UvA-DARE (Digital Academic Repository)

In search of video event semantics

Mazloom, M.

Publication date

2016

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Mazloom, M. (2016). *In search of video event semantics*. [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.

In Search of Video Event Semantics

In Search of Video Event Semantics

Masoud Mazloom

ISBN: 978-94-6182-717-3

Masoud Mazloom

In Search of Video Event Semantics

Masoud Mazloom

This book was typeset by the author using latex.

Printing: Off Page, Amsterdam

Copyright © 2016 by M. Mazloom.

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 978-94-6182-717-3

IN SEARCH OF VIDEO EVENT SEMANTICS

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 12:00 uur

door

Masoud Mazloom

geboren te Neishabour, Iran

Promotiecommissie

Promotor:	Prof. dr. 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. A. Hanjalic	TU Delft
	Prof. dr. A. G. Hauptmann	Carnegie Mellon University
	Prof. dr. M. Worring	Universiteit van Amsterdam
	dr. T. E. J. Mensink	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 358, at the Intelligent Systems Lab Amsterdam of the University of Amsterdam. This research is partially supported by the STW STORY project and the Dutch national program COMMIT/.

UNIVERSITEIT VAN AMSTERDAM

To My Son: Sina

CONTENTS

1	INTRODUCTION	9
1.1	List of publications	12
2	CONCEPTLET: SELECTIVE SEMANTICS FOR CLASSIFYING VIDEO EVENTS	15
2.1	Introduction	15
2.2	Related work	17
2.2.1	Concept Selection by Multimedia Retrieval	17
2.2.2	Concept Selection by Feature Selection	17
2.2.3	Contribution	19
2.3	Conceptlet	19
2.3.1	Preliminary	19
2.3.2	Cross-Entropy Formulation	20
2.3.3	Algorithm	21
2.4	Experimental setup	22
2.4.1	Data set	22
2.4.2	Implementation details	24
2.4.3	Experiments	25
2.5	Results	27
2.5.1	Influence of individual concepts	27
2.5.2	Influence of concept bank size	27
2.5.3	Conceptlets versus all concepts	29
2.5.4	Conceptlets versus other selections	33
2.6	Conclusions	35
3	ENCODING CONCEPT PROTOTYPES FOR VIDEO EVENT DETECTION AND SUMMARIZATION	37
3.1	Introduction	37
3.2	Encoding Concept Prototypes	40
3.2.1	Concept prototype learning	40
3.2.2	Video encoding	41
3.2.3	Event detection and summarization	43
3.3	Experimental Setup	44
3.3.1	Datasets	44
3.3.2	Implementation details	45
3.3.3	Experiments	45
3.4	Results	47
3.4.1	Few-example event detection	47
3.4.2	Zero-example event detection	48
3.4.3	Unsupervised event summarization	48
3.5	Conclusion	49
4	QUERYING FOR VIDEO EVENTS BY SEMANTIC SIGNATURES FROM FEW EXAMPLES	51
4.1	Introduction	51
4.2	Semantic Query Fusion	51
4.2.1	Semantic Signatures	51

Contents

4.2.2	Similarity Metrics	52
4.2.3	Query Fusion	53
4.3	Experimental Setup	53
4.3.1	Data Sets	53
4.3.2	Experiments	54
4.4	Results	56
4.5	Conclusions	59
5	TAGBOOK: A SEMANTIC VIDEO REPRESENTATION WITHOUT SUPERVISION FOR EVENT DETECTION	61
5.1	Introduction	61
5.2	Related Work	63
5.2.1	Representations from supervised concepts	63
5.2.2	Representations from weakly-supervised concepts	63
5.2.3	Representations from tags	64
5.2.4	Contributions	64
5.3	TagBook based video event detection	65
5.3.1	Problem Formalization	65
5.3.2	Two-Scenarios for Video Event Detection using TagBook	66
5.3.3	TagBook Construction by Content-Based Tag Propagation	67
5.4	Experiments	68
5.4.1	Datasets	68
5.4.2	Experiment 1: Finding a Good TagBook	70
5.4.3	Experiment 2: TagBook versus Others	72
5.5	Conclusions	75
6	SUMMARY AND CONCLUSION	77
6.1	Summary of In Search of Video Event Semantics	77
6.2	General conclusion	78
	Bibliography	85
	Samenvatting	87
	Acknowledgements	89

INTRODUCTION

Humans have the amazing ability to process and memorize events that happen during their lifetime instantaneously with near perfect recognition rates [21, 65]. The human brain is a memory machine that recollects episodes from an individual’s life, based on a combination of specific people, objects, and scenes experienced at a particular time and place [123]. When asked about events, which occurred in their lifetime, people remember and explain them by highlighting the semantic units of the events. For example for a wedding ceremony, they may mention the attendees, the wedding cake, the party location, etc. For memorizing and recollecting an event of interest, humans relate to the most relevant concepts of the event.

In our time, people not only memorize their own events, but they also share their events with others by posting them as videos and photos on social networks such as YouTube, Facebook, and Instagram. YouTube alone has more than one billion users each day. Millions of hours of video are uploaded and viewed. With such enormous collections of available events in the form of video content, manual processing and memorizing becomes prohibitive. It is therefore of great practical importance to build intelligent video search engines that automatically retrieve events as well as humans can perform that task.

This thesis contributes to event search in large multimedia collections using semantic video representations. Early work in video event search attempted to understand, to represent, and to model an event using low-level features derived from detectors for edges, corners, colors, and textures [41, 45, 85, 92, 115]. More recently, the low-level representation is learned from big data with deep convolutional neural networks [127, 129]. Such low-level representations are known to be effective in searching events when sufficient positive video examples are available to learn from. However, the applicability of the low-level representation is limited in the presence of only a few positive event video examples [68, 78], or when the only available information is a textual description of the event. Moreover, a low-level representation is unsuited to explain in human-understandable form what the most relevant concepts are for the event (nor has it the intention of doing so). We put forward that semantic representations are needed for events. This is leading to the fundamental question: How to represent events for video search?

The common tactic in event search using semantics, e.g. [1, 8, 9, 10], is to represent a video by the probability scores of a set of predefined concept detectors, see Figure 1. Then, an event detector is learned on top of the concept scores with the aid of event examples. The underlying concept detectors and their labeled examples are typically obtained from the TRECVID Semantic Indexing benchmark [110] and the ImageNet Large Scale Visual Recognition Challenge [7]. We follow the common convention in literature training concept detectors using visual features [51, 53, 64, 106, 121]. Then we apply concept detectors on unseen videos using their responses as the video representation for event detection [20, 28, 42, 81]. We note that in the common approach in the literature [20, 28, 42, 81], the representation contains all available concept responses, making it hard to pinpoint what concepts are most informative for each event under consideration. Rather than using all available concept detectors to represent a video, we ask the

Figure 1: Semantic representation for a video containing the event dog show. The weighted vector is obtained after applying concept detectors on the individual video frames, which are pooled over the entire video by the average operator. In this thesis we study semantic representation for video event search.

question:

What concepts matter for an event?

