	41.35	46.26
E009	41.58	47.94	E024	2.73	3.98
E010	22.69	28.49	E025	0.76	1.27
E011	11.58	16.77	E026	4.46	6.23
E012	31.86	39.38	E027	10.26	15.62
E013	61.59	68.25	E028	21.48	27.41
E014	43.87	52.33	E029	11.64	19.63
E015	27.43	35.34	E030	9.58	15.26
E021	12.68	19.53	E031	61.58	69.41
E022	5.86	8.77	E032	22.57	28.28
E023	41.35	46.26	E033	39.58	46.27
E024	2.73	3.98	E034	25.57	31.63
E025	0.76	1.27	E035	38.58	45.32
E026	4.46	6.23	E036	12.74	19.27
E027	10.26	15.62	E037	41.84	49.25
E028	21.48	27.41	E038	0.98	1.43
E029	11.64	19.63	E039	24.86	30.58
E030	9.58	15.26	E040	13.98	18.62
mean	25.32	30.95	mean	21.28	25.21

margin. We attribute this significant improvement to the coupling between detection and recounting: recounting helps detection by pruning irrelevant noisy concepts while detection directs recounting to more discriminative concepts.

3.4 Summary of This Chapter

We have proposed a novel joint training protocol to *simultaneously* conduct event detection and recounting. Based on a noisy semantic video representation, we couple a recounting model which aims at localizing the key evidences with a detection model which aims at enhancing the discriminative power. The recounting model assists detection by filtering out noisy irrelevant information while simultaneously the detection model guides recounting by directing it to the most discriminative

evidences, conceptual-wise and temporal-wise. The two models are optimized *jointly* hence benefit greatly from each other. To address the computational challenge due to operating on the shot level, we significantly improve the existing ADMM algorithm by proving closed-form solutions for the intermediate proximal updates. Augmented with the Uzawa linearization trick we are able to remove all inner loops in previous ADMM implementations, without affecting the convergence properties at all. The proposed method is tested on the large-scale and challenging TRECVID MEDTest 2013 and 2014 datasets, achieving very promising results in both detection and recounting. In the future we intend to incorporate domain knowledge and supplementary information such as text and audio.

Chapter 4

Event-Driven Concept Weighting for Zero-Exemplar Event Detection

In this chapter, we aim to detect complex events without any labeled training data for the event of interest. Following previous work on zero-shot learning Lampert et al. [2009]; Palatucci et al. [2009], we regard an event as a composition of multiple mid-level semantic concepts. These semantic concept classifiers are shared among events and can be trained using other resources. We then learn a skip-gram model Mikolov et al. [2013b] to assess the semantic correlation of the event description and the pre-trained vocabulary of concepts, based on which we automatically select the most relevant concepts for each event of interest. This step is carried out without any visual training data. This concept bundle view of an event aligns with the cognitive science literature, where humans are found to conceive objects as bundles of attributes Roach and Lloyd [1978]. The concept prediction scores on the testing videos are combined to obtain a final ranking of the presence of the event of interest. However, most existing zero-shot event detection systems aggregate the prediction scores of the concept classifiers with fixed weights. This clearly assumes all the predictions of a concept classifier share the same weight and fails to consider the differences in the classifier's prediction capability on individual testing videos. A concept classifier does in fact have different prediction capability on different testing videos, and some videos are correctly predicted while others are not. Therefore, instead of using a fixed weight for each concept classifier, a promising alternative is to estimate the specific weight for each testing video to alleviate the individual prediction errors from imperfect concept classifiers and achieve robust detection.

The problem of learning the specific weights of all the semantic concept classifiers for each testing video is challenging in the following aspects: First, it is unclear how to determine the specific weights for the unlabeled testing video since no label information can be used. Second, to obtain a robust detection result, we need to maximally ensure that positive videos have higher scores than negative videos in the

final ranking list. Note that the goal of event detection is to rank the positive videos before negative ones. To this end, we propose to learn the optimal weights for each testing video by exploring a set of online available videos which have free-form text descriptions of their content. We will additionally ensure that positive testing videos have the highest aggregated scores in the final result.

The main building blocks of the proposed approach for zero-example event detection can be described as follows. We first rank the semantic concepts for each event of interest using the skip-gram model, based on which the relevant concept classifiers are selected. The aggregation process can be cast as an information propagation procedure which propagates the weights learned on individual on-line available videos to the individual unlabeled testing videos, which ensures that visually similar videos have similar aggregated scores and offers the ability to infer weights for the testing videos. We then use the L_∞ norm infinite push constraint to minimize the number of positive videos ranked below the highest-scored negative videos, which ensures that positive videos mostly have higher aggregated scores than negative videos. In this way, we learn the optimal weights for each testing video and push positive videos to rank above negative videos as possible.

We summarize the contributions of this chapter as follows:

- We propose a novel event-driven concept weighting approach for zero example event detection to learn the optimal weights of related concept classifiers for each testing video by exploring a set of online available videos which have free-form text descriptions of their content.
- Infinity push SVM has been incorporated to ensure that most positive videos have the highest aggregated scores in the final prediction results.
- We conduct extensive experiments on three real video datasets (namely MEDTest 2014 dataset, MEDTest 2013 dataset and CCV_{sub}), and achieve state-of-the-art performance.

4.1 The Proposed Approach

In this chapter, we focus on the challenging zero-exemplar event detection problem in which we are given a sequence of unseen testing videos and also the event description, but no labeled training data for the event of interest. The goal is to rank the testing videos so that positive videos (those containing the event of interest) are ranked above negative videos. With this goal in mind, we first associate a query event with related semantic concepts that are pre-trained using other sources. We then aggregate the individual concept prediction scores using the proposed event-driven concept weighting approach.

4.1.1 Semantic Query Generation

Our work is built upon the observation that each event can be described by multiple *semantic concepts*. For example, the *marriage proposal* event can be attributed to several concepts, such as "ring" (object), "kissing" (action), "kneeling down" (action) and "cheering" (acoustic). Since semantic concepts are shared among different events and each concept classifier can be trained independently using data from other sources, zero-example event detection can be achieved by combining the relevant concept prediction scores. In contrast to the pioneer work in Lampert et al. [2009], which largely relies on human knowledge to decompose classes (events) into attributes (concepts), our goal is to automatically evaluate the semantic similarity between the event of interest and the concepts, based on which we select the relevant concepts for each event.

Events come with textual side information, *e.g.*, an event name or a short description. For example, the event *dog show* in the TRECVID MEDTest 2014 NIST [2014] is defined as "a competitive exhibition of dogs". With the availability of a pre-trained vocabulary of concept classifiers, we can evaluate the semantic correlation between the query event and individual concepts. We learn a skip-gram model Mikolov et al. [2013b] using the English Wikipedia dump[1]. The skip-gram model infers a D-dimensional vector space representation by fitting the joint probability of the co-occurrence of surrounding contexts on large unstructured text data, and places semantically similar words near each other in the embedded vector space. Thus it is able to capture a large number of precise syntactic and semantic word relationships. For short phrases consisting of multiple words (*e.g.*, event descriptions), we simply average its word-vector representations. After properly normalizing the respective word-vectors, we compute the cosine distance of the event description and all individual concepts, resulting in a correlation vector $\mathbf{w} \in [0,1]^m$, where w_k measures the priori relevance of the k-th concept and the event of interest. Based on the relevance vector, we select the most informative concept classifiers for each event.

4.1.2 Weak Label Generation

According to the NIST standard, we utilize the TRECVID MED *research* dataset to explore the optimal weights for the testing videos. All the videos in the *research* set come with a sentence of description, summarizing the contents contained in the videos. Note that none of the videos in the *research* set has event-level label information. We also collect a set of videos with free-form text descriptions of their content, which are widely available online at websites such as YouTube and NetFlix.

As the descriptions of the videos in both the *research* set and the online website are very noisy, we apply standard natural language processing (NLP) techniques to

[1] http://dumps.wikimedia.org/enwiki/

clean up the annotations, including removal of common stop words and stemming to normalize word inflections.

Similar to the steps in Semantic Query Generation (SQG), we measure the semantic correlation between the cleaned sentence and each concept description, and use it as a weak label for each concept. In the next section, we will learn the optimal aggregation weights for the individual testing video by exploiting the supervision information using the weak labels, which accounts for the differences in the prediction abilities of concept classifiers on the individual testing videos, and hences achieve robust aggregation results.

4.1.3 Event-Driven Concept Weighting

Let us assume we now have l videos with weak labels and u testing videos. We propose to learn an aggregation function $f_i(\mathbf{s}_i) = \mathbf{w}_i^T \mathbf{s}_i$ for each testing video ($i = 1, \cdots, l+u$), where $\mathbf{w} = [w_i^1, \cdots, w_i^m]^T$ is a non-negative aggregation weight vector with w_i^j being the aggregation weight of s_i^j. Clearly, it is straightforward for us to learn the optimal aggregation weights for the videos in the collected set based on the weak label information. However, it is challenging to derive the optimal aggregation weights for the *unlabeled* videos since no label information is available.

To achieve this goal, we build our model based on the local smoothness property in graph-based semi-supervised learning, which assumes that visually similar videos have comparable labels within a local region of the sample space Lai et al. [2015]; Liu et al. [2013a]. Exploring the local connectivity of data is a successful strategy for graph construction. The neighbors of video x_i can be defined as the k-nearest videos in the collection to x_i. In this work, we consider the probabilistic neighbors. Following the work in Nie et al. [2014], we learn the data similarity matrix by assigning the adaptive neighbors for each video based on local connectivity.

For the i-th video x_i, all the videos $\{x_1, x_2, \ldots, x_{l+u}\}$ can be connected to x_i as a neighbor with probability $\mathbf{a}_{ij}$. Usually, if $(\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2$ is smaller, a larger probability $\mathbf{a}_{ij}$ should be assigned between the video x_i and x_j. A natural method of determining the probabilities $\mathbf{a}_{ij}|_{j=1}^{l+u}$ is to solve the following problem:

$$\min_{\mathbf{W}, \mathbf{a}_i^T \mathbf{1}=1, 0 \le \mathbf{a}_i \le 1} \sum_{i,j=1}^{l+u} (\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{a}_{ij}, \tag{4.1}$$

where $\mathbf{s}_i$ is a vector with $\mathbf{s}_{ij}$ as the j-th element.

However, the problem Equation (4.1) has a trivial solution, only the nearest video can be the neighbor of the video x_i with probability 1 and no other videos can be the neighbors of x_i. On the other hand, if the following problem is solved without involving any distance information between the videos:

$$\min_{\mathbf{a}_i^T\mathbf{1}=1,0\leq\mathbf{a}_i\leq 1}\sum_{j=1}^{l+u}\mathbf{a}_{ij}^2, \tag{4.2}$$

the optimal solution is that all the data points can be neighbors of the video x_i with the same probability $\frac{1}{l+u}$, which can be seen as a prior in the neighbor assignment.

Combining Equation (4.1) and Equation (4.2), we can solve the following problem:

$$\min_{\mathbf{W},\mathbf{a}_i^T\mathbf{1}=1,0\leq\mathbf{a}_i\leq 1}\sum_{j=1}^{l+u}(\mathbf{w}_i^T\mathbf{s}_i-\mathbf{w}_j^T\mathbf{s}_j)^2\mathbf{a}_{ij}+\gamma\mathbf{a}_{ij}^2. \tag{4.3}$$

The second term in Equation (4.3) is a regularization and γ is the regularization parameter. Denote $d_{ij}^s = (\mathbf{w}_i^T\mathbf{s}_i - \mathbf{w}_j^T\mathbf{s}_j)^2$, and the problem Equation (4.3) can be written in vector form as:

$$\min_{\mathbf{a}_i^T\mathbf{1}=1,0\leq\mathbf{a}_i\leq 1}\|\mathbf{a}_i+\frac{1}{2\gamma}d_i^s\|_2^2. \tag{4.4}$$

We will verify that this problem can be solved with a closed form solution in Section 4.1.4.

We incorporate an infinite push loss function Rakotomamonjy [2012] to achieve a robust aggregation result. The goal of the infinite push loss function is to minimize the number of positive videos which are ranked below the highest scored negative videos. The infinite push loss function has shown promising performance in the event detection problem in Chang et al. [2015b]. In fact, the number of positive videos ranked below the highest scored negative videos equals the maximum number of positive videos ranked below any negative videos, hence, we define it as follows:

$$\ell(\{f_i\}_{i=1}^l;\mathfrak{P},\mathfrak{N})=\max_{j\in\mathfrak{N}}(\frac{1}{p}\sum_{i\in\mathfrak{P}}I_{f_i(\mathbf{s}_i^+)<f_j(\mathbf{s}_j^-)}), \tag{4.5}$$

where I is the indicator function whose value is 1 if $f_i(\mathbf{s}_i^+) < f_j(\mathbf{s}_j^-)$ and 0 otherwise. The maximum operator over j is equal to calculating the l_∞-norm of a vector consisting of n entries, each of which corresponds to one value based on j in the parentheses of Equation (4.5). By minimizing this penalty, positive videos tend to score higher than negative videos. This essentially ensures that positive videos have higher combined scores than the negative videos, leading to more accurate combined results.

For computational tractability, we upper bound the discrete 0-1 loss $I(\delta < 0)$ by the *convex* hinge loss $(1-\delta)_+$, where as usual $(\delta)_+ := \max(\delta, 0)$ is the positive part.

Since we usually pay more attention, if not exclusively, to the top of the rank list, we focus on minimizing the maximum ranking error among all negative exemplars $j \in \mathfrak{N}$:

$$\ell(\{f_i\}_{i=1}^{l}; \mathfrak{P}, \mathfrak{N}) = \max_{j \in \mathfrak{N}}(\frac{1}{p}\sum_{i \in \mathfrak{P}}(1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+), \tag{4.6}$$

Finally, the objective function can be written as:

$$\min_{\mathbf{W}, \mathbf{a}_i^T \mathbf{1}=1, 0 \leq \mathbf{a}_i \leq 1} \sum_{j=1}^{l+u}(\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{a}_{ij} + \gamma \mathbf{a}_{ij}^2 + \lambda \max_{j \in \mathfrak{N}}(\frac{1}{p}\sum_{i \in \mathfrak{P}}(1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+),$$
$$\text{s.t. } \mathbf{w}_i \geq 0, i = 1, \cdots, l+u.$$

4.1.4 Optimization

In this section, we develop an efficient alternative optimization algorithm to solve the problem Equation (4.7).

