UvA-DARE (Digital Academic Repository)

Storytelling machines for video search

Habibian, A.

Publication date

2016

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Habibian, A. (2016). *Storytelling machines for video search*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

STORYTELLING MACHINES FOR VIDEO SEARCH

AMIRHOSSEIN HABIBIAN

STORYTELLING MACHINES FOR VIDEO SEARCH

AMIRHOSSEIN HABIBIAN

ISBN 978-94-6182-715-9

Storytelling Machines for Video Search

Amirhossein Habibian

This book was typeset by the author using L^AT_EX 2_ε.

Copyright © 2016 by A. Habibian.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the author.

ISBN 978-94-6182-715-9

Storytelling Machines for Video Search

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K. I. J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 20 september 2016, te 14:00 uur

door

Amirhossein Habibian

geboren te Hamedan, Iran

Promotiecommissie

Promotor:	Prof. dr. ir. A. W. M. Smeulders	Universiteit van Amsterdam
Co-promotor:	dr. C. G. M. Snoek	Universiteit van Amsterdam
Overige leden:	Prof. dr. H. Afsarmanesh	Universiteit van Amsterdam
	Prof. dr. T. Gevers	Universiteit van Amsterdam
	Prof. dr. A. Hanjalic	TU Delft
	Prof. dr. A. Hauptmann	Carnegie Mellon University
	dr. C. Monz	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has been carried out within the graduate school ASCI, dissertation number 361, at the Intelligent Sensory Information Systems Lab of the University of Amsterdam. This research is partly supported by the STW STORY project and the Dutch national program COMMIT.

UNIVERSITEIT VAN AMSTERDAM

CONTENTS

1	INTRODUCTION	9
2	RECOMMENDATIONS FOR RECOGNIZING VIDEO EVENTS BY CONCEPT VOCABULARIES	15
2.1	Introduction	15
2.2	Related Work	16
2.2.1	Characterizing Concept Vocabulary Composition	16
2.2.2	Characterizing Vocabulary Concept Detectors	17
2.2.3	Research Questions	18
2.3	Human Event Description	18
2.4	Experimental Setup	19
2.4.1	Video Datasets	20
2.4.2	Concept Vocabulary	20
2.4.3	Event Recognition	22
2.5	Experiments	22
2.6	Results	24
2.6.1	Experiment 1: How many?	24
2.6.2	Experiment 2: What concept types?	25
2.6.3	Experiment 3: Which concepts?	27
2.6.4	Experiment 4: How accurate?	29
2.6.5	Experiment 5: How to normalize?	31
2.7	Recommendations	34
3	COMPOSITE CONCEPT DISCOVERY FOR ZERO-SHOT VIDEO EVENT DETECTION	37
3.1	Introduction	37
3.2	Related Work	39
3.3	Composite Concept Discovery	40
3.3.1	Zero-Shot Event Detection	40
3.3.2	Algorithm	41
3.3.3	Computational Efficiency	43
3.4	Experimental Setup	43
3.4.1	Datasets	43
3.4.2	Vocabulary Concept Classifiers	44
3.4.3	Experiments	44
3.5	Results	44
3.5.1	Zero-Shot Event Detection	44
3.5.2	OR-Composite vs AND-Composite	47
3.6	Conclusions	49
4	DISCOVERING SEMANTIC VOCABULARIES FOR CROSS-MEDIA RETRIEVAL	51
4.1	Introduction	51
4.2	Our Proposal	53
4.2.1	Discovering Semantic Annotations	54
4.2.2	Training Semantic Embeddings	55
4.3	Experimental Setup	57
4.3.1	Datasets	57

CONTENTS

4.3.2	Evaluation Protocol	59
4.3.3	Implementation Details	59
4.3.4	Experiments	59
4.4	Results	60
4.4.1	Discovering Semantic Vocabulary	60
4.4.2	Training Semantic Embeddings	62
4.4.3	Comparison to Other Alignments	63
4.5	Conclusion	63
5	VIDEOSTORY EMBEDDINGS RECOGNIZE EVENTS WHEN EXAMPLES ARE SCARCE	65
5.1	Introduction	65
5.2	Related Work	67
5.2.1	Representations for Event Recognition	67
5.2.2	Embedding Videos and their Descriptions	68
5.2.3	Zero-Example Event Recognition	68
5.3	Videostory Embeddings	69
5.3.1	Objective Function	70
5.3.2	VideoStory Fusion	72
5.3.3	VideoStory for Zero-Example Event Recognition	73
5.4	Experimental Setup	74
5.4.1	Datasets	74
5.4.2	Video Features	76
5.4.3	Implementation Details	76
5.5	Experiments	76
5.5.1	VideoStory for Few-Shot Event Recognition	76
5.5.2	VideoStory Fusion	80
5.5.3	VideoStory for Zero-Example Recognition	84
5.5.4	Comparison with the state-of-the-art	86
5.6	Conclusion	87
6	SUMMARY AND CONCLUSIONS	89
6.1	Summary	89
6.2	Conclusions	90
	Appendix A HARVESTING WEB VIDEOS AND THEIR DESCRIPTIONS	93
A.1	Web Spidering Algorithm	93
A.1.1	Quality Filters	93
A.1.2	VideoStory46K Dataset	94
A.2	Experiment on Description Quality and Quantity	96
	Appendix B STOP-FRAME REMOVAL IMPROVES WEB VIDEO CLASSIFICATION	97
B.1	Introduction	97
B.2	Stop-Frame Removal	98
B.2.1	Algorithm	99
B.3	Experimental Setup	100
B.3.1	Data Sets	100
B.3.2	Implementation Details	100
B.3.3	Experiments	101
B.4	Results	101
B.5	Conclusions	102

Appendix c VIDEO2SENTENCE AND VICE VERSA	105
c.1 Introduction	105
c.2 Concept Vocabulary	105
c.3 Demonstration	105
Summary	109
List of Publications	111
Bibliography	119
Samenvatting	121
Acknowledgements	123

INTRODUCTION

Storytelling is a magic of the human mind. It makes us able to see scenes we have never been to. A storyteller conveys a picture of a scene into words. The listener brings the words to life by imagining the scene, as if she is physically there observing. Storytelling can even bring us to places, which do not exist; to see fictitious creatures such as a dragon enjoying meatballs, an elephant flying, or a PhD-student who enjoys writing the introduction of his thesis.

Storytelling is a capacity, crucial to human communication. There is ample evidence that humans represent the physical world in their mind by a mental picture [16, 19, 37]. Mental pictures make us able to see the picture of a dog in a quasi-perceptual experience, when we only think about it. Mental pictures play a pivotal role in visual-spatial reasoning [16], thinking, and inventing ideas [19]. Although humans think by mental pictures, they also use the language of words to communicate thoughts. Hence, storytelling, as the capability to transpose mental pictures into words is crucial for communication between humans.

Storytelling is as old as our history. In its earliest forms, humans told stories as rough drawings carved on rock, as shown in Figure 1. The concise, lingual form of storytelling emerged after the development of language. Language enables humans to narrate their thoughts concisely by elaborating on details. There are even special forms of lingual expressions dedicated to create mental pictures. Artistic forms of language, such as poems and drama, are especially aiming at creating the most elaborate mental pictures in the mind of the listeners [50]. The Shahname is a long epic poem narrating ancient Iranian stories by describing characters, scenes, and events in great detail, as in Figure 2, so that the reader can imagine himself in the scene.

Figure 1.: Storytelling by drawing. Sketched 30,000-32,000 years ago in the Chauvet cave, France.

Figure 2.: *Storytelling by poem. Shahname consists of 60,000 verses narrating ancient Iranian stories.*

Initially, storytelling was used to communicate special events only, such as natural phenomena, battles, and ceremonies. In our time, the communication era, humans tend to communicate almost any event. Very soon most stories will be told through digital video. Daily routines on YouTube, private collections of holidays, and video diaries for policemen patrolling the streets are just a few examples of events being captured and narrated every day. Telling stories of videos in words makes them accessible as will soon be crucial for their communication. By the ever-increasing amount of videos, we need storytelling machines, which automatically understand events in videos and generate their stories in words.

Creating storytelling machines poses one of the grand challenges in artificial intelligence, involving computer vision, natural language processing and machine learning. The process breaks down in a few steps. The first step towards storytelling is to identify the concepts, which play a role in the story, such as actors, actions, objects, and scenes [137]. For this purpose, the machine should be able to identify a comprehensive vocabulary of concepts as they appear in the video. This thesis is dedicated to constructing such a vocabulary by studying the fundamental question: *What vocabulary is suited for machines to tell the story of a video?* As a leading case study throughout the thesis, we focus on describing web videos depicting events such as *marriage proposal*, *graduation ceremony*, and *changing a vehicle tire*.

The machine capacity to identify concepts such as *person*, *car*, or *refrigerator* in a video is achieved by supervised learning. The supervision will first specify a list of concepts. Subsequently, annotating a set of videos per concept as examples, and adding others as counter examples, will enable the training to recognize the concepts. A straightforward approach to provide supervision is by handcrafting, involving both manual specification and manual annotation of concepts.

Prior research mostly handcrafts the vocabulary by specifying an arbitrary set of concepts [79, 140] or by adapting the concepts from publicly available datasets, such as ImageNet [130] or TRECVID [119]. However, the handcrafted vocabularies are not necessarily effective for describing videos' stories as their concepts are specified rather ad hoc. In order to effectively handcraft the vocabulary, we need to answer the following fundamental questions: How many concepts to include in the vocabulary? What concepts to include in the vocabulary? How accurate should the concept detectors be? In short:

What are the best practices for handcrafting the vocabulary for video storytelling?

This research question is considered in Chapter 2. By relying on an extensive pool of manually specified concepts, we empirically study research questions related to the number, the type, the

specificity, and the quality of the concept detectors in handcrafted video vocabularies. From the analysis, we conclude that for effective storytelling the vocabulary should encompass over thousands of concepts from various types, including object, action, scene, people, animal, and attribute. Moreover, the vocabulary should include detectors for both generic concepts, *i.e.* *human*, *vegetation*, and *outdoor*, and specific concepts, *i.e.* *salmon*, *cheesecake*, and *sand castle*.

Creating such a handcrafted video vocabulary of concepts is labor intensive, especially for a universal storytelling machine, which should possess a vocabulary as large as humans do. Hence, developing a data-driven approach to automatic video vocabulary construction is crucial. The next three chapters of this thesis are dedicated to alleviating the manual labor to automatically create a comprehensive vocabulary.

Concept composition is a cognitive ability, which humans possess to expand their vocabulary [23]. Based on this ability, humans can identify an unknown concept by relating it to the other concepts which are known to him. As an example, when a human perceives a *hovercraft* for the first time, he is capable of identifying by decomposing the unknown object into a vehicle similar to a *car* and similar to a *boat*. Inspired by the mechanism of concept composition in humans, we aim for expanding machine vocabularies by studying the research question:

Can a machine learn novel video concepts by composition?

We study this question by proposing a learning algorithm in Chapter 3. Given a set of known concepts, from the handcrafted video vocabulary, our proposed algorithm learns new video concepts without additional human supervision. It composes new concepts by the Boolean logic connectives AND, OR, and XOR. Given the training examples for the concepts *bike*, *motorcycle*, and *ride*, our algorithm learns new composite concepts such as “*bike-OR-motorcycle*”, “*bike-AND-ride*”, and “(*bike-OR-motorcycle*)-AND-ride”. The training examples to learn the detector for a composite concept are inferred from their underlying primitive concepts, *i.e.* the positive examples for the concept “*bike-OR-motorcycle*” is made of the examples, which are positive for either *bike* or *motorcycle*. Composite concepts are learned by searching over all the possible concept compositions.

The composite concepts learned by our proposed algorithm resemble bi-concepts [17, 88] and visual phrases [33] as the co-occurrences of concepts, which are learned jointly, such as “*running dog*” and “*person in car*”. Bi-concepts and visual phrases are usually defined exhaustively over all the concept pairs [88] or they are selected manually [33]. In contrast, our composite concepts are learned automatically by searching for the concepts whose composition makes sense. It enables storytelling machines to expand their video vocabulary without additional human labor.

However, the composite concepts are still grounded in the handcrafted vocabulary, so cannot expand the vocabulary’s capacity dramatically. As a further step towards reducing the manual labor for video vocabulary construction, we research whether the vocabulary can be learned from weak supervision, namely from videos and their human written stories, see Figure 3. The research has a high practical value as there are huge amounts of videos and their corresponding stories available nowadays. The YouTube videos and their captions, movies and their subtitles, news footages and their reports are just a few examples of the videos and stories which are available for learning the vocabulary.

There is a large body of research on utilizing weak supervision from the web, such as socially tagged images and videos, to train concept detectors [11, 35, 84, 87, 158]. These works apply weakly supervised [11, 35], semi-supervised [84, 87], and multiple instance learning algorithms [158] to overcome the subjectiveness, incompleteness, and ambiguity of tags. Although the concept detectors trained by these algorithms are effective in predicting tags of images, they are unsuited for video stories. For storytelling, the concept detectors should be accurate

Figure 3.: Examples of videos and their captions from YouTube.

and comprehensive to describe the video content. We aim for learning such an accurate and comprehensive vocabulary by studying the research question:

Can a machine learn its video vocabulary from human stories?

This research question is considered in Chapter 4. We start from the observation that analyzing the human stories reveals the concepts that humans often use for storytelling. Moreover, the occurrences of concepts in stories can serve as weak supervision for training the concept detectors. For example, if the term *dog* occurs in the story of a video, the video probably depicts a *dog*, so its video can be used as a positive exemplar for training the detector.

From the observations and the preliminary experiments in this chapter, we arrive at the conclusion that human-written stories and their associated videos are precious resources for learning the vocabulary. This conclusion naturally raises the final research question of this thesis:

How to learn the video vocabulary from human stories?

We study this question in Chapter 5. We learn the vocabulary as an embedding from videos into their stories. We utilize the correlations between the terms to learn the embedding more effectively. More specifically, we learn similar embeddings for the terms, which highly co-occur in the stories, as these terms are usually synonyms. Furthermore, we extend our embedding to learn the vocabulary from various video modalities, including appearance, motion, and audio. The learned multimodal vocabulary makes the machine able to generate more natural stories by incorporating concepts from various modalities, *i.e.* the *laughing* and *singing* concepts from audio, and the *jumping* and *dancing* concepts from motion.

The multimodal embedding builds on prior research on learning the correspondences between different modalities. These works rely on canonical correlation analysis [127], multimodal topic models [14], or multimodal auto-encoders [112] to learn the cross-correlations between visual features and textual features from stories. We study the merit of our proposed embedding compared to these works in chapter 5.

The organization of the thesis is as follows: The four research questions of the thesis are addressed in Chapter 2 to Chapter 5. A summary and a conclusion of the thesis are provided in Chapter 6. In Appendix A, we discuss how the videos and their human stories, used for vocabulary learning in Chapter 4 and 5, are collected from the web. In Appendix B, we propose a pre-processing on web videos to remove the frames that are irrelevant for learning. Finally, Appendix C demonstrates a storytelling machine, which uses concept vocabularies to translate videos to stories, and vice versa.

Co-authorship and Roles

Recommendations for Recognizing Video Events by Concept Vocabularies,

- Presented in Chapter 2,
- Published in *Computer Vision and Image Understanding*, 2014,
- Amirhossein Habibian: All aspects of the paper,
- Cees G. M. Snoek: Insights and supervision.

Recommendations for Video Event Recognition using Concept Vocabularies,

- Presented in Chapter 2,
- Published in *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2013,
- Amirhossein Habibian: All aspects of the paper,
- Koen E. A. van de Sande: Guidance and technical advice,
- Cees G. M. Snoek: Insights and supervision.

Composite Concept Discovery for Zero-Shot Video Event Detection,

- Presented in Chapter 3,
- Published in *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2014,
- Amirhossein Habibian: All aspects of the paper,
- Thomas E. J. Mensink: Guidance and technical advice,
- Cees G. M. Snoek: Insights and supervision.

Discovering Semantic Vocabularies for Cross-Media Retrieval,

- Presented in Chapter 4,
- Published in *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2015,
- Amirhossein Habibian: All aspects of the paper,
- Thomas E. J. Mensink: Guidance and technical advice,
- Cees G. M. Snoek: Insights and supervision.

VideoStory Embeddings Recognize Events when Examples are Scarce,

- Presented in Chapter 5,
- Pending revision in *IEEE Trans. on Pattern Analysis and Machine Intelligence*,
- Amirhossein Habibian: All aspects of the paper,
- Thomas E. J. Mensink: Guidance and technical advice,
- Cees G. M. Snoek: Insights and supervision.

VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events,

- Presented in Chapter 5 and Appendix A,
- Awarded as the best paper in *Proc. ACM Int'l. Conf. Multimedia*, 2014,
- Amirhossein Habibian: All aspects of the paper,
- Thomas E. J. Mensink: Guidance and technical advice,
- Cees G. M. Snoek: Insights and supervision.

Video2Sentence and Vice Versa,

- Presented in Appendix B,
- Published in *Proc. ACM Int'l. Conf. Multimedia*, 2013,
- Amirhossein Habibian: All aspects of the paper,
- Cees G. M. Snoek: Insights and supervision.

Stop-Frame Removal Improves Web Video Classification,

- Presented in Appendix C,
- Published in *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2014,
- Amirhossein Habibian: All aspects of the paper,
- Cees G. M. Snoek: Insights and supervision.

RECOMMENDATIONS FOR RECOGNIZING VIDEO EVENTS BY CONCEPT VOCABULARIES

2.1 INTRODUCTION

We consider the problem of recognizing events in arbitrary web video, such as the ones depicted in Figure 4. Among the many challenges involved, resulting from the uncontrolled recording condition of web videos and the large variations in the visual appearance of events, probably one of the most fundamental questions in event recognition is what defines an event in video? The Oxford English dictionary defines an event as “anything that happens”. With such a broad definition it is not surprising that the topic has been addressed in the computer vision and multimedia retrieval community by many researchers from diverse angles [6, 7, 91, 113, 134, 160, 172].

In this paper, we study representations that contribute to defining events for automatic recognition. We are inspired by findings from cognition, where research has repeatedly shown that humans remember events by their actors, actions, objects, and locations [136]. Studying event representation based on such high-level concepts is now within reach because of the continued progress in supervised concept detection [141] and the availability of labeled training collections like the ones developed in benchmarks such as TRECVID [139], ImageNet [30] and several other venues [54, 108]. Different from concepts, which represent a single person, object, scene or action in videos, events are commonly defined as a more complex interaction of several persons, objects, and actions happening in a specific scene [104]. In this paper, we name the set of available concept detectors as the *vocabulary* and we study how to construct a vocabulary suited for effective recognition of events in video.

The state-of-the-art in event recognition represents a video in terms of low-level audiovisual features [66, 67, 70, 110, 114, 115, 149]. In general, these methods first extract from the video various types of static and/or dynamic features, *e.g.* color SIFT variations [159], MFCC [66], and Dense Trajectories [115]. Second, the descriptors are quantized and aggregated [115]. The robustness and efficiency of various low-level features for recognizing events are evaluated in [66, 106, 149]. Despite their good recognition performance, especially when combined together [66, 106, 110, 115], low-level features are incapable of providing an understanding of the semantic structure present in an event. Hence, it is not easy to derive how these event definitions arrive at their recognition. Therefore, essentially different representations are needed for events. We focus on high-level representations for event recognition.

Inspired by previous work in object recognition [85, 153], scene recognition [85, 128] and activity recognition [131], many have explored high-level representations for recognition of events [4, 47, 62, 90, 96, 100, 104, 180]. All these works follow a general pipeline consisting of three consecutive steps to arrive at a high-level video representation. First, frame extraction, where the video is decoded and a subset of frames is extracted. Second, concept detection, where each extracted frame is represented by a vector of predictions from vocabulary concept

Figure 4.: Examples of web videos depicting events. From top to bottom: “making a sandwich”, “marriage proposal”, “swimming” and “soccer”. Each event can be defined by its key concepts including actor, place, action and the involved objects.

detectors. Finally, video pooling, where the frame representations are averaged and aggregated into the video level representation. The obtained high-level representation is not only semantically interpretable, but is also reported to outperform the state-of-the-art low-level audiovisual features in recognizing events [104, 106]. Rather than training vocabulary concept detectors and event detectors separately, recent work aims for jointly learning the vocabulary concept and event detectors [2, 93, 95, 178]. In these works, the vocabulary concept detectors are trained to optimize the event detection, without explicitly optimizing the individual concept detector accuracy. As a consequence, the vocabulary concepts do not necessarily have a semantic interpretation needed to explain the video content. In this paper, we follow [62, 100, 104, 180] and train concept and event detectors separately.

Identifying a universal vocabulary of concepts suited for representing events is an important question that has been ignored in the literature. To the best of our knowledge, all the previous work on high-level representations for event recognition relies on an arbitrary set of concepts as the vocabulary. By contrast, we focus in this paper on characterizing the vocabulary which is most effective for representing events. We investigate the concept vocabulary from two perspectives: first by characterizing the composition, where we investigate *what* concepts should be included in the vocabulary. Second by characterizing the concept detectors, where we study *how* to create vocabulary concept detectors that are most suited for representing events. Before detailing our research questions, we discuss related work that we consider most relevant to these two perspectives.

2.2 RELATED WORK

2.2.1 Characterizing Concept Vocabulary Composition

Our study is inspired by the pioneering work of Hauptmann *et al.* [49] who focus on construction of concept vocabularies for broadcast news video retrieval. They examined how big the concept vocabulary should be and what concepts should be part of the vocabulary for effective shot retrieval. For this purpose, they used a pool of concepts to create and evaluate vocabularies under

different circumstances. In their work, the presence and absence of 320 human-annotated concepts was used as the main source for the investigations. To make the experiments more realistic they insert noise into the human annotations to simulate the behavior of automatic concept detectors. They concluded that 5,000 detectors with modest quality would be sufficient for general-purpose broadcast news video retrieval. Regarding the important question what concepts to include in the vocabulary, Hauptmann *et al.* [49] conclude that frequent concepts contribute more to overall news video retrieval performance than rare concepts, so they are preferred to be included in the vocabulary. However, it is not clear whether their conclusion generalizes to *event* recognition on the challenging domain of unconstrained web video.

In this paper, we start from the analysis by Hauptmann *et al.* [49] and adopt their research questions for event recognition. Our work is different with respect to the following five aspects. First, we focus exclusively on events, whereas [49] considers news use cases like *Find shots of U.S. Maps depicting the electoral vote distribution (blue vs. red state)* and *Find shots of Refugee Camps with women and children visible*. Second, our domain of study is unconstrained web video, rather than the highly structured broadcast television domain. Third, we place special emphasis on the importance of various concept types in representing events (*e.g.* objects, scenes, actions etc.), rather than considering all concepts equally important. Fourth, we evaluate retrieval accuracy on video-level rather than shot-level. Finally, in our analysis we do not rely on human concept annotations directly, but instead we use real detector predictions with varying levels of accuracy per concept. Using the real detectors to represent videos leads to surprising new findings, as we will show in the experiments.

2.2.2 Characterizing Vocabulary Concept Detectors

Automatic detection of concepts in videos is a well studied topic in computer vision and multimedia for which many algorithms have been proposed [53, 68, 142, 158]. These include descriptors, *e.g.* SIFT variations [159] and STIP [80], descriptor quantization strategies, *e.g.* Bag-of-Words, VLAD [58] and Fisher vector coding [132], the use of spatial pyramids [82] and various types of kernels to train classifiers, *e.g.* RBF, χ^2 and Histogram Intersection [98, 183]. Choosing among these options provides us with a wide range of concept detectors with varying accuracies and computational costs. In this paper, we investigate how to create vocabulary concept detectors that are most suited for representing events by considering detectors of varying accuracy.

The state-of-the-art in video concept detection employs an SVM classifier to train detectors. SVM predictions are real-valued numbers that could be positive or negative. To perform the subsequent processing steps on the prediction scores, they should be normalized. The general SVM normalization approach in the literature [62, 104, 180] is to fit a sigmoid function on top of the prediction scores to estimate the posterior probabilities of concept presence [89, 125]. The sigmoid function parameters are estimated from a held out partition of the concept detectors training data. In case the training and test data distributions differ, a common scenario when using pre-trained concept detectors for event recognition in arbitrary video, the detector reliability suffers [177]. Hence, the normalization should be executed with care. In this paper, we examine the influence of normalizing the predictions of vocabulary concept detectors on video event recognition accuracy. For this purpose, we consider several existing score normalizations [56].

We consider supervised normalization, which relies on labeled training data to fit the normalization function, *e.g.* Sigmoid normalization, and unsupervised normalization that does not require any labeled training data. To circumvent supervision, some unsupervised normalizations make assumptions about the distribution of scores. Z-score normalization, for example, assumes the scores have a Gaussian distribution, so the scores are normalized by shifting and scaling by their mean and standard deviation [56]. Others do not make any assumption about the distribution

of scores. For example the recent w-score normalization, which models the tails of score distributions by the Extreme Value Theory [133] from statistics and then uses the models to estimate the concept presence probabilities. We assess the influence of normalizing detectors in a concept vocabulary for event recognition.

2.2.3 Research Questions

Our study on the effectiveness of concept vocabularies for video event recognition, is directed by five research questions. The first three questions investigate the ideal concept vocabulary composition, while the last two questions consider the creation of the vocabulary concept detectors. A preliminary version of this study has been published in [47]. Here we put more emphasis on characterizing the vocabulary concept detectors. Our five research questions are:

RQ1 *How many concepts to include in the vocabulary?*

RQ2 *What concept types to include in the vocabulary?*

RQ3 *Which concepts to include in the vocabulary?*

RQ4 *How accurate should the concept detectors be?*

RQ5 *How to normalize the concept detectors?*

As humans remember events by the high level concepts they contain, *viz.*, actors, actions, objects, and locations [136], studying the characteristics of the concepts that humans use to describe events could be inspirational for automated event recognition. Therefore, before describing our experimental protocol to address the research questions, we first study the vocabulary that humans use to describe events in videos.

2.3 HUMAN EVENT DESCRIPTION

To analyze the vocabulary that humans use to describe events, we utilize a set of textual descriptions written by humans to describe web videos containing events. We process textual descriptions for 13,265 videos, as provided by the TRECVID 2012 Multimedia Event Detection task corpus [139]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by highlighting its dominant concepts. Figure 5 illustrates some videos and their corresponding textual descriptions.

After removing stop words and stemming, we end up with 5,433 distinct terms from the 13,265 descriptions making up a human vocabulary for describing events. Naturally, the frequency of these terms varies, as also observed by [49]. Most of the terms seldom occur in event descriptions. Whereas, only a few terms have high term-frequencies. To be precise, 50% of the terms occur once in the descriptions and only 2% occurs more than five times. Terms like *man*, *girl*, *perform* and *street* appear most frequent, while *bluefish*, *conductor*, *Mississippi* and *Bulgarian* are instances of less frequent terms. Looking into the vocabulary, we observe that the terms used in human descriptions can be mapped to five distinct concept types, as typically used in the computer vision and multimedia literature: *objects*, *actions*, *scenes*, *visual attributes* and *non visual concepts*. We manually assign each vocabulary term into one of these five types. After this exercise we observe that 44% of the terms refer to *objects*. Moreover, we note that a considerable number of objects are dedicated to various types of *animals* and *people*; *i.e.*, *lion*, and *teen*. About 21% of the terms depict *actions*, like *walking*. Approximately 10% of the concept types are about *scenes*, such as *kitchen*. *Visual attributes* cover about 13% of the terms; *i.e.*, *white*,

Figure 5.: Examples of videos and human-added textual descriptions, from which we study how humans describe events.

flat, and dirty. The remaining 12% of the terms belong to concepts, which are *not visually depictable*; i.e., poem, problem, and language. We summarize the statistics of our human event descriptions in Figure 6.

We observe that when describing video events, humans use terms with varying generalizations. Some terms are very specialized and refer to specific objects; like, salmon, cheesecake and sand castle. While other terms are more general and refer to a broader set of concepts; like human, vegetation and outdoor. We analyze the generalization of the vocabulary terms using their depth in the WordNet hierarchy [105]. In this hierarchy, the terms are structured based on their hypernym/hyponym relations, so the more specialized terms are placed at the deeper levels. Our study shows that the 5,433 vocabulary terms have an average depth of 9.07 ± 5.29 . The high variance in term depths indicates that the human vocabulary to describe events is composed of both specific and general terms.

To summarize, analyzing the available event descriptions, we observe that the vocabulary that humans use to describe events is composed of a few thousand words, derived from five distinct concept types: *objects*, *actions*, *scenes*, *visual attributes* and *non-visual concepts*. Moreover, we observe that the vocabulary contains both specific and general concepts. Strengthened by these observations about the human vocabulary for describing events, we design five experiments to answer our research questions on the ideal vocabulary for recognizing events in arbitrary web video.

2.4 EXPERIMENTAL SETUP

To answer the research questions raised in section 2.2.3, we create a rigorous empirical setting. First, we introduce the video datasets used to evaluate the event recognition experiments. Then

Figure 6.: We divide the human vocabulary for describing events into concept types containing objects, actions, scenes, attributes and non-visual concepts. Our analysis reveals that objects and actions constitute 65% of the human vocabulary when describing events.

we explain the pool of concept detectors, which we employ to create vocabularies. Finally, the pipeline used for event recognition using concept vocabularies is presented.

2.4.1 Video Datasets

For the event recognition experiments, we rely on two publicly available video collections: the TRECVID Multimedia Event Detection [139] and the Columbia Consumer Video [69] datasets.

TRECVID MED [139] consists of 13,274 event videos sampled from the TRECVID 2012 Multimedia Event Detection task corpus, as used in [47]. This dataset consists of over 400 hours of user-generated video with a large variation in quality, length and content. Moreover, it comes with ground-truth annotations at video level for 25 real-world events, including life events, instructional events, sport events, etc. Following [47], the dataset is divided into a training set (66%) and a test set (34%).

Columbia CV [69] contains 9,317 user-generated videos from YouTube. This dataset consists of over 210 hours of videos in total, where each video has an average length of 80 seconds. Moreover, the dataset contains ground-truth annotations at video level for 20 semantic categories, where 15 of them are events. The other 5 categories are objects and scenes, which are excluded from the dataset in our experiments: “bird”, “cat”, “dog”, “beach” and “playground”. We use the training and test set divisions as defined in [69].

We summarize the training and test set statistics for both video datasets per event in Table 1.

2.4.2 Concept Vocabulary

To create the vocabularies, we need a comprehensive pool of concept detectors. We build this pool of detectors using the human annotated training data from two publicly available resources: the TRECVID 2012 Semantic Indexing task [5, 139] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [10]. The former has annotations for 346 semantic concepts on

Table 1.: Number of positive videos in the datasets used in our experiments, split per event. The number of negative videos for each event are around 8,800 and 4,500 for the TRECVID MED and the Columbia CV dataset, respectively.