We address this question in Chapter 2, where we propose an algorithm that learns from examples what concepts in the video representation are most informative per event. We model finding the informative concepts from a large set of concept detectors as an importance-sampling problem. Our algorithm finds the optimal set of informative concepts using cross-entropy optimization. The selected representation is not only more discriminative than the state-of-the-art, but also makes more semantic sense than current solutions for events of interest, without being programmed to do so.

Having a selection of concepts in the video representations which are relevant to the event of interest is a necessary but not a sufficient condition for introducing a better understanding of the algorithm. We note that concept detectors trained from a collection of images have the appealing ability that they can be applied on each frame of a video. Consequently, we are able to follow the behavior of each concept in each video. However, the source of training data used in these concepts are still images, applying them on frames extracted from video results in an unreliable representation due to the domain mismatch between images and videos [14, 28, 81]. To counter the domain mismatch, video-based concept detectors are being proposed [8, 26, 27, 62, 69, 114], these works learn the concept detectors from video examples rather than image examples. All these methods compute a video representation by aggregating all its frames, whether relevant or irrelevant to the underlying concepts. The drawback of these works is that they fail to model concepts that only appear in a part of a video, leading to a sub-optimal representation. We therefore pose the question:

What concept frames matter for an event?

To address the question, we propose an algorithm in Chapter 3 that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations, and uses the obtained concept detectors for video representation. We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames from both positive and negative training videos of a target concept. Since the concept prototypes are a frame-level representation of concepts, we have the ability of mapping each frame of a video in a concept prototype space. This allows for event classification using less training examples and offers us to summarize an event video into its semantically most salient units.

Figure 2: Two videos depicting the event board trick. As their low-level visual properties are quite dissimilar, we study in this thesis whether a semantic representation is more suited to infer their similarity for video event search.

An important lesson learned from the first two research questions of the thesis, and also shown by concurrent work on few-example video event classification [11, 27, 68], is that the more semantic the video representation becomes, the less examples are needed to perform the event classification. Most of the works in few-example video event classification map each positive and negative video example of an event in a semantic space and then learn a model for ranking the videos at test time. Rather than learning a classifier from few-examples, we consider in Chapter 4 how semantic representations can be leveraged for event retrieval in a query-by-video example scenario. Leading to the question:

How to search events with concepts?

To address this question we propose a search engine for video events in Chapter 4. Different from the traditional query-by-visual-example paradigm [37, 71, 90, 97], which considers two videos similar as long as they exhibit identical patterns of color, texture, and shape, we compute the similarity of videos in a semantic space. We are inspired by the success of query-by-semantic-example in [99]. Such a semantic similarity seems more correlated with the measures of similarity adopted by humans for video comparison than similarity in low-level feature space. Figure 2 shows two videos that are similar for humans as they depict the event of a *board trick*, in spite if their different in low-level visual properties of color, shape, etc. By using the semantic representation, the retrieval operation is performed at a much higher level of abstraction. It can even generalize beyond the concepts it is build upon.

Despite the promise of semantic representations for video event search, obtaining representations remains a cumbersome process. Carefully labeled visual examples need to be collected, concept detectors need to be trained, and a carefully balanced semantic representation needs to be defined. The process can be simplified when the construction of individual concept detectors can be replaced by a rich semantic representation. We take inspiration from social image tagging, where many have shown the superiority of using model-free techniques instead of model-based methods for image tagging [5, 6, 23, 58]. Different from this work, which aims at assigning the most relevant tags to an image or video, we investigate in Chapter 5 how the tags themselves can be leveraged as a semantic video representation for event search. Leading to the question:

How to search events without concepts?

In answer to this question, we propose a new semantic video representation that is based on freely available social tagged videos only, without the need for training any intermediate concept detectors. We introduce a simple algorithm that propagates tags from a video's nearest neighbors, similar in spirit to the ones used for image retrieval [58], but redesign it for video event search.

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.

1.1 LIST OF PUBLICATIONS

The thesis is based on the following results:

- **Chapter 2** is based on “Searching informative concept banks for video event detection”, published in ACM International Conference on Multimedia Retrieval, 2013 [76], and the journal version of the paper “Conceptlets: Selective Semantics for Classifying Video Events”, published in IEEE Transactions on Multimedia, 2014 [75], by Masoud Mazloom, Efstratios Gavves, Koen van de Sande, and Cees Snoek.

Contribution of authors

Masoud Mazloom: all aspects

Efstratios Gavves: helped with designing the method

Koen van de Sande: helped with designing the method

Cees Snoek: supervision and insight

- **Chapter 3** is based on “Encoding concept prototypes for video event detection and summarization”, published in ACM International Conference on Multimedia Retrieval, 2015, by Masoud Mazloom, Amirhossein Habibian, Dong Liu, Cees Snoek, and Shih-Fu Chang [77].

Contribution of authors

Masoud Mazloom: all aspects

Amirhossein Habibian: helped with designing the method

Dong Liu: helped with designing the method

Cees Snoek: supervision and insight

Shih-Fu Chang: supervision and insight

- **Chapter 4** is based on “Querying for video events by semantic signatures from few examples”, published in ACM Conference on Multimedia, 2013 [78], and “On-the-fly video event search by semantic signatures”, published in ACM International Conference on Multimedia Retrieval, 2014 [25], by Masoud Mazloom, Amirhossein Habibian, and Cees Snoek.

Contribution of authors

Masoud Mazloom: all aspects

Amirhossein Habibian: helped with designing the method

Cees Snoek: supervision and insight

- **Chapter 5** is based on “Few-example video event retrieval using tag propagation”, published in ACM International Conference on Multimedia Retrieval, 2014 [79], and the

journal version of the paper “Tagbook: A semantic video representation without supervision for event detection”, published in IEEE Transactions on Multimedia, 2016 [80], by Masoud Mazloom, Xirong Li, and Cees Snoek.

Contribution of authors

Masoud Mazloom: all aspects

Xirong Li: helped with designing the method

Cees Snoek: supervision and insight

CONCEPTLET: SELECTIVE SEMANTICS FOR CLASSIFYING VIDEO EVENTS

2.1 INTRODUCTION

Automated understanding of events in unconstrained video has been a challenging problem in the multimedia community for decades [52]. This comes without surprise as providing access to events has great potential for many innovative applications [5, 42, 125]. Traditional classifiers represent an event by a carefully constructed explicit model [30, 35]. In [30], for example, Haering *et al.* propose a three-layer inference process to model events in wildlife video. In each layer event-specific knowledge is incorporated ranging from object-level motion, to domain-specific knowledge of wildlife hunting behavior. While effective for classifying hunting events, such a knowledge-intensive approach is unlikely to generalize to other problem domains. Hence, event representations based on explicit models are well suited for constrained domains like wildlife and railroad monitoring, but they are unable, nor intended, to generalize to a broad class of events in unconstrained video like the ones in Figure 3.