Optimizing W while fixing $\mathbf{a}$: We rewrite the optimization problem in Equation (4.7) as the following linearly-constrained problem:

$$\min_{\mathbf{W}} \sum_{j=1}^{l+u}(\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{a}_{ij} + \gamma \mathbf{a}_{ij}^2 + \lambda \max_{j \in \mathfrak{N}}(\frac{1}{p}\sum_{i \in \mathfrak{P}}(1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+),$$
$$\text{s.t. } \mathbf{w}_i \geq 0, i = 1, \cdots, l+u,$$

which is equivalent to:

$$\min_{\mathbf{W}} \sum_{j=1}^{l+u}(\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{a}_{ij} + \lambda \max_{j \in \mathfrak{N}}(\frac{1}{p}\sum_{i \in \mathfrak{P}}(1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+),$$
$$\text{s.t. } \mathbf{w}_i \geq 0, i = 1, \cdots, l+u,$$

With simple mathematical derivation, the objective function arrives at:

$$\min_{\mathbf{W}} (\Pi(\mathbf{W}))^T L(\Pi(\mathbf{W})) + \lambda \max_{j \in \mathfrak{N}}(\frac{1}{p}\sum_{i \in \mathfrak{P}}(1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+),$$
$$\text{s.t. } \mathbf{w}_i \geq 0, i = 1, \cdots, l+u,$$

where $\Pi(\mathbf{W})$ is a vector defined as $\Pi(\mathbf{W}) = ((\mathbf{W}^T \mathbf{S}) \circ \mathbf{I})\mathbf{1}$, $\circ$ denotes the Hadamard matrix product and $L = I - A$ is graph Laplacian. Recall that the minimization of the

Algorithm 3: Event-Driven Concept Weighting

Input: $\mathbf{S} \in \mathbb{R}^{m\times(l+u)}$, regularization λ, γ.
Output: $f_i(\mathbf{s}_i) = (\mathbf{w}_i^*)^T\mathbf{s}_i$, $i = l+1, \cdots, l+u$.
Initialize: $\mathbf{A} \in \mathbb{R}^{(l+u)\times(l+u)}$, $\mathbf{W}^0 > 0$, $\mathbf{e}^0$, $\alpha^0 = 0$, $\mu = 10^{-4}$.
Calculate $\mathbf{Q} \in \{-1, 0, 1\}^{(l+u)\times pn}$ based on the weak label information.
repeat
    Initialize $k = 0$.
    repeat
        Update $\mathbf{r}$ according to Equation (4.17).
        Update $\mathbf{W}^{k+1}$ according to Equation (4.16), and update the negative values to 0.
        Update $\mathbf{U}^{k+1}$ according to Equation (4.10).
        Update $\mathbf{r}$ according to Equation (4.17).
        Update $\mathbf{e}^-$ and $\mathbf{e}^+$ by solving Equation (4.20).
        Update $\mathbf{e}^{k+1}$ according to Equation (4.21).
        Update α_{k+1} according to Equation (4.15).
        k = k + 1.
    until *convergence*;
    Update $\mathbf{A}$ according to Equation (4.23).
until *convergence*;

first term ensures a smooth fusion score propagation over the graph structure, giving visually similar videos similar aggregation scores.

By defining

$$\mathbf{Q} = [(\mathbf{I}_p \bigotimes \mathbf{1}_{n\times 1}^T), (\mathbf{I}_n \bigotimes (-\mathbf{1}_{p\times 1}^T)), (\mathbf{0}_{u\times p} \bigotimes \mathbf{B})^T]^T, \tag{4.7}$$

the objective function can be rewritten as:

$$\begin{aligned} &\min_{\mathbf{W}} (\Pi(\mathbf{W}))^T L(\Pi(\mathbf{W})) + \lambda \max_{j\in\mathfrak{N}}(\frac{1}{p}\sum_{i\in\mathfrak{P}}(\mathbf{e}_{ij})_+), \\ &\text{s.t. } \mathbf{Q}^T\mathbf{U}\mathbf{1} + \mathbf{e} - \mathbf{1} = 0 \end{aligned}$$

where

$$\mathbf{e}_{ij} = 1 - (\mathbf{w}_i^T\mathbf{s}_i^+ - \mathbf{w}_j^T\mathbf{s}_j^-), \tag{4.8}$$

$$(\mathbf{e}_{ij})_+ = \max(\mathbf{e}_{ij}, 0) \tag{4.9}$$

and

$$\mathbf{U} = (\mathbf{W}^T\mathbf{S}) \circ \mathbf{I}. \tag{4.10}$$

The augmented Lagrangian of the above problem is:

$$\begin{aligned}&\mathcal{L}(\mathbf{W}, \mathbf{e}, \gamma, \mu) = (\Pi(\mathbf{W}))^T L(\Pi(\mathbf{W})) + \lambda \max_{j\in\mathfrak{N}}(\tfrac{1}{p}\textstyle\sum_{i\in\mathfrak{P}}(\mathbf{e}_{ij})_+) \\ &+\gamma^T(\mathbf{Q}^T\mathbf{U}\mathbf{1} + \mathbf{e} - \mathbf{1}) + \tfrac{\mu}{2}\|\mathbf{Q}^T\mathbf{U}\mathbf{1} + \mathbf{e} - \mathbf{1}\|_2^2,\end{aligned} \tag{4.11}$$

where γ is the vector of Lagrangian multipliers of the linear constraints and μ is a weighting parameter of the quadratic penalty. In the experiments, we empirically set μ as 10^{-4}. The objective function is equal to:

$$\begin{aligned}&\mathcal{L}(\mathbf{W}, \mathbf{e}, \alpha) = (\Pi(\mathbf{W}))^T L(\Pi(\mathbf{W})) \\ &+\lambda \max_{j\in\mathfrak{N}}(\tfrac{1}{p}\textstyle\sum_{i\in\mathfrak{P}}(\mathbf{e}_{ij})_+) + \tfrac{\mu}{2}\|\mathbf{Q}^T\mathbf{U}\mathbf{1} + \mathbf{e} - \mathbf{1} + \alpha\|_2^2,\end{aligned} \tag{4.12}$$

where $\alpha = \frac{\gamma}{\mu}$. At this point, the objective function can be solved by iteratively solving the saddle point of the augmented Lagrangian. At the k-th iteration, we need to solve the following three equations:

$$\mathbf{W}^{k+1} = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{e}^k, \alpha^k), \tag{4.13}$$

$$\mathbf{e}^{k+1} = \arg\min_{\mathbf{e}} \mathcal{L}(\mathbf{W}^{k+1}, \mathbf{e}, \alpha^k), \tag{4.14}$$

$$\alpha^{k+1} = \alpha^k + \mathbf{Q}^T\mathbf{U}^{k+1}\mathbf{1} + \mathbf{e}^{k+1} - \mathbf{1}. \tag{4.15}$$

The above three sub-problems can be solved by alternating optimization. First, the problem Equation (4.13) can be rewritten as:

$$\min_{\mathbf{W}} \mathcal{O}(\mathbf{W}) = \frac{\mu}{2}\|\mathbf{Q}^T\mathbf{U}^{k+1}\mathbf{1} - \mathbf{r}\|_2^2 + (\Pi(\mathbf{W}))^T L(\Pi(\mathbf{W})), \tag{4.16}$$

where

$$\mathbf{r} = \mathbf{1} - \mathbf{a}^k - \alpha^k. \tag{4.17}$$

The gradient of Equation (4.16) can be calculated as:

$$\frac{\bigtriangledown\mathcal{O}(\mathbf{W})}{\mathbf{W}_{ij}} = tr((\mu\mathbf{Q}(\mathbf{Q}^T\mathbf{U}\mathbf{1} - \mathbf{r})\mathbf{1}^T + 2\mathbf{L}\mathbf{U}\mathbf{1}\mathbf{1}^T)^T\frac{\partial\mathbf{U}}{\partial\mathbf{W}_{ij}}), \tag{4.18}$$

through which the problem can be solved by a conjugate gradient descent method.

Second, the problem Equation (4.14) can be solved by:

$$\min_{\mathbf{e}} \max_{j \in \mathfrak{N}} (\frac{1}{p} \sum_{i \in \mathfrak{P}} (\mathbf{e}_{ij})_+) + \frac{\mu}{2} \|\mathbf{e} - \mathbf{r}\|_2^2, \tag{4.19}$$

where $\mathbf{r} = \mathbf{1} - \alpha^k - \mathbf{Q}^T \mathbf{U}^{k+1} \mathbf{1}$. To solve the two nested max operators in $\frac{1}{p} \sum_{i \in \mathfrak{P}} (\mathbf{e}_{ij})_+$, the double trick can be used to convert the problem as:

$$\begin{array}{l} \min_{\mathbf{e}^+, \mathbf{e}^-} \ \frac{1}{2} \|\mathbf{e}^+ - \mathbf{e}^- - \mathbf{r}\|_2^2 + \max_{1 \le j \le n} (\frac{\lambda}{\mu p} \sum_{i \in \mathfrak{P}} \mathbf{e}_i^+) \\ \text{s.t. } \mathbf{e}^+ \ge 0, \mathbf{e}^- \ge 0, \end{array} \tag{4.20}$$

where

$$\mathbf{e} = \mathbf{e}^+ - \mathbf{e}^-. \tag{4.21}$$

This problem can be solved by iterative optimization by employing the optimization method in Rakotomamonjy [2012], where $\mathbf{e}^-$ has a closed-form solution while $\mathbf{e}^+$ can be solved by the Douglas-Rachford method Eckstein and Bertsekas [1992], which alternatively performs two proximal operators on the positive quadrant and the $\ell_{1,\text{inf}}$ mixed norm until convergence Quattoni et al. [2009].

Optimizing a while fixing W: In real-world application, the regularization parameter is very difficult to tune since the performance varies greatly as the regularization parameter changes from zero to infinite. In this section, we propose an effective method to determine the regularization parameter γ in Equation (4.4).

For each i, the Lagrangian function of Equation (4.4) is:

$$\mathcal{L}(\mathbf{a}_i, \eta, \beta_i) = \frac{1}{2} \|\mathbf{a}_i + \frac{1}{2\gamma_i} d_i^s\|_2^2 - \eta(\mathbf{a}_i^T \mathbf{1} - 1) - \beta_i^T \mathbf{a}_i, \tag{4.22}$$

where η and $\beta_i \ge 0$ are the Lagrangian multipliers.

According to the KKT condition, it can be verified that the optimal solution $\mathbf{a}_i$ should be:

$$\mathbf{a}_{ij} = (-\frac{d_{ij}^s}{2\gamma_i} + \eta)_+ \tag{4.23}$$

In practice, better performance could be achieved if we focus on the locality of the data. Therefore, it is preferable to learn a sparse $\mathbf{a}_i$, *i.e.*, only the k nearest neighbors of $\mathbf{a}_i$ would have the opportunity to connect to the i-th video x_i. Another advantage

of learning a sparse similarity matrix A is that the computation burden can be largely alleviated for subsequent processing.

Without loss of generality, suppose d_{i1}^s,d_{i2}^s,$\cdots$,d_{in}^s are ordered from small to large. If the optimal a_i has only k nonzero elements, then according to Equation (4.23), we know that $\mathbf{a}_{ik} \geq 0$ and $\mathbf{a}_{i,k+1} = 0$. Therefore, we have

$$\begin{cases} -\frac{d_{ik}^s}{2\gamma_i} + \eta > 0 \\ -\frac{d_{i,k+1}^s}{2\gamma_i} + \eta \leq 0 \end{cases} \tag{4.24}$$

According to Equation (4.23) and the constraint $\mathbf{s}_i^T \mathbf{1} = 1$, we have

$$\begin{array}{l} \sum_{j=1}^{k} (-\frac{d_{ij}^s}{2\gamma_i} + \eta) = 1 \\ \Rightarrow \eta = \frac{1}{k} + \frac{1}{2k\gamma_i} \sum_{j=1}^{k} d_{ij}^s \end{array} \tag{4.25}$$

We thus have the following inequality for γ_i according to Equation (4.24) and Equation (4.25):

$$\frac{k}{2} d_{ik}^s - \frac{1}{2} \sum_{j=1}^{k} d_{ij}^s < \gamma_i \leq \frac{k}{2} d_{i,k+1}^s - \frac{1}{2} \sum_{j=1}^{k} d_{ij}^s \tag{4.26}$$

To obtain an optimal solution $\mathbf{a}_i$ to the problem Equation (4.4) that has exact k nonzero values, we could therefore set γ_i as:

$$\gamma_i = \frac{k}{2} d_{i,k+1}^s - \frac{1}{2} \sum_{j=1}^{k} d_{ij}^s \tag{4.27}$$

The overall γ could be set to the main of γ_1, γ_2, $\cdots$, γ_n. That is, we could set γ as:

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} (\frac{k}{2} d_{i,k+1}^s - \frac{1}{2} \sum_{j=1}^{k} d_{ij}^s) \tag{4.28}$$

In the above equation, there is still one parameter k, which denotes the number of neighbors. In the experiment, we empirically set the number of neighbors as 5.

We summarize the optimization procedure in Algorithm 3.