TRECVID MED			Columbia CV		
Event	Train	Test	Event	Train	Test
Attempting board trick	98	49	Basketball	182	181
Feeding animal	75	48	Baseball	150	151
Landing fish	71	36	Soccer	161	162
Wedding ceremony	69	35	Ice Skating	192	193
Working wood working project	79	40	Skiing	197	196
Birthday party	121	61	Swimming	199	202
Changing vehicle tire	75	37	Biking	136	137
Flash mob gathering	115	58	Graduation	143	145
Getting vehicle unstuck	85	43	Birthday	158	160
Grooming animal	91	46	Wedding Reception	129	130
Making sandwich	83	42	Wedding Ceremony	111	110
Parade	105	50	Wedding Dance	174	176
Parkour	75	38	Music Performance	403	403
Repairing appliance	85	43	Non-Music Performance	345	346
Working sewing project	86	43	Parade	191	194
Attempting bike trick	43	22			
Cleaning appliance	43	22			
Dog show	43	22			
Giving directions location	43	22			
Marriage proposal	43	22			
Renovating home	43	22			
Rock climbing	43	22			
Town hall meeting	43	22			
Winning race without vehicle	43	22			
Working metal crafts project	43	22			

400,000 keyframes from web videos. The latter has annotations for 1,000 semantic concepts on 1,300,000 photos. The categories are quite diverse and include concepts from various types; *i.e.*, *object*, *scene* and *action*. Note that the training data is different from the TRECVID MED videos and their textual descriptions, which are used for studying the human vocabulary, as discussed in Section 2.3.

Leveraging the annotated data available in these datasets, we train 1,346 concept detectors in total. We follow the state-of-the-art for our implementation of the concept detectors. We use densely sampled SIFT, OpponentSIFT and C-SIFT descriptors [159] with Fisher vector coding [132]. The codebook used has a size of 256 words. As a spatial pyramid we use the full image and three horizontal bars [82]. The feature vectors representing the training images form the input for a fast linear Support Vector Machine [135].

As summarized in Figure 6, the concepts that humans use to describe events are derived from *object*, *action*, *scene*, *attributes* and *non-visual* concept types. It is hard to imagine that *non-visual* concepts can be detected by their visual features, so we exclude them from our study. With respect to the importance of the actors in depicting events [136], as well as their high frequency in human descriptions, we consider *people* and *animal* as extra concept types in our experiments. Inspired by this composition, we divide our concept pool by manually assigning each concept to one of the six types. Consequently, we end up with the following concept types: *object* containing

Figure 7.: Random training examples of the 1,346 concept detectors included in the overall vocabulary used in our experiments, grouped by their concept type.

706 concepts, *action* containing 36 concepts, *scene* containing 135 concepts, *people* containing 83 concepts, *animal* containing 338 concepts and *attribute* containing 48 concepts. Figure 7 provides an overview of the concept types and shows example instances.

2.4.3 Event Recognition

In the event recognition experiments, we follow the common pipeline as used in the literature [62, 100, 104, 180]. Unless noted otherwise we use the following implementation. We decode the videos by uniformly extracting one frame every two seconds. Then all the concept detectors are applied on the extracted frames. After concatenating the detector outputs, each frame is represented by a concept vector. Finally, the frame representations are pooled into a video level representation by averaging and normalizing as proposed in [133]. On top of this concept vocabulary representation per video, we use again a linear SVM classifier to train the event recognizers.

2.5 EXPERIMENTS

We perform five experiments to address our research questions. Each concept vocabulary used in the experiments is evaluated based on its performance in recognizing events using the datasets, pipeline and evaluation protocol described in section 2.4. Moreover, the vocabularies are all derived from the concept pool introduced in section 2.4.2.

- **Experiment 1: How many concepts to include in the vocabulary?** To study this question, we create several vocabularies with varying sizes and evaluate their performance for recognizing events. Each vocabulary is made of a random subset of the concept detectors from the concept pool. To compensate for possible random effects, all experiments are repeated 50 times and the results are averaged.
- **Experiment 2: What concept types to include in the vocabulary?** We look into this question by comparing two types of vocabularies: (i) *single type* vocabularies, where all

concepts are derived from one type and (ii) *joint type* vocabularies, where concepts are derived from all available concept types. We perform this experiment for six kinds of single type vocabularies: *object*, *action*, *scene*, *people*, *animal* and *attribute* types respectively.

To make the single type and joint type vocabularies more comparable, we force the vocabularies to be of equal size. We do so by randomly selecting the same number of concepts from the concept pool. All the experiments are repeated 500 times to balance possible random effects.

- **Experiment 3: Which concepts to include in the vocabulary?** In this experiment, we investigate whether the concept vocabulary for event recognition should be made of general concepts, specific concepts, or their mixture. We manually label and select two sets of general and specific concepts from the concept pool. The former contains 149 general concepts, *i.e.*, *vegetation*, *human* and *man made thing*, and the latter contains 619 specific concepts, *i.e.*, *religious figure*, *emergency vehicle* and *pickup truck*. The rest of the concepts, which are not clearly general or specific, are not involved in this experiment. Using these sets we compare three types of vocabularies: (i) a *general* vocabulary in which all the concepts are general, (ii) a *specific* vocabulary in which all the concepts are specific and (iii) a *mixture* vocabulary in which the concepts are randomly selected from both general and specific concept sets. We repeated this experiment for different vocabulary sizes and found that the results remained stable. The reported results are obtained for a vocabulary size of 70, averaged over 500 repetitions.
- **Experiment 4: How accurate should the concept detectors be?** We look into this question by decreasing the detector accuracies and measuring how the event recognition performance responds. To decrease the detector accuracies we follow two different approaches: the first approach trains less sophisticated concept detectors, and the second approach imposes noise into the concept prediction scores.

In the first approach, we train four versions of our vocabulary concept detectors at different levels of sophistication: (i) *100%-3SIFT-SP* is the most sophisticated version, where the detectors are implemented as described in section 2.4.2. In this version, detectors are trained on all available training data. (ii) *30%-3SIFT-SP* is similar to *100%-3SIFT-SP*, but the detectors are trained on a random subset of 30% of the available concept training examples. (iii) *30%-SIFT-SP* is similar to *30%-3SIFT-SP*, but does not include any color SIFT and only relies on standard intensity SIFT. (iv) *30%-SIFT* is the same as *30%-SIFT-SP*, but without using any spatial pyramid. The four versions of the detectors are trained for the 346 semantic concepts from the TRECVID Semantic Indexing dataset.

In the second approach, we make the concept detectors inaccurate by gradually imposing increasing amounts of noise into their predictions. The output of each concept detector, as an SVM classifier, is a real-valued number which is supposed to be larger than +1 and smaller than -1 for positive and negative samples. However in practice, the SVM only assigns these values to the samples which are confidently classified, while other samples are assigned to the unconfident area in between -1 and 1. Looking into the concept detector predictions, we observe that most of them are agglomerated in the unconfident area. The less accurate a concept detector is, the more samples are assigned to the unconfident area. To simulate the detector accuracy changes, we randomly select predictions and shift them towards the center of the unconfident area, which has the least decision confidence. We gradually increase the amount of noise and repeat the experiments 50 times to compensate for possible random factors.

Figure 8.: *Experiment 1: Increasing the vocabulary size improves the event recognition performance. This improvement is especially prevalent for small vocabularies containing less than 200 concept detectors.*

- Experiment 5: How to normalize the concept detectors?** In this experiment, we investigate the effect of normalizing concept vocabularies on video event recognition accuracy. We compare the representation obtained from un-normalized predictions with the representations obtained by applying several normalizations as introduced in section 2.2.2: supervised, unsupervised, assumption-based and assumption-free normalization. We apply sigmoid normalization as a supervised method and compare it with z-score and w-score, as instances of unsupervised normalizations. Moreover, to study the effect of making assumptions on the distribution of concept detector predictions we compare z-score, which assumes a Gaussian distribution, with w-score normalization, which is an assumption-free method.

Each experiment results in a ranking of the videos from both the test sets based on the probability that the video contains the event of interest. As the evaluation criterion for these ranked lists, we employ average precision (AP) which is in wide use for evaluating visual retrieval results [139]. We also report the average performance over all events as the mean average precision (MAP).

2.6 RESULTS

2.6.1 Experiment 1: How many?

As shown in Figure 8, adding more concept detectors to the vocabulary improves the event recognition performance. The improvement gain is particularly prevalent for small vocabularies. When increasing the vocabulary from 50 to 300, on TRECVID MED for example, the MAP increases from 0.125 to 0.221. The improvement is less prevalent when more than 1,000 detectors are part of the vocabulary. When increasing the vocabulary from 1,000 to 1,346 the absolute MAP improvement is only 0.012 on average. We observe similar behavior on Columbia CV. We speculate that the improvement comes from the finer gain partitioning of the event feature

Figure 9.: Experiment 2: Correlation between concept detector responses appears especially within a single concept type. Including too many concepts from the same type leads to a decreased event recognition performance (matrix smoothed for better viewing, results from TRECVID MED).

space, which in our case is caused by the concept annotations, but is also achievable along other means [178].

However, by looking into individual event recognition results we observe that not all events behave similar when increasing the vocabulary size. For some events, *i.e.* “flash mob gathering”, a relatively high average precision of 0.34 is obtained by including only 50 concepts. We observe that there are some concepts within the vocabulary which are very discriminative for this event *i.e.* group of people, dancing and people marching. In contrast, for some other events *i.e.* “giving directions to a location”, the event recognition performance is not improved by increasing the vocabulary size. Apparently, there is no concept in the vocabulary, which can effectively discriminate this event from others. It demonstrates that besides the vocabulary size, the relevance of vocabulary concepts should be considered.

The error bars plotted in Figure 8 indicate the variance in MAPs for various vocabularies. The variance demonstrates that with the same number of concept detectors, some vocabularies perform better than others. In the next two experiments, we study the characteristics of these optimal vocabularies.

Small vocabularies have poor performances in recognizing events. In addition, their effectiveness could be rapidly increased by adding a few more concept detectors. So, in general we recommend to include at least 200 concept detectors in the vocabulary.

2.6.2 Experiment 2: What concept types?

Tables 2 and 3 compare single type and joint type vocabularies for recognizing events. Comparing the MAPs for both datasets, we conclude that joint type vocabularies outperform single type vocabularies for all six concept types on average. It demonstrates that when creating the vocabulary, it is better to sample the concept detectors from diverse types. Hence, we need to

Table 2.: Experiment 2: Comparison of single type and joint type vocabularies for event recognition on the Columbia CV dataset. Each column pair compares a single and joint type vocabulary. To make the vocabularies more comparable within a concept type, we force them to be of equal size. Note that the number of concept detectors (in parenthesis) varies per concept type, so comparison across concept types should be avoided. The results demonstrate that for all the six concept types, joint type vocabularies outperform single type vocabularies on average. Best results per type denoted in bold.

Event	Concept Type											
	Object(670)		Action(34)		Scene(128)		People(78)		Animal(321)		Attribute(45)	
	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
Basketball	0.309	0.352	0.258	0.270	0.380	0.426	0.172	0.354	0.322	0.371	0.217	0.305
Baseball	0.159	0.233	0.151	0.217	0.180	0.205	0.172	0.279	0.161	0.233	0.144	0.158
Soccer	0.176	0.221	0.173	0.212	0.233	0.229	0.228	0.267	0.211	0.217	0.197	0.221
Ice Skating	0.275	0.296	0.248	0.200	0.204	0.284	0.264	0.364	0.168	0.296	0.128	0.228
Skiing	0.455	0.499	0.375	0.288	0.294	0.417	0.394	0.404	0.320	0.437	0.189	0.327
Swimming	0.573	0.597	0.220	0.256	0.328	0.405	0.338	0.346	0.542	0.516	0.149	0.288
Biking	0.234	0.201	0.106	0.116	0.183	0.210	0.109	0.162	0.128	0.205	0.096	0.131
Graduation	0.220	0.294	0.095	0.120	0.105	0.147	0.119	0.129	0.170	0.176	0.106	0.118
Birthday	0.295	0.303	0.164	0.228	0.219	0.242	0.217	0.202	0.275	0.286	0.189	0.193
Wedding Reception	0.172	0.179	0.126	0.144	0.168	0.162	0.149	0.122	0.166	0.170	0.144	0.133
Wedding Ceremony	0.225	0.276	0.172	0.195	0.153	0.239	0.226	0.175	0.130	0.268	0.165	0.194
Wedding Dance	0.414	0.433	0.195	0.198	0.375	0.340	0.231	0.286	0.399	0.407	0.218	0.227
Music Performance	0.413	0.418	0.231	0.240	0.314	0.338	0.305	0.316	0.375	0.387	0.268	0.280
Non-Music Performance	0.302	0.305	0.199	0.256	0.244	0.268	0.215	0.247	0.270	0.291	0.214	0.222
Parade	0.388	0.421	0.241	0.320	0.355	0.357	0.293	0.327	0.334	0.396	0.247	0.273
Mean	0.307	0.335	0.197	0.217	0.249	0.285	0.229	0.265	0.265	0.310	0.178	0.220

detect the objects, people, actions and scenes occurring in the video *jointly* to recognize the event properly. In other words, all of the concept types contribute to the recognition of events.

When we analyze individual event recognition results, we observe a few cases exist where a single type vocabulary outperforms the joint type because of the tight connection between the event description and specific concepts. For example, using a single type vocabulary made of *animals* only, we achieve a higher average precision for “*feeding animal*”, “*grooming animal*” and “*dog show*” events in comparison to a joint type vocabulary on the TRECVID MED dataset. Similarly, the “*Ice Skating*” and “*Skiing*” events from the Columbia CV dataset are recognized better by the action concepts than by the joint vocabulary. Nevertheless, joint type vocabularies do better than single type vocabularies on average. Therefore, we consider joint type vocabularies more suited for general purpose event recognition.

The performance difference between the single type and joint type vocabularies varies per concept type. For some types, like *animal*, the difference is substantial (0.158 vs. 0.239 and 0.310 vs. 0.265 on the TRECVID MED and the Columbia CV dataset respectively), while for others, like *action*, it is almost negligible (0.067 vs. 0.076 and 0.197 vs. 0.217 on the TRECVID MED and the Columbia CV dataset respectively). We attribute the performance difference to at least two reasons. First, our concept detectors are trained on a global image level, so they contain a considerable amount of contextual information. Consequently, some single types may contain a wide sample of contextual information including ‘semantic overlap’ from other concept types. The *action* pool, for example, may contain action detectors in varying scenes using various objects. Second, when creating several concept detectors for a similar type, it is likely that

Table 3.: Experiment 2: Comparison of single type and joint type vocabularies for event recognition on the TRECVID MED dataset, following the explanation in Table 2.

Event	Concept Type											
	Object(670)		Action(34)		Scene(128)		People(78)		Animal(321)		Attribute(45)	
	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
Attempting board trick	0.368	0.348	0.056	0.073	0.115	0.169	0.065	0.119	0.120	0.271	0.082	0.079
Feeding animal	0.035	0.044	0.029	0.074	0.024	0.042	0.040	0.041	0.073	0.045	0.055	0.037
Landing fish	0.337	0.423	0.055	0.076	0.157	0.246	0.074	0.182	0.323	0.360	0.054	0.111
Wedding ceremony	0.493	0.520	0.054	0.073	0.139	0.193	0.141	0.119	0.162	0.388	0.040	0.070
Working wood working project	0.194	0.203	0.029	0.040	0.074	0.101	0.118	0.072	0.116	0.167	0.032	0.048
Birthday party	0.264	0.277	0.098	0.099	0.115	0.174	0.138	0.131	0.139	0.239	0.058	0.095
Changing vehicle tire	0.171	0.174	0.034	0.054	0.073	0.105	0.036	0.076	0.054	0.153	0.043	0.052
Flash mob gathering	0.471	0.494	0.257	0.212	0.349	0.304	0.321	0.337	0.415	0.475	0.273	0.251
Getting vehicle unstuck	0.330	0.362	0.092	0.138	0.186	0.268	0.110	0.217	0.294	0.338	0.069	0.154
Grooming animal	0.126	0.149	0.033	0.070	0.129	0.147	0.075	0.080	0.146	0.127	0.075	0.068
Making sandwich	0.178	0.197	0.023	0.061	0.116	0.127	0.050	0.098	0.070	0.176	0.029	0.066
Parade	0.268	0.304	0.169	0.119	0.215	0.219	0.119	0.182	0.126	0.275	0.093	0.141
Parkour	0.398	0.432	0.023	0.063	0.150	0.234	0.034	0.147	0.089	0.356	0.031	0.074
Repairing appliance	0.244	0.323	0.063	0.078	0.192	0.224	0.086	0.126	0.104	0.259	0.100	0.083
Working sewing project	0.295	0.252	0.048	0.075	0.129	0.163	0.107	0.123	0.194	0.238	0.021	0.082
Attempting bike trick	0.480	0.502	0.264	0.076	0.250	0.245	0.037	0.171	0.129	0.392	0.031	0.096
Cleaning appliance	0.079	0.064	0.019	0.039	0.022	0.049	0.021	0.045	0.029	0.058	0.015	0.035
Dog show	0.500	0.534	0.093	0.102	0.423	0.455	0.114	0.236	0.555	0.512	0.116	0.122
Giving directions location	0.029	0.031	0.013	0.027	0.019	0.025	0.011	0.021	0.016	0.029	0.012	0.021
Marriage proposal	0.069	0.075	0.016	0.024	0.030	0.033	0.027	0.023	0.018	0.050	0.010	0.016
Renovating home	0.179	0.232	0.011	0.049	0.071	0.120	0.019	0.078	0.085	0.192	0.016	0.053
Rock climbing	0.347	0.375	0.027	0.092	0.217	0.176	0.101	0.173	0.309	0.322	0.063	0.104
Town hall meeting	0.424	0.456	0.059	0.099	0.270	0.244	0.116	0.172	0.266	0.379	0.158	0.115
Winning race without vehicle	0.139	0.147	0.082	0.061	0.075	0.101	0.069	0.081	0.088	0.138	0.073	0.060
Working metal crafts project	0.052	0.054	0.019	0.032	0.018	0.033	0.020	0.029	0.019	0.038	0.020	0.024
Mean	0.259	0.279	0.067	0.076	0.142	0.168	0.082	0.123	0.158	0.239	0.063	0.082

the detectors will be correlated, especially for the less diverse types, *e.g.* *People* and *Animal*. To clarify this observation we plot the correlation between concept detectors within a concept type in Figure 9. As shown in this figure, the highly correlated concepts tend to belong to the same concept type. Therefore, including too many concepts from the same type in a vocabulary, especially from the less diverse concept types like *animal* and *people*, leads to correlated concepts and should be avoided.

We recommend to make the vocabulary diverse by including concepts from various concept types and to limit the number of concepts for the less diverse types.

2.6.3 Experiment 3: Which concepts?

Tables 4 and 5 compare three types of vocabularies: specific, general and mixture. According to the MAPs on both datasets, the general vocabulary performs better than the specific vocabulary, but the mixture vocabulary is in both cases the best overall performer.

We observe that for a few events a specific vocabulary outperforms the others, *e.g.* “*repairing appliance*” and “*music performance*”. For these events, there are some specific and discriminative concepts available in the vocabulary. For example, the washing machine, refrigerator and

Table 4.: Experiment 3: Comparison of specific, general and mixture vocabularies for event recognition on the TRECVID MED dataset. The results demonstrate that the general vocabulary outperforms the specific vocabulary on average. The best results are obtained when the vocabulary consists of both general and specific concepts.

Event	Specific	General	Mixture
Attempting board trick	0.090	0.108	0.130
Feeding animal	0.041	0.042	0.045
Landing fish	0.113	0.107	0.139
Wedding ceremony	0.071	0.140	0.164
Working wood working project	0.083	0.065	0.073
Birthday party	0.078	0.135	0.138
Changing vehicle tire	0.058	0.062	0.071
Flash mob gathering	0.301	0.284	0.337
Getting vehicle unstuck	0.195	0.246	0.282
Grooming animal	0.064	0.079	0.081
Making sandwich	0.059	0.089	0.119
Parade	0.073	0.203	0.161
Parkour	0.104	0.226	0.210
Repairing appliance	0.111	0.098	0.101
Working sewing project	0.076	0.075	0.082
Attempting bike trick	0.044	0.080	0.090
Cleaning appliance	0.125	0.092	0.123
Dog show	0.219	0.178	0.230
Giving directions location	0.028	0.019	0.053
Marriage proposal	0.013	0.017	0.025
Renovating home	0.023	0.074	0.083
Rock climbing	0.178	0.156	0.194
Town hall meeting	0.064	0.226	0.158
Winning race without vehicle	0.102	0.102	0.117
Working metal crafts project	0.040	0.021	0.036
<i>Mean</i>	0.094	0.117	0.130

microwave concepts for “*repairing appliance*” and music stool, instrumental musician and acoustic guitar concepts for “*music performance*”. While the specific concepts may be distinctive for recognizing some events, the concepts typically occur in only few videos. Hence, they are absent in most videos and do not contribute much to generic event recognition. Therefore, if the vocabulary consists of specific concepts only, it will perform well in recognizing the events relevant to those concepts, but it will perform poor for other events. In contrast to the specific concepts, general concepts occur in a large numbers of videos. Although these concepts are not discriminative individually, taking several of them together into a vocabulary makes the event recognition better than using a specific vocabulary. Since it is able to simultaneously utilize distinctive specific concepts and general concepts, the best performance is obtained when the vocabulary contains a mixture of both specific and general concepts. We recommend to insert both general and specific concepts into the event recognition vocabulary.

Table 5.: Experiment 3: Comparison of specific, general and mixture vocabularies for event recognition on the Columbia CV dataset. Results and conclusions are similar as in Table 4.

Event	Specific	General	Mixture
Basketball	0.214	0.240	0.290
Baseball	0.130	0.273	0.167
Soccer	0.169	0.259	0.226
Ice Skating	0.215	0.204	0.222
Skiing	0.271	0.195	0.307
Swimming	0.324	0.163	0.531
Biking	0.125	0.298	0.177
Graduation	0.097	0.112	0.191
Birthday	0.158	0.256	0.222
Wedding Reception	0.112	0.121	0.129
Wedding Ceremony	0.124	0.147	0.181
Wedding Dance	0.263	0.301	0.316
Music Performance	0.313	0.297	0.305
Non-Music Performance	0.224	0.244	0.255
Parade	0.376	0.370	0.384
Mean	0.208	0.232	0.260

2.6.4 Experiment 4: How accurate?

Tables 6 and 7 demonstrate the effect of training less accurate vocabulary concept detectors on event recognition performance. Comparing *100%-3SIFT-SP* and *30%-3SIFT-SP* demonstrates the effect of using less examples to train vocabulary concept detectors. It shows that training concept detectors on 30% of the available training data does not substantially degrade the event recognition performance. More specifically on the TRECVID MED dataset, the performance is degraded only by a relative 8% in terms of MAP, and on the Columbia CV dataset the event

Figure 10.: Experiment 5: Event recognition performance is robust when small amounts of noise are inserted into the concept detectors of the vocabulary. The more accurate the concept detectors in a vocabulary, the higher the event recognition performance. However, adding more detectors with the same noise levels may be a more straightforward way to increase event recognition performance.

Table 6.: Experiment 4: Event recognition performance on the TRECVID MED dataset for four versions of vocabulary concept detectors with varying levels of robustness. The vocabularies include 346 semantic concepts trained on the TRECVID Semantic Indexing task 2012. More sophisticated concept detectors, using more training data, extra image descriptors, and spatial pyramids do not improve the event recognition performance substantially.

Event	100%-3SIFT-SP	30%-3SIFT-SP	30%-SIFT-SP	30%-SIFT
Attempting board trick	0.217	0.291	0.242	0.254
Feeding animal	0.045	0.042	0.075	0.043
Landing fish	0.231	0.237	0.309	0.374
Wedding ceremony	0.468	0.410	0.405	0.424
Working wood working project	0.118	0.080	0.131	0.101
Birthday party	0.119	0.136	0.148	0.144
Changing vehicle tire	0.156	0.179	0.088	0.084
Flash mob gathering	0.359	0.352	0.347	0.380
Getting vehicle unstuck	0.229	0.264	0.274	0.268
Grooming animal	0.158	0.116	0.212	0.133
Making sandwich	0.187	0.142	0.131	0.139
Parade	0.229	0.270	0.184	0.194
Parkour	0.437	0.344	0.317	0.328
Repairing appliance	0.247	0.206	0.191	0.266
Working sewing project	0.149	0.186	0.134	0.097
Attempting bike trick	0.402	0.379	0.283	0.298
Cleaning appliance	0.042	0.060	0.044	0.034
Dog show	0.417	0.342	0.401	0.341
Giving directions location	0.028	0.043	0.042	0.037
Marriage proposal	0.018	0.024	0.028	0.022
Renovating home	0.117	0.065	0.085	0.125
Rock climbing	0.271	0.178	0.175	0.251
Town hall meeting	0.388	0.265	0.179	0.217
Winning race without vehicle	0.082	0.064	0.104	0.043
Working metal crafts project	0.044	0.055	0.033	0.034
Mean	0.206	0.189	0.182	0.185

recognition performance is not degraded at all. Comparing 30%-3SIFT-SP and 30%-SIFT-SP demonstrates the effect of using fewer descriptor types in training the detectors. It shows that using only SIFT descriptors, rather than concatenation of SIFT, Opponent-SIFT and C-SIFT descriptors, degrades the MAP is only by a relative 4% and 5% for the TRECVID MED and the Columbia CCV datasets, respectively. Furthermore, comparing 30%-SIFT-SP and 30%-SIFT demonstrates that including spatial pyramids in the detectors does not improve the event recognition. To summarize, the results demonstrate that the more sophisticated detectors do not substantially improve the event recognition performance. Rather than training less sophisticated detectors we also perform the experiment with degrading the concept detector accuracies by imposing noise into their prediction scores. As expected, the results in Figure 10 demonstrate that event recognition performance degrades by adding more noise to the concept detector predictions in the vocabulary. When the noise amount is rather small, *i.e.*, up to 30%, the event recognition remains relatively robust. For a vocabulary containing 1,346 concepts, on the TRECVID MED for example, the relative performance drops by only 3% when the noise amount is 30%. When

Table 7.: Experiment 4: Repetition of the experiment explained in Table 6 on the Columbia CV dataset.

Event	100%-3SIFT-SP	30%-3SIFT-SP	30%-SIFT-SP	30%-SIFT
Basketball	0.469	0.532	0.490	0.489
Baseball	0.187	0.177	0.180	0.206
Soccer	0.435	0.467	0.466	0.499
Ice Skating	0.473	0.500	0.490	0.456
Skiing	0.494	0.507	0.457	0.485
Swimming	0.527	0.621	0.593	0.559
Biking	0.279	0.296	0.267	0.247
Graduation	0.207	0.185	0.177	0.179
Birthday	0.304	0.309	0.280	0.282
Wedding Reception	0.228	0.210	0.198	0.191
Wedding Ceremony	0.201	0.198	0.190	0.203
Wedding Dance	0.504	0.467	0.481	0.484
Music Performance	0.310	0.310	0.303	0.249
Non-Music Performance	0.277	0.286	0.267	0.264
Parade	0.497	0.494	0.473	0.495
Mean	0.359	0.371	0.354	0.353

50% noise is inserted into the concept detection results for the full vocabulary, the performance drops by 11%. It means that even if 50% of the detector predictions are distorted, the event recognition performance will be degraded by only 11%. We observe even more robust behaviour against the imposed noise on the Columbia CV. Interestingly, it implies that improving the current level of concept detector accuracy has at best a limited influence on the overall event recognition performance.

What is more, improving the detector accuracies has the same effect on event recognition performance as adding more detectors to the vocabulary. If we insert 50% noise into the vocabulary made of 50 concept detectors, on the TRECVID MED for example, the event recognition performance is 0.10 in terms of MAP. We may improve the accuracy by removing the noise again, or by adding 50 more (noisy) concept detectors to the vocabulary. In both cases the event recognition performance increases to 0.13 in terms of MAP. We observe similar behavior on Columbia CV. Considering the wide availability of large amounts of training data for concept detectors [54], adding more concept detectors seems to be more straightforward than improving the detector accuracies for event recognition vocabularies.

Our experiments confirm the observation by Hauptmann *et al.* [49]: effective video retrieval can be achieved even when concept detector accuracies are modest, if sufficiently many concepts are combined. As a conclusion, we recommend to increase the size of the concept vocabulary rather than improving the quality of the individual detectors.

2.6.5 Experiment 5: How to normalize?

The results of this experiment, are shown in Tables 8 and 9. Both tables demonstrate that the representation obtained from un-normalized detector predictions is outperformed by all the normalized representations. More specifically on the TRECVID MED, normalizing the detector predictions by sigmoid, z-score, and w-score normalization improves the event recognition performance, in terms of MAP, by 13%, 68% and 89%, where on Columbia CV the numbers

Table 8.: Experiment 5: Comparison of different methods for normalizing concept detector predictions on the Columbia CV dataset. Normalization improves the event recognition performance substantially.

Event	Un-normalized	Sigmoid [125]	Z-score [56]	W-score [133]
Basketball	0.208	0.384	0.573	0.560
Baseball	0.048	0.076	0.189	0.260
Soccer	0.143	0.241	0.465	0.439
Ice Skating	0.325	0.427	0.507	0.605
Skiing	0.286	0.446	0.584	0.574
Swimming	0.290	0.303	0.671	0.563
Biking	0.105	0.146	0.300	0.421
Graduation	0.102	0.082	0.282	0.273
Birthday	0.126	0.097	0.382	0.320
Wedding Reception	0.148	0.117	0.278	0.307
Wedding Ceremony	0.059	0.071	0.250	0.366
Wedding Dance	0.173	0.128	0.532	0.563
Music Performance	0.103	0.146	0.348	0.431
Non-Music Performance	0.103	0.157	0.326	0.222
Parade	0.142	0.179	0.522	0.421
<i>Mean</i>	0.157	0.200	0.414	0.422

are 27%, 164% and 169%, respectively. This substantial improvement is achieved because normalization boosts the event representation by making the predictions of different concept detectors comparable. Looking into the distribution of detector predictions, as illustrated in Figure 11, we observe that different detectors generate different predictions distributions, which are not directly comparable. For some detectors a prediction score might indicate absence of the concept, while for some other detectors exactly the same score might indicate concept presence. Normalizing the predictions makes the vocabulary concept detectors more comparable, which leads to a better event recognition.