Recently, other event classification solutions have started to emerge. Inspired by the success of bag-of-word representations for object and scene recognition [43, 119], several papers in the literature exploit this low-level representation for event classification [33, 41, 45, 83, 85, 92, 93, 115]. In [45] Jiang *et al.* show that robust event classification accuracy is feasible by combining bag-of-words derived from SIFT descriptors, with bag-of-words derived from both MFCC audio features and space-time interest points. Their idea of combining multi-modal bag-of-words is further extended by Natarajan *et al.* [85] and Tamrakar *et al.* [115], who adhere to a more is better approach to event classification by exhaustively combining various visual descriptors, quantization methods, and word pooling strategies. In [41, 83] the robustness and efficiency of various low-level features for event classification are compared. In challenging benchmarks like TRECVID’s multimedia event detection task [118] and Columbia University’s Consumer Video dataset [44] the bag-of-words representation has proven its merit with respect to robustness and generalization, but from the sheer number of highly correlated descriptors and vector quantized words, it is not easy to derive how these detectors arrive at their event classification. Moreover, events are often characterized by similarity in semantics rather than appearance. In this chapter we attempt to find a video representation able to recognize, and ultimately describe, events in arbitrary content. We argue that to reach that long-term goal a more semantic representation than bag-of-words is urged for.

Inspired by the success of semantic concept detectors such as ‘Car’, ‘Animal’, and ‘Indoor’ for image retrieval [99], object recognition [55, 117], action recognition [103], and video retrieval [31, 113] several papers in the event classification literature exploit a bank of concept detector scores as the video representation [20, 22, 28, 62, 66, 69, 81, 130]. Ebadollahi *et al.*, for the first time, explored the use of semantic concepts for classifying events [20]. For creating their bank-of-

Figure 3: Example videos for the events Assembling a shelter Board trick, and Birthday. Despite the challenging diversity in visual appearance, each event maintains specific semantics in a consistent fashion. This chapter studies whether a selective and descriptive event representation based on concept detectors can be learned from video examples.

concepts, they employed the 39 detectors from the Large Scale Concept Ontology [84]. Each frame in their broadcast news video collection is then represented as a vector describing the likelihood of the 39 concept detectors. To arrive at an event classification score they employ a Hidden Markov Model. Due to the availability of large lexicons of concept annotations [17, 84], several others have recently also explored the utility of bank-of-concept representations for event classification [22, 28, 62, 81]. In [81] Merler *et al.* argue to use all available concept detectors for representing an event. Based on a video representation containing 280 concept detector scores, and a support vector machine for learning, the authors show that competitive event classification results can be obtained on the challenging internet video clips from the TRECVID 2010 Multimedia event detection collection. In [28] Habibian *et al.* arrive at a similar conclusion as [81] using a concept bank consisting of 1,346 concepts for event classification on a portion of the TRECVID 2012 Multimedia event detection collection. We note that in all these works [20, 28, 81] the resulting event detector operates on all concepts simultaneously, making it hard to pinpoint what concepts are most informative for each event under consideration.

Rather than using as many concepts as one can obtain, Liu *et al.* [62] show that by characterizing events using only a small set of carefully selected concepts, competitive results are feasible as well. It means that we do not necessarily need a large set of concept detectors to represent events. Rather than exploiting prior knowledge to manually specify a concept-subset for each event, we aim to learn the most informative concepts for an event from examples. We are inspired by the concept bank approach to event representation [20, 22, 28, 62, 69, 81, 130], so we start with a set of concept detectors as well. However, instead of using all available concepts, we attempt

to learn from examples for a given event what concepts are most informative to include in its concept bank, which we call the *conceptlet*. Before detailing the contributions of our work, we first discuss related work on concept selection that we consider most relevant to this paper.

2.2 RELATED WORK

In our survey of related work, we consider concept selection in the context of the multimedia retrieval literature and the feature selection literature.

2.2.1 Concept Selection by Multimedia Retrieval

Concept selection has been studied extensively in the video retrieval literature [32, 60, 86, 88, 101, 112, 122]. These selections automatically translate an input query into a weighted list of concepts which are then used for the retrieval. In [86] Natsev *et al.* consider text-based, visual-based and result-based selections. Using these three algorithms they find three rankings of concepts and use them for selection. In [112] Snoek *et al.* use text and visual analysis to select the single best concept for a query. Concepts are ranked according to their similarity to the query using the vector space model [105]. In [122] Wei *et al.* propose a semantic space to measure concept similarity and facilitate the selection of concept detectors based on the cosine similarity between the concepts and the query in the semantic space. Compared to [112], their approach combines detector scores from multiple selected concepts. Li *et al.* in [60] are inspired by tf-idf, which weights the importance of a detector according to its appearance frequency. In [99] Rasiwasia *et al.* rank concepts based on the scores the detectors obtained on the visual query images. In [101] Rudinac *et al.* make a ranking of concepts based on the frequency, variance, and kurtosis of concepts in the video queries. Using these three criteria, they select concepts. We observe that, in general, concept selection in multimedia retrieval, ranks a bank of concepts using text and video analysis and selects the single best or multiple concepts from the top of the obtained ranking.

All these existing selections evaluate the concept detectors individually and optimize a ranking of concepts per query. However, none of them considers the co-occurrence between the selected concepts. One can reasonably expect that for the events *feeding an animal* and *grooming an animal*, the concept ‘cat’ is important, but to differentiate the two events ‘cat’ has to co-occur with either ‘food’ or ‘bathtub’. Rather than evaluating concepts individually, we aim in this chapter to evaluate subsets of selected concepts simultaneously. We strive to select a near optimal concept-subset for each event category. We propose an algorithm that learns from examples what concepts in a bank are most informative per event.

2.2.2 Concept Selection by Feature Selection

A second perspective on concept selection considers it as feature selection, as common in the machine learning literature. Feature selection reduces data dimensionality by removing the irrelevant and redundant features using either unsupervised or supervised approaches. An example of unsupervised feature selection in the context of event classification by concept detectors is the work by Gkalelis *et al.* [22] who propose Mixture Subclass Discriminant Analysis to reduce a bank-of-concepts consisting of 231 detector scores to a subspace best describing an event. Since the algorithm alters the original concept representation it can no longer describe the semantics of the events of interest. Different from their work, we focus here on the problem of supervised feature selection, where the class labels, in our case event labels, are known beforehand.

Supervised feature selections are commonly classified into three categories, depending on their integration into the classifier [24, 104]: filters, embedders and wrappers.