Table 4.1: Experiment results for 0Ex event detection on MEDTest 2014. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

	MEDTest 2014							
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	EDCW
E021	2.12	2.98	2.64	3.89	3.97	3.12	4.64	6.37
E022	0.75	0.97	0.83	1.36	1.49	1.15	1.48	2.85
E023	33.86	36.94	35.23	39.18	40.87	38.68	41.78	44.26
E024	2.64	3.75	3.02	4.66	4.92	4.11	4.87	6.12
E025	0.54	0.76	0.62	0.97	1.39	0.84	1.01	1.26
E026	0.96	1.59	1.32	2.41	2.96	1.96	2.65	4.23
E027	11.21	13.64	12.48	15.93	16.26	15.12	16.47	19.63
E028	0.79	0.67	1.06	1.57	1.95	1.72	2.25	4.04
E029	8.43	10.68	12.21	14.01	14.85	13.19	14.75	17.69
E030	0.35	0.63	0.48	0.91	0.96	0.36	0.48	0.52
E031	32.78	53.19	45.87	69.52	69.66	67.49	72.64	77.45
E032	3.12	5.88	4.37	8.12	8.45	7.54	8.65	11.38
E033	15.25	20.19	18.54	22.14	22.23	21.53	23.26	26.64
E034	0.28	0.47	0.41	0.71	0.75	0.53	0.76	0.94
E035	9.26	13.28	11.09	16.53	16.68	15.82	18.65	21.78
E036	1.87	2.63	2.14	3.15	3.39	2.88	3.76	5.47
E037	2.16	4.52	3.81	6.84	6.88	5.42	6.83	8.45
E038	0.66	0.74	0.58	0.99	1.16	0.85	1.12	2.89
E039	0.36	0.57	0.42	0.69	0.77	0.64	0.85	2.26
E040	0.65	0.98	0.72	1.57	1.57	1.24	1.76	3.12
mean	6.40	9.55	7.89	10.76	11.05	10.21	11.44	13.37

4.1.5 Out-of-sample Extension

We have thus far proposed a transductive framework to learn the optimal aggregation weights of the related concepts for each testing video. In this section, we propose an efficient way to obtain the optimal aggregation weights for an unseen video.

For each new video, we first use the low-level feature to search a set of nearest neighbors from all the videos in the original dataset. We then calculate the aggregation score of the unseen video by:

$$f(z) = \sum_{i=1}^{q} \frac{G(z, x_i)}{\sum_{i=1}^{q} G(z, x_i)} (\mathbf{w}_i^*)^T \mathbf{s}_i. \tag{4.29}$$

Hence, we obtain the prediction score for the unseen video.

4.2 Experiments

In this section, we conduct extensive experiments to validate the proposed **E**vent-**D**riven **C**oncept **W**eighting for large scale semantic video search, abbreviated as **EDCW**.

Table 4.2: Experiment results for 0Ex event detection on MEDTest 2013. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

	MEDTest 2013							
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	EDCW
E006	5.34	4.92	4.69	7.63	7.98	6.41	8.32	11.87
E007	1.06	1.14	0.86	1.76	2.34	1.44	1.62	3.15
E008	28.43	33.02	18.96	41.88	42.25	38.09	43.14	47.62
E009	13.61	13.43	13.12	15.48	16.11	14.68	16.29	21.46
E010	0.88	0.85	0.82	0.93	0.99	0.74	0.91	1.89
E011	7.42	7.68	7.39	7.91	8.15	6.58	7.36	11.53
E012	19.84	21.96	19.36	22.45	22.64	18.42	24.33	28.14
E013	0.63	0.52	0.91	2.14	2.69	1.61	2.82	5.93
E014	1.14	1.23	0.94	2.51	2.98	1.72	3.44	6.87
E015	1.32	1.41	1.44	1.49	1.75	0.47	0.54	1.15
E021	2.12	2.98	2.64	3.89	3.97	3.12	4.64	6.37
E022	0.75	0.97	0.83	1.36	1.49	1.15	1.48	2.85
E023	33.86	36.94	35.23	39.18	40.87	38.68	41.78	44.26
E024	2.64	3.75	3.02	4.66	4.92	4.11	4.87	6.12
E025	0.54	0.76	0.62	0.97	1.39	0.84	1.01	1.26
E026	0.96	1.59	1.32	2.41	2.96	1.96	2.65	4.23
E027	11.21	13.64	12.48	15.93	16.26	15.12	16.47	19.63
E028	0.79	0.67	1.06	1.57	1.95	1.72	2.25	4.04
E029	8.43	10.68	12.21	14.01	14.85	13.19	14.75	17.69
E030	0.35	0.63	0.48	0.91	0.96	0.36	0.48	0.52
mean	7.07	7.94	6.92	9.45	9.88	8.43	9.96	12.64

4.2.1 Experiment Setup

Dataset: To evaluate the effectiveness of the proposed approach, we conduct extensive experiments on the following three large-scale event detection datasets:

— TRECVID MEDTest 2014 dataset NIST [2014]: This dataset has been introduced by the NIST for all participants in the TRECVID competition and research community to perform experiments on. There are 20 events in total, the description of which can be found in NIST [2014]. We use the official test split released by the NIST, and strictly follow its standard procedure NIST [2014]. To be more specific, we detect each event *separately*, treating each of them as a binary classification/ranking problem.

— TRECVID MEDTest 2013 dataset NIST [2013]: The settings of MEDTest 2013 dataset is similar to MEDTest 2014, with 10 of their 20 events overlapping.

— Columbia Consumer Video dataset Jiang et al. [2011]: The official Columbia Consumer Video dataset contains 9,317 videos in 20 different categories, including scenes like “beach”, objects like “cat”, and events like “basketball” and “parade”. Since the goal of this work is to *search complex events*, we only use the 15 event categories.

Table 4.3: Experiment results for 0Ex event detection on CCV_{Sub}. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

	CCV_{sub}							
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	EDCW
1	26.84	25.98	25.53	28.16	28.54	27.49	28.32	30.12
2	18.67	17.74	18.94	20.26	21.16	24.17	24.68	24.45
3	16.68	17.16	17.85	18.56	19.35	19.53	20.85	22.06
4	28.16	28.98	30.26	31.42	31.87	31.94	32.92	33.08
5	31.27	29.86	31.84	32.15	32.84	32.14	32.89	33.27
6	29.53	31.13	31.46	31.98	32.54	33.26	34.19	33.84
7	21.19	21.45	21.98	22.05	22.47	25.37	25.84	25.75
8	15.52	14.75	16.38	16.94	17.26	19.92	20.08	19.72
9	10.26	10.87	11.95	13.07	14.42	15.75	16.42	16.48
10	16.64	17.33	18.12	18.86	19.59	18.97	19.84	22.46
11	13.98	14.52	15.17	16.43	16.88	18.29	19.37	21.69
12	18.15	18.87	19.49	20.66	21.34	24.09	24.68	25.52
13	15.72	17.13	17.95	18.59	20.04	21.57	22.49	23.66
14	13.47	14.86	15.39	16.17	17.58	18.13	19.46	20.47
15	9.64	10.38	11.49	12.09	12.48	15.54	16.02	15.39
mean	19.05	19.40	20.25	21.16	21.89	23.08	23.87	24.36

Table 4.4: Few-exemplar results on MED14 and MED13.

EDCW (0Ex)	13.37	12.64
IDT (10Ex)	13.92	18.08
EDCW (0Ex) + IDT (10Ex)	16.37	19.24

According to the standard of the NIST, each event is detected separately and the performance of event detection is evaluated using the mean Average Precision (mAP).
Concept Detectors: 3,135 concept detectors are pre-trained using TRECVID SIN dataset (346 categories) Jiang et al. [2014a]; Over et al. [2014], Google sports (478 categories) Jiang et al. [2014a]; Karpathy et al. [2014], UCF101 dataset (101 categories) Jiang et al. [2014a]; Soomro et al. [2012], YFCC dataset (609 categories) YFC; Jiang et al. [2014a] and DIY dataset (1601 categories) Jiang et al. [2014a]; Yu et al. [2014a]. The improved dense trajectory features (including trajectory, HOG, HOF and MBH) are first extracted using the code of Wang and Schmid [2013] and encode them with the Fisher vector representation Perronnin et al. [2010]. Following Wang and Schmid [2013], the dimension of each descriptor is first reduced by a factor of 2 and then use 256 components to generate the Fisher vectors. Then, on top of the extracted low-level features, the cascade SVM Graf et al. [2004] is trained for each concept detector.
Competitors: We compare the proposed approach with the following alternatives:
- Prim Habibian et al. [2014a]: Primitive concepts, separately trained.
- Sel Mazloom et al. [2013a]: A subset of primitive concepts that are more informative for each event.

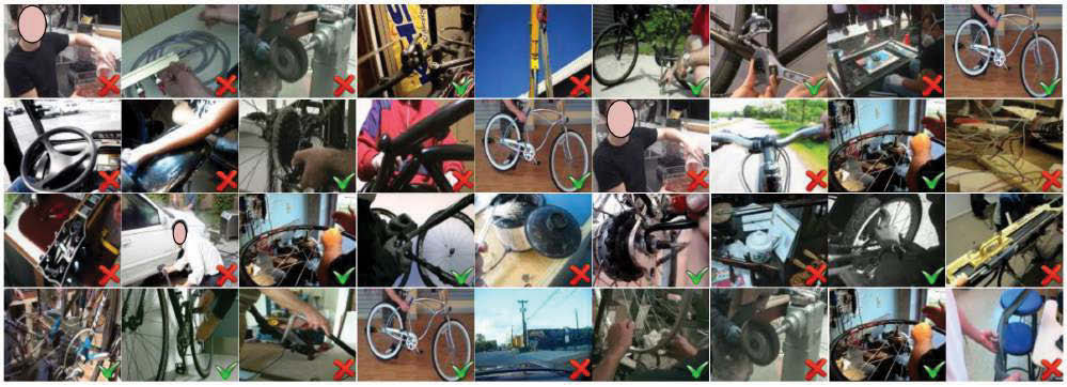


Figure 4.1: Top ranked videos for the event *non-motorized vehicle repair*. From top to below: OR, Fu, PCF and EDCW. True/false labels (provided by NIST) are marked in the lower-right of each frame.

- Bi Rastegari et al. [2013]: Bi-concepts discovered in Rastegari et al. [2013].
- OR Habibian et al. [2014a]: Boolean OR combinations of Prim concepts.
- Fu Habibian et al. [2014a]: Boolean AND/OR combinations of Prim concepts, w/o concept refinement.
- Bor: The Borda rank aggregation with equal weights on the discovered semantic concepts.
- PCF Chang et al. [2015a]: The pair-comparison framework is incorporated for zero-shot event detection.

4.2.2 Zero-exemplar event detection

We report the full experimental results on the TRECVID MEDTest 2014, MEDTest 2013 and CCV_{sub} in Table 5.1, Table 5.2 and Table 5.3. From the experimental results shown in Table 5.1, the proposed algorithm, **EDCW**, performs better than the other approaches with a large margin (13.37% vs 11.44% achieved by PCF). The proposed approach gets significant improvement on some vents, such as *Dog Show (E023)*, *Rock Climbing (E027)*, *Beekeeping (E031)* and *Non-motorized Vehicle Repair (E033)*. By analyzing the discovered concepts for these events, we find that their classifiers are very discriminative and reliable. For example, for the event *Rock Climbing*, we discovered the concepts "Sport climbing", "Person climbing" and "Bouldering", which are the most informative concepts for *Rock climbing* in the concept vocabulary. Figure 5.3 illustrates the top retrieved results on the *non-motorized vehicle repair* event. To save space, we only show the top 4 compared algorithms (OR, Fu, PCF and EDCW). It is clear that videos retrieved by the proposed EDCW are more accurate and visually coherent.

We make the following observations from the results shown in Table 5.1: 1). Sel

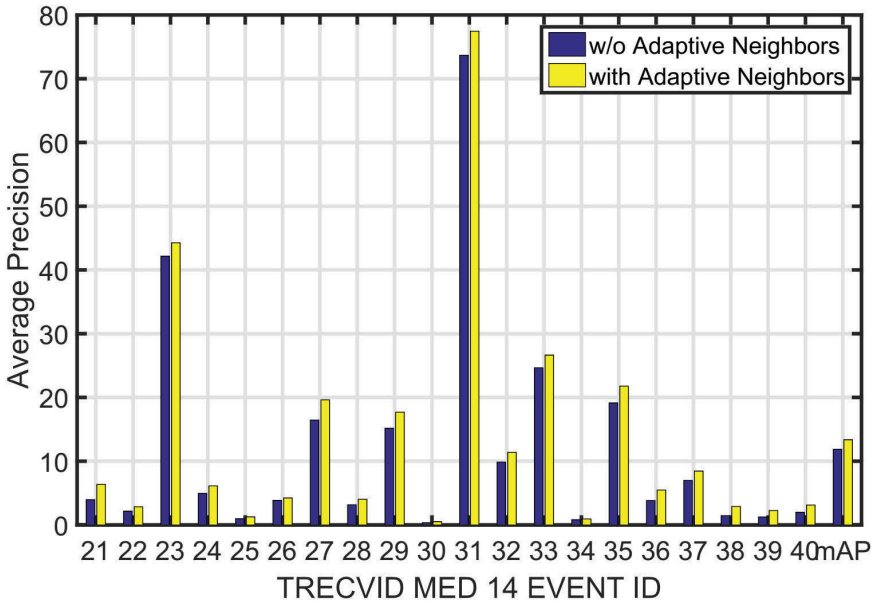


Figure 4.2: Performance comparison for the proposed framework w/o and with adaptive neighbors on the TRECVID MEDTest 2014 dataset. Results are presented in percentages. A larger mAP indicates better performance. The figures are best viewed in color.

significantly outperform Prim with mAP of 9.55% vs 6.40% on MEDTest 2014, which indicates that selecting the most discriminative concepts generally improves detection performance than naively using all the concepts. 2). Comparing the results of Bi, OR, Fu and Bor, we verify that the selected informative concept classifiers are not equally important for event detection. It is beneficial to derivate the weights for each concept classifier. By treating the concept classifiers differentially, better performance is achieved. 3). Comparing the results of EDCW with the other alternatives, we observe that learning optimal weights of concept classifiers for each testing video significantly improves performance of event detection. This confirms the importance of event-driven concept weighting.

We made similar observations from the results on the TRECVID MEDTest 2013 dataset and CCV_{sub} dataset in Table 5.2 and Table 5.3.