Comparing the performance of supervised sigmoid normalization, with unsupervised z-score and w-score normalizations, we observe that unsupervised methods are more effective in representing events. This contradicts the common practice in the literature to rely on sigmoid normalization *e.g.* [62, 104, 180]. As shown in Table 9 for the TRECVID MED dataset, using supervised score normalization we obtain an event recognition accuracy of 0.181, in terms of MAP. But with unsupervised normalization we achieve an MAP of 0.267 and 0.300 for z-score and w-score, respectively. Similarly the supervised normalization is substantially outperformed by unsupervised normalizations on the Columbia CV dataset. The lower performance of supervised normalization is mainly caused by the fact that it assumes the distribution of concept presence on training and test data are similar. But this assumption is violated when the concept detectors are applied, as a vocabulary, on arbitrary videos that could have different concept presence distribution from the detectors training data. For example, the concept `Military Vehicle` might have a high probability of presence in its training data but it might never be present in the event videos. The difference in concept presence distributions between concept detector training data and event videos degrades the normalization performance, leading to a less effective event representation. As Table 9 shows, despite its simplicity z-score normalization performs well in normalizing the

Table 9.: Experiment 5: Repetition of the experiment explained in Table 8 on the TRECVID MED dataset.

Event	Un-normalized	Sigmoid [125]	Z-score [56]	W-score [133]
Attempting board trick	0.173	0.186	0.300	0.417
Feeding animal	0.052	0.052	0.031	0.043
Landing fish	0.151	0.228	0.336	0.439
Wedding ceremony	0.113	0.115	0.476	0.569
Working wood working project	0.161	0.166	0.150	0.185
Birthday party	0.174	0.177	0.321	0.323
Changing vehicle tire	0.075	0.089	0.224	0.207
Flash mob gathering	0.400	0.426	0.406	0.500
Getting vehicle unstuck	0.319	0.308	0.396	0.391
Grooming animal	0.105	0.124	0.167	0.166
Making sandwich	0.125	0.130	0.185	0.213
Parade	0.227	0.240	0.236	0.323
Parkour	0.100	0.119	0.399	0.482
Repairing appliance	0.120	0.127	0.396	0.376
Working sewing project	0.203	0.211	0.282	0.351
Attempting bike trick	0.206	0.280	0.387	0.494
Cleaning appliance	0.060	0.049	0.061	0.070
Dog show	0.353	0.424	0.569	0.537
Giving directions location	0.050	0.053	0.053	0.035
Marriage proposal	0.029	0.033	0.032	0.075
Renovating home	0.127	0.135	0.239	0.234
Rock climbing	0.280	0.313	0.330	0.380
Town hall meeting	0.248	0.370	0.422	0.483
Winning race without vehicle	0.096	0.133	0.151	0.112
Working metal crafts project	0.034	0.037	0.137	0.095
<i>Mean</i>	0.159	0.181	0.267	0.300

detector outputs and achieves an event recognition accuracy of 0.267 in terms of MAP. We explain this by the observation that many concept detectors generate bell-shaped score distributions that could be modeled as Gaussian distribution. However, this Gaussian assumption is not valid for all the score distributions. Some concept detector distributions have high skewness and some others are not even bell-shaped, which violates the Gaussian distribution assumption, so degrades the normalization effectiveness. In our experiments, the best event recognition performance is obtained after applying the unsupervised and assumption-free w-score normalization. We explain it by two reasons. First, W-score, as an unsupervised normalization, does not suffer from the possible incompatibilities between the concept distributions in the concept detector training data and the event training data. Second, W-score does not make any assumption about the overall distribution of concept detector scores, leading to better generalization.

As a conclusion, we recommend to normalize the detector predictions in a concept vocabulary, preferably by unsupervised and assumption-free normalizations.

Figure 11.: Experiment 6: The distribution of detector predictions generated by four different concept detectors. Since the predictions have different ranges and distributions, they are not directly comparable. For example the prediction score -0.5 provides strong evidence about concept presence for Road and Building, while the same prediction indicates the concept absence for Person and Indoor. Hence, the predictions should be normalized before being used for representing events.

2.7 RECOMMENDATIONS

In this paper we study concept vocabularies for event recognition by characterizing the concept vocabulary composition and vocabulary concept detectors. We consider five research questions related to the number, the type, the specificity, the quality and the normalization of the detectors in concept vocabularies. From the analysis of our experiments using 1,346 concept detectors, two large public video datasets containing 40 events and a general event recognition pipeline, we arrive at the following five recommendations:

- **Recommendation 1:** In general, use vocabularies containing more than 200 concepts.
- **Recommendation 2:** Make the vocabulary diverse by including various concept types: *object*, *action*, *scene*, *people*, *animal* and *attributes*. However, selecting too many concepts from the same type, especially the less diverse concept types, leads to correlated concepts and should be avoided.
- **Recommendation 3:** Include both general and specific concepts into the vocabulary.
- **Recommendation 4:** Increase the size of the concept vocabulary rather than improve the quality of the individual detectors.
- **Recommendation 5:** Normalize the predictions of vocabulary concept detectors, preferably by an un-supervised and assumption-free normalization.

The recommendations may serve as guidelines to compose the appropriate concept vocabularies for future endeavors aiming for recognizing and, ultimately, explaining the semantics of complex events in video.

COMPOSITE CONCEPT DISCOVERY FOR ZERO-SHOT VIDEO EVENT DETECTION

3.1 INTRODUCTION

We address the problem of zero-shot event detection in video, where the goal is to detect complex events without the use of *any* visual training examples. Since there is no visual example available to train event classifiers, the common approach in zero-shot detection literature relies on textual specification of events to extract the event models. For this purpose, each event is modeled by specifying its relevant and irrelevant semantic concepts, which are identified from text. Afterward, test videos are ranked based on their similarity to the event model. For this purpose, every test video should be represented by its semantic features, such as transcripts from automatic speech recognition (ASR) [26, 109], text from video optical character recognition (VOCR) [109], and scores from concept classifiers [26, 180]. Using ASR and VOCR features leads to high precision in event detection, as these sources generally result in reliable textual descriptions. However, not all videos come with ASR and VOCR features, leading to low recall in event detection. In contrast, high recall is reported for zero-shot detection using concept classifiers, but the low accuracy of concept classifiers leads to poor precision in detecting events [26]. In this paper, we consider zero-shot event detection using concept-classifiers.

In zero-shot event detection by concept classifiers, videos are represented as the output of a vocabulary of concept classifiers. The common approach is to train a one-against-all classifier per concept in the vocabulary [26, 47, 104, 180]. Hence, the underlying assumption for these approaches is that the concepts in the vocabulary, and their annotations, are independent from each other. This strong assumption may not always be valid. Consider for example the concept classifiers “Bike” and “Motorcycle”, which are very likely to have a considerable overlap in the visual context in which they may appear. Consequently, we argue it is advantageous to exploit the visual (and semantic) consistency of various concept classifiers. Our intuition is that there are combinations of concepts, for which training one joint-classifier is more effective than separately training concept classifiers. Continuing the example, combining the positive annotations for “Bike” and “Motorcycle” into a single concept classifier may result in a more reliable classifier for zero-shot event detection. In addition, some of the concept combinations may be very descriptive and characteristic for an event, *i.e.* the combination of “Bike” and “Ride” better characterizes the event “attempting bike trick” than the “Bike” and “Ride” concepts individually. In this paper, we consider the interrelation between concepts before training vocabulary concept classifiers. Based on concept relations, we combine vocabulary concepts for each event so as to optimize the event detection accuracy.

Others have also investigated optimizing the vocabulary concepts for event detection [26, 100, 111]. In [100], for example, an iterative feature selection algorithm is proposed that learns from examples to select a subset of pre-trained concept classifiers that optimizes event detection

	Ride	Motorcycle	Bike	Bike-AND-Ride	Bike-OR-Motorcycle	Concept Annotations
0	0	1	0	1
1	0	1	1	1
1	1	0	0	1	

Figure 12.: Video examples for training vocabulary concept classifiers accompanied with their concept annotations. Positive and negative labels are denoted by 1 and 0. Primitive and composite concept annotations are shown in black and red colors. Composite concepts annotations are inferred from their primitive concepts.

accuracy. We also aim for optimizing the vocabulary concepts per event, but rather than selecting from rigid concept classifiers only, we introduce a flexible composition of classifiers by adapting their underlying training data. This bears similarity to recently introduced bi-concepts [88] and visual phrases [33] for concept detection. In these works, the concept combination is defined as the co-occurrence of annotations in training data. Learning concept detectors from these co-occurred annotations results in more effective and descriptive classifiers. We observe that bi-concepts combine concept annotations by logical “AND” relations. In this paper, we generalize this combination logic by also considering other logical relationships in particular logical “OR”, *i.e.* “Bike-OR-Motorcycle”. We call these logical combinations *composite concepts* and define them as the logical composition of *primitive concepts*, see Figure 12.

The main challenge in constructing composite concepts is to discover which primitive concepts should be combined together. This problem has been studied by Rastegari *et al.* [129], for discovering bi-concepts for image search. They discover bi-concepts by searching for concept pairs whose joint classifier is more accurate than individual concept classifiers. However, combining the concepts by only considering their classification accuracy might fail for event detection. For example, combining the concepts “Dog” and “Cat” might lead to a more accurate concept classifier. However, the “Dog” concept is individually more effective for detecting the event *dog show* and it loses its effectiveness when combined with other concepts. Different from the references, we propose an algorithm that automatically discovers composite concepts by jointly considering the accuracy of concept classifiers *and* their effectiveness for detecting events.

The main contribution of this paper are: First, we propose the notion of composite concepts for constructing concept vocabularies for zero-shot event detection. Second, we propose an algorithm to automatically discover composite concepts from a vocabulary of primitive concepts.

Third, our experiments on the challenging TRECVID Multimedia Event Detection 2013 corpus demonstrates the effectiveness of composite concepts for zero-shot event detection.

3.2 RELATED WORK

Representing videos as scores from a vocabulary of concept classifiers is shown to be promising for event detection [47, 62, 96, 104], especially when only few [102, 180] or no visual example of the events [26, 90] are available. In the references, the vocabulary concept classifiers are trained from a set of images or videos, which are annotated with respect to presence or absence of the concepts. The common approach is to train a separate classifier per concept. Different from the references, we train concept classifiers from combinations of vocabulary concept annotations, as composite concepts.

Optimizing the vocabulary concepts per event has recently attracted research attention. In [26, 100], for example, the vocabulary is optimized for each event by automatically selecting a subset of concepts per event. In [111] the Wordnet ontology is used to measure the relevance between each query and the available vocabulary concepts, in order to select the most relevant concepts. Dalton *et al.* [26] select the relevant concepts per event by considering concept dependencies modeled by a Markov Random Field. Mazloom *et al.* [100] rely on supervised feature selection to select the most discriminative concepts per event. Our paper is different from these works in the following two ways. First, sometimes two concepts are individually uninformative for an event but their composition is informative and should be kept in the vocabulary. Therefore, instead of excluding the uninformative concepts from the vocabulary, we search for a composition where the concepts are informative. Second, in addition to considering the informativeness of concepts for events, we also consider concept classification accuracy. Therefore, we include the concepts in the vocabulary which are not only informative for an event, but are also accurately classified.

Rather than combining the concept classifiers in the vocabulary a posteriori, one can also combine them a priori by considering concept interrelationships during training concept classifiers [17, 33, 88, 129, 182]. In [33, 88] the notion of “bi-concepts” or “visual phrase” are introduced as the co-occurrence of distinct concepts which correspond to a very characteristic appearance that makes their detection, as one concept, more effective. Bi-concepts have been used for various purposes: sentiment analysis [17], where the sentiment concept classifiers are trained as bi-concepts of adjective and nouns, *i.e.* cute dog. Image search [88], where co-occurring query concepts are trained together as bi-concepts. Object recognition [33], where co-occurring objects are trained as one bi-concept. Moreover, Rastegari *et al.* [129] proposes to automatically discover bi-concepts from query concepts for image search. They consider a combination of concepts as a bi-concept only if it is more effectively classified than its individual concepts.

Our paper is different from these works in the following ways. First, bi-concepts can be considered as one type of all possible concept compositions, where concepts are composed by “AND” relations. Differently, our proposed composite concepts can be discovered from any type of concept relations, including “OR”, “AND”, “XOR” etc. Second, Rastegari *et al.* [129] discover bi-concepts only based on their classification accuracy and ignore their effectiveness when classifying events. Differently, we discover composite concepts by jointly considering their classification accuracy *and* their effectiveness for recognizing events. As we will show in the experiments, this criteria has a high impact on performance of the composite concepts in event detection.

Figure 13.: Data flow diagram of our defined zero-shot event detection pipeline. The notations are explained in Section 3.3.1.

3.3 COMPOSITE CONCEPT DISCOVERY

We first formalize our zero-shot event detection settings. Then we propose our algorithm for discovering composite-concepts. Finally, we discuss about the computational costs of the proposed algorithm.

3.3.1 Zero-Shot Event Detection

In zero-shot event detection the goal is to detect an event, without using any visual examples of the event, see Figure 13 for an illustration. The detection problem is decomposed into two parts, first, low-level video features are mapped into an intermediate semantic representation, and second, the semantic representation is mapped to an event. In this work, we consider a set of concepts $C = \{c_1, \dots, c_l\}$ as the intermediate representation, and they cover a wide range of concepts including objects, actions, and scene concepts, such as the concepts recommended in [47].

To obtain the intermediate representation, we learn a mapping, $\mathcal{M}_c : x \mapsto \mathbf{a}$ from a d -dimensional video feature $x \in \mathbb{R}^d$ to an ℓ -dimensional binary vector $\mathbf{a} \in \{0, 1\}^\ell$, denoting the presence/absence of the concepts in C . $\mathcal{M}_c$ is learnt from a set of annotated videos, denoted as $\mathcal{A}^c = \{(x_i, \mathbf{a}_i)\}_{i=1}^I$, which we call the *video-concept annotations*. Using these annotation the mapping is obtained by training a classifier for each concept in C independently.

For the event detection, we learn a mapping, $\mathcal{M}_e : \mathbf{a} \mapsto e$ from the intermediate representation $\mathbf{a}$ to the event label $e \in \{0, 1\}$. In our work, $\mathcal{M}_e$ is obtained from a train set of textual descriptions of events, denoted by $\mathcal{A}^e = \{(\mathbf{a}_j, e_j)\}_{j=1}^J$, which we coin the *concept-event annotations*. This can be provided at category level or at instance level. For category level descriptions, there is just a single concept annotation per event. For instance level annotations, multiple concept-event annotations are provided for the same event denoting possibly a variety of concepts related to the event. We use instance level concept-event annotations, since these are usually easier to obtain. Using the annotation set $\mathcal{A}^e$, mapping $\mathcal{M}_e$ is obtained by training an event classifier from concept annotations as features.

Generalizing to category level annotations is straight forward.

Note that, the two mappings $\mathcal{M}_c$ and $\mathcal{M}_e$ are learned from a different train set, and that for training $\mathcal{M}_e$ no visual example is used. The performance of such a zero-shot detection framework, critically depends on the quality of the chosen intermediate representation.

3.3.2 Algorithm

In this section we detail our algorithm to learn a set of composite concepts for the zero-shot prediction of an event. We propose to use composite concepts as intermediate representation which are learned for a specific event e . Composite concepts are derived from the set of primitive concepts C , as a combination of two (or more) primitive concepts, with a logical operator. For simplicity, we only describe the use of the logical OR operator, but the described algorithm holds for the other operators as well.

Finding the composite concepts is reducible to a set partitioning problem. The goal is to find a set of composite concepts $\widehat{C}$, which is a division of primitive concepts C , as a union of non-overlapping and non-empty subsets. For a set C of ℓ primitive concepts, the total number of possible partitions is the Bell number $B_\ell = \sum_{k=0}^{\ell-1} \binom{\ell-1}{k} B_k$. Finding an optimal $\widehat{C}$ is an NP-hard problem, see e.g. [129], for which we propose a greedy approximation.

The accuracy of the event detection accuracy depends on two competing objectives, *concept predictability* and *event predictability*. Concept predictability measures the accuracy of the prediction of the concepts from the intermediate representation. Event predictability measures the discriminative power of the intermediate representation to detect an event. Based on these two criteria, we find the set of composite concepts $\widehat{C}$ for event e by maximizing:

$$S_e(\widehat{C}) = \lambda P_c(\widehat{C}) + (1-\lambda)P_e(\widehat{C}) \quad (1)$$

where P_c and P_e measure the concept predictability and event predictability, respectively, and will be described in detail below. Moreover, $\lambda \in [0, 1]$ is a parameter to balance between concept predictability and event predictability. This parameter can be optimized by cross-validation. However, when visual examples of events are not available for cross-validation, we pick $\lambda = 0.5$ to equally weight concept predictability and event predictability.

Our proposed greedy approximation is similar to hierarchical clustering. Starting from the set of primitive concepts $C_1 \leftarrow C$, in each iteration t we find two concepts n^* and m^* to be merged. The two concepts are selected based on the expected improvement of Eq. 1:

$$(m^*, n^*) = \arg \max_{m, n \in C_t, s.t. m \neq n} \Delta S_e(C_t, m, n), \quad (2)$$

where $\Delta S_e(C, m, n)$ denotes the difference between using C_t and the set where m and n are used as composite concept. We use n^* and m^* to define C_{t+1} , as:

$$C_{t+1} = C_t - C_{n^*} - C_{m^*} + (C_{n^*} \vee C_{m^*}). \quad (3)$$

The clustering algorithm is terminated at iteration t' when $\Delta S_e(C_t, m, n) < 0$. The final set of composite concepts $\widehat{C}$ used for this event is $\widehat{C} \leftarrow C_{t'}$. The Pseudo code of our clustering procedure is summarised in Algorithm 1.

Concept predictability

We measure the concept predictability of a set of concepts C , by their classification performance on a part of the train set. Therefore, we first learn the mapping $\mathcal{M}_c$ for each concept $c \in C$, using

```

input :  $C, \mathcal{A}^c$  and  $\mathcal{A}^e$ 
output :  $\widehat{C}$ 
 $C_1 \leftarrow C$ 
for  $t \leftarrow 1$  to  $\ell$  do
    compute  $\Delta S_e$  for each pair  $m, n \in C$  ( $m^*, n^*$ )  $\leftarrow \arg \max \Delta S_e$ 
    if  $\Delta S_e(m^*, n^*) > 0$  then
         $C_{t+1} = C_t - C_{n^*} - C_{m^*} + (C_{n^*} \vee C_{m^*})$ 
    else
        return  $\widehat{C} \leftarrow C_t$ 
    end
end

```

Algorithm 1: Pseudo code for the proposed algorithm to discover composite concepts for the event e . Notation conventions are detailed in Section 3.3.2.

the annotation $\mathcal{A}^c$. Then we evaluate this mapping on a hold out partition of training data, using the average precision measure, denoted as $P_{AP}(c)$. So, the concept predictability is given by:

$$P_c(C) = \frac{1}{|C|} \sum_{c \in C} P_{AP}(c). \quad (4)$$

In contrast to [129], this notion of concept predictability relies on the accuracy of the classifiers, rather than an estimation based on a geometric intuition.

For the greedy algorithm, we are interested in relative improvement of C_t when merging concepts m and n . Let $k = |C_t|$, and $C_{t'}$ denote the set C_t minus c_m and c_n , then:

$$\begin{aligned}
 \Delta P_c(C_t, m, n) &= \frac{1}{k-1} \sum_{c \in C_{t'}} P_{AP}(c) + \frac{1}{k-1} P_{AP}(c_m \vee c_n) \\
 &\quad - \frac{1}{k} \sum_{c \in C_{t'}} P_{AP}(c) - \frac{1}{k} (P_{AP}(c_m) + P_{AP}(c_n)) \\
 &\approx P_{AP}(c_m \vee c_n) - \frac{1}{2} (P_{AP}(c_m) + P_{AP}(c_n)), \quad (5)
 \end{aligned}$$

for the approximation in the final step, we assume that $\sum_{c \in C_{t'}} P_{AP}(c)$ has a value independent of the chosen m and n . For each iteration of the clustering algorithm, we train a classifier for each pair of concepts m and n , and compute the value of $\Delta P_c(C_t, m, n)$ by using Eq. 5.

Event predictability

The event predictability is measured by the quality of the mapping from the composite concepts to the event label, using the train set $\mathcal{A}^e$. More precisely, given a set of composite concepts C_t , we train an event classifier on the concept annotations $\mathbf{a}_j$ as features, apply the composition given by C_t , and using the labels e_j as desired outcome. Then the trained classifier is evaluated on a hold out partition of training data using average precision. The event predictability is given by:

$$P_e(C) = P_{AP}(e), \quad (6)$$

where C denotes the composition which should be applied on the concept annotations $\mathbf{a}$.

For the greedy algorithm, we evaluate the relative improvement of merging concepts m and n of set C_t , by:

$$\Delta P_e(C_t, m, n) = P_e(C_t, m, n) - P_e(C_t). \quad (7)$$

3.3.3 Computational Efficiency

For a vocabulary of n primitive concepts our algorithm needs to train $O(n^2)$ concept classifiers to discover composite concepts. In the first iteration $\frac{n(n-1)}{2}$ concept classifiers are trained. By storing the accuracies of the concept classifiers in the first iteration, we only have to train $n - 1$ new concept classifiers, in the second iteration, to compare the accuracy of the new composite concept to the existing $n - 1$ classifiers. Similarly $n - 2$ concept classifiers are trained in the third iteration and so on, which leads to an overall computational complexity of $O(n^2)$ in training concept classifiers.

Even though we use an efficient greedy approximation of the original learning problem, the computational complexity is still rather high. This is due to the high dimensional video features used for training the concept classifiers, as required by Eq. 5.

However, Eq. 5 shows that we are interested in the *performance differences* between concepts and composite concepts, not in the maximum achievable performance of these (composite) concepts. Visual classifiers, especially in a large-scale setting with many examples or concepts, can be effectively obtained by using stochastic gradient descent [124]. Stochastic gradient descent is an optimization algorithm, which is used to gradually train the classifiers by randomly passing through the training data. For good performance a high number of epochs are required, typically in the order of 100 – 500, however, for the performance difference we could use the classifier obtained after just a few epochs, since this is typically a good predictor of the expected performance.

Therefore, we use very early stopping (after 5 epochs) and a large value of the learning rate parameter, to ensure rapid learning from the given examples.

3.4 EXPERIMENTAL SETUP

3.4.1 Datasets

Video Data: We perform our experiments on the challenging TRECVID Multimedia Event Detection 2013 corpus, containing in total 51K arbitrary videos collected from the web. To the best of our knowledge this is the largest publicly available video corpus in the literature for event detection containing user-generated video with a large variation in quality, length and content. We perform our experiments on three partitions of videos: Research, MED test, and Kindred test including 10K, 27K, and 14K videos, respectively. Apart from the Research partition, the two other partitions come with ground truth annotation at video level for 20 event categories, such as *Marriage proposal*, *Attempting bike trick* and *Making sandwich*. In all our experiments, we followed the instruction provided by NIST [118]. We use the MED test and Kindred test partitions to report event detection results and use the Research partition to train vocabulary concept classifiers.

Video-Concept Annotations: Although the TRECVID Multimedia Event Detection 2013 corpus does not provide video level concept annotations, it comes with a textual summary for each video in the collection, describing what is happening in each video. We use these descriptions to automatically extract video-concept annotations by following the approach proposed in [11]. This approach starts by considering every frequent term in the text collection as a vocabulary concept and terms presence/absence in video descriptions as video-concept annotations. The obtained video-concept annotations are noisy, so [11] proposes to prune the extracted video-concept annotations by excluding some noisy concepts. For this purpose, a classifier is trained and evaluated per concept, then the concepts with low detection accuracy are considered as noise

and excluded from the vocabulary. We executed this procedure on the textual descriptions of the Research partition videos, which led to extracting 138 concepts and their video-concept annotations. The extracted concepts cover a wide range of semantics needed to represent events including, objects, actions, scenes and people related concepts. In our experiments, we rely on these extracted annotations as video-concept annotations. Our extracted video-concept annotations are available for download at: <http://www.mediamill.nl/datasets>.

Concept-Event Annotations: Following the standard practice in zero-shot detection, as discussed in Section 3.3.1, we require concept-event annotations to map the semantic video representation to an event. Similar to video-concept annotations, we rely on textual descriptions of events to extract concept-event annotations. For each event we use 10 textual descriptions from positive examples of the event, provided in the TRECVID Multimedia Event Detection 2013 corpus. Then each description is manually represented in terms of the 138 vocabulary concepts.

3.4.2 Vocabulary Concept Classifiers

Using the Research partition videos and video-concept annotations we train a classifier per vocabulary concept. As local descriptor we use MBH computed along the motion trajectories [162]. Fisher encoding is used to aggregate them followed by power normalization with $\alpha = 0.2$ as in [57]. This representation is shown to be state-of-the-art for recognizing events using single modality [164]. Better event detection accuracy is obtained by fusing multiple modalities [109] but it is beyond the scope of this paper. We use linear SVM to train the concept classifiers.

3.4.3 Experiments

1. Zero-Shot Event Detection: We evaluate the effectiveness of our composite concept vocabulary for zero-shot event detection. We compare the discovered OR-composite vocabulary with several baselines: i) primitive concepts vocabulary, which includes 138 concept detectors separately trained per primitive concept. ii) bi-concepts vocabulary, which includes bi-concepts discovered as proposed in [129]. iii) selected concepts vocabulary, which includes a subset of primitive concepts which are more informative per event [100]. Informative concepts are selected using mRMR [123] feature selection from concept-event annotations per event. We select the same number of concepts as in the OR-composite concepts vocabulary.

2. OR-Composite vs AND-Composite: We investigate the characteristics of AND-composite and OR-composite concept vocabularies by evaluating their performance for zero-shot event detection. Both the composite vocabularies are discovered as discussed in Section 3.3.2, by using “AND” and “OR” operators respectively to compose the concepts. Moreover, we compare their performance with a fused vocabulary containing both the discovered AND-composite and OR-composite concepts.

3.5 RESULTS

3.5.1 Zero-Shot Event Detection

The results of this experiment are shown in Table 10. It demonstrates that the OR-composite vocabulary outperforms the primitive vocabulary with mAP of 5.97% vs 4.00% for MED test set and mAP of 14.69% vs 10.47% for Kindred test set. For some events the improvement is considerable, *i.e. winning race without vehicle* and *flash mob gathering*, whose detection accuracy is increased from 2.67% to 10.98% and from 18.41% to 31.86% respectively, on MED test set.

Table 10.: Experiment 1: Comparing the effectiveness of various concept vocabularies in zero-shot event detection. Our proposed OR-composite concepts outperform present-day alternatives.

Event	MED test				Kindred test			
	Primitive	Selection [100]	Bi-Concepts [129]	OR-Composite	Primitive	Selection [100]	Bi-Concepts [129]	OR-Composite
Birthday party	5.30	4.97	4.72	7.55	6.51	4.84	6.49	9.64
Changing vehicle tire	0.96	0.97	0.80	1.81	1.21	1.20	1.33	1.12
Flash mob gathering	18.41	22.98	8.95	31.86	15.80	13.67	11.59	22.52
Getting vehicle unstuck	3.55	3.40	3.08	5.54	1.51	1.54	2.55	2.2
Grooming animal	0.91	0.91	0.86	0.91	17.11	17.06	10.91	17.06
Making sandwich	7.39	7.74	7.39	7.92	64.20	62.62	64.20	66.85
Parade	19.81	21.90	19.32	22.36	5.90	7.37	5.96	6.26
Parkour	0.60	0.50	0.91	2.09	1.01	0.77	1.00	4.31
Repairing appliance	1.06	1.24	0.88	2.49	10.81	12.31	5.92	40.08
Working sewing project	1.34	1.41	1.36	1.45	24.61	29.95	30.87	27.32
Attempting bike trick	1.09	1.05	0.62	2.02	1.30	1.33	5.86	2.78
Cleaning appliance	0.47	0.46	0.47	0.63	10.21	7.63	10.15	23.87
Dog show	0.10	0.11	0.25	0.11	0.61	0.72	0.29	0.65
Giving directions location	0.79	0.75	0.52	2.49	0.30	0.23	0.23	0.45
Marriage proposal	0.13	0.12	0.23	0.15	0.91	0.78	0.30	1.48
Renovating home	0.55	0.62	0.64	2.28	4.20	6.72	4.92	6.72
Rock climbing	13.96	14.23	13.94	14.60	30.80	30.70	26.58	34.87
Town hall meeting	0.52	0.96	0.55	1.47	0.71	1.62	0.63	0.37
Winning race without vehicle	2.67	3.11	3.09	10.98	4.21	5.70	9.05	15.36
Working metal crafts project	0.41	0.41	0.49	0.59	7.41	7.37	9.64	9.94
mean	4.00	4.39	3.45	5.97	10.47	10.71	10.42	14.69

By looking into the OR-composite concepts discovered for these events, we find some composite concepts whose classifiers are more accurate than their underlying primitive concept classifiers, as illustrated in Figure 15. Moreover, the quantitative comparison of concept classifier, as reported in Figure 14, demonstrates that OR-composite concepts have higher classification accuracies than their underlying primitive concepts.

We also compare OR-composites vocabulary with bi-concepts vocabulary, which are discovered as proposed in [129]. Our experiments demonstrate that the discovered bi-concepts have poor performance in zero-shot event detection. For some events, *i.e.* *birthday party* and *flash mob gathering*, bi-concepts vocabulary are even outperformed by the primitive concepts baseline. It

Figure 14.: Comparing accuracy of OR-composite concept classifiers and primitive concept classifiers. For each event, the discovered OR-composite concepts are indicated by squares and their underlying primitive concepts are indicated by dots. OR-composite concepts are more accurately classified than their primitive concepts.

Pool	Beach	Pool-OR-Beach	Pool	Beach	Pool-OR-Beach	Pool	Beach	Pool-OR-Beach
1	0	1	1	0	1	0	1	1
0	1	1	0	1	1	1	0	1

Figure 15.: Examples of the predictions made by primitive and composite concept classifiers. Blue and yellow rows denotes the ground-truth and concept classifier predictions, respectively. The primitive concept classifier predictions are incorrect, while the composite concept classifier predictions are correct.

is mainly because in [129], bi-concepts are discovered only based on their predictability and without consideration of their effectiveness for detecting events. However, for some concepts, although their composition leads to more predictable bi-concept, the derived bi-concept is less effective in detecting the event. For example in the *birthday party* event, two concepts “dancing” and “indoor” are composed as a bi-concept because detecting dancing in indoor scenes is more accurate. However, this bi-concept is incapable of detecting dancing in outdoor scenes, which leads to missing several birthday party videos which are outdoors. It demonstrates the importance of considering event predictability in discovering the composite concepts.

Our work is also comparable to concept selection [100], where the vocabulary is obtained by selecting a subset of informative concepts per event. As Table 10 shows, OR-composite vocabulary outperforms the concept selection vocabulary with mAP of 5.97% vs 4.39% for MED test set and 14.69% vs 10.71% for Kindred test set. For some events the improvement is substantial *i.e.* *winning race without vehicle* event, where OR-composites vocabulary outperforms

Figure 16.: Comparing accuracy of AND-composite and OR-composite concept classifiers. For each event, the discovered AND-composite and OR-composite concepts are indicated by green and black squares, respectively. AND-composite concepts are more accurately classified than OR-composite concepts, but the number of discovered AND-composite concepts are generally less than OR-composite concepts.