Filters [63, 96], evaluate each feature separately with a measure such as mutual information or the correlation coefficient between features and class label. Hence, filters ignore dependencies between a set of features, which may lead to decreased classification performance when compared to other feature selections. Moreover, filters are usually computationally efficient and they produce a feature set which is not optimized for a specific type of classifier. Finally, filters provide a feature ranking rather than an explicit best feature subset, which demands a cut off point that needs to be set during cross validation. A strong filter is the Minimum Redundancy Maximum Relevancy proposed by Peng *et al.* [96], which uses mutual information and correlation between features for selection. When applied for selecting concepts for representing events, this method selects concepts that are mutually far away from each other while they still have high correlation to the event of interest. The feature selection computes a score for each concept based on the ratio of the relevancy of the concept to the redundancy of the concepts in the concept bank. Then it provides a concept ranking and removes the low scoring concepts. However, there may exist some concepts which are ranked low when considered individually but are still useful when considered in relationship with other concepts. In fact, Habibian *et al.* [28] presented an analysis that showed effective video event classification can be achieved, even when individual concept detector accuracies are modest, if sufficiently many concepts are combined. Hence, instead of selecting concepts by relying purely on detector scores, as Minimum Redundancy Maximum Relevancy does, we prefer to be less sensitive to the performance of the concept detectors. If the presence of a concept, either an accurate or inaccurate one, improves the accuracy of an event classifier, we strive to maintain it.

Embedders [89, 116], consider feature selection within the classifier construction. Compared to filters, the embedders can better account for correlations between features. State-of-the-art embedders are L1 norm SVM methods such as L1-Regularized Logistic Regression proposed by Ng [89]. During constructing of a linear classifier this embedder penalizes the regression coefficients and pushes many of them to zero. The features which have non-zero regression coefficients are selected as the informative features. The algorithm is most effective when there are many more redundant features than training examples. Furthermore, by definition of sparsity, one should expect and target for minimizing the number of non-zero elements in the solution vector. This condition is equivalent to employing the L0 norm for regularization. The L0 norm is accompanied by non smooth derivatives, which cannot be minimized in a gradient descent based setting. As an approximation, in [89] the L0 norm is replaced with the L1 norm. However, the L1 norm does not necessarily return the most optimal sparse solution. In this chapter we attempt to solve the L0 problem directly and obtain a truly sparse, optimal solution. Not by using filters or embedders, but by a wrapper.

Wrappers [34, 108], search through the feature space and evaluate each feature subset by a classifier. To search the space of all feature subsets, a search algorithm is wrapped around the classification model. However, as the space of feature subsets grows exponentially with the number of features, heuristic methods are often used to conduct the search for an optimal subset. The advantages of wrappers are the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. Wrappers usually provide the proper feature set for that particular type of model. A common drawback of wrappers is that they have a higher risk of overfitting than other selections and are computationally more intensive. We demonstrate that the increased computation pays off for more accurate event classification. In our previous work [76], we propose a wrapper to find an informative concept-subset per event. For each keyframe in a video the most representative concepts are selected and eventually aggregated to video level. While effective, the algorithm cannot determine the optimal bank size and its

iterative procedure on frame-level is computationally demanding. In this chapter we address these two drawbacks. Inspired by the *wrappers* and our previous work [76], we attempt to find what concepts in a bank are most informative per event,

2.2.3 Contribution

We make three contributions in this chapter. First, we model selecting the conceptlet out of a large set of concept detectors as an importance sampling simulation. Second, we propose an approximate solution that finds the near optimal conceptlet using a cross-entropy optimization. Third, we show qualitatively that the found conceptlets make sense for the events of interest, without being programmed to do so. To the best of our knowledge no method currently exists in the literature able to determine the most informative concepts for video event classification, other than our initial version of this work [76]. Note especially the algorithmic difference with concept selection by multimedia retrieval [32, 86, 99, 101, 112, 122]. In the multimedia retrieval scenario the selected detector score is exploited directly for search. In our approach, the conceptlet is optimized for *learning* to classify an event. We study the behavior of conceptlets by performing several experiments on more than 1,000 hours of arbitrary internet video from the TRECVID Multimedia Event Detection tasks 2010 [118], 2012 [118], and Columbia’s Consumer Video dataset [44]. But before we report our experimental validation, we first introduce our algorithm which learns from video examples the conceptlet for video event classification.

2.3 CONCEPTLET

Our goal is to arrive at an event representation containing informative concept detectors only, which we call a *conceptlet*. However, we first need to define what is informative. For example, one can reasonably expect that for the event *feeding an animal*, concepts such as ‘food’, ‘animal’ or ‘person’ should be more important to create a discriminative event model, and thus informative. We start from a large bank of concept detectors for representing events. Given a set of video exemplars of an event category, the aim is to find a (smaller) conceptlet that accurately describes this event. In this Section we describe our algorithm for selecting the conceptlet for each event category.

2.3.1 Preliminary

We first introduce some basic notation. Suppose we have a concept bank consisting of m concepts, $C = \{c_1, \dots, c_m\}$ and C^n represents a concept-subset with length n , where ($n \ll m$). Given a set of exemplar videos, a conceptlet C^* for an event is sampled from the space D^n of all possible concept-subsets with size n , that is $C^n \in D^n$, to best describe the video exemplars of the event. For a subset C^n , a concept is selected according to the probability density function $p(\alpha_i; \cdot)$. Here, α_i denotes the binary variable that corresponds to whether concept c_i was selected or not, which is $\alpha_i = \{0, 1\}$. We denote the parameter that controls the probability of $\alpha_i = 1$, with θ_i , that is $p(\alpha_i; \theta_i)$.

The number of different concept-subsets with length n out of m concepts, *i.e.*, $|D^n|$, is equal to $\binom{m}{n}$, which grows combinatorially with increasing m and decreasing n . Thus, the probability of finding an informative conceptlet becomes very small. This inspires us to model the problem of finding the rare conceptlet, *i.e.*, C^* , in the space D^n , as an importance sampling problem [10]. We use the cross-entropy [100], proven to yield robust results in a variety of estimation and optimization problems [57, 72, 76] without depending too much on parameters and their initializa-

tion. As the cross-entropy requires only a small number of parameters, chances of overfitting are minimized during the process of finding the informative concepts during training. Moreover, convergence is relatively fast and a near-optimum solution is guaranteed [100].

Suppose that the sample of a random subset C^n is drawn from space D^n using the probability density function p . Since, every concept will either be sampled, or not, we assume function p to follow a one-trial binomial distribution:

$$p(\alpha_i, \theta_i) = \theta_i^{\alpha_i} (1 - \theta_i)^{1 - \alpha_i} . \quad (2.1)$$

Moreover, assume that there is a neighborhood $\epsilon \subset D^n$ containing the concept-subsets with size n , accurately describing the video exemplars. Let l be the probability of sampling a concept-subset C^n from the ϵ neighborhood. Each of the of concept-subsets C^n has a limited capacity of accurately representing the video exemplars. Let $f(C^n)$ be the score function which measures the capacity that the concept-subsets C^n accurately represents the video exemplars. Suppose s is the lowest score of all concept-subsets in the neighborhood ϵ , according to the score function f , i.e., $s = \min f(C^n)$, $C^n \in \epsilon$. Then, the concept-subset with probability l will be the informative conceptlet C^n for which

$$l = P_\theta(f(C^n) \geq s) \quad (2.2)$$

We approximate this probability with the expectation:

$$l = E_\theta I(f(C^n) \geq s) , \quad (2.3)$$

where $I(f(C^n) \geq s)$, is an indicator function, referring to the set of concept-subset C^n for which the condition $f(C^n) \geq s$ holds. The straightforward way for estimating l is to use conventional sampling methods, such as crude Monte Carlo. Since the space of all possible concept-subsets is huge, estimating the probability l of a concept-subset C^n in ϵ using the density function p is impractical.