4.2.3 Do Adaptive Neighbors help?

We further conduct an experiment to evaluate how much adaptive neighbors contribute to the detection performance. For the framework without adaptive neighbors, we consider the following objective function:

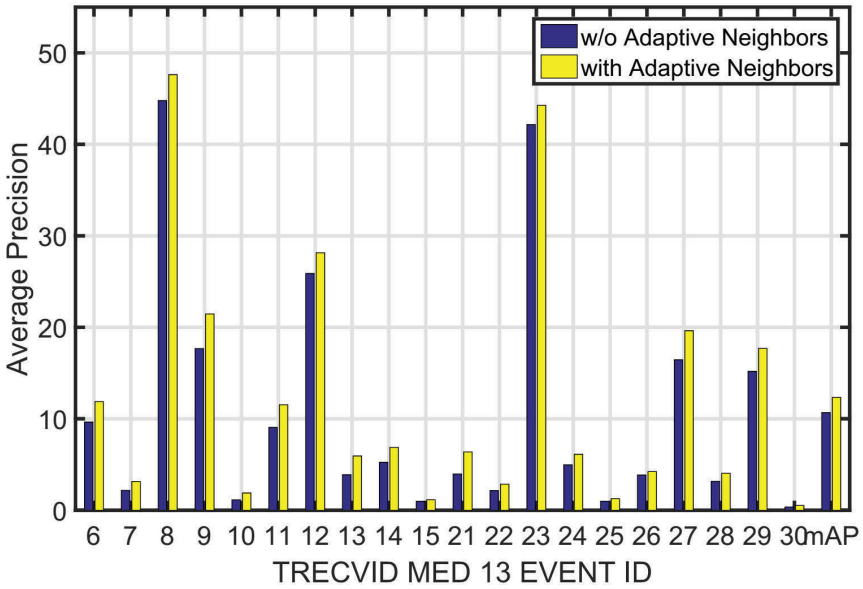


Figure 4.3: Performance comparison for the proposed framework w/o and with adaptive neighbors on the TRECVID MEDTest 2013 dataset. Results are presented in percentages. A larger mAP indicates better performance. The figures are best viewed in color.

$$\min_{\mathbf{W}} \sum_{j=1}^{l+u} (\mathbf{w}_i^T \mathbf{s}_i - \mathbf{w}_j^T \mathbf{s}_j)^2 \mathbf{a}_{ij} + \lambda \max_{j \in \mathfrak{N}} (\frac{1}{p} \sum_{i \in \mathfrak{P}} (1 - (\mathbf{w}_i^T \mathbf{s}_i^+ - \mathbf{w}_j^T \mathbf{s}_j^-))_+),$$
$$\text{s.t. } \mathbf{w}_i \geq 0, i = 1, \cdots, l+u.$$

We show the results on the three datasets in Figures 4.2 to 4.4. The results clearly demonstrate that the framework with adaptive neighbors consistently outperforms the framework without adaptive neighbors on all events, usually by a large margin. For example, using mAP as an evaluation metric, EDCW with adaptive neighbors achieves 13.37% while EDCW without adaptive neighbors only achieves 12.16%. We attribute this improvement to the novel adaptive neighbors.

4.2.4 Extension to few-exemplar event detection

The proposed zero-example event detection framework can also be used for few-exemplar event detection: we aggregate the concept classifiers and the supervised classifier using the event-driven concept weighting approach. In this section, experiments are conducted to demonstrate the benefit of this hybrid approach. Table 5.4

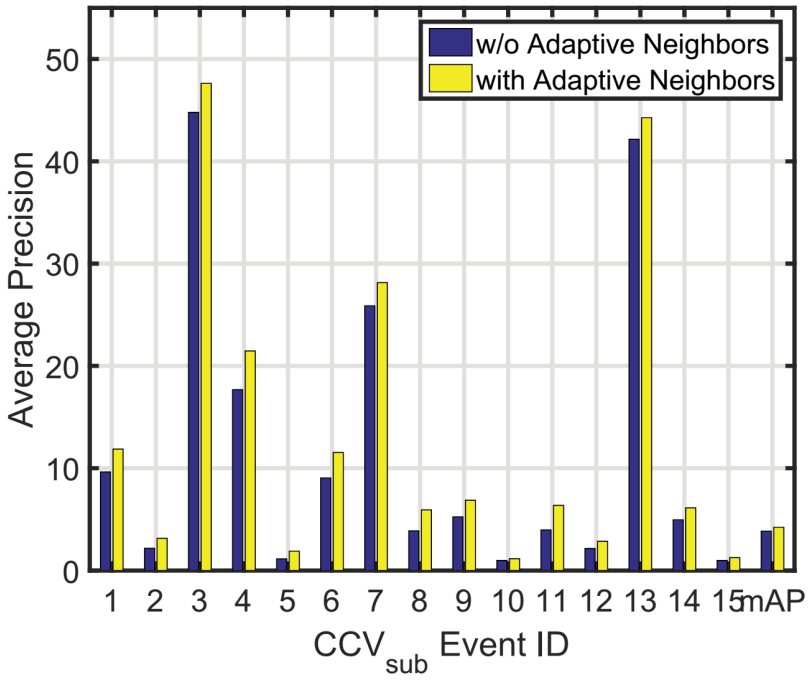


Figure 4.4: Performance comparison for the proposed framework w/o and with adaptive neighbors on the CCV_{sub} dataset. Results are presented in percentages. A larger mAP indicates better performance. The figures are best viewed in color.

summarizes the mAP on both the MEDTest 2014 and 2013 datasets, while Figure 5.4 compares the performance event-wise.

According to the NIST standard, we consider the 10 Ex setting, where 10 positive videos are given for each event of interest. The improved dense trajectories feature Wang and Schmid [2013] is extracted, on top of which an SVM classifier is trained. It is worthwhile to note that our EDCW which had no labeled training data can get comparable results with the supervised classifier (mAP 13.37% vs 13.92% on MEDTest 2014). This demonstrates that proper aggregation with optimal weights for each testing video can get promising results for event detection.

We compare EDCW with IDT event-wise in Figure 5.4. From the results we can see that our EDCW outperforms the supervised IDT on multiple events, namely E023, E024, E031 and E033. To be specific, on the event *Beekeeping* (E031), EDCW significantly outperform the supervised IDT (77.45% vs 33.92%). This is not supervising, since EDCW significantly benefited from the presence of informative and reliable concepts such as "apiary bee house" and "honeycomb" on the particular *Beekeeping* event.

Finally we combine EDCW with IDT to get the final few-example event detection result. This improves the performance from 13.92% to 16.98% on the MEDTest 2014

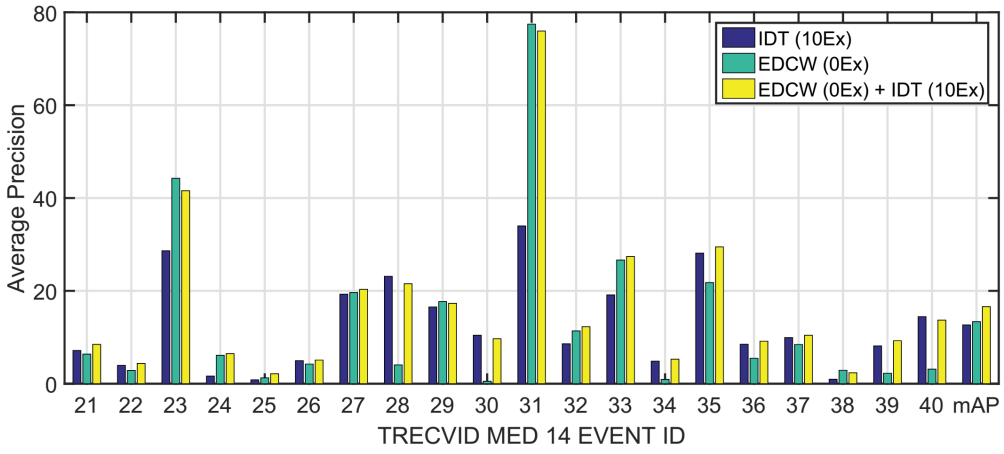


Figure 4.5: Performance comparison of IDT, EDCW, and the hybrid of IDT and EDCW. The figure is best viewed in color.

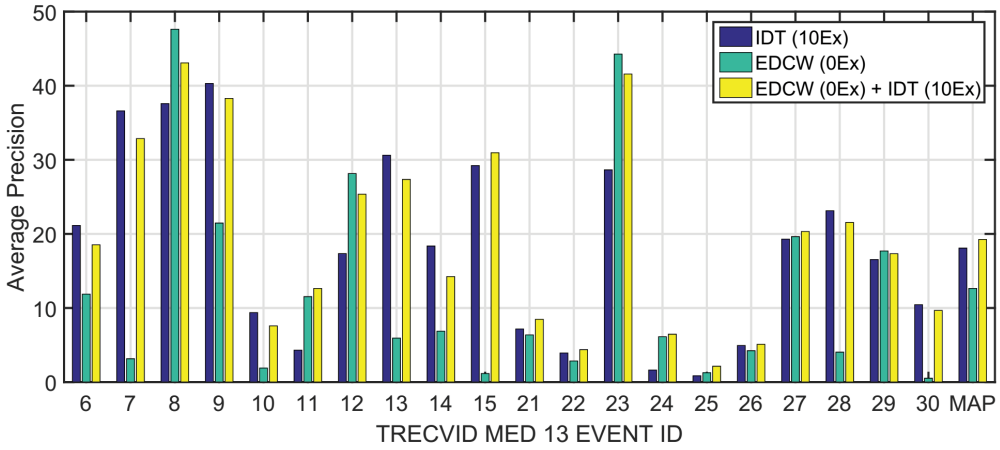
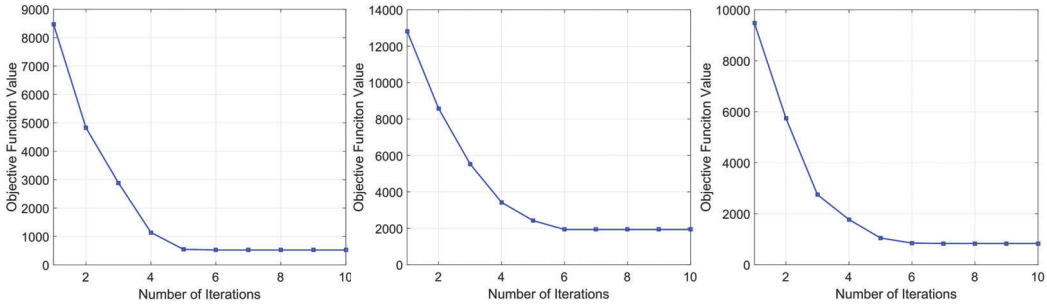


Figure 4.6: Performance comparison of IDT, EDCW, and the hybrid of IDT and WDCW. The figure is best viewed in color.

dataset. As expected, the gain obtained from such simple hybrid diminishes when combining with more sophisticated methods. Overall, the results clearly demonstrate the utility of our framework even in the few-exemplar setting.

4.2.5 Convergence Study

We solve the proposed objective function using an alternating approach. In practice, how fast the proposed algorithm converges is crucial for the whole computational efficiency. Hence, we conduct experiments to show the convergence curve of the proposed algorithm. As we have similar results on all the events of the datasets, we



(a) MEDTest 2014 (b) MEDTest 2013 (c) Columbia Consumer Video

Figure 4.7: Convergence curve of the objective function value in Equation (4.7) using Algorithm 3 on TRECVID MEDTest 2014, MEDTest 2013 and CCV_{sub} datasets. The figures show that the objective function value monotonically decreases until convergence by applying the proposed algorithm. The figures are best viewed in color.

randomly pick one event per dataset. All the parameters involved are fixed at 1.

Figure 4.7 shows the convergence curves. It can be seen that the objective function value converges within 10 iterations on the three datasets. The convergence experiments confirm the efficiency of the alternating algorithm.

4.3 Summary of This Chapter

To address the challenging task of zero-exemplar or few-exemplar event detection, we proposed to learn the optimal weights for each testing video by exploring the collected videos from other sources. Data-driven word embedding models were used to seek the relevance of the concepts to the event of interest. To further derivate the optimal weights of the concept classifiers for each testing video, we have proposed a novel event-driven concept weighting method by exploiting the textual information of the collected videos. Extensive experiments are conducted on three real video datasets. The experimental results confirm the efficiency of the proposed approach.

Chapter 5

Semantic Event Search using Differentiated Concept Classifiers

In this chapter we mainly focus on the semantic event search problem, where no example videos are provided for training whatsoever. Our system is built on the observation that an event is a composition of multiple mid-level concepts Chang et al. [2015a]; Lampert et al. [2009]; Palatucci et al. [2009]. These concepts are shared among events and can be collected from other sources (not necessarily related to the event search task). We then train a skip-gram language model Mikolov et al. [2013b] to automatically identify the most relevant concepts to a particular event of interest. For example, the most relevant concepts for the *marriage proposal* event might be "face," "ring," "kissing," "kneeling down," *etc*. Such concept bundle view of event also aligns with the cognitive science literature, where humans are found to conceive objects as bundles of attributes Roach and Lloyd [1978]. The concept scores on the test videos are combined to yield a final ranking of the presence of the event of interest. However, this approach, as well as most existing works on semantic event search Dalton et al. [2013]; Habibian et al. [2013, 2014a]; Rastegari et al. [2013]; Wu et al. [2014], ignore the fact that **not all concept classifiers are equally reliable**, especially when they are trained from other source domains. For example, "face" in video frames can now be reasonably accurately detected, but in contrast, the action "brush teeth" remains hard to recognize in short video clips. Consequently, a relevant concept can be of limited use or even misuse if its classifier is highly unreliable. Therefore, when combining concept scores, we propose to take their relevance, predictive power, and reliability all into account. This is achieved through a novel extension of the spectral meta-learner in Parisi et al. [2014], which provides a principled way to estimate classifier accuracies using purely *unlabeled* data. Figure 5.1 gives an overview of our entire system.