Table 11.: Experiment 2: OR-composite vocabularies outperform AND-composite vocabularies and the best result is obtained by fusing both compositions.

<i>Event</i>	MED test			Kindred test		
	AND-composite	OR-composite	Fusion	AND-composite	OR-composite	Fusion
Birthday party	5.60	7.55	7.55	6.63	9.64	9.64
Changing vehicle tire	1.73	1.81	1.83	1.06	1.12	1.08
Flash mob gathering	26.33	31.86	37.26	17.79	22.52	28.15
Getting vehicle unstuck	3.55	5.54	5.54	1.45	2.20	2.20
Grooming animal	1.02	0.91	0.91	21.89	17.06	17.06
Making sandwich	7.64	7.92	7.92	62.35	66.85	66.85
Parade	23.05	22.36	22.36	5.83	6.26	6.26
Parkour	1.80	2.09	2.18	3.37	4.31	3.92
Repairing appliance	1.52	2.49	2.51	19.80	40.08	48.21
Working sewing project	1.34	1.45	1.45	24.58	27.32	27.32
Attempting bike trick	1.61	2.02	2.16	1.69	2.78	2.39
Cleaning appliance	0.58	0.63	0.80	11.29	23.87	28.54
Dog show	0.13	0.11	0.11	0.22	0.65	0.58
Giving directions location	0.93	2.49	2.49	0.33	0.45	0.45
Marriage proposal	0.13	0.15	0.15	0.55	1.48	1.48
Renovating home	0.60	2.28	2.28	5.83	6.72	6.72
Rock climbing	14.37	14.60	14.68	28.33	34.87	40.72
Town hall meeting	0.52	1.47	1.47	0.73	0.37	0.37
Winning race without vehicle	6.50	10.98	13.59	8.63	15.36	19.25
Working metal crafts project	0.41	0.59	0.59	7.37	9.94	9.94
<i>mean</i>	4.97	5.97	6.39	11.49	14.69	16.06

concept selection by 10.98% versus 2.67% in terms of AP. Looking into the selected concepts for this event, we observe that although the selected concepts are informative for the event, their concept classifiers are inaccurate. More specifically, the accuracy of the classifiers for “running”, “jumping”, and “walking” concepts, which are among the 10 most informative concept for this event, are 3%, 7%, and 1%, respectively. It demonstrates the importance of considering concept classifiers predictability in discovering the composite concepts.

In summary, we conclude that OR-composite concepts outperform primitive concepts in zero-shot event detection. We explain it based on the observation that OR-composite concepts have more accurate concept classifiers compared to primitive concepts. Moreover, by comparing OR-composite concepts to bi-concepts and selected concepts, we demonstrate the importance of jointly considering concept predictability and event predictability in constructing the concept vocabularies.

3.5.2 OR-Composite vs AND-Composite

We report the results of this experiment in Table 11. It demonstrates that the OR-composite vocabulary obtains a higher event detection accuracy, 5.97% vs 4.97% for MED test and 11.49% vs 14.69% for Kindred test sets. However, by comparing concept classifier accuracies for AND-composite and OR-composite concepts, as shown in Figure 16, we observe that AND-composite concepts have more accurate concept classifiers, which contradicts with their lower event detection performance. More specifically, the concept classification accuracy for AND-composite and

Detected Videos	Composite Concepts
	Group-AND-Dance-AND-Shopping Celebrating-OR-Marching Performance-OR-Music People-OR-Girl Surprise-OR-Party
	Group-AND-Dance-AND-Shopping Band-OR-Singining Inside-OR-School Performance-OR-Music Surprise-OR-Party
	Group-AND-Dance -AND-Shopping Practice-OR-Gym Living-AND-Room Street-OR-Inside Performance-OR-Music

Figure 17.: Top three videos detected for flash mob gathering event. For each video, we report the five composite concepts which have contributed most to the zero-shot detection of the event. It demonstrates that both the AND-composite and OR-composite concepts are contributing to event detection.

OR-composite concepts are 22.45% and 16.19%, respectively in terms of mAP averaged over all concepts and all events. By comparing the number of discovered composite concepts we observe that there are more OR-composite concepts discovered than AND-composite concepts. More specifically, OR-composite vocabulary contains on average 40 OR-composite concepts per event, while AND-composite vocabulary contains on average 12 AND-composite items per event. We explain the higher classification accuracy of AND-composite concepts by the observation that they are more restricted and have less visual diversity in their training examples than OR-composites. For example, the AND-composite “bike-AND-ride” depicts a restricted situation where somebody is riding a bike. But the OR-composite “bike-OR-ride” includes positive examples from various riding actions, *i.e.* riding bike, horse, skateboard etc., as well as examples from bike in various situations, *i.e.* riding, repairing, parking etc. In contrast, AND-composite have much less positive examples, which restricts discovering many AND-composite concepts with enough training data to train the classifiers. Moreover, our experiments show that fusing AND-composite and OR-composite concepts in a fused vocabulary, leads to 7.0% and 9.3% relative improvements in MED test and Kindred test sets, respectively. It demonstrates that AND-composite and OR-composite concepts contain complementary information about the events and should both be included in the vocabulary. This hypothesis is validated by looking into the concepts which contribute most to the event detection, as shown in Figure 17. This figure shows that some of the contributing concepts are from OR-composite concepts while others are from AND-composites.

3.6 CONCLUSIONS

In this paper we propose the notion of composite concepts for zero-shot event detection using concept classifiers. Composite concepts are high order semantics obtained by combining primitive concepts by logical connectors, like “AND” and “OR”. We propose an algorithm to automatically discover composite concepts per event that jointly optimizes concept and event predictability. We demonstrate that our discovered composite concepts result in more accurate concept classifiers compared to their underlying primitive concepts, which improves zero-shot event detection accuracy. Moreover, by comparing AND-composite vs OR-composite concepts we observe that AND-composite concepts generally have more accurate concept classifiers. However, AND-composite concepts annotations are sparser than OR-composite concepts, which restricts the number of discovered AND-composite concepts. Moreover, we demonstrate that AND-composite and OR-composite are complementary in representing events and their fusion leads to further improvement. Finally, we argue that by training composite concept classifiers, we model the descriptive relationships between concepts explicitly inside the concept classifiers, rather than implicitly in the event classifier. We consider the approach promising for practical use when insufficient event training examples are available to learn concept relationships.

DISCOVERING SEMANTIC VOCABULARIES FOR CROSS-MEDIA RETRIEVAL

4.1 INTRODUCTION

We consider the problem of cross-media retrieval, where for an image query we search for the relevant text or vice versa. Initially, cross-media retrieval emphasized on simple queries made of few keywords or tags [14, 59], but recently they have addressed more complex retrieval problems like searching for an image based on a long article [8, 24, 34, 127, 170], or automatically finding the best sentences as the caption to describe a video [44].

The major challenge in cross-media retrieval is that the query and the retrieval set instances belong to different domains, so they are not directly comparable. In other words, the images are represented by visual feature vectors which have a different intrinsic dimensionality, meaning, and distribution than the textual feature vectors used for the sentences. As a solution, many works aim to *align* the two feature spaces so they become comparable. We discuss below related work based on the level of alignment of feature spaces used, as also illustrated in Figure 18.

Low-level Alignment Works based on low-level alignment aim to align the images and texts directly from the low-level feature spaces, *i.e.* the low-level visual descriptors and the individual words. Cross-media hashing [145, 184], canonical correlation analysis (CCA) [48] and its variants, such as cross-modal factor analysis [83] and Kernel-CCA [8, 24], are the main techniques for learning the low-level alignments. In a nutshell, these techniques extract the highly correlated features in the two feature spaces and use them to make a correlated representation of images and text. While encouraging results for cross-media retrieval using the low-level alignments have been reported, they suffer from two limitations. First, a low-level feature space is not the most effective place to find correlations, as the semantic gap is maximized. Second, the learned alignments are difficult to interpret, making it hard for a user to explain why a certain result was retrieved. **Mid-level Alignment** Works based on mid-level alignment first extract mid-level features from each modality. Then the alignment is learned between the two mid-level feature spaces. Multi-modal topic models [14, 167] and multi-modal autoencoders [34, 112, 144] are the two major trends for learning the mid-level alignments

Blei and Jordan [14] were the first to extend the latent Dirichlet allocation (LDA) to align text and images at the topic level. Despite their effectiveness in aligning the discrete visual features, such as the traditional bag-of-visual-words, applying the multi-modal topic models on modern visual features, such as Fisher vectors [132] or deep learned visual features [117], is not straightforward. Since probabilistic topic models rely on assumptions about the prior distribution of features, which are not well known for state-of-the-art visual descriptors, the multi-modal topic models cannot benefit from them.

Recently, there has been a trend of developing multi-modal autoencoders with deep architectures for learning the mid-level alignments [34, 112, 144]. These works learn the mid-level

Figure 18.: Level of alignments (gray) for cross-media retrieval using visual (orange) and textual (blue) features. This paper contributes to semantic alignment.

representation of images and texts by training autoencoders. Then they align the learned representation by CCA [112] or regression [144]. Recently, Feng *et al.* [34] propose to learn the alignments jointly with the autoencoders. These techniques are shown to be effective for learning the mid-level alignments from large training collections. However, when there is not enough training data available, the deep autoencoders might be overfitted due to their large number of parameters. Moreover, similar to the low-level alignments, the mid-level alignments are incapable of recounting why an image and text are similar.

Semantic Alignment Instead of explicitly learning the correspondences between the images and texts, either at low-level or mid-level features, Rasiswasia *et al.* [24, 127] propose to embed the images and texts into a mutual *semantic space*. In the semantic space, each image or text is represented in terms of the probabilities of being relevant to a pre-defined *vocabulary* of semantic concepts. By representing the images and texts as their concept probabilities, they have aligned representations which are directly comparable.

The semantic representations are obtained by following three consecutive steps: First, the vocabulary is defined by specifying its concepts. The vocabulary concepts should be diverse and comprehensive enough to provide a descriptive representation of images and texts. As the second step, a train set of image/text pairs are labeled as relevant or irrelevant to each vocabulary concept. Finally using the labeled train data, a set of visual and textual concept classifiers are trained to predict the vocabulary concepts on images and texts. Each visual concept classifier is trained on the image parts and the concept labels as a binary classifier *i.e.* binary SVM or logistic regression.

The textual concept classifiers are trained in a similar way, but on the textual data of the train set. After training the concept classifiers, each image or text is embedded into the semantic space by simply applying the visual and textual concept classifiers.

In [24, 127], Rasiswasia *et al.* rely on manual annotations for learning the vocabulary concept classifiers. More specifically, they manually specify the vocabulary concepts and also manually annotate each image/text pair as positive or negative with respect to each vocabulary concept. This requires a substantial amount of annotation effort, which is restrictive in creating descriptive vocabularies with a comprehensive set of concepts. Moreover, for each new dataset a new vocabulary, which is relevant and descriptive to the data at hand need to be defined manually. To overcome these problems, we propose a data-driven approach to automatically discover the vocabulary concepts and their annotations from the textual data, which are available in the dataset.

Semantic alignment is related to an attribute-based representation, commonly used for image and video classification from few training examples [11, 32, 42, 45]. In these works, the images or videos are represented as the outputs of attribute classifiers. However, the attribute classifier are applied with the purpose of enriching the image representations by transferring the knowledge from the attributes' training source. This is different from the intention behind the semantic alignment, where the concept classifiers are applied to align the textual and visual instances.

Contributions We propose a data-driven approach to cross-media retrieval using semantic alignment that automatically discovers the vocabulary concepts and their annotations from multimedia collections. Different from [24, 127], we do not pre-define the vocabulary and we do not require *any* human annotation effort to learn the concept classifiers. Moreover, we propose a learning framework for training the vocabulary concept classifiers from the discovered annotations. We experimentally show that our discovered vocabulary outperforms the state-of-the-art low-level, mid-level, and semantic alignments in three datasets of image/text and video/text pairs.

4.2 OUR PROPOSAL

The overall pipeline, which we follow for cross-media retrieval is shown in Figure 19. In the training phase, we learn two sets of visual and textual embeddings to map each visual or textual instance into a mutual semantic space. We learn the embeddings from a training collection of multimedia instances including images or videos and their corresponding textual descriptions. Wikipedia articles, which are made of textual articles and their illustrative images, and captioned videos, which are made of videos and their textual captions, are examples of the multimedia instances used in this paper.

As the first step for learning the embeddings, we automatically specify the vocabulary concepts by discovering a set of concepts, which can effectively describe the training collection. Moreover, we automatically annotate each multimedia instance in the train set as relevant or irrelevant with respect to the vocabulary concepts. Our method for automatically discovering the semantic annotations is detailed in Section 4.2.1. After extracting the semantic annotations we use them for training visual and textual embeddings, as detailed in Section 4.2.2.

In the retrieval phase (Figure 19-b), we use the learned visual and textual embeddings to embed the query and the test set instances into the semantic space. Then the cross-media retrieval is cast into a single media retrieval problem, where the traditional similarity measurements, such as cosine similarity, are applicable for the retrieval.

Figure 19.: The overall pipeline, which we follow for cross-media retrieval. The “discovering semantic annotations” and “training semantic embeddings”, as our main contributions, are detailed in Section 4.2.1 and Section 4.2.2

4.2.1 Discovering Semantic Annotations

Instead of manually specifying the vocabulary concepts and their annotations we propose to automatically discover them by applying a probabilistic topic model.

Probabilistic topic models are statistical models for discovering the hidden semantic structure in a collection of text [15]. They automatically discover sets of interrelated terms, as *topics*, based on co-occurrences of terms in the collection. Figure 20-a illustrates three examples topics extracted from a collection of text.

In addition to discovering the topics, probabilistic topic models also extract *topic portions* for each textual instance in the collection. For each text, the topic portions determine the relevance of the text to each topic, as a probability value between 0 and 1. In the example in Figure 20-b, the text includes the terms from all three topics but with different portions. The more terms from a topic occur in a text, the higher the topic portion value assigned to the topic.

We observe that topics and topic portions are suitable to serve as the vocabulary concepts and their annotations. The topics provide a comprehensive summary of the collection, so a vocabulary composed of topics has a high descriptiveness. Moreover, as experimentally validated in Section 4.4.1, combining the interrelated concepts into a more abstract concept, as performed by topic modeling, generally improves learning the concept classifiers. More specifically, the abstract concepts, such as “animal”, are in general more accurately predicted in images/videos

Figure 20.: Illustration of topics and topic portions. (a) Three examples of topics extracted from a text collection. (b) Topic portions estimated by LDA for a multimedia instance. Figure inspired from [13].

than the specific concepts, such as “goat”, “elephant”, and “tiger”, partly due to the amount of train data available per concept. Hence, a vocabulary composed of discovered topics serves as good concepts for the semantic representation.

To summarize, we apply the latent Dirichlet allocation (LDA) [15], as a widely used probabilistic topic model, on the textual instances in the train set. The number of topics is a parameter which is determined by cross-validation, as is detailed in Section 4.4.1. Then the discovered topics are considered as the vocabulary concepts. Moreover, the discovered topic portions are considered as the concept annotations and are used for training the visual and textual embeddings.

4.2.2 Training Semantic Embeddings

The semantic embeddings are defined as projections from visual or textual feature spaces into a mutual semantic space, by predicting the relevance of each instance to the vocabulary concepts. We train the visual and textual embeddings in a similar way. Hence, we first explain our learning algorithm for the visual embedding. Then we explain how the proposed algorithm can be applied for training the textual embedding.

Training Visual Embedding. We denote the visual embedding by $\Phi_{\mathbf{W}}(x) : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^k$ as the projection of each visual instance $x \in \mathbb{R}^{D_v}$ into its k -dimensional representation in the semantic space. The visual embedding is defined as:

$$\Phi_{\mathbf{W}}(x) = \mathbf{W}^\top x,$$

where $\mathbf{W} \in \mathbb{R}^{D_v \times k}$ is the visual prediction matrix stacking the weight vectors corresponding to each vocabulary concept classifier denoted as $w_c \in \mathbb{R}^{D_v}$.

Each vocabulary concept classifier w_c is trained on its train set $\mathcal{D}_c = \{(x_i, y_i), i = 1 \dots N\}$, where $x_i \in \mathbb{R}^{D_v}$ denotes the visual instances in the train set. For example, in a train set of Wikipedia articles, x_i refer to the visual features extracted from the article’s image. Moreover, y_i is the concept label denoting the relevance of the visual instance x_i to the concept c .

As proposed in Section 4.2.1, instead of manually annotating the concept labels, we automatically extract them from the topic portions estimated by LDA. More specifically, for each visual instance x_i we use the estimated topic portion of its corresponding text to extract the concept label y_i . However, the topic portions are estimated as continuous probability values between 0 and 1, so using them as labels for training the classifiers is not trivial. We consider three possible strategies for defining the concept labels based on the topic portions and discuss their applicability as follows:

		Value as Label	Binarized value as Label	Rank as Label
American Airlines Flight 11 was a passenger flight ...		0.99	1	1
The Adelaide O-Bahn Busway is a guided busway ...		0.92	1	2
Southwestern Ontario in Southwestern Ontario, several ...		0.83	1	3
⋮	⋮	⋮	⋮	⋮
Winter service vehicles are operated by both ...		0.01	-1	N-1
The birds usually live in southern parts of the ...		0.00	-1	N
		(a)	(b)	(c)

Figure 21.: Strategies for extracting concept labels from estimated topic portions.

- *Topic portion values as labels*, where the topic portion values are directly used as the concept labels (Figure 21-a). In this case, the concept labels are continuous, so the concept classifier is trained as a regressor. In other words, the w_c is trained to predict the exact value of the topic portions from the visual instances.
- *Binarized topic portion values as labels*, where the topic portion values are first quantized into binary labels, based on a threshold parameter. More specifically, the binary label is equal to 1, if the topic portion is higher than the threshold, otherwise the binary label is set to -1 (Figure 21-b). Then, the concept classifier w_c can be trained as a binary classifier using the binarized labels.
- *Topic portion ranks as labels*, where the visual instances are ordered based on their topic portion values and their rank is considered as the label, (Figure 21-c). In this case, the concept classifier is trained as a ranking function. In other words, the w_c is trained to rank the visual instances by predicting their relative concept relevancies.

Neither continuous values nor binarized labels are appropriate for training the concept classifiers from the estimated topic portions. Predicting the continuous label values from the images is difficult and ineffective, as we will show in the experiments. Although predicting binary labels is in general more simple than predicting continuous labels, a substantial amount of information in the topic portions can be lost by the binarization. Hence, we speculate that defining the topic portion ranks as labels and learning the concept classifiers as a learning to rank problem is the most effective solution for training the concept classifiers from the automatically discovered topic portions.

More formally, we first define each concept label $y_i \in \mathcal{D}_c$ by measuring its topic portion rank as follows:

$$y_i = \sum_{j=1}^N \mathbb{1}(\theta_i^c < \theta_j^c), \quad (8)$$

where $\mathbb{1}(\cdot)$ is the indicator function. Moreover, θ_i^c is the estimated topic portion value for instance i for topic c . After determining the labels, the concept classifier w_c can be learned by any learning to rank method. In this paper, we follow the rankSVM formulation [71], which learns the ranking function by minimizing the following objective function:

$$\frac{\lambda}{2} \|w_c\|^2 + \frac{1}{N} \sum_{i=1}^N \sum_{j=1, y_i < y_j}^N \max(0, 1 - w_c^\top(x_j - x_i)), \quad (9)$$

A man attempts to do an ollie but twists his ankle when he lands.

Woman makes peanut butter sandwich with chips in the kitchen.

Figure 22.: Two examples of the videos and their captions from the Captioned Videos dataset [119].

where λ is the regularizer penalty parameter. By minimizing this objective we learn to predict a higher value for the j^{th} visual instance than the i^{th} visual instance, if its label has a higher rank. We minimize the Eq. (9) for each concept c independently using stochastic gradient descent [3], note that the ranking of the documents will be different for each topic.

Training Textual Embedding. While we could have used the LDA model directly as the textual embedding, we propose to yield a textual embedding which is more aligned with the visual embedding. The textual embedding are trained in a similar fashion as the visual embeddings, however instead of using the visual features the concept classifiers are now trained on the textual features.

For each vocabulary concept c , the train set $\mathcal{D}_c = \{(\mathbf{x}_i, y_i), i = 1 \dots N\}$ is defined, where $\mathbf{x}_i \in \mathbb{R}^{D_t}$ denotes the textual instances in the train set. For example, in a train set of Wikipedia articles, $\mathbf{x}_i$ refer to the textual features extracted from the article’s text. Moreover, y_i are the concept labels, which are defined as the topic portion ranks as determined by Eq. (8). Note, that for each concept c the labels y_i are identical for both the visual embedding and the textual embedding, yielding a desired alignment between the two embeddings. Based on the train set $\mathcal{D}_c$, the textual classifier of the concept $\mathbf{w}_c \in \mathbb{R}^{D_t}$ is trained by minimizing the objective function of Eq. (9).

4.3 EXPERIMENTAL SETUP

4.3.1 Datasets

We use three multimedia datasets in our experiments, as summarized in Table 12. The first two datasets are made of image/text pairs, while the third set includes video/text pairs.

1. Wikipedia [127]. This dataset is provided by Rasiswasia *et al.* [127] and includes 2,866 Wikipedia articles from 10 categories. Each article is made of few paragraphs in text explaining the article, as well as an illustrative image. We adopt the protocol in [34] and use 2,173 articles as train set, 231 articles as validation set, and the remaining 462 articles as test set.

2. Wikipedia++. We expand the Wikipedia dataset [127] by collecting more articles from more categories. Our dataset, which we name as Wikipedia++, is collected by exactly following the same procedure as used for creating the Wikipedia dataset [127]. More specifically, we collected the articles from the Wikipedia “featured articles”. Each featured article is categorized by Wikipedia into one of 30 categories. Excluding scarce categories, with less than 50 articles,

Table 12.: Statistics of the Wikipedia++ and the Captioned Videos datasets used in our experiments. The statistics of the Wikipedia dataset are available in [127]. Our created Wikipedia++ dataset is publicly available at <http://www.mediamill.nl>.

Wikipedia++		Captioned Videos [119]	
Category	Size	Category	Size
Animal	1,454	Attempting board trick	160
Art & architecture & archaeology	695	Feeding animal	161
Biology	477	Landing fish	119
Business & economics & finance	234	Wedding ceremony	123
Culture & society	326	Working wood working project	141
Education	271	Birthday party	343
Engineering & technology	231	Changing vehicle tire	221
Geography & places	1,758	Flash mob gathering	303
History	793	Getting vehicle unstuck	211
Literature & theatre	453	Grooming animal	218
Media	566	Making sandwich	255
Meteorology	345	Parade	322
Music	400	Parkour	213
Physics & astronomy	573	Repairing appliance	196
Religion & mysticism & mythology	212	Working sewing project	196
Royalty & nobility & heraldry	583	Attempting bike trick	130
Sport & recreation	700	Cleaning appliance	130
Transport	638	Dog show	130
Video gaming	401	Giving directions location	130
Warfare	1,507	Marriage proposal	129
		Renovating home	130
		Rock climbing	130
		Town hall meeting	130
		Winning race without vehicle	130
		Working metal crafts project	130
Total	12,617		4,481

we end up with the 20 article categories listed in Table 12. We split each article into sections, based on its section headings, and assign each image in the article to the section in which it was placed. This leads to a set of short and focused articles, containing a single image. Then the dataset is pruned by excluding the sections without any image. The final corpus contains 12,617 articles. We randomly split the articles into three partitions: 50% of the articles as train set, 25% as validation set, and 25% as test set. We make our collected Wikipedia++, which is four times larger than the Wikipedia dataset [127], publicly available.

3. Captioned Videos [119]. This dataset is part of the TRECVID Multimedia Event Detection corpus [119]. It consists of user-generated videos accompanied with human provided textual captions. The dataset contains videos depicting 25 complex events, including life events, instructional events, and sport events. The textual caption of each video summarizes what happens in the content, as shown in Figure 22. We use the videos, which are assigned to one of the 25 event categories. It leads to 4,481 videos and their captions, where 50% of them are used as train set, 25% as validation set and the remaining 25% as test set.

4.3.2 Evaluation Protocol

We perform our cross-media retrieval experiments by exactly following [34, 127]. The multimedia instances in the train set and validation set are used for learning the visual and textual embeddings. The test set instances are used to perform two cross-media retrieval experiments: *i) Visual Query* experiments, where each image or video in the test set is considered as a query, for which we rank the textual instances in the test set by measuring their similarities. *ii) Textual Query* experiments, where each textual instance in the test set is considered as a query, for which we rank the images or videos in the test set by measuring their similarities. The similarities are measured by first projecting the instances into the semantic space, by the trained visual and textual embeddings, then we use the normalized correlation as the similarity metric [127].

The retrieval performance is evaluated using the mean average precision (mAP) metric, as in [34, 127]. More specifically, for each query and its R top retrieved instances, the AP is measured as:

$$\frac{1}{M} \sum_{r=1}^R p(r) \cdot rel(r),$$

where M is the total number of relevant instances in the retrieved set, $p(r)$ is the precision at r , and $rel(r)$ is a binary value determining the relevance of the r^{th} ranked instance. If the retrieved instance and the query has the same category label the $rel(.)$ is one, otherwise it is zero. Finally, mAP is obtained by averaging the measured AP values over all the queries. Similar to [34] we report the mAP@50 ($R = 50$) in all the experiments.

4.3.3 Implementation Details

Features. We use the deep learned visual descriptors [117] as *image features*. Each image is fed into a pre-trained convolutional neural network and the output of the second fully connected layer is considered as a 4,096 dimensional feature vector. The convolutional neural network has an AlexNet architecture [76]. It is pre-trained on all the 15,293 categories in ImageNet dataset, for which there are at least 50 positive examples available. The *video features* are obtained by first extracting the video frames by uniformly sampling the frames every two seconds. Then each video frame is represented by deep learning features, which are extracted in the same way as the image features. Afterward, each video is represented by the average pooling of its frames as a 4,096 dimensional feature vector. As the *text features* we use the term histograms. More specifically, a dictionary of 3,000 high-frequency terms are extracted per dataset. Then each textual instance is represented as the histogram of its terms with respect to the dictionary [34].

Learning parameters. We learn all the vocabulary concept classifiers by the stochastic gradient descent solver [3]. The learning rate parameter η , the regularization penalty parameter λ , and the number of epochs are empirically set to 0.01, 0.001, and 100, respectively.

Topic model parameters. We extract the topics by MALLET implementation of LDA [103]. The Dirichlet prior parameters α and β are empirically set to 1 and 0.1. Moreover, the optimal number of topics for each dataset is obtained by cross-validation over 5, 10, 20, 50, 100, 500, and 1000 topics.

4.3.4 Experiments

1. Discovering Semantic Vocabulary. We evaluate the effectiveness of our automatically discovered concept labels, as proposed in Section 4.2.1, for training semantic vocabularies. We

Table 13.: Experiment 1. Evaluating the discovered vocabulary. The discovered vocabulary is more effective and does not require manual supervision during training of concept classifiers.

Dataset	Visual Query		Textual Query	
	Pre-defined [24]	Discovered	Pre-defined [24]	Discovered
Wikipedia	0.431	0.450	0.491	0.516
Wikipedia++	0.377	0.418	0.493	0.511
Captioned Videos	0.528	0.642	0.627	0.732

compare two concept vocabularies: First, our proposed *discovered vocabulary*, which is trained on the automatically extracted concept labels, as detailed in Section 4.2. Second, the *pre-defined vocabulary* baseline [24], which is trained on manually provided concept labels. In this baseline, each category in the train set is considered as a vocabulary concept, and the concept labels come from the category labels, which are manually provided per instance. Please note that our proposed discovered vocabulary does not need *any* manual category labels in the train set, so we ignore all of them when training the discovered vocabulary. Each vocabulary is evaluated by its cross-media retrieval accuracy.

2. Training Semantic Embeddings. We evaluate our proposed strategy for training the semantic embeddings from the discovered vocabulary labels, as detailed in Section 4.2.2, by comparing three concept vocabularies: First, our proposed vocabulary of *concept rankers*, where the concept classifiers are trained as rankers on the topic portion ranks as labels. The rankers are trained based on the rankSVM formulation [71]. Second, a baseline vocabulary of *concept binary classifiers*, where the concept classifiers are trained as binary classifiers on the binarized topic portion as labels. The binary classifiers are trained based on the binary SVM formulation. Third, a baseline vocabulary of *concept regressors*, where the concept classifiers are trained as regressors on the topic portion values as labels. The regressors are trained based on the ridge regression formulation. To be consistent, we use the stochastic gradient descent solver for training all the rankers, binary classifiers, and regressors with the same parameter settings as detailed in Section 4.3.3.

3. Comparison to Others Alignments. We investigate the effectiveness of our discovered vocabulary for cross-media retrieval by comparing it with three state-of-the-art baselines: *i) CCA* [127], as a low-level alignment, which uses CCA to align the two modalities. *ii) Correspondence AE* [34], as a recent mid-level alignment, which uses correspondence autoencoders with deep architectures to align the textual and visual features. *iii) pre-defined vocabulary* [24], as a semantic alignment, which uses a pre-defined vocabulary of concept classifiers to embed the textual and visual instances in the mutual semantic space. For the CCA and pre-defined vocabulary baselines, we run the author’s implementation on our features. For the correspondence AE baseline the numbers are exactly reported from [34].

4.4 RESULTS

4.4.1 Discovering Semantic Vocabulary

The results are shown in Table 13. The proposed discovered vocabulary outperforms the pre-defined vocabulary consistently for both cross-media retrieval tasks on all the three datasets. It

Figure 23.: Eight examples of the topics extracted from the Wikipedia++ dataset. Each topic is represented by its five most relevant terms in the same color. The topics in (a) and (b) are more specific than the predefined concepts “sport & recreating” and “warfare”, respectively.

demonstrates that the vocabulary concept labels can be effectively extracted from the textual instances in the train set without any manual supervision.

We provide two reasons to explain the better performance of the discovered vocabularies over the predefined vocabularies: First, many of the manually defined vocabulary concepts are very general, *i.e.* the concepts “warfare” and “sport & recreation” in the *Wikipedia++*. These general concepts have a large diversity in their training examples, which undermines the accuracy of their concept classifiers. In contrast, the discovered vocabulary defines the vocabulary concepts by topic modeling, so is able to define more specific concepts by extracting an appropriate number of topics. Figure 23 shows some examples of the specific concepts, which are discovered by topic modeling.