2.3.2 Cross-Entropy Formulation

An alternative way is based on importance sampling simulation. To illustrate, suppose a different probability density function h exists, which draws samples from neighborhood ϵ with high probability. Using h has the advantage of drawing more concept-subsets C^n from ϵ . Indeed, h is used as an importance sampling density function to estimate the expectation of l , denoted $\hat{l}$, using a likelihood ratio estimator. More precisely, for N concept-subsets C^n samples, $\hat{l}$ is equal to:

$$\hat{l} = \frac{1}{N} \sum_{r=1}^N I(f(C_r^n) \geq s) \frac{p(C_r^n; \theta)}{h(C_r^n)} , \quad (2.4)$$

where C_r^n denotes the r^{th} concept-subset with size n . The expectation $\hat{l}$ is then optimally estimated when the right side of Eq. 2.4 is equal to l , which means the value of expression inside sigma has to be equal to l , $I(f(C_r^n) \geq s) \cdot \frac{p(C_r^n; \theta)}{h(C_r^n)} = l$. For this reason the value of density function $h(C_r^n)$ has to be equal to:

$$h(C_r^n) = \frac{I(f(C_r^n) \geq s) p(C_r^n; \theta)}{l} . \quad (2.5)$$

Since Eq. 2.5 depends on the unknown quantity l , an analytical solution is impossible. Instead, the solution is to be found in an iterative approximation. Let us assume that there exists an optimal conceptlet C^* , controlled by the parameter vector θ^* .

Using C^* , the maximum score with respect to a specific video event classification accuracy is given by s^* . We denote this theoretical conceptlet state as $\langle C^*, \theta^*, s^* \rangle$ and all other estimated concept-subsets C^n as $\langle \hat{C}^n, \hat{\theta}, \hat{s} \rangle$. The goal is to find $\langle \hat{C}^*, \hat{\theta}^*, \hat{s}^* \rangle$ which best approximates

Table 1: The proposed algorithm which models finding a conceptlet for video event classification as a cross-entropy optimization.

INPUT:	Number of iterations (T), samples (N), size of concept bank (m), percentage of best performing concepts (ρN), index of events ($event$), labeled event examples, k different sizes of concept-subsets (n)= $\{n_1, n_2, \dots, n_k\}$ for finding the optimum conceptlet per event
OUTPUT:	Conceptlet per event
1.	for each $event$
2.	$Max = -1$
3.	for each $i = 1, \dots, k$
4.	Initialize $\Theta^{(0)} : \Theta^{(0)} = n(i) / m$
5.	for $t = 1, \dots, T$
6.	Sampling of concept-subset: Generate N samples $\{C_1^{n(i)(t)}, \dots, C_N^{n(i)(t)}\}$ by using current parameter $\Theta^{(t-1)}$.
7.	Adaptive updating of score s': Find the ρN samples that perform best given the score function $f(\cdot)$ and the labeled examples. Sort the samples in descending order by performance: $s_1 \geq \dots \geq s_{\rho N}$ and update $s' = s_{\rho N}$.
8.	Adaptive updating of parameter vector $\Theta^{(t)}$: Based on the best concept samples from step 7, update parameter set $\Theta^{(t)}$ by using Eq. 2.7
9.	end
10.	$C^* \leftarrow \Theta^{(T)}$
11.	if $f(C^*) \geq Max$
12.	$Max = f(C^*)$
13.	Conceptlet = C^*
14.	end
15.	end
16.	Return Conceptlet
17.	end

$\langle C^*, \theta^*, s^* \rangle$, i.e., the theoretical optimal conceptlet. In order to reach the goal state, $\langle C^*, \theta^*, s^* \rangle$, we generate multiple $\langle \hat{C}^n, \hat{\theta}, \hat{s} \rangle$ at each iteration. At each iteration, the concept-subsets C^n that perform best are used to update the search parameters θ . The iterations gradually converge to neighborhood ϵ with high probability. To guarantee convergence towards the goal state, the distance between p and h should be decreased after each iteration. This is achieved by adapting the importance sampling density function h via updating the parameters θ of the iteration's best performing subsets. A particularly convenient measure of distance between two densities h and p is the *Kullback-Leibler distance*, which is also termed the *cross-entropy* between h and p .

The cross-entropy is defined as:

$$D_{CE}(h, p) = \int h(x) \ln \frac{h(x)}{p(x)} dx. \quad (2.6)$$

Given that the sampling distributions p and h of concept-subsets follow a one-trial binomial distribution, the cross-entropy between the density function h and density function p is reduced for:

$$\hat{\theta}_i^t = \frac{1}{\rho N} \sum_{r=1}^N I(f(C_r^n) \geq \hat{s}_t) C_{r,i}^n, \quad (2.7)$$

where $\hat{\theta}_i^t$ denotes the probability of concept i in iteration t and $C_{r,i}^n = \{0, 1\}$ denotes the existence of concept i in the r^{th} concept-subset C^n . The parameter $\hat{\theta}_i^t$ directly shows the impact of concept i in video event classification at iteration t . Larger $\hat{\theta}_i^t$ makes the presence of concept i in the optimal solution more likely. The parameter ρN , $\rho \in (0, 1)$ defines the percentage of best performing concepts C^n scoring higher than s taken into account during each iteration.

2.3.3 Algorithm

The sample and search strategy for concept-subsets at iteration t has three steps:

- (1) **Sampling of concept-subsets C^n .** Based on the current parameter values θ^{t-1} , sample N concept-subsets C^n using $p(\cdot; \theta^{t-1})$ that is:

$$\{C_1^{n(t)}, \dots, C_N^{n(t)}\} \sim p(\cdot; \theta^{t-1}) . \quad (2.8)$$

- (2) **Adaptive updating of score s^t .** At iteration t , evaluate each of sample $C_j^{n(t)}$ using score function $f(\cdot)$ and find the ρN samples $C_j^{n(t)}$ that scored best on the $f(\cdot)$. After having sampled N concept-subsets and sorted them in descending order by performance: $s_{(1)} \geq \dots \geq s_{(N)}$, the smallest score value is used as the next iterations' reference score s^t , namely $\hat{s}^t = s_{[\rho N]}$. All samples $C_j^{n(t)}$ taken into account should perform at least as good as $\hat{s}^t$.

- (3) **Adaptive updating of parameter vector θ^t .** Given the ρN good performing samples $C_j^{n(t)}$ found in step 2, the updated parameter set θ^t is estimated as a function of the parameter vectors of these samples using Eq. 2.7. Informative concept-subsets are best captured by the concepts represented by high value of θ^t .