Contributions. To summarize, we make the following contributions in this chapter:

- To account for the unreliability of concept classifiers, we propose to use the warped

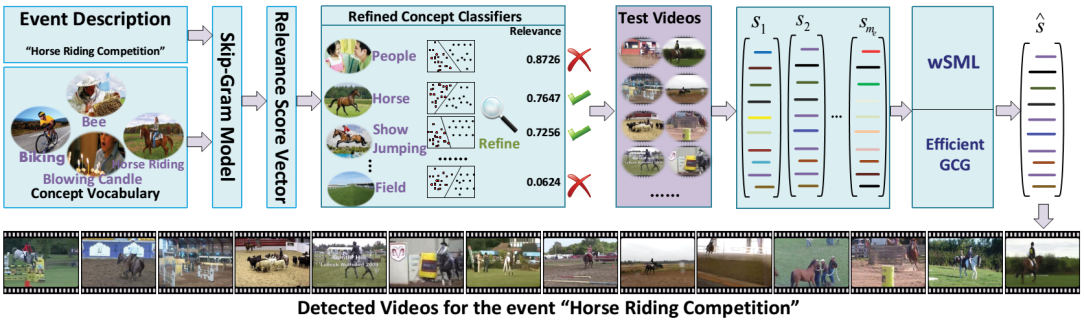


Figure 5.1: The proposed framework for large-scale semantic event search (§5.1), illustrated on the particular *horse riding competition* event. The relevance of concept classifiers to the event of interest are measured using the skip-gram language model (§5.1.2), followed by some further refinements (§5.1.3). To account for their reliability, the concept scores are combined through the warped spectral meta-learner (§5.1.6) and solved using the efficient GCG algorithm (§5.1.7).

spectral meta-learner to estimate the concept accuracies and combine them in a principled and purely unsupervised manner (§5.1.5 and §5.1.6).

- We provide an efficient implementation based on the recent generalized conditional gradient (§5.1.7), which is the key to conduct event search in large-scale video datasets.
- We conduct experiments on three real video datasets (MEDTest 2014, MEDTest 2013 and CCV_{sub}), and achieve state-of-the-art performances (§5.2).

5.1 Semantic Event Search

In this work we mainly consider the semantic event search problem, where the learning algorithm is asked to rank unlabeled test videos according to their likelihood of containing a certain event of interest, for instance, *birthday party*. The significant challenge here is that we are not given any labeled training data.

5.1.1 Concept Classifiers

Without labeled training data, we can no longer train a supervised statistical classifier but resort to rule based learning. The key observation is that each object class can be described as the composition of a set of *semantic concepts*, *i.e.*, middle-level interpretable attributes. For example, the event *marriage proposal* can be described as the composition of multiple objects (*e.g.*, ring, faces), scene (*e.g.*, in a restaurant), and actions (*e.g.*, talking, kneeling down). Since concepts are shared among many

different classes (events) and each concept classifier can be trained independently on datasets from other sources, semantic event search can be achieved by combining the *relevant* concept classification scores, even in the absence of event labeled training data. Different from the pioneering work Lampert et al. [2009], which largely relied on human knowledge to decompose classes (events) into attributes (concepts), we seek below an automated way.

5.1.2 Semantic Concept Relevance

Events come with short textual information, *e.g.*, an event name or a short description. For example, the event *dog show* in the TRECVID MEDTest 2014 NIST [2014] is defined as "a competitive exhibition of dogs." We exploit this textual information by learning a relevance score between the event description and the pre-trained concept (attribute) classifiers. Since the concept classifiers are trained without any event label information, the relevance score makes it possible to share information between the concept space and the event space. More precisely, we pre-train a skip-gram model Mikolov et al. [2013b] using the English Wikipedia dump[1]. The skip-gram model infers a D-dimensional vector space representation of words by fitting the joint probability of the co-occurrence of surrounding contexts on large unstructured text in the embedding vector space. Thus it is able to capture a large number of precise syntactic and semantic word relationships. For short phases consisting of multiple words (*e.g.*, event descriptions), we simply average its word-vector representation. After properly normalizing the respective word-vectors, we compute the cosine distance of the event description and all individual concepts, resulting in a relevance vector $\mathbf{w} \in [0,1]^m$, where w_k measures a priori relevance of the k-th concept and the event of interest. Similar approaches have appeared before in *e.g.* Mensink et al. [2014]; Norouzi et al. [2014]; Wu et al. [2014].

5.1.3 Concept Pruning and Refining

In the above we have introduced the relevance score vector $\mathbf{w} \in [0,1]^m$ that measures the similarity between the m concepts and the event of interest. We further prune and refine these weights for the following reasons: 1). Some concepts, although relevant to the event of interest, may not be very discriminative (low predictive power). For example, the concept *people* is relevant to the event *Birthday party*, but it appears almost in every video hence does not provide much discriminative power. 2). Some concepts may not be very reliable, possibly because they are trained on different domains. In the experiments, we use the (unlabeled) MED 2014 Research dataset[2] to

[1] http://dumps.wikimedia.org/enwiki/

[2] This adheres strictly to the NIST standard: "research set may be used for training concepts and assigning importance weights," see the online doc (Q35) at `http://www.nist.gov/itl/iad/`

crudely refine the concepts as follows: We first compute a similarity score between the concept names and the text description of each video in the research dataset, which acts as a *concept label*, *i.e*. the likelihood of each video to contain a particular concept. Then we run concept classifiers on each video in the research dataset, and use the aforementioned concept labels to compute the average precisions. Concepts with low precision or low predictive power (such as concept *people*) are then dropped. Importantly, our procedure does not require any manual annotation on the research dataset.

5.1.4 Combine the Classifier Ensemble

Suppose for event e we have selected m concepts[1], each with a weight $w_i \in [0, 1],\ i = 1, \ldots, m$. Then, for any test video $\mathbf{v}$, the i-th concept classifier generates a confidence score $s_i(\mathbf{v}) \in [-1, 1]$. Since different concept classifiers result in different confidence scores, we need a principled way to combine them, preferably also taking their relevance $\mathbf{w}$ into account. This can be treated as an ensemble learning problem, and there are many different ways to approach it. For instance, we can use each concept classifier i to induce a total ordering among n test videos, namely,

$$\text{video } k \text{ ranked above video } l \iff s_i(\mathbf{v}_k) > s_i(\mathbf{v}_l). \tag{5.1}$$

Then we can use rank aggregation techniques Dwork et al. [2001]; Ye et al. [2012b] to combine the resulting ranks. A very intuitive and straightforward approach is to use the weighted score vector

$$\mathbf{s} = \textstyle\sum_{i=1}^{m} w_i \mathbf{s}_i \tag{5.2}$$

and its induced ranking as in (5.1). This is known as the Borda count in social choice theory, and has been explored in Akata et al. [2013]; Mensink et al. [2014]; Norouzi et al. [2014] when no labeled training examples are given. In our later experiments, Borda works reasonably well. However, rank aggregation techniques can still be suboptimal, because the concept classifiers are obtained from other domains thus their accuracy on the test domain differs a lot. This motivates us to consider a recent approach that explicitly estimates the inaccuracy using *unlabeled* data.

5.1.5 Spectral Meta-Learning

Assuming for a moment that each score vector is binary, *i.e*. $s_i(\mathbf{v}) \in \{-1, 1\}$. We assume that the videos $\mathbf{v}$ are i.i.d. samples from an unknown distribution. The accuracy

mig/upload/MED_MER-FAQ_2014-07-30.pdf

[1]Different events may use different concepts. For notational clarity, throughout we omit the dependence on the event e.

of the i-th concept classifier is defined as follows[2]:

$$p_i = \Pr(s_i(\mathbf{v}) = 1 | y = 1), \tag{5.3}$$
$$n_i = \Pr(s_i(\mathbf{v}) = -1 | y = -1), \tag{5.4}$$
$$\pi_i = (p_i + n_i)/2, \tag{5.5}$$

where y is the true event label of the test video $\mathbf{v}$, and $\pi_i \in [0, 1]$ is the average accuracy. Since we do not have labeled data, it is not immediately clear how we can estimate π_i.

The following assumption is standard for estimating classifier accuracy using *unlabeled* data Dawid and Skene [1979]; Jaffe et al. [2015]; Parisi et al. [2014]:

Assumption 1 (Conditional Independence)**.**

$$\Pr(s_i(\mathbf{v}), s_j(\mathbf{v}) | y) = \Pr(s_i(\mathbf{v}) | y) \cdot \Pr(s_j(\mathbf{v}) | y) \tag{5.6}$$

In other words, given the label y, the classifiers make independent predictions. In our setting, the concept classifiers are trained from different sources, therefore the conditional independence assumption is reasonable.

Based on the conditional independence assumption, the following key observation is made in Parisi et al. [2014]:

Lemma 5.1. *Let $b = \Pr(y = 1) - \Pr(y = -1)$ be the class imbalance, $\mu_i = \mathbb{E}_{\mathbf{v}}(s_i(\mathbf{v}))$ be the mean prediction of the i-th concept classifier, and the population covariance matrix*

$$Q_{ij} = \mathbb{E}_{\mathbf{v}}[(s_i(\mathbf{v}) - \mu_i)(s_j(\mathbf{v}) - \mu_j)]. \tag{5.7}$$

Then, under the conditional independence assumption,

$$Q_{ij} = \begin{cases} 1 - \mu_i^2, & i = j \\ (2\pi_i - 1)(2\pi_j - 1)(1 - b^2), & i \neq j \end{cases}. \tag{5.8}$$

Crucially, from Lemma 5.1 we see that, except the diagonals, the population matrix Q arises from a rank-1 matrix, whose leading eigenvector $\mathbf{u}$ satisfies

$$u_i \propto (2\pi_i - 1). \tag{5.9}$$

[2]We implicitly assume that the scores are *positively* related to the label.

This immediately leads to a principled way to estimate the accuracies π_i (up to a scale factor), since the covariance matrix Q can be easily estimated using unlabeled data. Consider the sample covariance matrix

$$\hat{Q}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(s_i(\mathbf{v}_k) - \hat{\mu}_i)(s_j(\mathbf{v}_k) - \hat{\mu}_j), \tag{5.10}$$

where $\hat{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} s_i(\mathbf{v}_k)$. Clearly, $\hat{Q}$ is an unbiased estimator of the population covariance matrix Q, and it can be shown that $\|\hat{Q} - Q\| = O_p(\frac{1}{\sqrt{n}})$. Therefore, for a large number of unlabeled data, we can estimate the accuracy π_i by solving the following problem:

$$\min_{R\succeq 0} \quad \tfrac{1}{2}\sum_{i\neq j}(\hat{Q}_{ij} - R_{ij})^2 \tag{5.11}$$

$$\text{s.t.} \quad \operatorname{rank}(R) = 1. \tag{5.12}$$

Note that it is important to exclude the diagonals of Q. Indeed, as shown in Parisi et al. [2014], the leading eigenvector of Q is a *biased* estimator of the accuracy π_i, and the bias depends on the number of classifiers m and the class imbalance b.

Unfortunately, (5.11) is a non-convex problem hence may be hard to solve. Instead, we turn to the following alternative, which uses the trace (since R is constrained to be positive semidefinite) as a convex surrogate for the nonconvex rank constraint:

$$\min_{R\succeq 0} \ \tfrac{1}{2}\sum_{i\neq j}(\hat{Q}_{ij} - R_{ij})^2 + \lambda \operatorname{tr}(R). \tag{5.13}$$

The regularization constant λ controls the desired rank of the optimal solution. Parisi et al. [2014] proposed to solve (5.13) using generic semidefinite programming (SDP) toolboxes, which unfortunately do not scale very well. In Section 5.1.7 we will provide a much faster $O(m^2)$ time algorithm.

After solving R from (5.13), we extract the accuracy π_i from its leading eigenvector $\mathbf{u}$. Now the question is can we combine the classifiers more smartly by taking their accuracy into account? The answer is yes, and traces back to Dawid and Skene [1979], who considered the maximum likelihood estimator:

$$y^* = \operatorname{sign}\left[\sum_{i=1}^{m}(s_i(\mathbf{v})\log\alpha_i + \log\beta_i)\right], \tag{5.14}$$

$$\alpha_i = \frac{p_i n_i}{(1-p_i)(1-n_i)}, \quad \beta_i = \frac{p_i(1-p_i)}{n_i(1-n_i)}. \tag{5.15}$$

To get α and β from the accuracy π, Parisi et al. [2014] considered Taylor expansion of the MLE at the most inaccurate setting $p_i = n_i = 1/2$. This yields the spectral meta-learner (SML):

$$\hat{y} = \text{sign}\left[\sum_{i=1}^{m} s_i(\mathbf{v})(2\pi_i - 1)\right] \approx \text{sign}\left[\sum_{i=1}^{m} s_i(\mathbf{v})u_i\right], \tag{5.16}$$

where recall that $\mathbf{u}$ is the leading eigenvector of the minimizer R of (5.13). Interestingly, the spectral meta-learner is essentially a weighted majority voting rule, where the weights come from the estimates of the accuracy. Intuitively, it gives more weight to classifiers whose estimated accuracy is high, and vice versa. We note that it is possible to construct the meta-learner using more sophisticated tensor approaches Jaffe et al. [2015].

5.1.6 Specialization and Extension

In this section we specialize the spectral meta-learner above to our semantic event search framework.

Probabilistic classifiers. Recall that we obtain m concept classifiers from other domains and apply them to n unlabeled test videos, resulting in the score vectors $\mathbf{s}_i \in [-1, 1]^n, i = 1, \ldots, m$. The theory in section 5.1.5 requires $\mathbf{s}_i$ to be binary, but this can be easily addressed by treating each score vector $\mathbf{s}_i$ as *probabilistic* classifiers, namely, we classify the k-th test video as positive with probability $s_i(\mathbf{v}_k)$, independently of everything else. Under this interpretation we can still derive Lemma 5.1, the sample covariance $\hat{Q}$, and the spectral meta-learner as before, without the need of thresholding the score vectors.