As the second reason, the pre-defined vocabularies include a limited number of concepts, which are not enough to effectively represent all of the instances. More specifically, for the *Wikipedia*, *Wikipedia++*, and *Captioned Videos* datasets, the pre-defined vocabulary includes 10, 20, and 25 concepts as pre-defined in the dataset. In contrast, our discovered vocabulary is able to discover higher number of vocabulary concepts by extracting more topics from the textual instances. We further investigate the impact of increasing the number of vocabulary concepts by extracting more topics in Figure 24. It demonstrates that extracting more topics generally leads to more comprehensive and descriptive concept vocabulary, which is more effective for cross-media retrieval. However, after a certain vocabulary size, extracting more topics has a negative effect on the vocabulary. We explain it by the fact that when we extract too many topics from a small collection of texts, the extracted topics might be overfitted to the random co-occurrences of terms. It is also observable in Figure 24, where there is a relation between the dataset size and the

Figure 24.: Effect of increasing the number of vocabulary concepts by extracting more topics from text. Extracting more topics generally improves the vocabulary by making it more descriptive. However, extracting too many topics is prone to overfitting, which degrades the effectiveness of the trained vocabulary.

Table 14.: Experiment 2. Evaluation of the three strategies for training vocabulary classifiers from the discovered concept labels. The most effective vocabulary is obtained by training ranking functions on topic portion ranks as labels.

Dataset	Visual Query			Textual Query		
	Concept Binary Classifiers	Concept Regressors	Concept Rankers	Concept Binary Classifiers	Concept Regressors	Concept Rankers
Wikipedia	0.416	0.412	0.450	0.497	0.487	0.516
Wikipedia++	0.342	0.401	0.418	0.406	0.495	0.511
Captioned Videos	0.453	0.594	0.642	0.522	0.706	0.732

optimal number of topics. The optimal number of topics for *Wikipedia++*, as the largest dataset in our experiments, is 100 that is higher than for the other two datasets.

As the conclusion, the results demonstrate that by discovering the concept labels from the textual instances, we not only alleviate the manual labeling but also train a more effective vocabulary of concepts for cross-media retrieval.

4.4.2 Training Semantic Embeddings

The results are shown in Table 14. The vocabulary of concept rankers consistently outperforms the vocabularies of concept binary classifiers and concepts regressors for the both cross-media retrieval tasks on all the three datasets. It demonstrates that defining the concept labels as the topic portion ranks and training ranking functions on them is the most effective strategy for learning from the automatically discovered concept labels.

The lowest performing vocabulary is obtained with the concept binary classifiers, which are trained as binary SVMs on the binarized topic portions as labels. We explain it by the fact that lots of information in the topic portions are lost by binarization. More specifically, all the instances above the binarization threshold are equally considered as positive examples without consideration of their relevancy degrees encoded in topic portions. The low performance of the concept binary classifiers implies the importance of the relevancy degrees for training the vocabulary.

Figure 25.: Top three retrievals for a caption query (left) and a video query (right) in the Captioned Videos dataset. The semantic representations are shown in the gray box. The common concepts between the query and the retrieved instances, as highlighted by colors, explain why the query and the retrieved instances are similar. Each highlighted concept is represented by its three most relevant terms in the bottom.

Table 15.: Experiment 3. Comparison of our discovered vocabulary with three state-of-the-art alignments. The discovered vocabulary consistently outperforms the alternatives.

Dataset	Visual Query				Textual Query			
	CCA [127]	Correspondence AE [34]	Pre-defined [24]	Discovered	CCA [127]	Correspondence AE [34]	Pre-defined [24]	Discovered
Wikipedia	0.348	0.335	0.431	0.450	0.359	0.368	0.491	0.516
Wikipedia++	0.352	N.A.	0.377	0.418	0.412	N.A.	0.493	0.511
Captioned Videos	0.475	N.A.	0.528	0.642	0.545	N.A.	0.627	0.732

The concept regressors vocabulary performs better than the binary classifiers vocabulary, but it is still outperformed by the vocabulary of concept rankers. As an explanation, we speculate that the concept regressors does not binarize the topic portions, so are not suffered by the information lost as in the binary classifiers. However, predicting the exact value of the topic portions from the visual features, as targeted by the concept regressors, is generally very hard. Hence the trained concept regressors might be inaccurate. This drawback is relaxed by the concept rankers by predicting the relative orders between the topic portions instead of their exact values. As a conclusion, training the concept rankers is the most effective strategy, compared to the alternatives, for training concept classifiers from the automatically discovered labels.

4.4.3 Comparison to Other Alignments

The results are shown in Table 15. Our discovered vocabulary consistently outperforms the state-of-the-art alternatives for the both cross-media retrieval tasks on all the three datasets. It demonstrates the effectiveness of our proposed vocabulary for cross-media retrieval.

Furthermore, the results demonstrate that the both pre-defined and discovered vocabularies substantially outperform the CCA and correspondence autoencoders baselines. We explain the relatively low performance of the correspondence autoencoders by the fact that this baseline uses a deep architecture, which generally requires large amount of training data to be effectively learnt. However, the largest train set in our experiments includes 6K examples, which seems to be not large enough. Moreover, the low performance of CCA baseline, as a low-level alignment, validates that learning the correspondences between the two modalities directly from the low-level features is not effective, since the semantic gap is maximized.

Besides their higher retrieval accuracies, the semantic vocabularies provide interpretable representation of instances, which make it possible to explain why two visual and textual instances are similar, as shown in Figure 25. In summary, the discovered vocabulary is not only effective for cross-media retrieval, but also recounts why the instances are retrieved.

4.5 CONCLUSION

We propose a data-driven approach for cross-media retrieval by automatically learning its underlying semantic vocabulary, rather than specifying and annotating the vocabulary as commonly done. We demonstrate that the textual instances in cross-media collections are a rich source of semantics, which can be utilized to (weakly) supervise the concept classifier training. More specifically, we demonstrate that probabilistic topic models are effective tools to extract the underlying semantic structures from a collection of text. Moreover, we experimentally show that learning to rank is an effective strategy for learning the classifiers from the text-driven annotations. Our experiments show that the discovered vocabulary outperform the state-of-the-art alternatives for cross-media retrieval. These conclusions may generalize for any other problem, where the textual descriptions are served as labels for concept classifier training.

VIDEOSTORY EMBEDDINGS RECOGNIZE EVENTS WHEN EXAMPLES ARE SCARCE

5.1 INTRODUCTION

This paper strives to recognize events such as *parking a vehicle*, *flash mob*, and *renovating a home* in web video content. A problem of increasing importance in a world that is swiftly adapting to video communication. The leading computer vision and multimedia retrieval solutions for this challenging problem, *e.g.* [67, 110, 149, 163, 175], all learn to recognize events with the aid of many labeled video examples. However, as events become more and more specific, we consider it unrealistic to assume that ample examples to learn from will be commonly available. In practical use only a handful of video examples, an event name and an event definition may be present, such as the ones in Figure 26. We aim for event recognition when video examples are scarce or even completely absent.

Recognizing events from few or zero examples imposes constraints on the video representation. End-to-end deep learning of the video representation [60, 73, 152] demands too many video examples. In [73] for example, Karpathy *et al.* exploit more than 1 million YouTube videos and their sport category labels to learn the video representation. Alternatively, Xu *et al.* [175] show how a very deep convolutional neural network (CNN) [138] intended for image classification can be leveraged as representation for event recognition in video. They use responses from intermediate layers of the network to represent frames, which are aggregated over the video using VLAD encoding [58]. When combined with a linear SVM, excellent results on the leading NIST TRECVID event detection benchmarks [119] are reported for scenarios where many and few examples are available. The CNN video representation outperforms more traditional video encodings such as improved dense trajectories [163, 164] and multimedia representations combining appearance, motion and audio features [109, 114, 181]. However, both the learned and engineered representations are incapable, nor intended to, recognize events when examples are completely absent. We propose a video representation that can leverage any underlying feature, be it a CNN, improved dense trajectories and/or audio features, while being capable of few-example *and* zero-example event recognition with state-of-the-art accuracy.

The key to few- and zero-example event recognition is to add meaning to the video representation. Inspired by the success in image classification [32, 79, 121], many rely on the prediction scores made by a set of pre-trained attribute classifiers [12, 29, 45, 55, 65, 92, 94, 96, 99, 104, 176]. In [45] Habibiian *et al.* study the properties of 1,346 attribute classifiers trained from ImageNet [130] and TRECVID [119] for representing and recognizing events in web video. Rather than pre-specifying and manually labeling each individual attribute in advance, the attributes can also be learned on top of imagery and (weak) labels harvested from the web [22, 25, 171, 179]. To assure visual predictability of the discovered attribute detectors a common tactic is to leverage

Event Name: Winning a race without a vehicle

Definition: An individual (or more) succeeds in reaching a pre-determined destination before all other individuals, without vehicle assistance or assistance of a horse or other animal. Racing generally involves accomplishing a task in less time than other competitors. The only type of racing considered relevant for the purposes of this event is the type where the task is traveling to a destination, completed by a person(s) without assistance of a vehicle or animal. Different types of races involve different types of human ...

Figure 26.: Video exemplars (top) and the textual definition (bottom) of the event winning a race without a vehicle to illustrate one of the events studied in this paper. Following the NIST TRECVID evaluation guidelines [119], the textual definition is for zero-example recognition, and the ten provided video exemplars are for few-example recognition of the event.

part of the harvested data for validation [11]. However, a drawback is that many attribute labels rarely occur. For these infrequent labels only a limited number of positive examples are available, which leads to a biased estimation of their classification reliability. As a consequence, many of the discovered attributes might be overfitted to their small training set and do not generalize well for new videos. We also discover our semantic representation from the web, but rather than selecting individual, and often unreliable, classifiers per attribute label, we prefer to combine labels automatically into more predictable attributes. By combining labels, more training examples are available and a more robust representation is obtained, without losing descriptive ability.

In this paper we present a semantic representation learning algorithm for videos. Instead of relying on pre-specified attribute labels, we learn the representation from freely available web videos and their descriptions. To this end, we propose an embedding between the video features and their textual descriptions, which is learned by utilizing the correlations between the words in the descriptions. We learn the embedding by minimizing a joint objective function balancing the descriptiveness and predictability of the learned video representation. Our proposed embedding is able to leverage the multiple modalities which coexist in video to learn a more reliable semantic representation. We call our video representation *VideoStory*, since it strives to encode the story of a video from its description.

A preliminary version of this article appeared as [42]. The current version adds *i)* representation learning for multiple video modalities, *ii)* representation learning for zero-example event recognition, *iii)* an additional TRECVID video data set, revised experiments, and improved baselines, all using a recent video CNN feature [148, 175], and *iv)* a new related work section, which will be discussed next.

5.2 RELATED WORK

In this section we focus on three directions of related work we deem most closely connected to ours.

5.2.1 Representations for Event Recognition

Until recently most event recognition methods exploit video representations based on densely extracted low-level visual features, such as HOG/HOF [81], or MBH [116, 164], often combined with audio features like MFCC features [70, 110, 149]. Currently, most methods extract frame-based deep convolutional neural network (CNN) features [76, 138, 148], using the responses from intermediate layers of the CNN, which is pre-trained on ImageNet images [130], see *e.g.* [22, 65, 107, 175, 181]. To obtain per video descriptors the local/frame-based descriptors are aggregated by their mean, by using the Fisher vector [132] or VLAD encoding [58]. Despite the fact that these low-level based video representations obtain state-of-the-art event recognition performance, they suffer from two drawbacks. First, because of their high-dimensionality, training effective event classifiers on these representations require a sufficient number of training examples to prevent overfitting. Second, all these representations are incapable of providing a semantic interpretation of the video, which is crucial for zero-example recognition.

Semantic Representations To obtain a semantic representation for videos, inspiration is obtained from describing images with attributes [36, 79], and the video is represented by its attribute prediction scores. Creating the training data for a set of task-specific attribute classifiers manually involves lots of annotation effort, which is restrictive. Therefore, often public available datasets, such as ImageNet [130] and TRECVID [119], are used to train the attribute classifiers [45, 94, 99]. Although this overcomes the need for (additional) manual annotation, the attributes of these datasets are not necessarily descriptive for event recognition.

To tune the attributes for the task at hand, several works aim to automatically discover the attributes from web images/videos and their textual descriptions [22, 25, 171, 179]. For example Wu *et al.* [171], start by selecting the most frequent/relevant terms from a set of provided event descriptions as attributes. Then, inspired by Berg *et al.* [11], they use Internet search engines like Google and YouTube to gather positive examples. They assure visual predictability of the discovered attribute detectors by cross-validation. Similarly the work of Ye *et al.* [22] relies on the website WikiHow to obtain event descriptions, and the visual predictability of the selected terms is ensured by keeping only those terms which are present in existing image classification datasets. We refer to these methods as *term attributes* since they all discover the attributes from the terms in the descriptions.

Despite their effectiveness in obtaining training data, term attributes suffer from three drawbacks. First, descriptions have a long-tail distribution, therefore many terms rarely occur and these will not have enough positive examples to train reliable visual classifiers. Second, term attributes are selected mostly based on their visual prediction accuracy, while for effective event recognition the selected term attributes should also be descriptive for the target events. Third, contextual information is lost, since term attributes are learned independently by binary classifiers.

In this paper, we also learn the representation from videos and their descriptions. But, rather than selecting individual, and often unreliable, term attributes, we learn the entire representation by automatically combining the terms through embedding. In our proposed embedding, the correlation between the terms are utilized to learn a more effective representation, which is predictable *and* descriptive.

5.2.2 *Embedding Videos and their Descriptions*

To learn correspondences between the visual domain and textual descriptions different embedding methods have been proposed for various purposes *i.e.* image annotation [14], image classification [2, 169], image captioning [72, 161], video to text translation [28, 174] and cross-modal retrieval [24, 34, 112].

Canonical correlation analysis (CCA) [52] is the classical unsupervised manner to relate different modalities and can be used for cross-modal retrieval [24]. CCA is the multimodal generalization of PCA, and can be computed as a generalized eigenvalue problem on the cross-covariance matrix between the visual and textual features. It finds a sequence of uncorrelated projections in which the cross-correlation between the modalities is maximized. This is not necessary suited to obtain a discriminative representation, as is also validated by our experiments.

Probabilistic topic models, such as the seminal *correspondence latent Dirichlet allocation* [14] and its extensions [27, 126], learn correspondences by extracting a set of correlated visual and textual topics from imagery and their captions. Despite their effectiveness for captioning images and videos [28], these methods are not designed to learn a representation for the purpose of recognition. Moreover, we note that by design these models are only applicable on discrete features, and therefore they cannot leverage the state-of-the-art video features used in our paper.

Embeddings for multi-class image classification, such as [2, 169] find a low-dimensional subspace in which multi-class classification is performed. The subspace is found jointly with the multi-class classifier by minimizing a classification loss [169], or a zero-shot classification loss [2]. In contrast to this multi-class image classification setting, we are interested in finding descriptions modeled as a multi-label video classification problem. Moreover, we argue that solely optimizing for classification is not sufficient to obtain a predictable and descriptive video embedding.

Recently deep neural network architectures have been proposed to learn multimodal correspondences for retrieval [34, 112] and captioning [72, 161]. Notably the multimodal recurrent neural networks capture the sequential ordering between the words in image captions and generate more accurate image captions, compared to the probabilistic topic models. However, training deep neural networks generally requires large amounts of training examples, while our purpose is to learn from few- and zero-examples.

5.2.3 *Zero-Example Event Recognition*

In zero-example event recognition the goal is to recognize an event, only based on a given textual *event definition*, without using any video examples. The event definition is usually provided in the form of a title and description, see Figure 26. This zero-example setting is beyond the conventional zero-shot image recognition of objects and scenes [32, 79, 122], where often a training set of related classes is available as well as pre-specified class-to-attribute mappings. Recently, this challenging event recognition problem has attracted a lot of attention because of its high practical value and the corresponding TRECVID benchmark task initiated by NIST [119]. The common approach for zero-example event recognition is to represent videos and the event queries using a semantic representation, and to rank all video representations based on the cosine similarity with the event query.

To represent the video mostly attributes and term-attributes [22, 25, 63, 64, 171, 179] are exploited. Extensions include combining attributes using logical operators [41], using video-segments to detect attributes [101], and adjustment of attribute scores using an ontology structure [65]. In addition to attributes, automatic speech recognition and optical character recognition have been considered to enrich the semantic video representations [26, 65, 171].

Table 16.: Summary of the core notation used for VideoStory.

Notation	Description
N	Number of videos
M	Number of unique terms in descriptions
D	Dimensionality of low-level feature
J	Number of low-level features to fuse
k	Dimensionality of VideoStory embedding
$\mathbf{X} \in \mathbb{R}^{D \times N}$	Matrix of low-level video features
$\mathbf{Y} \in \{0, 1\}^{M \times N}$	Matrix of binary term vectors
$\mathbf{W} \in \mathbb{R}^{D \times k}$	VideoStory visual projection
$\mathbf{A} \in \mathbb{R}^{M \times k}$	VideoStory textual projection
$\mathbf{S} \in \mathbb{R}^{k \times N}$	VideoStory embedding
$\mathbf{H} \in \mathbb{R}^{M \times M}$	Diagonal matrix with per-term weights
$\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i$	The column representing the i -th video

To answer the event query, usually term matching between the semantic video representation and the event definition is performed. The retrieval can be enriched by using contextual information, such as word embeddings and term co-occurrences [20, 179]. For the final ranking, the cosine similarity can be extended by pseudo relevance feedback mechanisms, such as self-paced ranking [63]. This has been used to improve the zero-shot event recognition by re-ranking [20, 64, 65]. Our representation learning is orthogonal to these efforts and can be joined with them to further improve the event recognition performance.

5.3 VIDEOSTORY EMBEDDINGS

Our goal is to learn a representation function $f : \mathcal{X} \rightarrow \mathcal{S}$, which maps each low-level video representation $\mathbf{x}_i \in \mathcal{X}$ into the semantic representation $\mathbf{s}_i \in \mathcal{S}$. The representation function is trained on a collection of videos and their semantic descriptions, which we use in the form of binary term vectors $\mathbf{y}_i \in \mathcal{Y}$.

In the common attribute learning approach for learning the representation, the semantic representations are defined to be in the same space as the term vectors ($\mathcal{S} = \mathcal{Y}$). However, in practice, the term vectors are intrinsically noisy and highly sparse, which undermines their effectiveness as labels to train the representation function. Therefore, we propose to learn the semantic representation on a lower dimensional projection of the term vectors, which are less sparse and less noisy.

We formulate the representation function as an embedding, coined as VideoStory, which is learned by balancing two compelling forces in a joint optimization:

1. *Descriptiveness*, to preserve the information encoded in the video descriptions as much as possible, and
2. *Predictability*, to ensure that the representation could be effectively recognized from video content.

We first introduce the VideoStory embedding based on visual features (Section 5.3.1), then we generalize our embedding to fuse any multimedia feature (Section 5.3.2), and finally we gear the embedding towards zero-example recognition (Section 5.3.3). We summarize our notation conventions in Table 16.

5.3.1 Objective Function

We start from a dataset of videos, represented by video features $\mathbf{X}$, and their textual descriptions, represented by binary term vectors $\mathbf{Y}$, indicating which terms are present in each video description. Then, our VideoStory representation is learned by minimizing:

$$L_{VS}(\mathbf{A}, \mathbf{W}) = \min_{\mathbf{S}} L_d(\mathbf{A}, \mathbf{S}) + L_p(\mathbf{S}, \mathbf{W}), \quad (10)$$

where $\mathbf{A}$ is the textual projection matrix, $\mathbf{W}$ is the visual projection matrix, and $\mathbf{S}$ is the VideoStory embedding. The loss function L_d corresponds to our first objective for learning a descriptive VideoStory, and the loss function L_p corresponds to our second objective for learning a predictable VideoStory. The VideoStory embedding $\mathbf{S}$ interconnects the two loss functions. To the best of our knowledge this joint embedding is novel.

Descriptiveness. For the L_d function, we use a variant of regularized Latent Semantic Indexing (LSI) [166]. This objective minimizes the quadratic error between the original video descriptions $\mathbf{Y}$, and the reconstructed translations obtained from $\mathbf{A}$ and $\mathbf{S}$:

$$L_d(\mathbf{A}, \mathbf{S}) = \frac{1}{2} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{A} \mathbf{s}_i\|_2^2 + \lambda_a \Omega(\mathbf{A}) + \lambda_s \Psi(\mathbf{S}), \quad (11)$$

where $\Psi(\cdot)$ and $\Omega(\cdot)$ denote regularization functions, and $\lambda_a \geq 0$ and $\lambda_s \geq 0$ are regularizer coefficients. We use the squared Frobenius norm for regularization, which is the matrix variant of the ℓ_2 regularizer, *i.e.* $\Omega(\mathbf{A}) = \frac{1}{2} \|\mathbf{A}\|_F^2 = \frac{1}{2} \sum_i \|\mathbf{a}_i\|_2^2 = \frac{1}{2} \sum_{ij} a_{ij}^2$, the sum of the squared matrix elements. Similarly for the VideoStory matrix $\Psi(\mathbf{S}) = \frac{1}{2} \|\mathbf{S}\|_F^2$.

The main difference with regularized Latent Semantic Indexing [166] is their ℓ_1 regularizer, $\Omega(\mathbf{A}) = \sum_i \|\mathbf{a}_i\|_1$, which enforces sparsity in the textual projection $\mathbf{A}$. However, with our larger representation (typically we use a dimensionality of k between 512 and 2,048 in our experiments compared to only $k = 20$ in [166]) and fewer number of unique terms (around 10K, compared to 100K), enforcing sparsity is not necessary for good performance.

Note that other textual embeddings can be formulated similar to Eq. (11), when appropriate regularization functions $\Omega(\cdot)$ and $\Psi(\cdot)$ are used. For example using $\Omega(\cdot) = \|\cdot\|_1$ enforces sparsity [51, 97, 166], or in the extreme case that $\mathbf{A}$ is constrained such that each column has a single non-zero value, the objective becomes very close to methods that select the best single term labels [11], and when $\mathbf{A}$ is enforced to preserve the taxonomical term relations, the objective resembles taxonomy embedding [155, 168].

Predictability. The L_p function measures the occurred loss between the VideoStory $\mathbf{S}$ and the embedding of video features using $\mathbf{W}$. Since the VideoStory $\mathbf{S}$ is real valued, as opposed to a binary or multi-class encoding, we can not rely on standard classification losses such as the hinge-loss used in SVMs. Therefore, we define L_p as a regularized regression, similar to ridge regression:

$$L_p(\mathbf{S}, \mathbf{W}) = \frac{1}{2} \sum_{i=1}^N \|\mathbf{s}_i - \mathbf{W}^\top \mathbf{x}_i\|_2^2 + \lambda_w \Theta(\mathbf{W}), \quad (12)$$

where we use (again) the Frobenius norm for regularization of the visual projection matrix $\mathbf{W}$, $\Theta(\mathbf{W}) = \frac{1}{2} \|\mathbf{W}\|_F^2$, and λ_w is the regularization coefficient.

Joint optimization. To handle large scale datasets and state-of-the-art high-dimensional visual features, *e.g.* Fisher vectors [132] on video features [164] or deep learned representations [76], we employ a Stochastic Gradient Descent (SGD) [18] optimization, summarized in Algorithm 2. The number of passes over the datasets (*epochs*) and the step-size η are hyper-parameters of SGD.

```

input :  $X, Y, k, \eta$  (step-size),  $m$  (max-epochs)
output :  $W$  and  $A$ 

 $A$ , and  $S \leftarrow$  SVD decomposition of  $Y$ 
 $W \leftarrow$  random (zero-mean)

for  $e \leftarrow 1$  to  $m$  do
  for  $i \leftarrow 1$  to  $N$  do
    Pick a random video-description pair  $(x_t, y_t)$ 
    Compute gradients w.r.t.  $A, W$  and  $s_t$ 
    Update parameters:
       $A \leftarrow A - \eta_t \nabla_A L_{VS}$  see Eq. (13)
       $W \leftarrow W - \eta_t \nabla_W L_{VS}$  see Eq. (14)
       $S \leftarrow s_t - \eta_t \nabla_{s_t} L_{VS}$  see Eq. (15)
  end
end
return:  $W$  and  $A$ 

```

Algorithm 2: Pseudocode for learning the VideoStory

The VideoStory objective function, as given in Eq. (10), is convex with respect to matrix A and W when the embedding S is fixed. In that case, the joint optimization is decoupled into Eq. (11) and Eq. (12), which are both reduced to a standard ridge regression for a fixed S . Moreover, when both A and W are fixed, the objective in Eq. (10) is convex *w.r.t.* S . Therefore we use standard SGD by computing the gradients of a sample *w.r.t.* the current value of the parameters, and we minimize S jointly with A and W .

Lets denote a randomly sampled video and description pair at step t by (x_t, y_t) , and let s_t denote the current VideoStory embedding of sample t . The gradients of Eq. (10) for this sample *w.r.t.* A, W and s_t are given by:

$$\nabla_A L_{VS} = -(y_t - A s_t) s_t^\top + \lambda_a A, \quad (13)$$

$$\nabla_W L_{VS} = -x_t (s_t - W^\top x_t)^\top + \lambda_w W, \text{ and} \quad (14)$$

$$\nabla_{s_t} L_{VS} = -A^\top (y_t - A s_t) + (s_t - W^\top x_t) + \lambda_s s_t. \quad (15)$$

The effect of joint learning the descriptiveness and the predictability, becomes clear in Eq. (15), where both the textual projection matrix A and visual projection matrix W contribute to learning the VideoStory embedding S . This embedding S is subsequently used to obtain the textual projection A matrix, in Eq. (13), and the visual projection W matrix, in Eq. (14). This leads to the VideoStory embedding, which is both descriptive, by preserving the textual information, and predictable, by minimizing the visual prediction loss.

The parameters A, S , and W can be initialized by random numbers with zero-mean. However, we experimentally observed that initializing the A and S matrices by singular value decomposition (SVD) of the term vectors Y , speeds up the convergence.

After training the visual and textual projection matrices, they are used to predict the VideoStory representation and the term vector of each video. In the case that both a video x_i and description y_i are given, we could obtain the VideoStory representation by returning s_i from Eq. (10), while keeping both A and W fixed. However, in practice most videos are not provided with a description. Therefore, we use

$$s_i = W^\top x_i, \quad (16)$$

Figure 27.: VideoStory prediction: From the low-level video features the VideoStory representation and the term vector are predicted using the visual projection matrix W and the textual projection matrix A .

to predict our VideoStory representation from the low-level video features x_i . Moreover, using the predicted representation s_i , the term vectors for each unseen video are predicated as follows:

$$\hat{y}_i = A s_i, \quad (17)$$

where the terms with the highest values are most relevant for this video, see the illustration in Figure 27

5.3.2 VideoStory Fusion

Videos are inherently multimodal. In general any video contains appearance, motion, and audio cues and sometimes even textual information in the form of subtitles or speech recognition scripts. Fusing the different modalities is typically achieved by early-fusion, *i.e.* fusion at the level of the representations, and late-fusion, *i.e.* fusion at the level of prediction scores [143]. Both fusion strategies have been shown to be effective for understanding complex events as well, *e.g.* [70, 110, 151, 171]. We propose $\text{VideoStory}^{\mathcal{F}}$, which extends the VideoStory embedding by learning the semantic representation from multiple modalities.

A straightforward approach to learn the multimodal semantic representation is by fusing multiple VideoStory embeddings, which are independently trained per modality. More specifically, given the video descriptions and low-level features from one modality, a VideoStory embedding is trained as detailed in Algorithm 2. Then the learned embeddings from all modalities are fused to obtain the multimodal representation. Despite its simplicity, training the VideoStory embeddings

per modality is not optimal as it ignores the interrelations between the various modalities for learning the embeddings.

We aim for learning the embeddings jointly over all the video modalities. Our intuition is that the semantic representation is more effective if it is predictable from *all* the modalities rather than from each individual modality. For *VideoStory^F* we adjust the predictability loss, of Eq. (12), to incorporate a weighted combination of the predictability from all J modalities as follows:

$$L_p^{\mathcal{F}}(\mathbf{S}, \mathcal{W}) = \sum_{j=1}^J L_p(\mathbf{S}, \mathbf{W}^j) \quad (18)$$

where $\mathbf{S}$ is the multimodal VideoStory embedding, and $\mathcal{W} = \{\mathbf{W}^j, j = 1 \dots J\}$ is a set of projection matrices from the J modalities. Each feature projection matrix $\mathbf{W}^j \in \mathbb{R}^{D_j \times k}$ projects the low-level feature $\mathbf{x}_i^j \in \mathbb{R}^{1 \times D_j}$ extracted from the video into the VideoStory representation $\mathbf{s}_i$. Note that, instead of weighing each modality equally, a term $\gamma_j \geq 0$ could be used to weight the importance of each modality, if sufficient training examples are available for cross-validation of these weighting parameters.

The objective function Eq. (10) is still convex with respect to the parameters $\mathbf{S}$, $\mathbf{A}$, and $\mathbf{W}^j$ when the other parameters are fixed. However, the gradient with respect to $\mathbf{s}_t$, Eq. (15) becomes:

$$\nabla_{\mathbf{s}_t} L_{VS} = -\mathbf{A}^\top (\mathbf{y}_t - \mathbf{A} \mathbf{s}_t) + \left(\mathbf{s}_t - \sum_j \mathbf{W}^{j\top} \mathbf{x}_t^j \right) + \lambda_s \mathbf{s}_t. \quad (19)$$

It can be seen that all the modalities are jointly contributing to learn the multimodal VideoStory representations $\mathbf{S}$.

After training the textual and feature projection matrices, they are used to extract the multimodal VideoStory representation. Each feature projection matrix $\mathbf{W}^j$ predicts the VideoStory representation based on its underlying modality as follows:

$$\mathbf{s}_i^j = \mathbf{W}^{j\top} \mathbf{x}_i^j. \quad (20)$$

The final VideoStory representation is the concatenation of the per modality embeddings: $\mathbf{s}_i = [\mathbf{s}_i^1, \dots, \mathbf{s}_i^J]$.

5.3.3 VideoStory for Zero-Example Event Recognition

The objective function for learning the VideoStory embedding for few-example recognition is not necessarily optimal for the zero-example setting. In Section 5.3.1, the descriptiveness loss L_d is defined as the overall error in reconstructing *all* the terms from the VideoStory representations, see Eq. (11). With this definition, the descriptiveness loss is biased toward the more frequent terms, as minimizing their reconstruction error leads to a higher decrease in the overall error. Consequently, the terms which are infrequent in the descriptions might be discarded, which degrades their prediction accuracy from video features. This undermines the effectiveness of our representation learning for zero-example setting, where the accuracy in predicting the terms from video is crucial.