In the first step, we sample the concept-subsets C^n based on the parameters from iteration $t - 1$. The second step aims at keeping at each iteration the top performing concept-subsets C^n , sampled in the first step. Finally, the parameter vector θ^t is updated according to Eq. 2.7 in the third step, in a way that the distance $D_{CE}(h, p)$ is reduced. Updating parameters θ using Eq. 2.7 is equivalent to finding the frequency of a concept in the top performing concept-subsets C^n at iteration $t - 1$. It means that the probability of those concepts that together improve the event classification accuracy are increased after each iteration. Repeating these three steps for each iteration leads the search towards the conceptlet $\langle C^*, \theta^*, s^* \rangle$ in neighborhood ϵ . The selection process is illustrated in Figure 4.

In the above analysis, θ plays an important role. Since θ controls the binomial distribution p , the initial values of the parameter vector $\theta^{t=0}$ regulates the size of the conceptlet. Typically, these initial values are set to the same value for all concepts. Lower initial values lead to smaller concept-subsets. Moreover, due to the randomness of sampling, the exact size of the final conceptlet is not known a priori. As a result, other than uniformly setting θ to a constant for all concepts, different values can be assigned to favor a certain subset of the concepts. In general, different sources of prior knowledge will lead to different concept-subsets. Thus the initialization of θ will influence both the size and the informativeness of the resulting conceptlet.

For the purpose of event classification, the score function $f(\cdot)$ typically needs labeled training data to quantify the accuracy of various concept-subsets C^n . To do so, we split the training data into a training and validation set. An event classifier is then learned from the conceptlet in the training set and validated on the validation set. We use average precision to reflect the accuracy on the validation set. To find the optimum conceptlet per event, we also change the initialization value of θ by considering different values of n , the size of the concept-subsets. Our supervised selection algorithm for obtaining the conceptlets is summarized in Table 1.

2.4 EXPERIMENTAL SETUP

2.4.1 Data set

We investigate the effectiveness of conceptlets for video event classification by performing four experiments on three large datasets of challenging real-world web video for event classification: the TRECVID 2010 Multimedia Event Detection dataset [118], the partition of the TRECVID

Figure 4: Concept selection for an event using the values of the parameter vector θ at several iterations. At the beginning θ has a uniform low value, indicating that all concepts have the same low probability to be selected. After a few iterations, some concepts emerge as more probable for selection than others. After 20 iterations spikes are clearly visible, implying that the corresponding concepts are considered in the conceptlet.

2012 Multimedia Event Detection dataset [118] used in [28], and the Columbia Consumer Video dataset [44].

TRECVID MED 2010 [118] contains 3,465 internet video clips with over 115 hours of user generated videos content. The dataset contains ground truth for three event categories: *Assembling a shelter*, *Batting a run* and *Making a cake*. We train and evaluate our event classifiers on the train and test set that consist of 1,723 and 1,742 video examples respectively.

MediaMill MED 2012 [28] is a partition of the TRECVID 2012 Multimedia Event Detection dataset [118] defined by Habibiian *et al.* [28]. It consist of 1,500 hours of unconstrained videos provided in MPEG-4 format taken from the web with challenges such as high camera motions, different view points, large intra class variation and poor quality with varying resolution. The dataset comes with ground-truth annotations at video level for 25 real-world complex events, such as *Attempting a board trick*, *Flash mob gathering*, *Town hall meeting*, etc. Following the setup of [28], we extract two partitions consisting of 8,840 and 4,434 videos from the annotated part of the development set. In this chapter we use the first partition as the train set, on which we train our event classifiers, and we report all results on the second partition.

Table 2: Number of positive videos in the TRECVID 2010 MED, MediaMill 2012 MED, and Columbia CV datasets used in our experiments, split per event. The number of negative videos for each event are around 1,600, 8,800, and 4,500, respectively.

TRECVID MED 2010			MediaMill MED 2012			Columbia CV		
Event	Train	Test	Event	Train	Test	Event	Train	Test
<i>Assembling a shelter</i>	50	48	<i>Board trick</i>	98	49	<i>Basketball</i>	182	181
<i>Batting a run</i>	52	50	<i>Feeding animal</i>	75	48	<i>Baseball</i>	150	151
<i>Making a cake</i>	58	48	<i>Landing fish</i>	71	36	<i>Soccer</i>	161	162
			<i>Wedding ceremony</i>	69	35	<i>Ice skating</i>	192	193
			<i>Wood working</i>	79	40	<i>Skiing</i>	197	196
			<i>Birthday party</i>	121	61	<i>Swimming</i>	199	202
			<i>Changing vehicle tire</i>	75	37	<i>Biking</i>	136	137
			<i>Flash mob gathering</i>	115	58	<i>Graduation</i>	143	145
			<i>Getting vehicle unstuck</i>	85	43	<i>Birthday</i>	158	160
			<i>Grooming animal</i>	91	46	<i>Wedding reception</i>	129	130
			<i>Making sandwich</i>	83	42	<i>Wedding ceremony</i>	111	110
			<i>Parade</i>	105	50	<i>Wedding dance</i>	174	176
			<i>Parkour</i>	75	38	<i>Music performance</i>	403	403
			<i>Repairing appliance</i>	85	43	<i>Non-Music performance</i>	345	346
			<i>Working on sewing project</i>	86	43	<i>Parade</i>	191	194
			<i>Attempting bike trick</i>	43	22			
			<i>Cleaning an appliance</i>	43	22			
			<i>Dog show</i>	43	22			
			<i>Giving directions to location</i>	43	22			
			<i>Marriage proposal</i>	43	22			
			<i>Renovating home</i>	43	22			
			<i>Rock climbing</i>	43	22			
			<i>Town hall meeting</i>	43	22			
			<i>Winning race without vehicle</i>	43	22			
			<i>Working on metal crafts project</i>	43	22			

Columbia CV [44] consists of 9,317 user-generated YouTube videos with over 210 hours of content. The dataset is annotated with 20 semantic categories, where 15 of them are events, such as *Basketball*, *Ice skating*, *Birthday* etc. As we focus exclusively on events, the five object and scene categories in this dataset are excluded from our experiment. We use the split suggested by the authors which consists of 4,625 train videos and 4,637 test videos.

Table 2 summarizes the statistics of the training and test sets per event for the three video datasets. For a visual impression of characteristic event examples we refer to Figure 3 showing two examples for the events *Assembling a shelter*, *Board trick*, and *Birthday* in the TRECVID 2010 MED, the MediaMill 2012 MED, and the Columbia CV dataset.