Warping functions. Next, we wish to incorporate the relevance vector $\mathbf{w}$ that we constructed in section 5.1.2 and refined in section 5.1.3. To see why this is desirable, let us first note that Lemma 5.1 applies to *any* classifiers, as long as they satisfy the conditional independence assumption. More precisely, for transformations f_i that do not depend on the unseen test video $\mathbf{v}$ or its unknown label y, we can consider the "warped" classifiers

$$t_i(\mathbf{v}) = f_i(s_i(\mathbf{v})), \quad i = 1, \ldots, m. \tag{5.17}$$

Clearly, the warped classifiers $\mathbf{t}$ are conditionally independent if and only if the original classifiers $\mathbf{s}$ are so. Therefore Lemma 5.1 still holds, and we can construct the sample covariance matrix

$$\hat{Q}_{ij}^f = \frac{1}{n-1}\sum_{k=1}^{n}(t_i(\mathbf{v}_k) - \hat{\mu}_i^f)(t_j(\mathbf{v}_k) - \hat{\mu}_j^f), \tag{5.18}$$

where as before $\hat{\mu}_i^f = \frac{1}{n}\sum_{k=1}^{n} t_i(\mathbf{v}_k)$. The spectral meta-learner for the warped classifiers is thus given as:

$$\hat{y}^f = \text{sign}\Big[\sum_{i=1}^{m} f_i(s_i(\mathbf{v}))u_i\Big] \tag{5.19}$$

where $\mathbf{u}$ is the leading eigenvector of R, the minimizer of (5.13) where we use $\hat{Q}_{ij}^f$ instead.

Warped spectral meta-learner. Straightforward as it is, the extension using different warping functions f_i can lead to a significant performance improvement. This is because the accuracy of the spectral meta-learner $\hat{y}$ in (5.16) depends on the accuracies of the base classifiers s_i: SML is a smart way to combine the base classifiers, but we should not expect it to improve the accuracy much if the base classifiers are themselves near random. After all, garbage in garbage out. The warping functions f_i provide an extremely simple way to adjust the base classifiers. Since the relevance vector $\mathbf{w}$ we constructed in Section 5.1.2 provides a crude assessment of the relevance between the concept classifiers and the event of interest, we consider the following warped concept classifiers:

$$\mathbf{t} = (w_1 s_1, \ldots, w_m s_m), \tag{5.20}$$

although other warping functions can similarly be used. Intuitively, the weight w_i is the *a priori* co-occurrence frequency of the i-th concept and the event of interest while s_i is the confidence *likelihood* of detecting the i-th concept. As we will see in the experiments, this simple warping trick significantly improves the performance.

Few exemplars. The warped spectral meta-learner above can also be applied for few-exemplar event detection, where few (say 10) labeled training videos are provided. In this case, we can train an additional classifier (or few) using the provided labeled videos. Due to the small training size, the accuracy of the resulting supervised classifier is likely also low. We combine the supervised classifier with the concept classifiers but give it the maximum weight $w = 1$. Then we apply the warped spectral learner to get the final prediction.

Algorithm 4: The warped SML algorithm.

Construct concept classifiers $\mathbf{s}$ and relevance vector $\mathbf{w}$.
Apply warping $\mathbf{t} = (f_1(s_1), \ldots, f_m(s_m))$.
Assemble the sample covariance $\hat{Q}^f$.
Set $U_1 = \mathbf{0}$.
for $t = 1, 2, \ldots$ **do**
    $R \leftarrow U_t U_t^\top$;
    $G_{ij} \leftarrow \begin{cases} 0, & i = j \\ R_{ij} - \hat{Q}_{ij}, & i \neq j \end{cases}$;
    $\mathbf{u} \leftarrow$ leading eigenvector of $-G$;
    $(a_t, b_t) \leftarrow \arg\min_{a,b \geq 0} \frac{1}{2} \sum_{i \neq j} (aR_{ij} + bu_i u_j - \hat{Q}_{ij})^2 + \lambda(a \operatorname{tr}(R) + b)$;
    $U_{\text{init}} \leftarrow (\sqrt{a_t} U_{t-1},\ \sqrt{b_t}\mathbf{u})$;
    $U_t \leftarrow$ local minimizer of (5.24), initial with U_{init} ;
$\mathbf{u} \leftarrow$ leading eigenvector of R ;
Rank test videos using (5.19).

5.1.7 Optimization using GCG

Lastly, we provide a fast algorithm for solving the semidefinite program (5.13). This is crucial if we want to combine a large number of concept classifiers.

We use the generalized conditional gradient (GCG, *a.k.a* Frank-Wolfe) algorithm in Zhang et al. [2012], with essential modifications to take the positive semidefinite constraint into account. In each iteration, GCG first computes the gradient

$$G = \nabla_R \left[\tfrac{1}{2} \textstyle\sum_{i \neq j} (R_{ij} - \hat{Q}_{ij})^2 \right]. \tag{5.21}$$

Then it finds a rank-1 update

$$\mathbf{u} = \arg \min_{\|\mathbf{z}\|_2 \leq 1} \mathbf{z}^\top G \mathbf{z}, \tag{5.22}$$

which amounts to the leading eigenvector of $-G$. This step takes into account the trace regularizer, and is essentially its dual operator (spectral norm). Finally, GCG augments the previous iterate R with the new rank-1 update:

$$R \leftarrow a \cdot R + b \cdot \mathbf{u}\mathbf{u}^\top, \tag{5.23}$$

where the positive coefficients a, b are found by line search.

To accelerate convergence, we consider the following smooth *unconstrained* problem:

Table 5.1: Experiment results for semantic event search on the TRECVID MEDTest 2014 dataset. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

	MEDTest 2014											
ID	Sel	Bi	OR	Fu	Bor	Bor+	SML	SML+	wBor	wBor+	wSML	wSML+
E021	2.98	2.64	3.89	3.97	2.67	3.46	4.29	5.14	3.12	4.64	5.37	6.48
E022	0.97	0.83	1.36	1.49	0.76	0.82	1.01	1.58	1.15	1.48	1.85	2.43
E023	36.94	35.23	39.18	40.87	35.65	37.22	38.19	41.73	38.68	41.78	43.26	50.55
E024	3.75	3.02	4.66	4.92	2.98	3.26	3.84	4.02	4.11	4.87	5.12	5.69
E025	0.76	0.62	0.97	1.39	0.52	0.75	0.92	1.06	0.84	1.01	1.26	1.43
E026	1.59	1.32	2.41	2.96	1.03	1.84	2.48	3.11	1.96	2.65	3.23	3.74
E027	13.64	12.48	15.93	16.26	12.52	12.73	13.96	15.62	15.12	16.47	18.63	20.56
E028	0.67	1.06	1.57	1.95	0.75	1.46	2.51	3.28	1.72	2.25	3.04	4.56
E029	10.68	12.21	14.01	14.85	9.64	10.25	11.93	13.48	13.19	14.75	16.69	18.84
E030	0.63	0.48	0.91	0.96	0.21	0.32	0.38	0.45	0.36	0.48	0.52	0.67
E031	53.19	45.87	69.52	69.66	54.29	61.82	65.75	70.43	67.49	72.64	76.45	82.86
E032	5.88	4.37	8.12	8.45	4.69	5.23	7.31	8.96	7.54	8.65	10.38	11.65
E033	20.19	18.54	22.14	22.23	17.66	18.71	19.49	22.04	21.53	23.26	25.64	28.93
E034	0.47	0.41	0.71	0.75	0.32	0.48	0.69	0.87	0.53	0.76	0.94	1.25
E035	13.28	11.09	16.53	16.68	12.74	14.95	16.28	19.49	15.82	18.65	20.78	25.39
E036	2.63	2.14	3.15	3.39	1.98	2.29	2.92	3.85	2.88	3.76	4.47	5.36
E037	4.52	3.81	6.84	6.88	3.26	4.19	5.33	6.41	5.42	6.83	7.45	8.68
E038	0.74	0.58	0.99	1.16	0.57	0.74	1.26	1.93	0.85	1.12	1.89	2.67
E039	0.57	0.42	0.69	0.77	0.45	0.58	0.83	1.02	0.64	0.85	1.26	1.73
E040	0.98	0.72	1.57	1.57	0.86	1.23	1.57	1.98	1.24	1.76	2.12	2.56
mean	9.55	7.89	10.76	11.05	8.18	9.12	10.05	11.33	10.21	11.44	12.52	14.32

$$\min_{U} \tfrac{1}{2} \sum_{i \neq j} ((UU^\top)_{ij} - \hat{Q}_{ij})^2 + \lambda \|U\|_{\mathsf{F}}^2, \tag{5.24}$$

which, unlike the original problem (5.13), is nonconvex. But we can combine the global GCG algorithm with a local fast solver for the nonconvex problem (5.24). The intuition is that both (5.13) and (5.24) share the same set of global minimizers, and by combining them we gain both global optimality and local fast convergence, especially because the latter nonconvex problem has no constraint at all. We summarize the entire procedure in Algorithm 4. Following a similar argument as in Zhang et al. [2012], we can prove that Algorithm 4 converges globally to an ϵ-optimal solution of (5.13) in at most $O(1/\epsilon)$ steps. In practice, once we arrive at the true rank of the minimizer, the local solver (*e.g.* lbfgs) for the nonconvex problem (5.24) usually finds the solution at once. The per-step time complexity is $O(m^2)$ since the most time-consuming step is computing the rank-1 update in (5.22).

5.2 Experimental Results

In this section we conduct thorough experiments to validate our warped spectral meta-learner for both the semantic event search and few-exemplar event detection tasks.

Table 5.2: Experiment results for semantic event search on the TRECVID MEDTest 2013 dataset. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

	MEDTest 2013											
ID	Selec	Bi	OR	Fu	Bor	Bor+	SML	SML+	wBor	wBor+	wSML	wSML+
E006	4.92	4.69	7.63	7.98	4.86	7.12	10.76	12.48	6.41	8.32	13.17	17.23
E007	1.14	0.86	1.76	2.34	0.87	1.16	1.79	2.25	1.44	1.62	2.43	3.84
E008	33.02	18.96	41.88	42.25	31.87	35.24	38.49	45.36	38.09	43.14	48.82	53.45
E009	13.43	13.12	15.48	16.11	11.26	12.41	14.27	15.13	14.68	16.29	17.84	19.62
E010	0.85	0.82	0.93	0.99	0.42	0.50	0.87	1.03	0.74	0.91	1.42	1.67
E011	7.68	7.39	7.91	8.15	4.39	4.98	5.47	7.12	6.58	7.36	8.46	9.83
E012	21.96	19.36	22.45	22.64	14.68	15.21	19.83	22.44	18.42	24.33	34.18	36.59
E013	0.52	0.91	2.14	2.69	0.87	1.05	1.87	2.06	1.61	2.82	3.66	4.84
E014	1.23	0.94	2.51	2.98	1.12	1.84	2.27	3.29	1.72	3.44	4.11	5.93
E015	1.41	1.44	1.49	1.75	0.19	0.39	0.62	0.98	0.47	0.54	0.93	1.26
E021	2.98	2.64	3.89	3.97	2.67	3.46	4.29	5.14	3.12	4.64	5.37	6.48
E022	0.97	0.83	1.36	1.49	0.76	0.82	1.01	1.58	1.15	1.48	1.85	2.43
E023	36.94	35.23	39.18	40.87	35.65	37.22	38.19	41.73	38.68	41.78	43.26	50.55
E024	3.75	3.02	4.66	4.92	2.98	3.26	3.84	4.02	4.11	4.87	5.12	5.69
E025	0.76	0.62	0.97	1.39	0.52	0.75	0.92	1.06	0.84	1.01	1.26	1.43
E026	1.59	1.32	2.41	2.96	1.03	1.84	2.48	3.11	1.96	2.65	3.23	3.74
E027	13.64	12.48	15.93	16.26	12.52	12.73	13.96	15.62	15.12	16.47	18.63	20.56
E028	0.67	1.06	1.57	1.95	0.75	1.46	2.51	3.28	1.72	2.25	3.04	4.56
E029	10.68	12.21	14.01	14.85	9.64	10.25	11.93	13.48	13.19	14.75	16.69	18.84
E030	0.63	0.48	0.91	0.96	0.21	0.32	0.38	0.45	0.36	0.48	0.52	0.67
mean	7.94	6.92	9.45	9.88	6.86	7.61	8.79	10.08	8.43	9.96	11.64	13.46

5.2.1 Speed comparison on synthetic data

We first verify the efficiency of the GCG Algorithm 4. We randomly generate m score vectors $\mathbf{s}_i \in \mathbb{R}^n, i = 1, \ldots, m$, and vary m from $m = 2$ to $m = 100$ (largest we were able to try). As can be seen from Figure 5.2, the running time of the naive SDP implementation (using YALMIP) increased sharply with the number of concepts. In comparison, the running time of our GCG implementation remains negligible (when achieving the same stopping criteria). It is clear that without our efficient GCG implementation, it is impossible to apply the (warped) spectral meta-learner to the large video datasets in the next section.

5.2.2 Experiment setup on real datasets

Datasets. We run experiments on three real datasets:

- MED14: The TRECVID MEDTest 2014 dataset NIST [2014] is collected by the NIST for all participants in the TRECVID competition. There are in total 20 events, whose description can be found in NIST [2014]. We use the official test split released by the NIST, and strictly follow its standard procedure NIST [2014]. In particular, we detect each event *separately*, treating each of them as a binary classification/ranking problem.