We extend the VideoStory embedding to learn a video representation, which is effective for zero-example event recognition. Our extension, which we coin as *VideoStory₀*, minimizes the reconstruction error of the terms with respect to their importance for describing the events, rather than their frequency in the VideoStory train data. For this purpose, we use a *term sensitive* descriptiveness loss, as follows:

$$L_d^{\text{ts}}(\mathbf{A}, \mathbf{S}) = \frac{1}{2} \sum_{i=1}^N \|\mathbf{H}^{\frac{1}{2}} (\mathbf{y}_i - \mathbf{A} \mathbf{s}_i)\|_2^2 + \lambda_a \Omega(\mathbf{A}) + \lambda_s \Psi(\mathbf{S}), \quad (21)$$

where $\mathbf{H} \in \mathbb{R}^{M \times M}$ is a diagonal matrix, denoting the importance of each term for describing the events. By setting a relatively high value for h_{jj} for term j , its reconstruction error is more penalized compared to the other terms. Hence the term is expected to be more precisely reconstructed.

We determine the term importance matrix $\mathbf{H}$ by relying on the presence/absence of terms in the textual event definitions. Our assumption is that the terms, which are present in event definitions are more important than the absent terms. More precisely, we set each element of the importance matrix h_{jj} to α , if the term j is present, and $1 - \alpha$ if the term j is absent in the event definitions. α is a balancing parameter between 0 and 1, which should be higher than 0.5 to assign more importance to the present terms. We empirically set this parameter to 0.75 in all our zero-example event recognition experiments. The importance matrix can be extracted either for all events jointly or separately per event. We opt for the latter in our experiments, *i.e.* extracting $\mathbf{H}^e$ based on the description for event e , since in the literature the events are usually treated separately.

After training the visual and textual projection matrices, we follow the standard pipeline for zero-example event recognition: First, each test video is represented by predicting its term vector $\hat{\mathbf{y}}_i^e$ based on Eq. (17), using the representation learned using $\mathbf{H}^e$. Second, we translate the textual event definition into the event query, denoted as $\mathbf{y}^e \in \mathbb{R}^M$, by matching the terms in the event definition with the M unique terms in the VideoStory training data. Finally, the ranking is obtained by measuring the similarity between the video representations and the event query based on the cosine similarity:

$$s_e(\mathbf{x}_i) = \frac{\mathbf{y}^{e\top} \hat{\mathbf{y}}_i^e}{\|\mathbf{y}^e\| \|\hat{\mathbf{y}}_i^e\|}. \quad (22)$$

Multimodal fusion. Finally, we can leverage the multimodal features for zero-example event recognition by combining the multimodal predictability loss and the term sensitive descriptiveness in a joint objective:

$$L_{\text{VS}}(\mathbf{A}, \mathbf{W}) = \min_{\mathbf{S}} L_d^{ts}(\mathbf{A}, \mathbf{S}) + L_p^{\mathcal{F}}(\mathbf{S}, \mathbf{W}). \quad (23)$$

We coin the learned video representation $\text{VideoStory}_0^{\mathcal{F}}$, since it is learned on multiple video modalities for zero-example event recognition.

5.4 EXPERIMENTAL SETUP

5.4.1 Datasets

We first introduce the dataset used for learning the VideoStory embeddings. Then, we detail the video datasets by which we evaluate the event recognition experiments.

VideoStory Learning

In all the experiments, we learn the VideoStory embeddings on the **VideoStory46K** [42] dataset. This collection encompass 45,826 videos harvested from YouTube, with a total length of 743 hours. Every video comes with a short title caption provided by the user who has uploaded the video, as shown in Figure 28. There are 19,159 unique terms in the captions, most of them occurring infrequently. We filter out the terms occurring only once as they generally are misspelled terms, numbers, or noisy terms. It provides us with 9,828 unique terms, which are used in our experiments.

Figure 28.: Example videos and title captions from the VideoStory46K dataset [42], which we use for VideoStory representation learning.

Event Recognition Evaluation

We perform our event recognition experiments on the challenging TRECVID Multimedia Event Detection corpus [146] and the Columbia Consumer Video collection [69]. These corpora contain more than 42K videos in total, including user generated web videos with a large variation in quality, length and content.

TRECVID Multimedia Event Detection (MED) [146]. This dataset is introduced by NIST as a benchmark for event recognition. We perform our experiments on the two latest releases of the dataset, referred to as *MED 2013* and *MED 2014*. Each dataset includes videos from 20 complex events categories, but with 10 events overlapping as listed in Table 20. Each dataset includes three labeled video partitions: Event Kit training, Background training, and test set MED including 200, 5K, and 27K videos, respectively. Apart from the videos, a textual definition is provided per event, which explicates the event as unformatted plain text, such as the one shown in Figure 26.

We perform our few-example and zero-example event recognition experiments by exactly following the *10Ex* and *0Ex* evaluation procedure outlined by the NIST TRECVID event recognition task [119]. It means that, in the few-example event recognition experiments, training data for each event is composed of 10 positive videos from the Event Kit training data along with about 5K negative videos from the Background training data. The event recognition results for each event classifier are reported on the 27K videos from test set MED. In our zero-example event recognition experiments, we rely on the provided textual event definitions to create an event query vector. Then the event recognition performances are reported on the test set MED.

Columbia Consumer Video (CCV) [69]. This dataset contains 9,317 user-generated videos from YouTube including over 210 hours of videos in total. The dataset contains ground truth annotations at video-level for 20 semantic categories, *i.e.* *wedding reception* and *music performance*. We use the standard partitioning of the dataset, but we use only 10 positive examples per event in the training data. These 10 are selected based on alphabetical order of the respective video names, we ignore the remaining positive examples in the train set. We report event recognition results on

There is also a PROGRESS set with 98K videos, but this partition is for blind testing by NIST only.

the standard test partition. We denote our redefinition of the CCV dataset for few-example event recognition as CCV_{10} .

5.4.2 Video Features

To cancel out accidental effects of the choice for the underlying features, we consider the same set of appearance, motion and audio features for all our experiments, the various baselines, and our VideoStory variants. The appearance features are used for all the experiments. The motion and audio features are added for the experiments involving the multimodal fusion.

Appearance. We adopt the *video CNN* representation [175] as appearance features for event recognition, but found the very deep network of [148] to perform slightly better than [138]. For each video the frames are extracted by uniformly sampling every two seconds. Then, the CNN descriptors are extracted per frame as the 1K dimensional responses from the last fully connected layer (pool₅) of the Google Inception network [148]. We train the network on the 15,293 ImageNet categories with more than 200 examples, using the Caffe toolbox [61]. The final video CNN features are obtained by aggregating the frame descriptors over each video by VLAD encoding [58] with a codebook size of 20, resulting in a 20,480 dimensional vector.

Motion. We use the MBH descriptors along the motion trajectories [164] as motion features. The extracted 288-dimensional descriptors are reduced to 128 dimensions using PCA and are then aggregated per video using a Fisher vector [132], with 128 Gaussians resulting in a 32,768 dimensional vector. Each Fisher vector is power normalized, with $\alpha = 0.2$, as in [57].

Audio. We extract MFCCs descriptors [1] over a 10ms window. The descriptors consist of 13 values, 30 coefficients and the log-energy, along with their derivatives and the second derivatives. The MFCC descriptors are aggregated by Fisher vectors using a Gaussian Mixture Model with 256 components, resulting in a 46,080 dimensional vector.

5.4.3 Implementation Details

We learn the VideoStory embeddings by using 75% of the VideoStory46K dataset for training and 25% for validation to set the hyper-parameters of our model ($\lambda_w, \lambda_a, \lambda_s$) and of SGD (number of epochs, η). As the validation criterion we rely on the objective function value, when using $S = W^T X$.

For few-example event recognition, the event classifiers are trained as binary SVM with RBF kernels, which are shown to be effective for semantic representations [104]. Following [175, 181] we set the SVM regularization and the RBF kernel parameters by a default value of 1, as the train set is not big enough to reliably estimate them by cross-validation.

As evaluation criteria we follow the standard convention in the literature [69, 119] by relying on the average precision (AP) per event, and we report the mean average precision (mAP) for overall accuracy.

5.5 EXPERIMENTS

5.5.1 VideoStory for Few-Shot Event Recognition

Effect of Embedding

We first quantify the impact of embeddings for learning the representation by comparing VideoStory with term attribute baselines. The baselines learn the representations directly from terms

Figure 29.: Effect of embedding. The VideoStory embedding outperforms the term attribute and term attribute-f baselines, which are directly learned from the terms without embedding.

without any embedding. We evaluate all video representations for few-example event recognition using a dimensionality varying from 32 to 8,192.

1. **Term attributes.** This representation is extracted by following the tradition of predicting relevant individual terms from the video descriptions *e.g.* [22, 25, 171, 179]. A linear SVM classifier is trained per term. The classifiers which have the highest prediction accuracy, based on a 2-fold cross-validation [11], are selected as term attributes.

2. **Term attributes-f.** This baseline is similar to the previous baseline, but rather than using cross-validation to select the term attributes it simply selects the terms with the highest *frequency* in the descriptions.

Results. The results in Figure 29 demonstrate that VideoStory embeddings outperform the term attribute and term attribute-f baselines on all the three test sets.

Term attributes, which relies on the estimated reliability of individual term classifiers, performs worst. This representation suffers from two drawbacks. First, many of the visual terms are very specific and therefore incapable of characterizing the events of interest *i.e.* necklace, suitcase, and earring. Although these terms can be accurately predicted from videos, they are incapable of providing a characteristic representation for event recognition. Second, many of the terms rarely occur in video descriptions. Hence, only a limited number of positive examples are available to learn the term classifiers, which leads to a biased estimation of their reliability. Consequently, many of the discovered visual terms overfit to their small training set and do not generalize well for new videos.

The drawbacks of term attributes are relaxed by simply relying on the most frequent terms. We observe the most frequent terms usually refer to characteristic attributes of events which are frequently used by humans when describing a video, *i.e.* car, girl, and kid. Moreover, because

Figure 30.: Effect of embedding on computation efficiency. Learning the VideoStory embedding is more efficient than the term attribute and term attribute-f baselines, as it learns one classifier per group of terms rather than individual classifiers per term.

Figure 31.: *VideoStory vs Other Embeddings. VideoStory embedding outperforms the CCA and the Description embedding baselines on all three test sets. The Description embedding is the closest competitor, but it suffers from embedding correlated terms which are visually dissimilar. CCA, uses the same objective function to learn the visual and textual embeddings, which is suboptimal due to intrinsic differences between the visual and textual features.*

of their large number of positive examples, the trained visual classifiers from the term attributes-f baseline are in general more reliable.

The VideoStory embedding represents the terms in a reduced-dimensional space, where correlated terms are usually combined together. Combining correlated terms leads to less correlation between dimensions of the learned representation. Moreover, as the positive examples for all correlated terms are combined, it provides more positive video examples to train visual classifiers, often leading to better accuracy.

We also compare the learning time for VideoStory versus the term-attribute baselines on a CPU Intel Xeon E5-2690@2.90GHZ with 256GB of memory. As Figure 30 shows, learning the VideoStory embedding is more efficient than both baselines. The term attributes baseline is least efficient as it trains individual classifiers for all the 9,828 terms within the collection, before selecting the most accurate ones. Term attributes-f is more efficient, as it only trains the classifiers for the most frequent terms. Our proposed VideoStory embedding has the fastest learning time as it learns the visual embeddings for the combination of terms rather than individual terms.

Besides its effectiveness for few-example event recognition, the VideoStory embedding also improves the representation learning time by reducing the term redundancies.

VideoStory vs Other Embeddings

We compare the effectiveness of our proposed VideoStory embedding with two alternative embeddings:

1. **CCA embedding.** This baseline learns the textual and visual projections by CCA [24], which maximizes the cross-correlation between the video features and descriptions. We experimentally observed that the embedding is even more effective when we PCA-reduce the video features to a dimensionality of 1,024 before learning the CCA embedding.

2. **Description embedding.** Similar to the VideoStory, this embedding is learned by minimizing the descriptiveness and predictability losses, but in two disjoint steps: The textual projection is first learned based on the regularized Latent Semantic Indexing [166], as in Eq. (11). Then the visual projection is learned separately, by minimizing the error for predicting the embedded descriptions from the video features based on ridge regression, as in Eq. (12).

Results. The results in Figure 31 demonstrate that the VideoStory embedding outperforms the CCA and the Description embedding baselines on all the three test sets for a dimensionality larger than 256.

Table 17.: *VideoStory vs other Representations for few-example event recognition. Our proposed VideoStory outperforms the alternatives on all three test sets.*

	Low-Level	Attributes	VideoStory
MED 2013	28.1	22.5	32.4
MED 2014	21.8	17.2	24.2
CCV₁₀	50.0	48.8	54.8

We explain the gain over CCA by the fact that CCA is a symmetric embedding, it learns both the textual and visual projections in the same way. However, the textual and visual features have different distributions and properties, which may require different objective functions to learn their projections. This is achieved by our proposed VideoStory embedding as it relies on two separate LSI and ridge regression loss functions to learn the textual and visual projections, respectively.

We explain the improvement over the Description embedding by the fact that combining the terms based on textual correlation only does not necessarily imply that the corresponding video is visually correlated as well. As it happens, the terms *puppy* and *kid* have a high correlation in the descriptions but are visually dissimilar. Combining these two terms together, as is done by Description embedding, undermines the accuracy of the classifiers predicting them from videos. In contrast, in a VideoStory the correlated terms are combined only if their combination improves their classifier prediction. It prevents the combination of correlated terms which are visually dissimilar.

From now on we use a fixed 2,048 dimensional VideoStory representation, which is optimal based on Figure 29.

VideoStory vs other Representations

We evaluate the VideoStory by comparing it with state-of-the-art video representations for few-example event recognition:

1. **Low-Level.** In this baseline, the event classifiers are trained directly on the low-level video representations, without extracting a semantic representation. We rely on the video CNN features of [175], based on our implementation as detailed in Section 5.4.2.
2. **Attributes.** This representation is obtained by adopting the public ImageNet dataset as the source for training attribute classifiers as proposed in [45]. However, instead of training SVM classifiers on bag-of-words encoding of color SIFT descriptors [45], we upgrade the attributes by training a deep CNN with Google Inception architecture on the 15,293 ImageNet categories as detailed in Section 5.4.2.

Results. Table 17 shows that VideoStory outperforms the state-of-the-art attributes and low-level video representations on all three test sets.

By comparing the VideoStory and the attribute representation we observe a higher event recognition accuracy of 32.4 vs 22.5 for the MED 2013, 17.2 vs 21.8 for the MED 2014, and 54.8 vs 48.8 for the CCV₁₀ test set. We explain it by the fact that attribute baseline relies on the ImageNet categories as attributes. However, many of these pre-specified categories are not semantically relevant for the events of interest. For example, many of the ImageNet categories are devoted to specific animal species, which are not generally characteristics of the events. In contrast, the VideoStory embedding is automatically derived from the VideoStory46K dataset,

Figure 32.: From 1 to 100 examples. When the number of event exemplars are limited the VideoStory embedding outperforms the low-level representation. By increasing the number of training examples their difference becomes more subtle.

which includes many descriptions relevant to events. Note, in contrast to previous work, our low-level representation outperforms the attribute representation, indicating the strong low-level features used in this work.

The results further demonstrate that the VideoStory outperforms the low-level representation. As an explanation, we speculate that the low-level representation is prone to overfitting due to its high dimensionality. More specifically, in the low-level baseline, the event classifiers are trained on the 20,480 dimensional low-level features only from 10 positive exemplars. This may lead to overfitting as a result of the curse of dimensionality. In contrast, the VideoStory representation transfers the semantics from descriptions into the video representation to alleviate the overfitting as a sort of regularization.

From 1 to 100 examples. For further investigation, we gradually increase the number of positive examples from 1 to 100, and evaluate the accuracy of the event classifiers trained on both the low-level and VideoStory representation. The positive examples are selected randomly and the results are reported by averaging over 10 repetitions to compensate for the random effect. As Figure 32 shows, when the number of event train examples is less than 10, the VideoStory representation outperforms the low-level representation. However, by increasing the number of training examples to 100, the difference in the performances becomes more subtle, which confirms the hypothesis.

5.5.2 VideoStory Fusion

We evaluate the effectiveness of the VideoStory fusion (VideoStory ^{$\mathcal{F}$}) proposed in Section 5.3.2 for few-example event recognition. We perform the experiments using the appearance, motion, and audio modalities, as detailed in Section 5.4.2.

The additional event exemplars are selected from the 100Ex evaluation procedure provided in the MED dataset

Table 18.: Effect of fusion. Fusing more modalities to learn VideoStory^F leads to a more effective semantic representation.

Event	MED 2013				MED 2014				CCV ₁₀			
	Appearance	+ Motion	+ Audio	Event	Appearance	+ Motion	+ Audio	Event	Appearance	+ Motion	+ Audio	
Birthday party	37.1	38.8	43.4	Attempting a bike trick	9.0	9.4	8.8	Basketball	65.6	67.9	68.6	
Changing a vehicle tire	64.7	65.5	67.2	Cleaning an appliance	11.1	13.7	16.0	Baseball	58.6	59.6	59.5	
Flash mob gathering	55.3	63.8	65.0	Dog show	83.7	88.0	89.7	Soccer	58.7	64.5	64.7	
Getting a vehicle unstuck	59.3	65.2	65.3	Giving directions to a location	0.5	1.4	1.0	Ice Skating	70.2	74.5	75.1	
Grooming an animal	24.3	29.4	28.2	Marriage proposal	0.3	0.5	0.6	Skating	79.4	81.8	81.9	
Making a sandwich	16.7	18.9	21.3	Renovating a home	11.5	11.9	12.9	Swimming	78.9	81.8	81.7	
Parade	33.9	44.7	44.0	Rock climbing	13.8	14.2	13.7	Biking	67.3	69.0	70.2	
Parkour	61.3	72.8	72.2	Town hall meeting	40.1	36.3	42.4	Cat	66.2	68.3	70.8	
Repairing an appliance	44.5	49.7	57.4	Winning a race without a vehicle	22.1	26.8	26.7	Dog	65.4	68.9	68.6	
Working on a sewing project	47.2	48.0	46.0	Working on a metal crafts	15.0	17.9	20.6	Bird	33.9	35.9	36.4	
Attempting a bike trick	8.8	9.3	8.9	Beekeeping	54.1	53.5	48.1	Graduation	25.3	27.6	25.1	
Cleaning an appliance	10.5	13.2	15.5	Wedding shower	19.7	31.5	40.3	Birthday	54.2	56.2	61.9	
Dog show	81.6	84.9	85.9	Non-motorized vehicle repair	65.3	66.0	66.5	Wedding Reception	17.3	18.3	19.5	
Giving directions to a location	0.6	1.0	0.9	Fixing musical instrument	25.4	31.0	44.0	Wedding Ceremony	45.6	53.5	58.4	
Marriage proposal	0.3	0.4	0.5	Horse riding competition	40.5	37.9	35.6	Wedding Dance	51.3	58.6	61.7	
Renovating a home	11.4	12.1	13.8	Felling a tree	12.4	16.2	22.3	Music Performance	41.0	41.6	51.1	
Rock climbing	13.7	14.9	14.4	Parking a vehicle	17.3	19.5	21.9	NonMusic Performance	30.6	35.6	36.4	
Town hall meeting	40.2	38.9	44.8	Playing fetch	1.3	1.3	1.4	Parade	49.4	64.0	64.0	
Winning a race without a vehicle	21.9	28.8	27.8	Tailgating	32.9	34.5	37.7	Beach	74.5	73.9	72.9	
Working on a metal crafts	15.2	18.0	20.4	Tuning musical instrument	8.7	8.3	15.1	Playground	63.4	61.9	58.1	
<i>mAP</i>	32.4	35.9	37.1	<i>mAP</i>	24.2	26.0	28.3	<i>mAP</i>	54.8	58.2	59.3	

Table 19.: Comparison with other fusions. The VideoStory $\mathcal{F}$ outperforms the alternative fusion tactics.

	Early fusion	Late fusion	VideoStory early fusion	VideoStory late fusion	VideoStory $\mathcal{F}$
MED 2013	32.6	33.8	33.8	33.6	37.1
MED 2014	27.0	27.0	27.0	26.1	28.3
CCV₁₀	56.3	55.8	56.9	55.4	59.3

Effect of fusion

We study the impact of fusing multiple modalities for learning the video representation. We start from using only the appearance features (**Appearance**) and gradually add the motion (**+ Motion**) and audio features (**+ Audio**) to learn the video representation by our proposed VideoStory $\mathcal{F}$.

Results. Table 18 shows that the more modalities are fused, the more effective video representation is learned by VideoStory $\mathcal{F}$. It confirms that incorporating more modalities for estimating the predictability loss leads to a more effective representation learning.

By looking into the results, we observe that some events are benefiting more from adding the auditory modality, *i.e.* *birthday party* and *music performance*. For these events, there are some distinctive terms which are more effectively predicted from the audio features, *i.e.* *singing*, *clapping*, and *piano*. Similarly, some other events are more improved by including motion features, *i.e.* *parkour* and *parade*, for which some distinctive terms such as *jumping*, *rolling*, and *running* are well predictable from the motion features. It demonstrates that different modalities are complementary for predicting the terms, so they are all required to effectively predict the video descriptions.

Comparison with other fusions

We evaluate our proposed VideoStory $\mathcal{F}$ by comparing it with the following fusion baselines:

1. **Early fusion.** This baseline fuses the modalities by simply concatenating the low-level features from all the three modalities into a longer feature vector. Then the event classifiers are trained and applied on the concatenated low-level features. This simple baseline is shown to be very competitive to the more complicated multiple kernel learning [151].
2. **Late fusion.** This baseline fuses the modalities at the level of event classification scores. A separate event classifier is learned by training an SVM on low-level video features per modality. For each test video, the final event detection score is obtained by averaging the detection scores predicted by each classifier, as evaluated in [110].
3. **VideoStory early fusion.** This baseline learns one VideoStory embedding on the concatenation of various low-level video features, using the standard VideoStory objective of Eq. (10). For each video, a semantic representation is extracted by applying the learned feature projection on the concatenation of the low-level video features. Then, the event classifiers are trained and applied on the semantic video representations.
4. **VideoStory late fusion.** This baseline learns three separate VideoStory embeddings, one per modality, using the standard VideoStory objective of Eq. (10). For each video, a semantic representation is obtained by concatenating the three VideoStory representations, which are predicted from each modality. Then, the event classifiers are trained and applied on the semantic video representations.

Results. The results are reported in Table 19. The VideoStory $\mathcal{F}$ outperforms the alternative fusion

baselines on all the three test sets. We explain the better performance of VideoStory $\mathcal{F}$ compared to the early fusion and late fusion by the fact that both baselines train the event classifiers directly from low-level video features. However, in VideoStory $\mathcal{F}$ the event classifiers are trained on the semantic VideoStory representations, which are more effective than the low-level features in general, as shown in Section 5.5.1 for the appearance features.

In the VideoStory late fusion the embeddings are learned *separately* per modality. In contrast, VideoStory $\mathcal{F}$ relies on all the video modalities jointly to estimate the predictability loss, which in general is more reliable than the per modality predictability estimations. For further investigation, we visualizing the learned textual projection matrices $\mathbf{A}$ in Figure 33. As shown in this figure, the VideoStory $\mathcal{F}$ prevents some undesirable combination of terms that happen when the predictability losses are estimated separately per modality. For example, in the VideoStory which is learned only on audio features, the terms *laugh*, *cheer*, *bark*, *dog*, and *woof* are all combined as they have similar auditory features (see the right plot). However, the VideoStory $\mathcal{F}$ does not combine the *laugh* and *cheer* with the *bark*, *dog*, and *woof* terms, as these terms are different in the

Figure 33.: Effect of learning the embeddings jointly over all the modalities by comparing VideoStory $\mathcal{F}$ with the VideoStory late fusion baseline. Each textual projection matrix is visualized by plotting $\mathbf{A} \times \mathbf{A}^T$, which reveals the learned term combinations. The VideoStory $\mathcal{F}$ learn a more reasonable combination of terms compared to the VideoStory late fusion, where the embeddings are learned disjointedly per modality.

Figure 34.: Unseen video examples and their term vectors predicted by the VideoStory (middle) and the VideoStory₀ embeddings (bottom). The size of each term indicates its prediction confidence score. The VideoStory₀ more effectively predicts the indicative terms of the event.

appearance and motion features. Hence, the VideoStory ^{$\mathcal{F}$} learns a more reliable combination of terms, which leads to a more effective video representation.

VideoStory early fusion learns a single feature projection matrix on the concatenated video features. However, the low-level features from different modalities have a different intrinsic dimensionality, distribution, and meaning, which aggravates the learning from their concatenation. In contrast, the VideoStory ^{$\mathcal{F}$} learns separate feature projection matrices per modality, where each feature projection is optimized based on the features from one modality. It alleviates learning the feature projections, which leads to a more effective video representation.

To conclude, the VideoStory ^{$\mathcal{F}$} effectively fuses the features from various video modalities by embedding them into a mutual semantic representation learned jointly over all the modalities.

5.5.3 VideoStory for Zero-Example Recognition

We evaluate the effectiveness of the VideoStory₀ and its multimodal fusion VideoStory₀ ^{$\mathcal{F}$} , proposed in Section 5.3.3, for zero-example event recognition by comparing with the following video representations:

1. **Term attributes-f.** Similar to the few-example experiments in Section 5.5.1, this representation is extracted by the frequent term attributes, which are learned directly from terms without any embedding.
2. **VideoStory.** This video representation is learned by the VideoStory embeddings, proposed in Section 5.3.1, for few-example event recognition. More specifically, this representation is learned based on the LSI, in Eq. (11), as the descriptiveness loss, and the unimodal predictability, in Eq. (12), as the predictability loss.

Results.

The results are reported in Table 20. The three embedding based representations, VideoStory, VideoStory₀, and VideoStory₀ ^{$\mathcal{F}$} , outperform the term attributes-f representation on both the MED 2013 and MED 2014 test sets. It confirms the effectiveness of embeddings for learning the semantic representations for zero-example event recognition. The most effective representation is learned by the VideoStory₀ ^{$\mathcal{F}$} , which rely on the term sensitive descriptiveness loss (L_d^{ts}), and the multimodal predictability loss ($L_p^{\mathcal{F}}$).

By comparing the term attributes-f and VideoStory, we observe that although the VideoStory utilizes the term correlations by embedding, its improvement over the term attributes-f is subtle,

Table 20.: Effect of embedding for zero-example event recognition. The results demonstrate that the term sensitive descriptiveness loss (L_d^{ts}) and the multimodal predictability loss (L_p^f) are both effective and complementary for zero-example event recognition.

Event	MED 2013					MED 2014				
	Term attributes-f	VideoStory	VideoStory ₀	VideoStory ₀ ^f	Event	Term attributes-f	VideoStory	VideoStory ₀	VideoStory ₀ ^f	
Birthday party	14.5	24.6	30.3	37.4	Attempting a bike trick	4.9	8.8	7.1	5.9	
Changing a vehicle tire	17.7	43.9	42.1	38.2	Cleaning an appliance	11.9	8.2	9.5	12.3	
Flash mob gathering	15.2	14.5	22.4	33.8	Dog show	22.8	4.0	5.8	5.4	
Getting a vehicle unstuck	28.9	40.2	36.0	32.2	Giving directions to a location	0.4	0.6	0.6	1.1	
Grooming an animal	17.4	18.7	28.0	23.4	Marriage proposal	0.5	0.3	0.4	0.7	
Making a sandwich	14.8	19.4	21.1	17.1	Renovating a home	4.8	5.2	6.8	6.3	
Parade	20.0	17.6	26.4	38.2	Rock climbing	0.8	1.6	1.4	1.0	
Parkour	15.9	26.1	28.0	40.3	Town hall meeting	15.5	1.9	9.2	9.2	
Repairing an appliance	41.3	39.8	41.4	46.3	Winning a race without a vehicle	5.8	9.4	8.4	7.6	
Working on a sewing project	23.8	30.8	36.4	36.0	Working on a metal crafts	7.5	1.5	4.7	4.8	
Attempting a bike trick	5.0	8.8	7.1	5.9	Beekeeping	0.4	0.3	0.6	0.9	
Cleaning an appliance	10.7	8.2	9.1	12.5	Wedding shower	1.3	2.3	3.3	3.3	
Dog show	23.5	4.0	5.8	5.9	Non-motorized vehicle repair	8.4	33.2	44.2	46.9	
Giving directions to a location	0.5	0.6	0.7	0.8	Fixing musical instrument	3.1	4.7	7.9	25.5	
Marriage proposal	0.5	0.3	0.4	0.7	Horse riding competition	2.3	7.7	12.8	13.7	
Renovating a home	4.8	5.2	6.8	6.4	Felling a tree	3.9	7.7	6.5	6.8	
Rock climbing	0.8	1.6	1.4	1.0	Parking a vehicle	0.3	2.8	2.3	2.9	
Town hall meeting	15.5	1.9	9.2	9.6	Playing fetch	3.6	2.2	2.9	3.3	
Winning a race without a vehicle	5.8	9.4	8.4	7.9	Tailgating	0.6	0.2	0.2	0.4	
Working on a metal crafts	7.3	1.6	4.9	7.0	Tuning musical instrument	0.9	0.5	0.9	2.0	
mAP	14.2	15.9	18.3	20.0	mAP	5.0	5.2	6.8	8.0	

Table 21.: Comparison with the state-of-the-art. VideoStory embeddings set a new state-of-the-art for the both few- and zero-example event recognition.

Few-example on MED 2013			Zero-example on MED 2013		
Habibian <i>et al.</i> [42]	MM 2014	19.6	Ye <i>et al.</i> [179]	MM 2015	9.0
Nagel <i>et al.</i> [107]	BMVC 2015	21.8	Chang <i>et al.</i> [20]	IJCAI 2015	9.6
Li <i>et al.</i> [86] [†]	ICCV 2013	23.7	Mazloom <i>et al.</i> [101]	ICMR 2015	11.9
Tang <i>et al.</i> [150] [†]	CVPR 2012	26.8	Wu <i>et al.</i> [171]	CVPR 2014	12.7
Sun <i>et al.</i> [147] [†]	CVPR 2014	28.7	Jiang <i>et al.</i> [63]	AAAI 2015	12.9
Chang <i>et al.</i> [21]	MM 2015	30.9	Jiang <i>et al.</i> [65]	MM 2015	18.3
This paper: VideoStory		32.4	This paper: VideoStory ₀		18.3
This paper: VideoStory ^F		37.1	This paper: VideoStory ₀ ^F		20.0

[†] Based on implementation by [21]

especially on the MED 2014 test set. As an explanation we highlight that in the VideoStory baseline, the descriptiveness loss is defined based on LSI. As discussed before, the LSI loss treats all the terms equally when measuring the reconstruction error. Hence this loss is biased toward minimizing the reconstruction error for the frequent terms. As a consequence, the VideoStory embedding generally is not accurate in predicting the terms, which are infrequent in its training data. It degrades the effectiveness of the VideoStory embedding for zero-example event recognition, where the accuracy in predicting some specific terms is crucial.