2.4.2 Implementation details

Concept Bank In the TRECVID 2010 MED dataset we represent each video by a histogram of the output of 280 concept detectors, defined and provided by Merler *et al.* [81]. From the videos one frame is extracted every two seconds and represented as a histogram of 280 concept detectors scores. Then, the histograms are aggregated using average-pooling to arrive at a representation per video. For representing the videos in the MediaMill 2012 MED and Columbia CV datasets we use

a concept bank that consists of 1,346 concept detectors. The 1,346 concept detectors are trained using the training data for 346 concepts from the TRECVID 2012 Semantic Indexing task [110] and for 1,000 objects from the ImageNet Large Scale Visual Recognition Challenge 2011 [7]. Although some of the detector names overlap, we prefer to keep all 1,346 as their training data is different. The detectors are trained using a linear SVM atop a standard bag-of-words of densely sampled color SIFT [119] with Fisher vector coding [106] and spatial pyramids [53]. The 1,000 concepts from ImageNet are trained one versus all. The negative examples for each concept from the TRECVID 2012 Semantic Indexing task are the positive examples from other concepts and several examples without label. We compute concept detector scores per video frame, which are extracted once every two seconds. By concatenating and normalizing the detector outputs, each frame is represented by a concept score histogram of 1,346 elements. Finally the concept score histograms are aggregated into a video-level representation by average-pooling, which is known to be a stable choice for video classification [81].

Event classification As we focus on obtaining an informative representation for video event classification, we are for the moment less interested in the accuracy optimizations that may be obtained from various kernel settings [4, 19, 126]. Hence, we train for each event a one-versus-all linear support vector machine [107] and an approximated histogram intersection kernel map [120]. We find the optimal parameter settings using 5-fold cross-validation.

Cross entropy parameters After initial testing on small partitions of the data, we set the parameters of our algorithm to find the conceptlets for each event as follows: number of iterations $T = 20$, number of concept samples in each iteration $N = 1,000$, and a percentage of best performing concept samples $\rho = 0.1$, leaving 100 best performing concept samples per iteration for updating the sampling parameters. For finding the best conceptlet size per event, we consider various sizes of n , *i.e.*, the size of concept-subsets, during 5-fold cross-validation within the training data only.

Evaluation criteria For both the objective function $f(\cdot)$ in our conceptlet algorithm, as well as the final event classification evaluation, we consider as criterion the average precision (AP), which is a well known and popular measure in the video retrieval literature [110]. We also report the average performance over all events as the mean average precision (MAP).

2.4.3 Experiments

In order to establish the effectiveness of conceptlets for video event classification, we perform four experiments.

Experiment 1: Influence of individual concepts To evaluate the maximum effect of individual concept detectors on event classification accuracy, we perform an oracle experiment by simply evaluating each individual concept detector as if it was an event classifier. We evaluate all individual concepts on all events. Then we sort the list of concepts by their classification accuracy for each of the events in the three datasets.

Experiment 2: Influence of concept bank size To assess the effect of a growing number of concepts in a bank on video event classification performance, we randomly sample a concept-subset from our concept bank. For TRECVID 2010 MED we randomly select concepts from the concept bank with 280 concepts defined by Merler *et al.* [81] with a step size of 10. For both MediaMill 2012 MED and Columbia CV dataset we randomly select concepts from our 1,346 concept bank with a step size of 100. Each video in our dataset is then represented in terms of the detector scores from the concepts in this random subset. To cancel out the accidental effects of randomness, we repeat this procedure 20 times for each subset size.

Experiment 3: Conceptlets versus all concepts In this experiment we compare our proposed conceptlets to a bank based on all available concept detectors [28, 81]. As the baseline, we

Table 3: *Experiment 1. Influence of individual concepts on video event classification accuracy. We list the five best concepts for three events per dataset, together with the number of positive training examples used to train the concept detectors. Note the semantic correspondence between good performing concepts and events. Concepts in italics are also automatically selected by the conceptlet algorithm in experiment 3.*

Assembling a shelter			Batting a run			Making a cake		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
<i>Snow scene</i>	0.158	1,138	<i>Baseball cricket</i>	0.326	1,000	<i>Cake</i>	0.141	230
<i>Outdoors</i>	0.121	1,000	<i>Hockey</i>	0.214	1,998	<i>Food</i>	0.126	794
<i>Mountain scene</i>	0.105	3,972	<i>Diamond</i>	0.202	1,000	<i>Table desk</i>	0.097	1,000
<i>Forest</i>	0.103	12	<i>Running</i>	0.184	896	<i>Building</i>	0.053	2,354
<i>Water scene</i>	0.082	2,746	<i>Suit</i>	0.176	1,000	<i>Room</i>	0.052	2,90

Board trick			Wedding ceremony			Flash mob gathering		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
<i>Skating</i>	0.194	1,300	<i>Church</i>	0.396	1,300	<i>Crowd</i>	0.280	2,341
<i>Road</i>	0.171	1,096	<i>Altar</i>	0.324	1,300	<i>3 or more people</i>	0.214	2,099
<i>Snow</i>	0.162	1,013	<i>Gown</i>	0.306	1300	<i>People marching</i>	0.205	624
<i>Snowplow</i>	0.123	540	<i>Groom</i>	0.288	1,280	<i>Street battle</i>	0.202	1,300
<i>Ski</i>	0.119	1,096	<i>Suit</i>	0.251	1,300	<i>Meeting</i>	0.186	340

Basketball			Swimming			Parade		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
<i>Basketball</i>	0.488	1,300	<i>Swimming</i>	0.698	1,300	<i>People marching</i>	0.318	624
<i>Throw ball</i>	0.485	811	<i>Swimming pool</i>	0.621	1,300	<i>Urban scenes</i>	0.155	1,403
<i>Throwing</i>	0.432	1,300	<i>Underwater</i>	0.432	1,300	<i>Police van</i>	0.150	1,300
<i>Indoor sport venue</i>	0.355	1,300	<i>Stingray</i>	0.227	1,300	<i>3 or more people</i>	0.138	2,099
<i>Gym</i>	0.337	153	<i>Waterscape/Waterfront</i>	0.211	604	<i>Streets</i>	0.135	1,300

represent each video in TRECVID 2010 MED as a 280D vector of detector scores [81] and each video in MediaMill 2012 MED and Columbia CV datasets as a 1,346D vector of detector scores [28] (see Section 2.4.2). For finding the conceptlet per event, we apply the cross-entropy optimization as described in Section 2.3.3 on the training set only. To find the best conceptlet size, we vary parameter n . For events in the TRECVID 2010 MED we consider n in the range [10, 20, ..., 100]. In the MediaMill 2012 MED and Columbia CV datasets, we consider values of n in the range [10, 20, ..., 100, 200, 300, 400, 500]. We train an event detector on the found conceptlet and report its performance on the (unseen) test set.

Experiment 4: Conceptlets versus other selections In this experiment we compare conceptlets obtained with our cross-entropy algorithm to conceptlets obtained from state-of-the-art feature selection algorithms: Minimum Redundancy Maximum Relevancy [96] and L1-Regularized Logistic Regression [89]. To select the concepts per event by Minimum Redundancy Maximum Relevancy, at first we rank all concepts. Then we conduct a 5-fold cross validation on the training set with a varying number of selected concepts ranging from 10 to 1,000 with a step size of 10. In L1-Regularized Logistic Regression, since the regularization parameter controls the sparsity of concepts, we conduct a 5-fold cross validations on the training set by varying this parameter from 1 to 100 with step 5 to select the concepts per event. For both feature selections we train an event classifier with a linear SVM on the selected concepts and report its performance on the (unseen) test set.