Table 5.3: Experiment results for semantic event search on the CCV_{sub} dataset. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

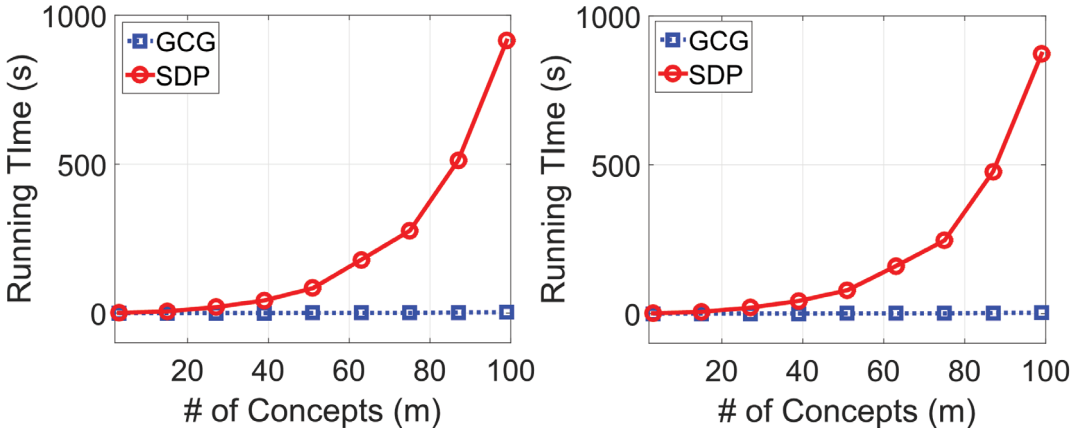
	CCV_{sub}											
ID	Selec	Bi	OR	Fu	Bor	Bor+	SML	SML+	wBor	wBor+	wSML	wSML+
1	25.98	25.53	28.16	28.54	26.36	26.98	27.34	28.12	27.49	28.32	29.86	31.23
2	17.74	18.94	20.26	21.16	20.43	20.86	22.55	23.12	24.17	24.68	24.96	25.28
3	17.16	17.85	18.56	19.35	18.54	19.25	19.87	20.36	19.53	20.85	21.64	22.87
4	28.98	30.26	31.42	31.87	30.98	31.27	32.76	33.19	31.94	32.92	33.48	34.16
5	29.86	31.84	32.15	32.84	32.03	32.92	33.27	33.82	32.14	32.89	33.63	34.89
6	31.13	31.46	31.98	32.54	31.63	33.05	33.47	34.12	33.26	34.19	34.88	34.95
7	21.45	21.98	22.05	22.47	22.64	23.33	24.04	24.58	25.37	25.84	26.11	26.95
8	14.75	16.38	16.94	17.26	16.96	17.85	18.29	19.53	19.92	20.08	20.17	20.83
9	10.87	11.95	13.07	14.42	13.84	14.66	15.34	16.18	15.75	16.42	16.85	17.32
10	17.33	18.12	18.86	19.59	21.24	21.98	17.56	18.43	18.97	19.84	21.13	23.65
11	14.52	15.17	16.43	16.88	16.12	17.43	17.92	18.54	18.29	19.37	21.12	22.38
12	18.87	19.49	20.66	21.34	20.13	22.42	23.16	24.45	24.09	24.68	25.62	26.81
13	17.13	17.95	18.59	20.04	18.13	19.95	21.14	22.08	21.57	22.49	23.78	24.26
14	14.86	15.39	16.17	17.58	16.39	17.64	17.98	18.36	18.13	19.46	20.84	21.37
15	10.38	11.49	12.09	12.48	12.37	13.89	15.18	16.43	15.54	16.02	16.59	16.91
mean	19.40	20.25	21.16	21.89	21.19	22.23	22.66	23.42	23.08	23.87	24.71	25.59

- MED13 NIST [2013]: Similar to MED14. Note that 10 of its 20 events overlap with those of MED14.
- CCV_{sub}: The official Columbia Consumer Video dataset Jiang et al. [2011] contains 9,317 videos in 20 semantic classes, including scenes like "beach," objects like "cat," and events like "baseball" and "parade." Since our goal is to *detect events*, we only use the 15 event categories.

We evaluate the performance using the mean Average Precision (mAP). Parameters of all compared algorithms are similarly tuned by grid search.

Concept detectors. We pre-trained 3,135 concept detectors, using TRECVID SIN dataset (346 categories), Google sports (478 categories) Karpathy et al. [2014], UCF101 dataset (101 categories) Soomro et al. [2012], YFCC dataset (609 categories) YFC and DIY dataset (1601 categories) Yu et al. [2014a]. We first extracted the improved dense trajectory features (including trajectory, HOG, HOF and MBH) using the code of Wang and Schmid [2013] and encode them with the Fisher vector representation Perronnin et al. [2010]. Following Wang and Schmid [2013], we first reduce the dimension of each descriptor by a factor of 2 and then use 256 components to generate the Fisher vectors. Then, on top of the extracted low-level features, we trained the cascade SVM Graf et al. [2004] for each concept detector. Using these concept detectors we obtain a 3,135-dimensional score vector for each video.

Competitors. We compare the following algorithms:



(a) λ = 1e-3 (b) λ = 1e-2

Figure 5.2: Efficiency comparison between GCG and SDP.

- Sel Mazloom et al. [2013a]: A subset of primitive concepts that are more informative for each event.
- Bi Rastegari et al. [2013]: Bi-concepts discovered in Rastegari et al. [2013].
- OR Habibian et al. [2014a]: Boolean OR combinations of Prim concepts.
- Fu Habibian et al. [2014a]: Boolean AND/OR combinations of Prim concepts.
- Bor: The Borda rank aggregation in (5.2), with equal weights on the discovered semantic concepts.
- Bor+: Borda rank aggregation with equal weights on the refined semantic concepts.
- wBor: Borda rank aggregation with relevance weights on the discovered semantic concepts.
- wBor+: Borda rank aggregation with relevance weights on the refined semantic concepts.
- SML: Spectral meta-learner (5.16) on discovered semantic concepts.
- SML+: SML on refined semantic concepts.
- wSML: Warped SML (5.19) (with warping function (5.20)) on discovered semantic concepts.
- wSML+: Warped SML on refined semantic concepts.

The last eight methods are first proposed here. The refined concepts are subset of discovered semantic concepts, after dropping inaccurate, low predictive, and irrelevant ones.

Note that we did not compare with approaches that use multiple modalities of features, *e.g*. Jiang et al. [2014b]; Wu et al. [2014], since we only considered the visual feature. In future work we plan to exploit speech and OCR information.

5.2.3 Semantic Event Search

We report the experimental results on the TRECVID MEDTest 2014 dataset in Table 5.1, the experimental results on the TRECVID MEDTest 2013 dataset in Table 5.2 and the result on the CCV_{sub} dataset in Table 5.3. We first consider the semantic event search setting where no labeled training video is available. As is clear from Tables 5.1 to 5.3, the proposed methods (last eight columns) compare favorably against existing alternatives (first four columns), with a large margin obtained by the most sophisticated method wSML+ (14.32% vs the second best 10.76% achieved by OR) . The improvements are particularly impressive on some events, including *Dog Show (E23)*, *Rock Climbing (E27)*, *Beekeeping (E31)* and *Non-motorized vehicle repair (E33)*. By further looking into the discovered semantic concepts for these events, we find they all benefit greatly from relevant classifiers that are discriminative and reliable. For example, for the *Beekeeping* event, the performance of wSML+ significantly relied on concepts such as "apiary bee house" and "honeycomb", which turn out to be the most reliable concepts for *Beekeeping* in our concept vocabulary. Figure 5.3 shows the top 9 retrieved videos for the *Beekeeping* event. It is clear that videos retrieved by the proposed wSML+ are more accurate and visually coherent.

From Table 5.1 we make the following observations:

- Comparing the columns w/o the "+" suffix, we can see that concept refinement generally improves performance than naively using all concepts. This confirms the importance of using data-driven word embeddings to eliminate irrelevant and non-discriminative concepts.
- Comparing the Border and SML variants we verify the great benefit of using a principled method such as SML to combine classifiers. By taking into account the accuracy, albeit being estimated using unlabeled data, SML achieves better performance than simple majority voting.
- Comparing the columns w/o the "w" prefix, we observe that warping through the relevance significantly improves performance. This confirms the necessity to improve base classifiers, and illustrates the limitation of SML.

Similar conclusions can also be made from the results on MEDTest 2013 dataset and CCV_{sub} dataset.

5.2.4 Few-exemplar event detection

As mentioned in Section 5.1.6, our semantic event search framework can also be used for few-exemplar event detection: We simply combine the concept classifiers and the supervised classifiers using the warped spectral meta-learner (with maximum weight for the latter). In this section, we demonstrate the benefit of this hybrid approach. Table 5.4 summarizes the mAP on both the MEDTest 2014 and 2013 datasets, while Figure 5.4 compares the performance event-wise.

Figure 5.3: Top ranked videos for the event *Beekeeping*. From top to below: Sel (AP: 53.19), Bi (AP: 45.87), OR (AP: 69.52), and wSML+ (AP: 82.86). True/false labels (provided by NIST) are marked in the lower-right of each frame.

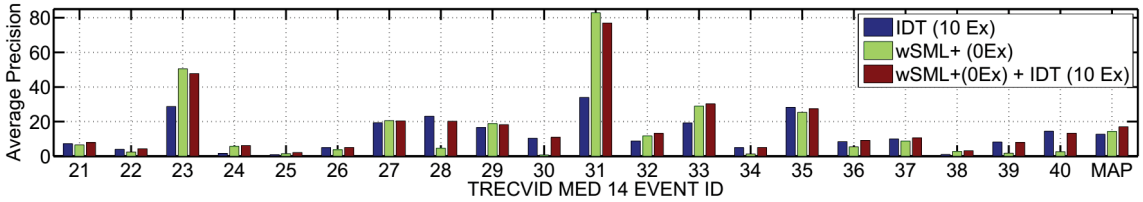


Figure 5.4: Performance comparison of IDT, wSML+, and the hybrid of IDT and wSML+.

As a baseline (denoted as IDT), we considered training an SVM classifier using the improved dense trajectory features Wang and Schmid [2013] on 10 positive examples. Interestingly, this supervised classifier performed even slightly worse than our wSML+ which had no access to labeled data at all (mAP 13.92% vs 14.32%). This clearly demonstrates that a large pool of unsupervised but relevant concept classifiers, after proper correction of their accuracy, can outperform supervised classifiers trained with limited positive samples. As expected, with more training exemplars (100Ex), IDT can achieve a much higher 27.6% mAP on MEDTest 2014.

Figure 5.4 gives a more detailed view of comparison: the supervised IDT is largely outperformed by our wSML+ on events E23, E24, E31, and E33, while the converse is observed on events E28, E30, E39, E40. In particular, on the event *Beekeeping* (E31), wSML+ improved IDT more than 2x (82.86% vs 33.9%). As mentioned before, this is because wSML+ significantly benefited from the presence of informative and reliable concepts such as "apiary bee house" and "honeycomb" on the particular *Beekeeping* event.

Finally we combine IDT with wSML+ as described before. This again significantly

Table 5.4: Few-exemplar results on MED14 and MED13.

IDT (10Ex)	13.92	18.08
wSML+ (0Ex) + IDT (10Ex)	16.98	19.65
CNN (10Ex) Xu et al. [2015]	24.46	29.84
wSML+ (0Ex) + CNN (10Ex)	25.82	31.05

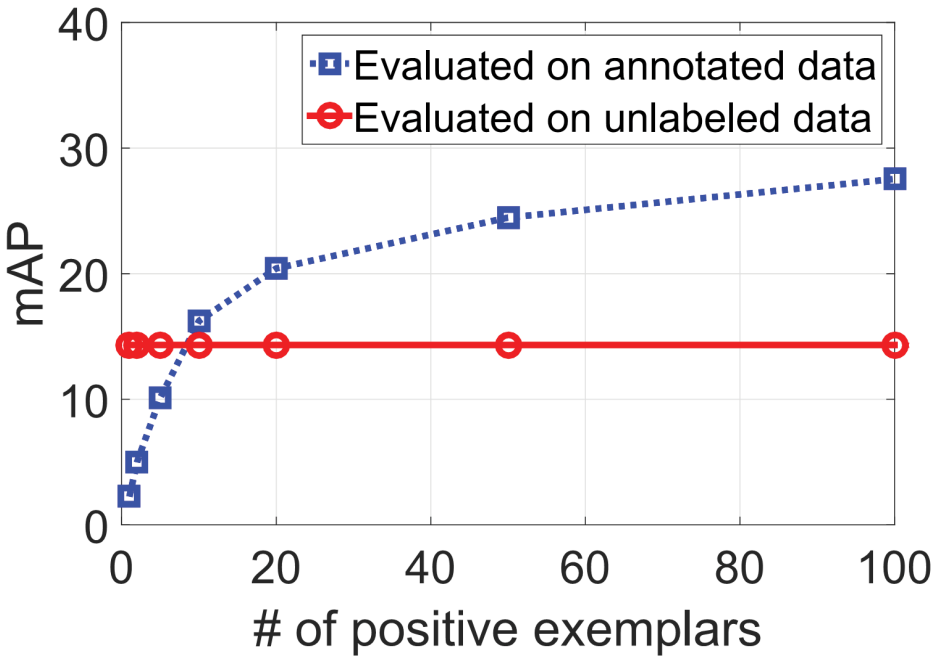


Figure 5.5: mAPs with increasing number of annotated pairs.

improves the performance from 13.92% to 16.98% on the MEDTest 2014 dataset. We also tried to combine with the state-of-the-art algorithm in Xu et al. [2015], and increased its performance from 24.46% to 25.82%. As expected, the gain obtained from such simple hybrid diminishes when combining with more sophisticated methods. Overall, the results clearly demonstrate the utility of our framework even in the few-exemplar setting.

5.2.5 Annotated vs. unannotated data

We also compared the unsupervised SML approach with a supervised approach as follows. For each event we randomly select k labeled data from the MED14 training set, which are used to estimate the accuracy of the concept classifiers (*i.e.*, estimate p_i and n_i in (5.3)). Then we plug the estimates to the MLE (5.14) and obtain predictions on the test set. As can be seen from Figure 5.5, the unsupervised SML approach is advantageous when roughly 8 or less labeled training videos are used to estimate the

concept accuracies. This experiment again demonstrates the utility of our method when the number of labeled training data is limited.

5.3 Summary of This Chapter

To address the challenging task of semantic event search and few-exemplar event detection, we proposed to leverage on concept classifiers collected from other sources. Data-driven word embedding models were used to seek the relevance of the concepts to the event of interest. To further account for the unreliability of the concept classifiers, we extended the recent spectral meta-learner that combines the classifiers based on a principled estimate of their accuracies using unlabeled data. Efficient implementations were provided and promising experimental results were obtained on three real video datasets.