The drawback of the VideoStory baseline is addressed by VideoStory₀, which relies on the term sensitive descriptiveness loss. This loss minimizes the reconstruction error for the terms, which are indicative of the event, even if those terms are not frequent in the VideoStory train data. As a result, it can predict the distinctive terms of an event more accurately, compared to the VideoStory, which improves the event recognition accuracies from 15.9 to 18.3 on MED 2013, and from 5.2 to 6.8 on MED 2014. In Figure 34, we compare the term vectors, which are predicted by VideoStory and VideoStory₀ for three video examples. It shows that the VideoStory₀ more effectively predicts the indicative terms of the event, although they are not frequent in the train data.

The most effective representation is learned by the VideoStory₀^F. It indicates that the term sensitive descriptiveness loss (L_d^{ts}) and the multimodal predictability loss (L_p^F) are both effective and complementary when learning semantic video representations for zero-example event recognition.

5.5.4 Comparison with the state-of-the-art

We evaluate the merit of the proposed VideoStory by comparing it with several other recent works on few- and zero-example event recognition. Since in most papers the results are reported only on the MED 2013 test set, we limit our comparisons to this test set.

The results are reported in Table 21. It shows that our proposed representation learning sets a new state-of-the-art for the both few- and zero-example event recognition. It should be noted that these works rely on modeling the temporal aspects of the events [21, 86, 147, 150], using larger train set for representation learning [20, 63, 65], and query expansion [20, 171] to improve the event recognition. These improvements can be also applied together with our representation learning for a more effective event recognition. Very recently Jiang *et al.* [65] improved their results for

zero-example recognition from 18.3 to 20.8 after adding re-ranking by pseudo relevance feedback, we expect a similar gain for VideoStory_0 and $\text{VideoStory}_0^{\mathcal{F}}$.

5.6 CONCLUSION

In this paper, we attack the problem of event recognition in video when examples are scarce. We propose the *VideoStory* embedding that learns a semantic video representation from a set of videos and their textual descriptions by minimizing a joint objective function balancing term descriptiveness and video predictability losses. As a result, the terms which are correlated in the descriptions are combined together to improve their video predictability. Besides its effectiveness for few-example event recognition, the *VideoStory* embedding also improves the representation learning time by reducing redundancies during training.

In addition, we propose the $\text{VideoStory}^{\mathcal{F}}$ embedding with a multimodal predictability loss learned jointly over video appearance, motion, and audio features. The different modalities are complementary for predicting the terms, so they are all required for a richer video description. Moreover, embedding the heterogeneous video features into a mutual semantic space leads to few-example event recognition that is more effective than traditional fusion tactics.

We also propose the VideoStory_0 embedding to recognize an event in video from just a textual query in the form of an event description. This embedding relies on a term sensitive descriptiveness loss to learn a more accurate representation for the terms, which are indicative of the event. Finally, the $\text{VideoStory}_0^{\mathcal{F}}$ embedding demonstrates that the term sensitive descriptiveness loss and the multimodal predictability loss are both effective and complementary to learn semantic video representations for zero-example event recognition.

By its abilities to improve predictability upon any underlying video feature while at the same time maximizing semantic descriptiveness, *VideoStory* leads to state-of-the-art accuracy for both few- and zero-example recognition of events in video. We consider *VideoStory*'s ability to generate human interpretable representations for previously unseen videos most appealing, as it opens up new connections with natural language processing and computational linguistics for describing and querying videos.

SUMMARY AND CONCLUSIONS

6.1 SUMMARY

This thesis studies learning concept vocabularies for describing video content. Starting from manually specifying the vocabulary, in chapter 2, we gradually automate vocabulary construction by concept compositions, in chapter 3, and finally by learning from human stories, in chapter 4 and 5. Before revisiting the research questions posed in the introduction, we will give a brief overview of each chapter as follows:

Chapter 2: Recommendations for Recognizing Video Events by Concept Vocabularies.

Representing videos using vocabularies composed of concept detectors appears promising for generic event recognition. While many have recently shown the benefits of concept vocabularies for recognition, studying the characteristics of a universal concept vocabulary suited for representing events is ignored. In this chapter, we study how to create an effective vocabulary for arbitrary-event recognition in web video. We consider five research questions related to the number, the type, the specificity, the quality and the normalization of the detectors in concept vocabularies. A rigorous experimental protocol using a pool of 1,346 concept detectors trained on publicly available annotations, two large arbitrary web video datasets and a common event recognition pipeline allow us to analyze the performance of various concept vocabulary definitions. From the analysis we arrive at the recommendation that for effective event recognition the concept vocabulary should *i)* contain more than 200 concepts, *ii)* be diverse by covering *object, action, scene, people, animal* and *attribute* concepts, *iii)* include both general and specific concepts, *iv)* increase the number of concepts rather than improve the quality of the individual detectors, and *v)* contain detectors that are appropriately normalized. We consider the recommendations for recognizing video events by concept vocabularies the most important contribution of the chapter, as they provide guidelines for future work.

Chapter 3: Composite Concept Discovery for Zero-Shot Video Event Detection.

We consider automated detection of events in video without the use of any visual training examples. A common approach is to represent videos as classification scores obtained from a vocabulary of pre-trained concept classifiers. Where others construct the vocabulary by training individual concept classifiers, we propose to train classifiers for combination of concepts composed by Boolean logic operators. We call these concept combinations *composite concepts* and contribute an algorithm that automatically discovers them from existing video-level concept annotations. We discover composite concepts by jointly optimizing the accuracy of concept classifiers *and* their effectiveness for detecting events. We demonstrate that by combining concepts into composite concepts, we can train more accurate classifiers for the concept vocabulary, which leads to improved zero-shot event detection. Moreover, we demonstrate that by using different logic operators, namely “AND”, “OR”, we discover different types of composite concepts, which are complementary for zero-shot event detection. We perform a search for 20 events in 41K web videos from two test sets of the challenging TRECVID Multimedia Event Detection 2013 corpus.

The experiments demonstrate the superior performance of the discovered composite concepts, compared to present-day alternatives, for zero-shot event detection.

Chapter 4: Discovering Semantic Vocabularies for Cross-Media Retrieval. This chapter proposes a data-driven approach for cross-media retrieval by automatically learning its underlying semantic vocabulary. Different from the existing semantic vocabularies, which are manually pre-defined and annotated, we automatically discover the vocabulary concepts and their annotations from multimedia collections. To this end, we apply a probabilistic topic model on the text available in the collection to extract its semantic structure. Moreover, we propose a learning to rank framework, to effectively learn the concept classifiers from the extracted annotations. We evaluate the discovered semantic vocabulary for cross-media retrieval on three datasets of image/text and video/text pairs. Our experiments demonstrate that the discovered vocabulary does not require *any* manual labeling to outperform three recent alternatives for cross-media retrieval.

Chapter 5: VideoStory Embeddings Recognize Events when Examples are Scarce. This chapter aims for event recognition when video examples are scarce or even completely absent. The key in such a challenging setting is a semantic video representation. Rather than building the representation from individual attribute detectors and their annotations, we propose to learn the entire representation from freely available web videos and their descriptions using an embedding between video features and term vectors. In our proposed embedding, which we call *VideoStory*, the correlations between the terms are utilized to learn a more effective representation by optimizing a joint objective balancing descriptiveness and predictability. We show how learning the VideoStory using a multimodal predictability loss, including appearance, motion and audio features, results in a better predictable representation. We also propose a variant of VideoStory to recognize an event in video from just the important terms in a text query by introducing a term sensitive descriptiveness loss. Our experiments on three challenging collections of web videos from the NIST TRECVID Multimedia Event Detection and Columbia Consumer Videos datasets demonstrate: i) the advantages of VideoStory over representations using attributes or alternative embeddings, ii) the benefit of fusing video modalities by an embedding over common strategies, iii) the complementarity of term sensitive descriptiveness and multimodal predictability for event recognition without examples. By its abilities to improve predictability upon any underlying video feature while at the same time maximizing semantic descriptiveness, VideoStory leads to state-of-the-art accuracy for both few- and zero-example recognition of events in video.

6.2 CONCLUSIONS

In this thesis, we study a fundamental question for developing storytelling machines: *what vocabulary is suited for machines to tell the story of a video?* We start by manually specifying the vocabulary concepts and their annotations. In order to effectively handcraft the vocabulary, we empirically study *what are the best practices for handcrafting the vocabulary for video storytelling?* From our analysis, we conclude that for an effective storytelling the vocabulary should encompass over thousands of concepts from various types, including object, action, scene, people, animal, and attribute. Moreover, the vocabulary should include the detectors for both generic concepts and specific concepts, which are trained and normalized in an appropriate way.

Creating such a handcrafted vocabulary of concepts is labor intensive. We alleviate the manual labor by addressing the next research question: *can a machine learn novel video concepts by composition?* We propose an algorithm, which learns new concepts by composing the vocabulary concepts by Boolean logic connectives. We demonstrate that concept composition is an effective trick to infer the annotations, needed for training new concept detectors, without additional

human supervision. The learned composite concepts enrich the vocabulary by modeling various relationships between the concepts, such as co-occurrences, abstractions, and exclusions.

As a further step towards reducing the manual labor for vocabulary construction, we investigate the question *can a machine learn its vocabulary from human stories?* We demonstrate that the human-written stories and their associated videos are precious resources for learning the vocabulary. By analyzing the human stories using topic models, we effectively extract the concepts that humans use for describing videos. Moreover, we show that the occurrences of concepts in stories can be effectively used as weak supervision to train concept detectors.

Finally, we address the question *how to learn the vocabulary from human stories?* We formulate vocabulary construction as learning a multimodal embedding between visual features and terms from stories. The embedding is learned by minimizing a joint objective function balancing term descriptiveness and video predictability losses. As a result, the terms which are correlated in the stories are combined together to improve their video predictability. Furthermore, we demonstrate how the multimodal embeddings can learn concept detectors from various video modalities, *i.e.* the *laughing* concept from audio, and the *dancing* concept from motion. As a final conclusion, we believe that by training a multimodal embedding, on easy to obtain videos and their stories *i.e.* from YouTube, machines can automatically learn comprehensive vocabularies including tens of thousands of concepts.

This thesis contributes in constructing a comprehensive vocabulary, which classifies thousands of concepts in a video. The recognized concepts play a role as the observable nouns, verbs, and adjectives of the story. To generate concise stories comparable to human performance, the machine still needs to learn other skills. For example, it needs to put the detected concepts into a syntactically correct sentence. Recently, recurrent neural networks with a deep long short-term memory architecture, have shown some promise towards this challenge [72, 173]. As another future direction, modeling the temporal and dynamic interaction between the concepts makes the machine able to convey the video stories over time, such as a girl blows candles *before* cutting a cake. To this end, we need to train the concept detectors to recognize concepts at segment or frame levels [12, 78, 101]. In addition, by detecting the spatial location of concepts in video frames, we can express spatial relations and interactions between concepts, as it clarifies whether a person is inside a car or in front of it. We close this thesis by noting that learning the vocabulary, as the primary focus of this thesis, is only the first step towards developing general storytelling machines. Understanding and expressing emotions through stories, telling humorous, artistic and inspiring stories are just few examples of the future steps needed to develop storytelling machines that rival the magic of the human mind.

HARVESTING WEB VIDEOS AND THEIR DESCRIPTIONS

In this appendix, we detail the procedure for collecting the train data needed for learning the VideoStory embeddings detailed in Chapter 5. We also assess the impact of the quality and quantity of the harvested videos and descriptions on the learned VideoStory embeddings by comparison with expert provided video descriptions.

A.1 WEB SPIDERING ALGORITHM

Rather than describing the video content manually, we opt to harvest both the videos and the descriptions from the web. Video sharing web sites, such as YouTube and Vimeo, provide a rich and varied source of videos and user provided descriptions, such as their title captions and comments. Although video title captions do not necessarily correspond to the visual content of the videos, we will show that by harvesting a large number of these captioned videos and applying a set of quality filters we obtain reliable video descriptions.

We start from an initial pool of descriptions, as the collection seeds, and iteratively collect videos and their title captions from YouTube. For the collection seeds, we rely on 3,000 sentence descriptions from the training partition of the NIST TRECVID HAVIC corpus [146]. Then each description within the pool is queried to YouTube and the 25 most relevant videos are retrieved, based on YouTube’s textual similarity search. Every retrieved video is passed through a set of quality filters. The videos which pass all the filters are added to the collection and their title captions are added to the description pool. We iteratively repeat this procedure until enough videos are collected. We will first detail our quality filters before providing the statistics of our harvested video and description dataset.

A.1.1 *Quality Filters*

Event Filter Events are generally described by their actors, actions, and possible involved objects [40]. Hence we assume that a description of an event video should contain actors, actions and objects. For this purpose, we parse the grammatical structure of title captions using a probabilistic context free grammar parser [74]. Then we accept a video only if its caption includes verbs, subjects, and objects.

Visualness Filter There are many terms in title captions which do not refer to visually depictable attributes, such as buy, God, and genius. These attributes are not recognizable by present-day visual classifiers, so should not be included in the collection. For this filter we evaluate the visualness of caption terms. Rather than relying on visual features [11, 31, 77], which are expensive to extract and limit the scalability, we evaluate the visualness of each term in the title caption by measuring its similarities to the ImageNet synsets [30] in the WordNet hierarchy. We measure the similarity of each synset pair by following [9], which finds overlaps between the

Figure 35.: Terms from the VideoStory46K dataset occurring in more than 500 title captions of the harvested YouTube videos.

glosses of two synsets as well as their directly linked synsets. Finally, we define the visualness of a title caption by averaging the visualness of all its terms. The captions whose visualness exceeds a threshold of 0.5 are accepted by our harvesting procedure.

Reality Filter A considerable amount of YouTube videos are related to celebrities, TV series, and movie trailers. We observe these professional videos are typically semantically dissimilar to the event videos which we are interested in. Moreover, they often infringe intellectual property rights. Therefore, we prefer to filter out the corresponding videos and title captions. Our reality filter relies on a list of keywords from Wikipedia, which provides an extensive index of celebrity, TV series and movie names. We exclude the videos whose description matches any of the keywords from the list.

Temporal Filter Our last filter rests on the assumption that short videos better match their title captions, compared to long videos. It is because long videos usually contain a broad set of attributes that are typically not specified completely by their captions. Hence, we only retrieve the YouTube videos which are shorter than 120 seconds.

A.1.2 VideoStory46K Dataset

Following the proposed procedure, including all quality filters, we harvest 45,826 videos from YouTube. The videos have an average length of 58.4 seconds and the whole collection contains 743 hours of videos. Every video comes with a short title caption provided by the user who has uploaded the video. Every caption is made of 7.7 individual terms on average, with a standard deviation of 1.8 terms. There are 19,159 unique terms in the captions, most of them occurring infrequently in the collection, *i.e.* 50% of the terms occur only once in the collection, and only 0.4% of the terms occur more than 500 times (see Figure 36.). Some examples of these frequent

wikipedia.org/wiki/List_of_American_television_series

Figure 36.: Cumulative distribution function (CDF) for term frequencies in the title captions of the VideoStory46K dataset. The y-axis shows the cumulative probability that a term occurs in the caption, the x-axis denotes the term frequency tf in logarithmic scale. It shows that most of the terms rarely occur in captions e.g. 50% (resp. 84%) of the terms only occur in 1 (resp. 10) captions.

terms are shown as a tag cloud in Figure 35. Our dataset of videos and their descriptions, which we call VideoStory46K, is available for download at <http://www.mediamill.nl>. Illustrative examples from the dataset are shown in Figure 37.

a man explains how to repair ice maker in kitchen.

a man replacing stove heating element in kitchen.

Brittany and Brianna fix their computer fan.

Figure 37.: Example videos and title captions from our VideoStory46K dataset.

Figure 38.: Description quality and quantity. As expected, the more reliable the correspondence between description and video content, the better the result. The event recognition accuracy obtained by ExpertSentences10K can be approached, and even improved, by simply harvesting more descriptions from the web.

A.2 EXPERIMENT ON DESCRIPTION QUALITY AND QUANTITY

We assess the influence of the quality and quantity of the videos and descriptions that we use as input to learn our VideoStory. We compare the VideoStory learned from the VideoStory46K dataset with two baselines.

1. **ExpertSentences10K**, includes 10K videos from the Research partition of the NIST TRECVID HAVIC corpus [44, 146]. Each video in this collection comes with an expert written description per video. The descriptions are made of a few sentences written by a team of 60 expert annotators with the purpose of summarizing the visual content of the videos. Consequently, there is always a strong correspondence between a description and its video in this collection.

2. **VideoStory10K**, includes 10K random videos and descriptions from the VideoStory46K dataset, which we collected as discussed in Section A.1. This collection includes the same number of videos and captions as the ExpertSentences10K dataset, but the captions are generally of lower quality for event recognition because of the non-expert descriptions and the fact that video captions on YouTube do not necessarily correspond to the visual content.

Experimental Setup. The experiments are performed on three test sets: MED and Kindred from the TRECVID Multimedia Event Detection [146] and the Columbia Consumer Video (CCV) [69], by following the standard data partitioning [42]. We rely on the MBH descriptors along the motion trajectories [164] to extract the low-level video representation as detailed in Section 5.4.2. The learned VideoStory embeddings are evaluated by their effectiveness of event recognition as in Section 5.5.1.

Results. The results are presented in Figure 38. As expected, the more reliable the correspondence between the description and the video content it describes, the better the result. A VideoStory learned from ExpertSentences10K always performs better than an embedding learned from the same amount of descriptions from VideoStory10K. Yet it should be noted that the expert descriptions are generally unavailable or hard to obtain. Interestingly, the event recognition accuracy obtained by expert descriptions can be approached, and even improved (for both the test set MED and test set Kindred) by simply harvesting more descriptions from the web. When considering all results, the complete VideoStory46K dataset is the best choice overall. The embedding reduces the influence of noisy video descriptions, especially when the number of input videos and descriptions are large. It demonstrates the value of user generated videos and descriptions as an unlimited, free, yet precious resource for constructing an effective VideoStory.

STOP-FRAME REMOVAL IMPROVES WEB VIDEO CLASSIFICATION

B.1 INTRODUCTION

Classifying semantic concepts in videos is limited by acquisition of adequate training data. Manual labeling of training data is expensive, so recent efforts utilize web videos for automatic creation of training data [75, 156]. These works simply download videos from sharing sites, like YouTube or Internet Archive, and use the user provided tags per video as the ground truth annotations. Apart from the fact that user provide tags may be uncontrolled, ambiguous, and overly personalized, web videos may contain various frames, which are irrelevant to the provided video category.

Including irrelevant frames during training degrades classifier accuracy. This problem has been addressed by several works [39, 75, 156, 165]. Ulges *et al.* [156] propose a probabilistic framework for learning from web videos in the presence of irrelevant frames. They model the relevance of frames as a latent random variable, which is estimated by expectation maximization during training. Gu *et al.* [39] cast the problem as multiple instance learning, in which training data are grouped as bags of instances and each bag is allowed to contain irrelevant frames besides relevant frames. They group video frames as bags of instances and propose a kernel function to learn from the bags. These methods do not make any assumption about the distribution of irrelevant frames between videos, which might be useful in recognizing some kinds of irrelevant frames.

Another approach to detect irrelevant frames is to employ outlier detection techniques [120, 185]. They identify irrelevant frames as outliers, which are dissimilar to the majority of video frames. Different appearance and temporal features are employed to measure the similarity between video frames, such as color histograms and motion feature. However, these methods only rely on the visual similarities of video frames without considering the semantics of videos. In contrast, we speculate that semantics of videos can be served as a novel prior information for identifying irrelevant frames.

Looking into web videos, we observe considerable number of frames that may randomly appear in any video, *i.e.* blank, dark and over exposed frames. These frames are randomly distributed between videos from different semantic categories. In this paper, we address this subset of irrelevant frames and utilize their random distribution to explicitly identify them within web videos. We call them *stop-frames*, which is inspired from stop-words concept in document retrieval [38]. Stop-words such as “the”, “a”, “is” and “of” are terms that commonly occur in every document and have no discriminating ability. Removing stop-words is a common pre-processing step in document retrieval. Analogous to stop-words, we define stop-frames as those frames that occur in many videos without being correlated to any particular category. Some examples of stop-frames are shown in Figure 39.

Figure 39.: Stop-frame examples which we aim to identify and remove automatically, to arrive at more reliable video concept and event classifiers.

Stop-words are usually detected based on a pre-specified list of words, *i.e.* prepositions. Similarly, a natural approach to detect stop-frames is to pre-specify the categories of stop-frames and to train a visual detector per category [120], such as blank frame detector and motion blur detector. However, due to the large diversity of the visual domain it is impossible to pre-specify all the stop-frame categories. Hence, we aim for automatically detecting stop-frames without pre-specifying them.

The main contributions of this paper are: (i) introducing stop-frames as a frequent sample of irrelevant frames in web videos, and (ii) proposing a simple algorithm for their detection. This algorithm can be considered as a pre-process, which explicitly removes stop-frames and can be coupled with other existing methods to handle irrelevant frames in web videos. Experiments demonstrate that using our method to remove stop-frames, improves classification accuracy by 60% and 24% in terms of mean average precision for event and concept detection benchmarks, respectively.

B.2 STOP-FRAME REMOVAL

Stop-frames are produced independent to the video content. They may be caused by different reasons: (i) some are made because of amateur recording and uncontrolled conditions, in which web videos are taken. *i.e.* dark and over exposed frames, which are made because of ill illumination conditions, blurred frames, which are made if camera is shaken during recording, and blank frames, which are made when camera is extremely occluded by moving objects. (ii) Another group are a consequence of editing web videos. The edited videos contain blank frames, sometimes with overlayed text or a logo, which do not necessarily provide a visual clue about the semantic category. (iii) Moreover, some frames are affected by typical encoding problems that may occur while down sampling web videos and ruin their visual information. These circumstances randomly happen and affect videos with different content.

Stop-frames may randomly appear in any video, so it is harder to recognize their category in comparison with the informative frames, which usually occur in a particular category. There-

fore, stop-frames are more probable to be misclassified than informative frames. Based on this observation, we hypothesize that stop-frames are the frames commonly misclassified by classifiers.

B.2.1 Algorithm

The inputs to our proposed algorithm are a group of semantic classifiers C_i trained on web videos so as to predict if a frame f belongs to the category i . Semantic categories are arbitrary selected and the classifiers are trained on the frames extracted from web videos. Also, frame labels are inherited from video level annotations. We use these classifiers to identify the stop-frames within the training data.

Applying C_i on a frame f , will predict if f belongs to the category i . The classifiers are applied on training video frames with known video level categories, so we can evaluate their predictions. For example, suppose C_i is trained to recognize the video frames representing the concept *basketball* and we apply it on a frame f from the *swimming* category. We know that f is misclassified, if C_i predicts that it belongs to basketball category. Let us define a binary random variable $M_{C_i,f}$ to denote f is misclassified by C_i . In addition we define a binary random variable S_f indicating whether frame f is a stop-frame. We utilize the classifier predictions ($M_{C_i,f}$) to estimate the probability that a frame f is a stop-frame ($P(S_f)$). According to the Bayes rule we formulate $P(S_f)$ as:

$$P(S_f) = \frac{P(S_f | M_{C_1,f}, \dots, M_{C_n,f}) \cdot P(M_{C_1,f}, \dots, M_{C_n,f})}{P(M_{C_1,f}, \dots, M_{C_n,f} | S_f)} \quad (24)$$

Assuming that the semantic classifiers are independent from each other, equation 1 can be reformulated as equation 2. To hold this assumption, the semantic categories should be distinct and be selected independently.

$$P(S_f) = \prod_{i=1}^n \frac{P(S_f | M_{C_i,f}) \cdot P(M_{C_i,f})}{P(M_{C_i,f} | S_f)} \quad (25)$$

Stop-frames are uniformly distributed within all categories, so we assume that all classifiers might misclassify them with the same probability. Based on this assumption, we replace $P(M_{C_i,f} | S_f)$ with a constant.

Stop-frames are more probable to be misclassified, but classifiers may misclassify informative frames too. The more accurate a classifier is, the less informative frames it misclassifies. In other words, the more accurate a classifier is, the more probable its misclassified frames are stop-frames. Therefore, we have approximated $P(S_f | M_{C_i,f})$ with accuracy of C_i in terms of average precision (AP). Putting all together, equation 2 is reformulated as the following.

$$P(S_f) \propto \prod_{i=1}^n AP(C_i) \cdot P(M_{C_i,f}) \quad (26)$$

In order to determine $P(M_{C_i,f})$ we apply C_i on f which predicts the posterior probability that f belongs to category i ($P(i | f)$). For the cases that f does not belong to category i , $P(M_{C_i,f})$ is equal to $P(i | f)$. Otherwise it is equal to $1 - P(i | f)$.

In summary, to find stop-frames within training videos, we determine $P(S_f)$ for all video frames according to equation 3. In this equation, stop-frames are identified as the frames commonly misclassified by semantic classifiers. After determining $P(S_f)$ we remove the frames with the highest probabilities. The number of frames which should be removed is a parameter

that represents the number of stop-frames in the data set. Overestimating or underestimating this parameter degrades the classification accuracy. Therefore, the parameter is estimated by maximizing the classifier accuracy on validation data.

B.3 EXPERIMENTAL SETUP

B.3.1 Data Sets

We evaluate our stop-frame remover on two web video collections: the 2011 TRECVID Multimedia Event Detection corpus [139] and the YouTube 22 concepts from Ulges *et al.* [157].

TRECVID’s 2011 multimedia event detection corpus is a large publicly available collection of web videos. It contains 38,387 web videos, totaling 1,299 hours, collected from various web videos hosting sites. This dataset consists of training and test samples for 15 events. Each event is a complex activity occurring in unconstrained conditions. The dataset is released in three parts: event kit, development and opaque collections. Event kit contains the positive exemplar videos for all 15 events. In our experiments, we use it as the training set. For each event, we use its exemplar videos as the positive samples and the other 14 events’ videos as the negatives. The development collection includes the test data for five events. The test set for the other ten events are included in the opaque video collection. In our experiments, we use the development collection as the test set.

The YouTube 22 concepts dataset is prepared by the German Research Center for Artificial Intelligence [157]. It contains web videos for 22 visual concepts downloaded from YouTube. The concepts include activities (*e.g.* riot, sailing), objects (*e.g.* cat, helicopter), and scenes (*e.g.* desert, beach). For each concept 100 videos are downloaded totaling 194 hours. In our experiment, we divide the videos equally into training and test sets. For each concept, we use the videos from the other 21 concepts, as the negative examples.

B.3.2 Implementation Details

Video Representation: Each video is segmented into its shots, based on the significant changes in opponent color histograms within a window of 12 frames. For each shot, the middle frame in addition to i-frames distributed around it, are extracted as the key frames. Each key frame is represented by bag-of-words encoding of SIFT, Opponent-SIFT and RGB-SIFT descriptors extracted at Harris-Laplace keypoints and dense sampled points.

Video Classification: We employ SVM classifiers with fast histogram intersection kernel to learn concepts and event categories. The classification is performed at frame level, which means that classifiers are trained and tested on the extracted key frames. To arrive at a decision at video level, we use max pooling over all classified frames. To evaluate the classifiers accuracy, we use the average precision criterion, as a good combination of precision and recall [139].

Stop-frame Detection: We detect the stop-frames based on the algorithm proposed in Section B.2.1. As semantic classifiers C_i , we rely on the trained event and concept classifiers. In other words, for event and concept detection experiments we rely on the 15 and 22 event and concept classifiers, respectively, as C_i .

Figure 40.: Experiment 1. Using more classifiers results in a better identification of stop-frames and consequently an improved concept and event detection in web video. Good stop-frame identification is achieved by relying on the output of five (random) classifiers.

B.3.3 Experiments

We perform two experiments to evaluate the effectiveness of our method in improving the classification accuracy by removing the stop-frames. Each experiment is performed for both data sets.

Experiment 1: How many classifiers to use? In our method, stop-frames are detected by applying a number of semantic classifiers on frames. The question arises how many classifiers are needed? To answer this question, we compare the stop-frames detected by a varying number of classifiers. We start from one classifier and incrementally add more classifiers. Classifiers are selected randomly and for more robustness, the experiment is repeated for five different random selections of classifiers. To compare different cases, we exclude the stop-frames detected by each group of classifiers before training.

Experiment 2: Does stop-frames removal improve video classification? This experiment examines the effect of stop-frame removal in web video classification. We compare two cases: (i) a baseline, in which the classifier is trained on all key frames extracted from the training videos, and (ii) stop-frames removal, in which the stop frames are detected by our proposed algorithm and excluded before training the classifier.

B.4 RESULTS

Experiment 1: How many classifiers to use? We present the results of experiment 1 in Figure 40. As can be observed, the answer to the question how many classifiers to use is that in general using more classifiers to detect the stop-frames lead to a more confident decision. While using more classifiers is better in general, the results also indicate that more than five classifiers

Figure 41.: Experiment 2. Left figure: Classification results for the YouTube 22 Concepts data set. Right figure: Classification results for the TRECVID MED 2011 data set. Removing stop-frames improves the classification accuracy for almost all classes in both data sets.

does not significantly change the results. It demonstrates that we can efficiently remove the stop-frames with five classifiers.

Experiment 2: Does stop-frames removal improve video classification? We visualize the results of experiment 2 in Figure 41. Stop-frame removal improves the video classification accuracy. The mean average precision increases with 60% from 0.15 to 0.24, for TRECVID MED 2011, and with 24% from 0.41 to 0.51 for YouTube 22 Concepts. A considerable improvement given the simplicity of our approach.

Looking at the individual results, we find that our method improves accuracy for almost all classes but with varying degree. For the categories with poor classification accuracy, like the event *feeding an animal*, the improvement gained by removing stop-words is minor. In addition some categories in the YouTube 22 concepts data set, like *concert*, *crash*, *desert*, *explosion*, *helicopter*, *race* and *soccer* contain some key frames visually similar to the stop-frames. For example, the *concert* contains dark scenes with tiny flash spots in the stage. *Desert*, *soccer* and *helicopter* are dominated by frames showing a background of sands, green football field or the sky without any object. These informative frames are easily confused with stop-frames and excluded from training, which degrades the classifier accuracy. Figure 42 shows some samples of these frames.

For some concepts like *Eiffel tower*, *second life*, *swimming*, *video blog* and some events like *wood working* and *wedding ceremony* the improvements are substantial. Looking into their frames, we observe they contain more stop-frames in comparison with the other categories. In *video blog* and *wedding ceremony*, a lot of blank frames exist, which mostly emerge because of editing the videos. Also in *wood working*, videos are taken from the close distances from the objects so a lot of frames are completely occluded. In addition for *second life*, whose source videos contain only computer graphics, many blank frames occur without any informative clue regarding the video content. In summary, this experiment demonstrates that using our proposed algorithm to remove the stop-frames, the classifiers accuracies are improved for almost all semantic categories.