Chapter 6

Conclusion and Future Work

In this section, we first conclude the entire thesis, and then elaborate the possible trends for future research.

6.1 Conclusion

This thesis addresses the problem of complex event detection in unconstrained videos. According to the number of positive training exemplars, MED can be grouped into 100Ex (complex event detection), 10Ex (few-exemplar event detection) and 0Ex (zero-exemplar event detection). Since semantic concept detectors are easy to obtain in the real-world, we designed various approaches to use semantic concepts to boost the performance of MED. Specifically, we proposed a semantic pooling approach for 100Ex, a joint framework of event detection and recounting for 10Ex, two semantic event search approach for 0Ex.

In **SP** (Chapter 2), we proposed to prioritize the video shots using a novel notion of semantic saliency based on the observation that not all video shots are equally relevant to an event of interest. Through a suitable isotonic regularizer we designed the "informed" nearly-isotonic SVM classifier (NI-SVM) that is able to exploit the carefully constructed ordering information. An efficient proximal gradient implementation, with new and closed-form proximal steps, is developed. We further extend NI-SVM to the multi-class setting to perform event recognition. Extensive experiments on three real video datasets are conducted to validate the proposed algorithm on video analysis tasks such as event detection, recognition and recounting.

In **JEDaR** (Chapter 3), we proposed a novel joint training protocol to simultaneously conduct event detection and recounting. Based on a noisy semantic video representation, we couple a recounting model which aims at localizing the key evidences with a detection model which aims at enhancing the discriminative power.

The recounting model assists detection by filtering out noisy irrelevant information which simultaneously the detection model guides recounting b directing it to the most discriminative evidences, conceptual-wise and temporal-wise. The two models are optimized jointly hence benefit greatly from each other. To address the computational challenge due to operating on the shot level, we significantly improve the existing ADMM algorithm by proving closed-form solutions for the intermediate proximal updates. Augmented with the Uzawa Linearization trick we are able to remove all inner loops in previous ADMM implementations, without affecting the convergence properties at all. The proposed method was tested on the large-scale and challenging TRECVID MEDTest 2013 and 2014 datasets, achieving very promising results in both detection and recounting.

To address the challenging task of zero-exemplar event detection, in **EDCW** (Chapter 4), we proposed to learn the optimal weights for each testing video by exploring the collected videos from other sources. Data-driven word embedding models were used to seek the relevance of the concepts to the event of interest. To further derivate the optimal weights of the concept classifiers for each testing video, we proposed a novel event-driven concept weighting method by exploiting the textual information of the collected videos. The experimental results confirm the efficiency of the proposed approach.

To further account for the unreliability of the concept classifiers, in **DCC** (Chapter 5), we extended the recent spectral meta-learner that combines the classifiers based on a principled estimate of their accuracies using unlabeled data. Efficient implementations were provided and promising experimental results were obtained on three real video datasets.

6.2 Future Work

We have studied how semantic concept detectors help MED in terms of 100Ex, 10Ex and 0Ex. Our work suggests that proper usage of semantic concepts does help improve the overall performance of MED. Hence, in the future, we will continue our research on this direction with the following possible pursuits:

- **Fine-Grained Evidence Recounting**: In the current framework, the minimal evidence is frame-level. Intuitively, fine-granularity benefits from learning the particular object of interest. However, this will result in increasing computational burden. Hence, it would be interesting to study on novel approaches that are capable of learning fine-grained evidence for multimedia event detection.
- **Curriculum**: Curriculum Learning has attracted much research attention. It aligns with the cognitive science literature, where humans and animals are found to start

with learning easier examples, and then gradually take more complex examples into consideration. In the future, we plan to incorporate this idea into the evidence recounting protocol, which learns the evidence from simple to complex.

- **Temporal Information**: In the current framework for 0Ex, we did not consider temporal information. Inspired by the great success of temporal information in 100Ex and 10Ex, we plan to conduct semantic event search while taking temporal information into consideration. Multiple instance learning (MIL) might be an interesting toolkit for solving this problem.
- **More Applications**: It will be valuable if we can apply the proposed approach to other practical applications. Although the experimental results reveal that by properly utilizing semantic concepts with the proposed algorithm promising results have been achieved in MED, we believe that our algorithms are also applicable to some other applications, such as social media analysis, time-series problems, and *etc*.

References

The YFCC dataset. http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67.

Shivani Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SDM*, pages 839–850, 2011.

Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

Robin Aly, Relja Arandjelovic, Ken Chatfield, Matthijs Douze, Basura Fernando, Zaid Harchaoui, and Kevin McGuinness et al. The axes submissions at Trecvid 2013. 2013.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. John Wiley & Sons, 1972.

Subhabrata Bhattacharya, Felix X. Yu, and Shih-Fu Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *ICMR*, 2014.

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

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2): 459–494, 2014a.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming, Series A*, 146:459–494, 2014b.

Ethem F. Can and R. Manmatha. Modeling concept dependencies for event detection. In *ICMR*, 2014.

Liangliang Cao, Yadong Mu, Apostol Natsev, Shih-Fu Chang, Gang Hua, and John R. Smith. Scene aligned pooling for complex video recognition. In *ECCV*, 2012.

Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, Eric P. Xing, and Yaoliang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015a.

Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Alexander G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, 2015b.

Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016.

Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001a.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001b.

Stéphane Clinchant and Florent Perronnin. Textual similarity with a bag-of-embedded-words model. In *ICTIR*, 2013.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Yin Cui, Dong Liu, Jiawei Chen, and Shih-Fu Chang. Building A large concept bank for representing events in video. *CoRR*, abs/1403.7591, 2014.

Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013.

P. Das, Chenliang Xu, R.F. Doell, and J.J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013.

P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29(1):1–65, 2001.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, 2001.

Jonathan Eckstein and Dimitri P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.*, 55:293–318, 1992.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981.

Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, 2015.

Nikolaos Gkalelis and Vasileios Mezaris. Video event detection using generalized subclass discriminant analysis and linear support vector machines. In *ICMR*, 2014.

Hans Peter Graf, Eric Cosatto, Léon Bottou, Igor Durdanovic, and Vladimir Vapnik. Parallel support vector machines: The cascade SVM. In *NIPS*, 2004.

Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009a.

Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry. S. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009b.

Amirhossein Habibian, Koen E.A. van de Sande, and Cees G.M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.

AmirHossein Habibian, Thomas Mensink, and Cees G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014a.

AmirHossein Habibian, Thomas Mensink, and Cees G. M. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 2014b.

A. Hauptmann, Rong Yan, Wei-Hao Lin, M. Christel, and Howard Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

Alexander G. Hauptmann. Lessons for the future from a decade of informedia video analysis research. In *CIVR*, 2005.

Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50 (2):700–709, 2012.

Sung Ju Hwang, Fei Sha, and Kristen Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011.

Nazli Ikizler-Cinbis and Stan Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

Ariel Jaffe, Boaz Nadler, and Yuval Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *AISTATS*, 2015.

Lu Jiang, Alexander G. Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, 2012.

Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, and Alexander G. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014a.

Lu Jiang, Teruko Mitamura, Shoou-I Yu, and Alexander G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014b.

Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, and Alexander G. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. In *ACM MM*, 2015.

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel P. W. Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.

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

Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. 1987.

Kuan-Ting Lai, Dong Liu, Ming-Syan Chen, and Shih-Fu Chang. Recognizing complex events in videos by learning key static-dynamic evidences. In *ECCV*, 2014.

Kuan-Ting Lai, Dong Liu, Shih-Fu Chang, and Ming-Syan Chen. Learning sample specific weights for late fusion. *IEEE Transactions on Image Processing*, 24(9): 2772–2783, 2015.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

Weixin Li, Qian Yu, Ajay Divakaran, and Nuno Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013.

Dong Liu, Kuan-Ting Lai, Guangnan Ye, Ming-Syan Chen, and Shih-Fu Chang. Sample-specific late fusion for visual category recognition. In *CVPR*, 2013a.

Gaowen Liu, Yan Yan, Chenqiang Gao, Wei Tong, Alexander G. Hauptmann, and Nicu Sebe. The mystery of faces: Investigating face contribution for multimedia event detection. In *ICMR*, 2014.

Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Video event recognition using concept attributes. In *WACV*, 2013b.

Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet S. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013c.

Zhigang Ma, Yi Yang, Nicu Sebe, Kai Zheng, and Alexander G. Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia*, 15(7):1628–1637, 2013a.

Zhigang Ma, Yi Yang, Zhongwen Xu, Nicu Sebe, and Alexander G. Hauptmann. We are not equally negative: fine-grained labeling for multimedia event detection. In *ACM MM*, 2013b.

Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, N. Sebe, and A.G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013c.

Zhigang Ma, Yi Yang, Nicu Sebe, and Alexander G. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1789–1802, 2014.

Masoud Mazloom, Efstratios Gavves, Koen E. A. van de Sande, and Cees Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013a.

Masoud Mazloom, Amirhossein Habibian, and Cees G.M. Snoek. Querying for video events by semantic signatures from few examples. In *ACM MM*, 2013b.

Masoud Mazloom, AmirHossein Habibian, Dong Liu, Cees G. M. Snoek, and Shih-Fu Chang. Encoding concept prototypes for video event detection and summarization. In *ACM ICMR*, 2015.

Thomas Mensink, Efstratios Gavves, and Cees G. M. Snoek. COSTA: co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013b.

Bingbing Ni, Yong Pei, Pierre Moulin, and Shuicheng Yan. Multilevel depth and image fusion for human activity detection. *IEEE Trans. Cybernetics*, 43(5):1383–1394, 2013.

Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *KDD*, 2014.

NIST. The Trecvid MED 2013 dataset. `http://nist.gov/itl/iad/mig/med13.cfm`, 2013.

NIST. The Trecvid MED 2014 dataset. `http://nist.gov/itl/iad/mig/med14.cfm`, 2014.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Queenot. Trecvid 2014 an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2014.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111:1253–1258, 2014.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

Xueming Qian, Xian-Sheng Hua, Yuan Yan Tang, and Tao Mei. Social image tagging with diverse semantics. *IEEE Trans. Cybernetics*, 44(12):2493–2508, 2014.

Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for $l_{1,\text{infinity}}$ regularization. In *ICML*, 2009.

Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, 2010.

Alain Rakotomamonjy. Sparse support vector infinite push. In *ICML*, 2012.

Mohammad Rastegari, Ali Diba, Devi Parikh, and Ali Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013.

Eleanor Roach and Barbara B Lloyd. Cognition and categorization. 1978.

Ralph Tyrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 1998.

Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, 2009.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

Ling Shao, Xiantong Zhen, Dacheng Tao, and Xuelong Li. Spatio-temporal laplacian pyramid coding for action recognition. *IEEE Trans. Cybernetics*, 44(6):817–827, 2014.

Kimiaki Shirahama, Marcin Grzegorzek, and Kuniaki Uehara. Multimedia event detection using hidden conditional random fields. In *ICMR*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012.

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

Chen Sun, Brian Burns, Ram Nevatia, Cees Snoek, Bob Bolles, Gregory K. Myers, Wen Wang, and Eric Yeh. ISOMER: informative segment observations for multimedia event recounting. In *ICMR*, 2014.

Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

Chun Chet Tan, Yu-Gang Jiang, and Chong-Wah Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *ACM MM*, 2011a.

Chun Chet Tan, Yu-Gang Jiang, and Chong-Wah Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *ACM MM*, 2011b.

Kevin Tang, Fei-Fei Li, and Daphne Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

Kevin Tang, Bangpeng Yao, Fei-Fei Li, and Daphne Koller. Combining the right features for complex event recognition. In *ICCV*, 2013.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *CoRR*, 2015.

Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

Chun-Yu Tsai, Michelle L. Alexander, Nnenna Okwara, and John R. Kender. Highly efficient multimedia event recounting from user semantic preferences. In *ICMR*, 2014.

Christos Tzelepis, Nikolaos Gkalelis, Vasileios Mezaris, and Ioannis Kompatsiaris. Improving event detection using related videos and relevance degree support vector machines. In *ACM MM*, 2013.

Arash Vahdat, Kevin Cannons, Greg Mori, Sangmin Oh, and Ilseo Kim. Compositional models for video event detection: A multiple kernel learning latent variable approach. In *ICCV*, 2013.

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

Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *ACM MM*, 2010.

Gang Wang and David A. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009.

Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014.

Zhongwen Xu, Ivor W. Tsang, Yi Yang, Zhigang Ma, and Alexander G. Hauptmann. Event detection using multi-level relevance labels and multiple features. In *CVPR*, 2014.

Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *CVPR*, 2015.

Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. Feature grouping and selection over an undirected graph. In *KDD*, 2012.

Yang Yang and Mubarak Shah. Complex events detection using data-driven concepts. In *ECCV*, 2012.

Yi Yang, Zhigang Ma, Zhongwen Xu, Shuicheng Yan, and Alexander G. Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013.

Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, D. T. Lee, and Shih-Fu Chang. Joint audio-visual bi-modal codewords for video event detection. In *ICMR*, 2012a.

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

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

Ehsan Younessian, Teruko Mitamura, and Alexander G. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ICMR*, 2012.

Qian Yu, Jingen Liu, Hui Cheng, Ajay Divakaran, and Harpreet S. Sawhney. Semantic pooling for complex event detection. In *ACM MM*, 2013.

Shoou-I Yu, Lu Jiang, and Alexander G. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*, 2014a.

Shoou-I Yu, Lu Jiang, Zexi Mao, Xiaojun Chang, Xingzhong Du, Chuang Gan, Zhenzhong Lan, Zhongwen Xu, Xuanchong Li, Yang Cai, et al. Informedia@ trecvid 2014 med and mer. In *NIST TRECVID Video Retrieval Evaluation Workshop*, 2014b.

Shoou-I Yu, Lu Jiang, Zhongwen Xu, Zhenzhong Lan, Shicheng Xu, Xiaojun Chang, and et al. Informedia@TRECVID 2014 MED and MER. 2014c.

Yaoliang Yu. On decomposing the proximal map. In *NIPS*, 2013.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006a.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006b.

Xinhua Zhang, Yaoliang Yu, and Dale Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, 2012.