B.5 CONCLUSIONS

We identify stop-frames in web videos as those frames, which are uniformly distributed between videos from all categories and do not correlate with any particular semantic category (see Figure 39). We propose a simple algorithm to identify and remove stop-frames in web video. We identify stop-frames as the frames, which are commonly misclassified by multiple semantic

Figure 42.: Samples for the frames incorrectly determined as stop-frames.

classifiers. In our experiments, we demonstrate that by removing stop-frames we can easily improve classification accuracy by 60% and 24% in terms of mean average precision for event and concept detection benchmarks. Our proposed algorithm, as a pre-processing step, can be coupled with all the systems that use web videos as their training data.

VIDEO2SENTENCE AND VICE VERSA

c.1 INTRODUCTION

All of a sudden, video stories are everywhere. For sharing daily routines on YouTube, for private recollection of holidays, or for new professional markets. Anyone who must work with large amounts of video data is overwhelmed by its volume and the lack of tools to search and interpret the material easily. We aim to ease both retrieval and interpretation by translating a video to a sentence, similar to [27], and a sentence to video.

We consider semantic vocabularies for representing both video and human-provided sentences in a cross-media setting. Representing text and images by their semantic concepts has been shown to be beneficial for captioned-image retrieval [127]. At the same time, representing video by semantic concepts has been shown to be advantageous for video event recognition [47, 100, 104, 154]. In [104], for example, Merler *et al.* arrive at a robust high-level representation of video events using 280 concept detectors, which outperforms a low-level audiovisual representation. Our novelty is to demonstrate the capabilities of semantic representations for translating arbitrary web video into their lingual description, and vice versa. See Figures 43 and 44.

c.2 CONCEPT VOCABULARY

We demonstrate capabilities for translating arbitrary web video into a textual sentence and the other way around.

Data set We rely on the web video corpus from the TRECVID 2012 Multimedia Event Detection task [139]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by highlighting its dominant concepts [146]. In addition, we rely on the publicly available 1,346 concept detector scores for this dataset, which we provided in [47].

Multimedia Translation We map the sentences and video into a unified semantic space and then find their cross-media relations within this space. To map the texts and videos into the semantic space, we apply two sets of concept detectors which are trained so as to detect the semantic concepts on both videos and texts. Hence the textual detectors correspond to the labels of the visual detectors. The cross-media retrieval metric simply minimizes the distance between joint-vocabulary probability vectors [127]. We will show how a vocabulary of concept detectors can be exploited for effective translation at a meaningful level.

c.3 DEMONSTRATION

During our demonstration we will focus on two multimedia translation use cases:

Figure 43.: *Sentence2Video*: We automatically project an input sentence into a cross-media concept vocabulary, which is also used for representing video. The joint vocabulary allows us to translate a sentence to a video, and vice versa.

1. **Video2Sentence**: where we retrieve the sentences that best describe the visual content of the query video.
2. **Sentence2Video**: where we retrieve the videos that best illustrate the semantic content of the query sentence.

In addition, we will exhibit novel applications of this capability that aggregates the translated information for recounting. Taken together, the search engine provides a means to collect video examples illustrating a sentence and it provides a textual explanation of what happens in a video.

(a) User selects video to translate.

(b) System returns translations.

Figure 44.: Video2Sentence: Illustrative example for user interaction in our multimedia search engine. In (a) a user selects a candidate video for translation. In (b) the retrieved sentence translations are provided to the user together with the concept vocabulary of the query video. The tag cloud aggregates the sentences into a more precise translation.

SUMMARY

Storytelling Machines for Video Search

This thesis studies learning concept vocabularies for describing video content. Starting from manually specifying the vocabulary, in chapter 2, we gradually automate vocabulary construction by concept compositions, in chapter 3, and finally by learning from human stories, in chapter 4 and 5. We give a brief overview of each chapter as follows:

Chapter 2: Recommendations for Recognizing Video Events by Concept Vocabularies.

Representing videos using vocabularies composed of concept detectors appears promising for generic event recognition. While many have recently shown the benefits of concept vocabularies for recognition, studying the characteristics of a universal concept vocabulary suited for representing events is ignored. In this chapter, we study how to create an effective vocabulary for arbitrary-event recognition in web video. We consider five research questions related to the number, the type, the specificity, the quality and the normalization of the detectors in concept vocabularies. A rigorous experimental protocol using a pool of 1,346 concept detectors trained on publicly available annotations, two large arbitrary web video datasets and a common event recognition pipeline allow us to analyze the performance of various concept vocabulary definitions. From the analysis we arrive at the recommendation that for effective event recognition the concept vocabulary should *i)* contain more than 200 concepts, *ii)* be diverse by covering *object*, *action*, *scene*, *people*, *animal* and *attribute* concepts, *iii)* include both general and specific concepts, *iv)* increase the number of concepts rather than improve the quality of the individual detectors, and *v)* contain detectors that are appropriately normalized. We consider the recommendations for recognizing video events by concept vocabularies the most important contribution of the chapter, as they provide guidelines for future work.

Chapter 3: Composite Concept Discovery for Zero-Shot Video Event Detection. We consider automated detection of events in video without the use of any visual training examples. A common approach is to represent videos as classification scores obtained from a vocabulary of pre-trained concept classifiers. Where others construct the vocabulary by training individual concept classifiers, we propose to train classifiers for combination of concepts composed by Boolean logic operators. We call these concept combinations *composite concepts* and contribute an algorithm that automatically discovers them from existing video-level concept annotations. We discover composite concepts by jointly optimizing the accuracy of concept classifiers *and* their effectiveness for detecting events. We demonstrate that by combining concepts into composite concepts, we can train more accurate classifiers for the concept vocabulary, which leads to improved zero-shot event detection. Moreover, we demonstrate that by using different logic operators, namely “AND”, “OR”, we discover different types of composite concepts, which are complementary for zero-shot event detection. We perform a search for 20 events in 41K web videos from two test sets of the challenging TRECVID Multimedia Event Detection 2013 corpus. The experiments demonstrate the superior performance of the discovered composite concepts, compared to present-day alternatives, for zero-shot event detection.

Chapter 4: Discovering Semantic Vocabularies for Cross-Media Retrieval. This chapter proposes a data-driven approach for cross-media retrieval by automatically learning its underlying semantic vocabulary. Different from the existing semantic vocabularies, which are manually pre-defined and annotated, we automatically discover the vocabulary concepts and their annotations

from multimedia collections. To this end, we apply a probabilistic topic model on the text available in the collection to extract its semantic structure. Moreover, we propose a learning to rank framework, to effectively learn the concept classifiers from the extracted annotations. We evaluate the discovered semantic vocabulary for cross-media retrieval on three datasets of image/text and video/text pairs. Our experiments demonstrate that the discovered vocabulary does not require *any* manual labeling to outperform three recent alternatives for cross-media retrieval.

Chapter 5: VideoStory Embeddings Recognize Events when Examples are Scarce. This chapter aims for event recognition when video examples are scarce or even completely absent. The key in such a challenging setting is a semantic video representation. Rather than building the representation from individual attribute detectors and their annotations, we propose to learn the entire representation from freely available web videos and their descriptions using an embedding between video features and term vectors. In our proposed embedding, which we call *VideoStory*, the correlations between the terms are utilized to learn a more effective representation by optimizing a joint objective balancing descriptiveness and predictability. We show how learning the VideoStory using a multimodal predictability loss, including appearance, motion and audio features, results in a better predictable representation. We also propose a variant of VideoStory to recognize an event in video from just the important terms in a text query by introducing a term sensitive descriptiveness loss. Our experiments on three challenging collections of web videos from the NIST TRECVID Multimedia Event Detection and Columbia Consumer Videos datasets demonstrate: i) the advantages of VideoStory over representations using attributes or alternative embeddings, ii) the benefit of fusing video modalities by an embedding over common strategies, iii) the complementarity of term sensitive descriptiveness and multimodal predictability for event recognition without examples. By its abilities to improve predictability upon any underlying video feature while at the same time maximizing semantic descriptiveness, VideoStory leads to state-of-the-art accuracy for both few- and zero-example recognition of events in video.

LIST OF PUBLICATIONS

- A. Habibian, T. Mensink, and C. G. M. Snoek. VideoStory Embeddings Recognize Events when Examples are Scarce. Pending revision *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- A. Habibian, T. Mensink, and C. G. M. Snoek. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proc. Annual Conf. of the Advanced School for Comput. and Imag.*, 2015, **(Best paper award)**.
- A. Habibian, T. Mensink, and C. G. M. Snoek. Discovering Semantic Vocabularies for Cross-Media Retrieval. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2015.
- M. Mazloom, A. Habibian, D. Liu, C. G. M. Snoek, and S. F. Chang. Encoding Concept Prototypes for Video Event Detection and Summarization. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2015.
- A. Habibian, and C. G. M. Snoek. Vocabularies for Video Event Detection. In *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2015.
- A. Habibian, T. Mensink, and C. G. M. Snoek. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proc. ACM Int'l. Conf. Multimedia*, 2014, **(Best paper award)**.
- A. Habibian, T. Mensink, and C. G. M. Snoek. Composite Concept Discovery for Zero-Shot Video Event Detection. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2014.
- A. Habibian, and C. G. M. Snoek. Stop-Frame Removal Improves Web Video Classification. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2014.
- A. Habibian, M. Mazloom, and C. G. M. Snoek. On-the-Fly Video Event Search by Semantic Signatures. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2014.
- C. G. M. Snoek, K. E. A. van de Sande, D. Fontijne, S. Cappallo, J. van Gemert, A. Habibian, T. Mensink, P. Mettes, R. Tao, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2014 Semantic Video Search Engine. In *the 12th TRECVID Workshop*, 2014.
- G. K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. E. A. van de Sande, A. W. M. Smeulders, and C. G. M. Snoek. Evaluating Multimedia Features and Fusion for Example-based Event Detection. In *Machine Vision and Applications*, 2014.
- A. Habibian, and C. G. M. Snoek. Recommendations for Recognizing Video Events by Concept Vocabularies. *Comput. Vision and Image Understanding*, 2014.
- C. G. M. Snoek, K. E. A. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S. Pintea, R. Tao, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2013 Semantic Video Search Engine. In *the 11th TRECVID Workshop*, 2013.
- A. Habibian, and C. G. M. Snoek. Video2Sentence and Vice Versa. In *Proc. ACM Int'l. Conf. Multimedia*, 2013.

List of Publications

- M. Mazloom, A. Habibian, and C. G. M. Snoek. Querying for Video Events by Semantic Signatures from Few Examples. In *Proc. ACM Int'l. Conf. Multimedia*, 2013.
- A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for Video Event Recognition using Concept Vocabularies. In *Proc. ACM Int'l. Conf. Multimedia Retrieval*, 2013.
- C. G. M. Snoek, K. E. A. van de Sande, A. Habibian, S. Kordumova, Z. Li, M. Mazloom, S. Pintea, R. Tao, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2012 Semantic Video Search Engine. In *the 10th TRECVID Workshop*, 2012.
- A. Habibian, and C. G. M. Snoek. Stop-Frame Removal Improves Web Video Classification. In *Proc. Annual Conf. of the Advanced School for Comput. and Imag.*, 2012, **(Best poster award)**.

BIBLIOGRAPHY

- [1] CMU sphinx open source toolkit for speech recognition. <http://cmusphinx.sourceforge.net/>. Accessed: 2015-08-03.
- [2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.
- [3] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *IEEE Trans. PAMI*, 2014.
- [4] T. Althoff, H. O. Song, and T. Darrell. Detection bank: An object detection based video representation for multimedia event recognition. In *ACM MM*, 2012.
- [5] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *European Conference on Information Retrieval*, 2008.
- [6] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. Multimedia*, 2002.
- [7] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 2011.
- [8] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *ICMR*, 2014.
- [9] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, 2003.
- [10] A. Berg, J. Deng, S. Satheesh, H. Su, and L. Fei-Fei. Imagenet large scale visual recognition challenge 2011.
- [11] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.
- [12] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *CVPR*, 2014.
- [13] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 2012.
- [14] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.
- [15] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.
- [16] N. Block. Mental pictures and cognitive science. *The Philosophical Review*, 1983.
- [17] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.
- [18] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.
- [19] W. A. Brant. *Mental Imagery and Creativity: Cognition, Observation and Realization*. 2013.
- [20] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015.
- [21] X. Chang, Y.-L. Yu, Y. Yang, and A. G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, 2015.
- [22] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014.
- [23] B. Cohen and G. L. Murphy. Models of concepts. *Cognitive Science*, 1984.
- [24] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans. PAMI*, 2014.
- [25] Y. Cui, D. Liu, J. Chen, and S.-F. Chang. Building a large concept bank for representing events in video. *arXiv:1403.7591*, 2014.
- [26] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013.
- [27] P. Das, R. K. Srihari, and J. J. Corso. Translating related words to videos and back through latent topics. In *WSDM*, 2013.

Bibliography

- [28] P. Das, C. Xu, R. Doell, and J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013.
- [29] A. Dehghan, H. Idrees, and M. Shah. Improving semantic concept detection through the dictionary of visually-distinct elements. In *CVPR*, 2014.
- [30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [31] J. Dodge et al. Detecting visual text. In *NAACL*, 2012.
- [32] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.
- [33] A. Farhadi and M. A. Sadeghi. Phrasal recognition. *IEEE Trans. PAMI*, 2013.
- [34] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM MM*, 2014.
- [35] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google’s image search. In *ICCV*, 2005.
- [36] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.
- [37] J. A. Fodor. *The language of thought*. 1975.
- [38] C. Fox. A stop list for general text. In *ACM SIGIR*, 1989.
- [39] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance kernel for video concept detection. In *ACM MM*, 2007.
- [40] S. Guadarrama et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.
- [41] A. Habibian, T. Mensink, and C. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.
- [42] A. Habibian, T. Mensink, and C. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 2014.
- [43] A. Habibian, T. Mensink, and C. Snoek. Discovering semantic vocabularies for cross-media retrieval. In *ICMR*, 2015.
- [44] A. Habibian and C. Snoek. Video2sentence and vice versa. In *ACM MM*, 2013.
- [45] A. Habibian and C. Snoek. Recommendations for recognizing video events by concept vocabularies. *Computer Vision and Image Understanding*, 2014.
- [46] A. Habibian and C. Snoek. Stop-frame removal improves web video classification. In *ICMR*, 2014.
- [47] A. Habibian, K. van de Sande, and C. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.
- [48] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2004.
- [49] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Trans. Multimedia*, 2007.
- [50] D. Herman. *Storytelling and the Sciences of Mind*. 2013.
- [51] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR*, 1999.
- [52] H. Hotelling. Relation between two sets of variates. *Biometrika*, 1936.
- [53] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2011.
- [54] B. Huet, T.-S. Chua, and A. Hauptmann. Large-scale multimedia data collections. *IEEE MultiMedia*, 2012.
- [55] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.
- [56] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 2005.
- [57] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.
- [58] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. PAMI*, 2012.
- [59] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.

- [60] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. PAMI*, 2013.
- [61] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.
- [62] L. Jiang, A. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, 2012.
- [63] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.
- [64] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.
- [65] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. In *ACM MM*, 2015.
- [66] Y.-G. Jiang. Super: Towards real-time event recognition in internet videos. In *ICMR*, 2012.
- [67] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, 2013.
- [68] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. Multimedia*, 2010.
- [69] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.
- [70] Y.-G. Jiang, X. Zeng, G. Ye, D. Ellis, S.-F. Chang, S. Bhattacharya, and M. Shah. Columbia-UCF TRECVID 2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *TRECVID*, 2010.
- [71] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, 2002.
- [72] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [73] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.
- [74] D. Klein and C. Manning. Accurate unlexicalized parsing. In *ACL*, 2003.
- [75] S. Kordumova, X. Li, and C. Snoek. Evaluating sources and strategies for learning video concepts from social media. In *CBMI*, 2013.
- [76] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [77] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *ACL*, 2013.
- [78] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.
- [79] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.
- [80] I. Laptev. On space-time interest points. *IJCV*, 2005.
- [81] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
- [82] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [83] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *ACM MM*, 2003.
- [84] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *IJCV*, 2010.
- [85] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.
- [86] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013.
- [87] X. Li, C. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Trans. Multimedia*, 2009.

Bibliography

- [88] X. Li, C. Snoek, M. Worring, and A. Smeulders. Harvesting social images for bi-concept search. *IEEE Trans. Multimedia*, 2012.
- [89] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt’s probabilistic outputs for support vector machines. *Machine Learning*, 2007.
- [90] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.
- [91] X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *ICMR*, 2011.
- [92] Z. Ma. *From concepts to events: A progressive process for multimedia content analysis*. PhD thesis, University of Trento, 2013.
- [93] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM MM*, 2012.
- [94] Z. Ma, Y. Yang, N. Sebe, and A. G. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. *IEEE Trans. PAMI*, 2014.
- [95] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Trans. Multimedia*, 2013.
- [96] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.
- [97] J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *FnTCGV*, 2012.
- [98] S. Maji, A. C. Berg, and J. Malik. Efficient classification for additive kernel SVMs. *IEEE Trans. PAMI*, 2013.
- [99] M. Mazloom, E. Gavves, and C. Snoek. Conceptlets: Selective semantics for classifying video events. *IEEE Trans. Multimedia*, 2014.
- [100] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.
- [101] M. Mazloom, A. Habibian, D. Liu, C. Snoek, and S.-F. Chang. Encoding concept prototypes for video event detection and summarization. In *ICMR*, 2015.
- [102] M. Mazloom, A. Habibian, and C. Snoek. Querying for video events by semantic signatures from few examples. In *ACM MM*, 2013.
- [103] A. K. McCallum. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
- [104] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia*, 14, 2012.
- [105] G. Miller et al. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.
- [106] G. K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. van de Sande, A. Smeulders, and C. Snoek. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 2014.
- [107] M. Nagel, T. Mensink, and C. Snoek. Event fisher vectors: Robust encoding visual diversity of visual streams. In *BMVC*, 2015.
- [108] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 2006.
- [109] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo. BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems. In *TRECVID*, 2013.
- [110] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.
- [111] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *CIVR*, 2006.
- [112] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.
- [113] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.
- [114] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. A. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *MVA*, 2014.
- [115] D. Oneata, M. Douze, J. Revaud, J. Schwenninger, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. McGuinness, S. Chen, N. O’Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, F. Basura, and T. Tuytelaars. AXES at trecvid 2012: KIS, INS, and MED. In *TRECVID*, 2012.

- [116] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.
- [117] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
- [118] P. Over et al. TRECVID 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2012.
- [119] P. Over, J. Fiscus, G. Sanders, et al. TRECVID 2013-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2013.
- [120] P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In *TRECVID*, 2007.
- [121] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.
- [122] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.
- [123] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. PAMI*, 2005.
- [124] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012.
- [125] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999.
- [126] D. Putthividhy, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010.
- [127] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.
- [128] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Trans. PAMI*, 2012.
- [129] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013.
- [130] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2014.
- [131] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.
- [132] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3), 2013.
- [133] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012.
- [134] A. Scherp, R. Jain, M. Kankanhalli, and V. Mezaris. Modeling, detecting, and processing events in multimedia. In *ACM MM*.
- [135] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 2011.
- [136] J. M. Shipley and T. F. Zack, editors. *Understanding Events*. Oxford Series in Visual Cognition. 2008.
- [137] T. F. Shipley and J. Zacks. *Understanding events: From perception to action*. 2008.
- [138] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [139] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *ACM MIR*, 2006.
- [140] J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *ICME*, 2003.
- [141] C. Snoek and A. Smeulders. Visual-concept search solved? *Computer*, 2010.
- [142] C. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2008.
- [143] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005.
- [144] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

Bibliography

- [145] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, 2013.
- [146] S. Strassel et al. Creating havic: Heterogeneous audio visual internet collection. In *LREC*, 2012.
- [147] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, 2014.
- [148] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [149] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.
- [150] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.
- [151] K. Tang, B. Yao, L. Fei-Fei, and D. Koller. Combining the right features for complex event recognition. In *ICCV*, 2013.
- [152] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010.
- [153] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.
- [154] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *ICMR*, 2011.
- [155] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, 2005.
- [156] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *CIVR*, 2008.
- [157] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. A system that learns to tag videos by watching youtube. In *ICVS*, 2008.
- [158] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning automatic concept detectors from online video. *Computer Vision and Image Understanding*, 2010.
- [159] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. PAMI*, 2010.
- [160] J. Varadarajan, R. Emonet, and J. Odobez. Bridging the past, present and future: Modeling scene activities from event relationships and global rules. In *CVPR*, 2012.
- [161] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL*, 2015.
- [162] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.
- [163] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *IJCV*, 2015.
- [164] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.
- [165] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang. Automatic video annotation by semi-supervised learning with kernel density estimation. In *ACM MM*, 2006.
- [166] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing. In *ACM SIGIR*, 2011.
- [167] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *ACM MM*, 2014.
- [168] K. Q. Weinberger and O. Chapelle. Large margin taxonomy embedding for document categorization. In *NIPS*, 2009.
- [169] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.
- [170] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. Cross-media semantic representation via bi-directional learning to rank. In *ACM MM*, 2013.
- [171] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014.
- [172] L. Xie, H. Sundaram, and M. Campbell. Event mining in multimedia streams. *Proceedings of the IEEE*, 2008.

- [173] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2015.
- [174] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015.
- [175] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *CVPR*, 2015.
- [176] Y. Yan, Y. Yang, H. Shen, D. Meng, G. Liu, A. Hauptmann, and N. Sebe. Complex event detection via event oriented dictionary learning. In *AAAI*, 2015.
- [177] J. Yang and A. Hauptmann. (un) reliability of video concept detection. In *CIVR*, 2008.
- [178] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *ECCV*, 2012.
- [179] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*, 2015.
- [180] E. Younessian, T. Mitamura, and A. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ICMR*, 2012.
- [181] S.-I. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, et al. Informedia@ TRECVID 2014 MED and MER. In *TRECVID*, 2014.
- [182] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua. Learning concept bundles for video search with complex queries. In *ACM MM*, 2011.
- [183] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.
- [184] J. Zhou, G. Ding, and Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In *ACM SIGIR*, 2014.
- [185] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *ICIP*, 1998.

Verhalende Computers voor het Zoeken naar Video's

Dit proefschrift bestudeert het leren van het vocabulaire van concepten voor het beschrijven van de inhoud van video's. Beginnend met het handmatig specificeren van het vocabulaire van concepten, in hoofdstuk 2, gaan we het vocabulaire geleidelijk automatiseren bepalen met behulp van samengestelde concepten, in hoofdstuk 3, en ten slotte leren we de set van concepten van geschreven verhalen, in hoofdstuk 4 en 5. Voordat we de onderzoeksvragen nogmaals beschrijven geven we een beknopt overzicht van de hoofdstukken :

Hoofdstuk 2: Aanbevelingen voor het Herkennen van Video Gebeurtenissen met behulp van een Vocabulaire van Concepten. Video's beschrijven met een vocabulaire van concept detectoren lijkt een veelbelovende aanpak voor het herkennen van algemene gebeurtenissen. Terwijl velen onlangs de voordelen van een vocabulaire van concepten voor herkenning hebben aangetoond, is het bestuderen van de kenmerken van een universeel vocabulaire van concepten, geschikt voor het representeren van gebeurtenissen, genegeerd. In dit hoofdstuk onderzoeken we hoe een effectief vocabulaire van concepten voor het herkennen van willekeurige gebeurtenissen in webvideo's te creëren.

We beschouwen vijf onderzoeksvragen gerelateerd aan het aantal, het type, de specificiteit, de kwaliteit en de normalisering van de detectoren in het vocabulaire van concepten. Een zeer gedegen experimenteel protocol, gebruik makend van een set van 1,346 concept detectors getraind op openbaar beschikbare annotaties, twee grote webvideo datasets met willekeurige gebeurtenissen en een gemeenschappelijke gebeurtenis classificatiesysteem stellen ons in staat om de prestaties van de verschillende definities van vocabulaires van concepten te analyseren. Uit de analyse distilleren we de aanbeveling dat voor effectieve herkenning van gebeurtenissen, het vocabulaire van concepten zou *i)* meer dan 200 concepten moeten bevatten, *ii)* divers moeten zijn door het beschrijven van voorwerpen, acties, scènes, mensen, dieren en attribuu concepten, *iii)* zowel algemene als specifieke concepten moeten bevatten, *iv)* eerder meer concepten moeten beslaan dan het verbeteren van de kwaliteit van de individuele detectoren en *v)* moeten bestaan uit detectoren die goed genormaliseerd zijn. Wij beschouwen deze aanbevelingen voor het herkennen van video gebeurtenissen met behulp van een vocabulaire van concepten als de belangrijkste bijdrage van dit hoofdstuk, omdat zij richtlijnen voor de toekomstig onderzoek bevatten.

Hoofdstuk 3: Achterhalen van Samengestelde Concepten voor het Herkennen van Video Gebeurtenissen zonder Voorbeelden. We beschouwen het automatische herkennen van gebeurtenissen in video's zonder gebruik te maken van visuele trainingsvoorbeelden. Een gebruikelijke aanpak is om video's te representeren met de classificaties scores verkregen uit een vocabulaire van voor getrainde conceptdetectoren. Waar anderen het vocabulaire verkrijgen door individuele concept detectoren te gebruiken, stellen wij voor om detectoren te trainen voor combinaties van concepten op basis van Booleaanse logische operatoren. We noemen deze combinaties samengestelde concepten en stellen een algoritme voor dat deze automatisch ontdekt uit bestaande concept annotaties per video. We ontdekken samengestelde concepten door het gezamenlijk optimaliseren van de juistheid van de concept detectoren en hun effectiviteit voor het detecteren van gebeurtenissen. We tonen aan dat door het combineren van concepten tot samengestelde concepten, we betere detectoren kunnen trainen voor de concepten in het vocabulaire, wat leidt tot verbeterde detectie van gebeurtenissen zonder voorbeelden. Bovendien tonen we aan dat verschillende logische operatoren, namelijk "EN", "OF", verschillende soorten samengestelde concepten ontdekken, die complementair zijn voor het herkennen van gebeurtenissen zonder voorbeelden. We zoeken naar 20 gebeurtenissen in 41K internetvideo's komend van twee test-sets van de uitdagende TRECVID Multimedia Event Detection 2013 dataset. Deze experimenten tonen de verbeterde nauwkeurigheid van de ontdekte samengestelde concepten tegenover hedendaagse alternatieven voor het herkennen van gebeurtenissen zonder voorbeelden.

Hoofdstuk 4: Achterhalen van Semantische Vocabulaires voor Media Overschrijdend Zoeken. Dit hoofdstuk beschrijft een data-gedreven aanpak voor media overschrijdend zoeken door het automatisch leren van een onderliggende semantisch vocabulaire. Anders dan de bestaande semantische vocabulaires,

die vooraf handmatig worden gedefinieerd en geannoteerd, ontdekken wij automatisch het vocabulaire en de annotaties uit multimedia verzamelingen. Hiervoor laten wij een probabilistisch thema model los op de tekst in de verzameling om de semantische structuur te extraheren. Daarnaast maken we gebruik van een leren-om-te-ordenen model om de concept detectoren effectief te leren van de gevonden annotaties. We evalueren de gevonden semantische vocabulaires voor media overschrijdend zoeken op drie datasets bestaande uit beeld/tekst en video/tekst paren. Onze experimenten tonen aan dat de gevonden vocabulaires geen handmatige annotaties vereisen en de nauwkeuriger zijn dan drie recente alternatieven voor media overschrijdend zoeken.

Hoofdstuk 5: VideoStory Inbeddingen Herkennen Gebeurtenissen wanneer Voorbeelden Schaars zijn. Dit hoofdstuk bestudeert de herkenning van gebeurtenissen wanneer videovoorbeelden schaars zijn of zelfs geheel afwezig. De sleutel in zo'n uitdagend scenario is een semantische videorepresentatie. In plaats van een representatie gebaseerd op individuele attributen detectoren en hun annotaties, stellen wij voor om een representatie te leren van vrij beschikbare webvideo's en hun beschrijvingen met behulp van een inbeddingsfunctie tussen video kenmerken en term vectoren. In de door ons voorgestelde inbedding, die we VideoStory noemen, zijn de correlaties tussen de termen gebruikt om een betere representatie te leren door het optimaliseren van een gemeenschappelijke doelfunctie, balancerend tussen beschrijfbaarheid en voorspelbaarheid. We laten zien dat VideoStory geleerd met een multimodale doelfunctie, gebruik makend van uiterlijke-, bewegings- en audio-kenmerken, nog een betere representatie leert. We stellen ook een variant van VideoStory voor die gebeurtenissen herkent van alleen de belangrijkste termen uit een tekstuele zoekopdracht door gebruik te maken van een per term gewogen doelfunctie. Onze experimenten op drie uitdagende collecties van webvideo's van de NIST TRECVID Multimedia Event Detection en Columbia Consumer Video datasets tonen: *i)* de voordelen van VideoStory ten opzichte van representaties gebaseerd op attributen of alternatieve inbeddingfuncties, *ii)* het voordeel van het fuseren van video modaliteiten door een inbeddingsfunctie ten opzichte van gebruikelijke strategieën, *iii)* de aanvullende waarde van zowel de per term gewogen doelfunctie en de multimodale doelfunctie voor het herkennen van gebeurtenissen zonder voorbeelden. Met haar mogelijkheid om de voorspelbaarheid te verbeteren van enige onderliggende video karakteristieken, en op hetzelfde moment de semantische beschrijfbaarheid te maximaliseren, leidt VideoStory tot de best bekende nauwkeurigheid voor het herkennen van gebeurtenissen in video's wanneer zowel slechts een aantal voorbeelden of zelfs geen enkel voorbeeld beschikbaar is.

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Cees Snoek, you have been a tremendous supervisor and friend for me. Thank you for the continuous support of my research, for your patience, motivation, and your great guidance. Your guidance helped me in all the time of research and writing of this thesis.

I would also like to thank Thomas Mensink, you have taught me scientific and critical thinking skills. You have successfully challenged me to go beyond simply scratching the surface, and I wish to thank you for helping me to grow as a research scientist. I am grateful to my promotor Arnold Smeulders, for all your insightful advices and discussions during my study. Your scientific mindset and influential personality have been always inspiring to me.

An special thanks goes to Dennis Koelma, you have been supporting me throughout my research technically and scientifically. You have been always open to my questions and enthusiastic to share your ideas and insights with me and all the students. Furthermore, my sincere thanks also goes to Masoud Mazloom, you have been a great labmate and a close friend for me. I am thankful to you and all other colleagues for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years.

Lastly, and most importantly, I wish to thank my family. My parents, who bore me, raised me, supported me, taught me, and loved me. My beloved wife who spent sleepless nights with me and was always my support in the moments when there was no one to answer my queries. To them I dedicate this thesis.