2.5 RESULTS

2.5.1 Influence of individual concepts

We show the results of experiment 1 in Table 3. We observe that the best detectors per event also make sense, most of the time. When we consider the event *Wedding ceremony*, for example, the best possible concepts are ‘Church’, ‘Altar’, ‘Gown’, ‘Groom’ and ‘Suit’. For the event *Making a cake*, concepts like ‘Cake’, ‘Food’, ‘Table desk’, ‘Building’ and ‘Room’ are the oracle choice. However, for the event *Batting a run* we find an irrelevant concept in the top of the concept ranking: ‘Hockey’. We explain this by the fact that ‘Hockey’ shares many low-level visual characteristics with *Baseball e.g.*, both sports are played on a green field. It is also interesting to note that some of the relevant concept detectors obtain the good event classification accuracy by having only a few positive training examples, consider for example ‘Forest’ for the event *Assembling a shelter*, which has only 12 positive examples. This result shows that there are individual concepts that are more discriminative and descriptive than others for representing events in internet video.

2.5.2 Influence of concept bank size

We plot the results of experiment 2 on the three datasets in Figure 5. As expected the event classification accuracy increases when more and more concept detectors are part of the bank.

For the TRECVID 2010 MED dataset (Figure 5(a)), the increase in event classification accuracy is close to linear up to approximately 40 (random) concept detectors, afterwards it saturates to the end value of 0.361 MAP when using all 280 available concept detectors. Interestingly, the plot reveals that there exist an outlier concept-subset, containing only 70 concepts, which performs better than using all 280 concepts (compare the MAP of 0.389 with the maximum MAP of 0.361 when using all concepts). This result shows that some concept-subsets are more informative than others for video event classification. The results on the other two datasets confirm this conclusion. For the MediaMill 2012 MED dataset (Figure 5(c)), there is an outlier concept-subset, containing only 800 concepts, which performs better than using all 1,346 concepts (compare the MAP of 0.312 with the maximum MAP of 0.292 when using all concepts). Also for the Columbia CV dataset (Figure 5(e)), we find that there is an outlier concept-subset, containing only 600 concepts, which performs better than using all 1,346 concepts (compare the MAP of 0.531 with the maximum MAP of 0.507 when using all concepts). These results indicate much is to be expected from a priori search for the conceptlet of an event.

When we zoom in on individual events the connection between concept-subsets and event definitions can be studied. We inspect the box plot also for all the individual events of the three datasets (data not shown). The plots reveal several positive outliers using just a small number of concepts in the subset. Figures 5(b)(d)(f) detail the box plot for the specific events *Batting a run*, *Landing a fish*, and *Wedding ceremony*. For event *Batting a run* (Figure 5(b)) we perceive an outlier subset with an AP of 0.590 containing only 50 randomly selected concepts (compare to the maximum of 0.553 when using all 280 concepts). For event *Landing a fish* (Figure 5(d)) the box plot reveals that there exist a subset, containing only 400 concepts, which performs better than using all 1,346 concepts (compare the top of the whisker at 400 concepts, with an MAP of 0.489 with the maximum MAP of 0.433 when using all concepts). Also for the event *Wedding ceremony* (Figure 5(f)) we observe an outlier subset with an AP of 0.500 containing only 500 randomly selected concepts (compare to the maximum of 0.473 when using all 1,346 concepts). The results of experiment 2 on three datasets with two different concept banks show that, in general, the event classification accuracy increases with the number of semantic concepts in the

(a) TRECVID 2010 MED all 3 events

(b) Event: Batting a run

(c) MediaMill 2012 MED all 25 events

(d) Event: Landing a fish

(e) Columbia CV all 15 events

(f) Event: Wedding ceremony

Figure 5: Experiment 2. Influence of concept bank size in (a) TRECVID 2010 MED, (c) MediaMill 2012 MED, and (e) Columbia CV: Event classification accuracy increases with the number of concepts in the bank, but the variance suggests that some concept-subsets are more informative than others. (d), (b) and (f) Influence of concept bank size for the particular events: Landing a fish, Batting a run, and Wedding ceremony. For these events a small subset outperforms the bank using all available concepts. Indicating that much is to be expected from a priori search for the most informative conceptlet for an event.

bank. However, it also shows that some concept-subsets are more informative than others for specific events, and this may result in improved event classification accuracy.

Figure 6: Experiment 3. Conceptlets versus all concepts. A conceptlet outperforms a bank containing all available concept detectors for the large majority of event categories when using either a linear or a non-linear SVM for video event classification.

2.5.3 Conceptlets versus all concepts

We plot the result of experiment 3 in Figure 6. For the large majority of event categories, conceptlets with selective semantics are better than using all available concepts.

Figure 7: Conceptlets for various events as automatically selected from event video examples by our algorithm. Font size correlates with automatically estimated informativeness. Note that the algorithm finds concepts that make sense, most of the time, without being programmed to do so.

On the TRECVID 2010 MED dataset (Figure 6(a)), we achieve a 0.483 MAP in event classification by conceptlets, where the result is 0.361 / 0.421 MAP when using all 280 concepts [81]. Conceptlets obtain a relative improvement of 34.0% over the linear SVM and 14.7% over the non-linear SVM with only 83 concepts per event on average. We observe a considerable improvement for all three events using only a fraction of the available concept detectors (90, 70, and 90). Figure 7(a)(b) shows the conceptlets for the events *Batting a run* and *Making a cake*. Our algorithm selects concepts such as ‘Baseball’, ‘Cricket’, ‘Field’, ‘Running’, and ‘Sport’ that make sense for the event *Batting a run*, without being programmed to do so. However, the conceptlet also contains some irrelevant semantic concepts, such as ‘Hockey and ‘Soccer’, which share several visual characteristics to the event (see Figure 8). Similar conclusions hold for the event *Making a cake*.

On the MediaMill 2012 MED dataset (Figure 6(b)), our conceptlets reach to a 0.329 MAP in event classification, where using all 1,346 concepts results in 0.292 / 0.317 [28]. A relative improvement of 13.0% for the linear SVM and 3.7% for the non-linear SVM using about 245 concepts on average per event. Conceptlets obtain a considerable improvement for events such as *Landing fish*, *Dog show* and *Flash mob gathering* using only 300, 200, and 40 of the concept

Figure 8: Example images used for training the concept detectors: ‘Baseball’, ‘Hockey’, ‘Candle’ and ‘Abacus’ (top to bottom). The visual similarity between ‘Baseball’ and ‘Hockey’ causes that our algorithm mistakenly selects ‘Hockey’ into the conceptlet for the event *Batting in run*. Likewise, ‘Abacus’ results in a high probability in the videos containing Birthday events since the color beads of the abacus are visually similar to typical Birthday objects such as candles and balloons.

detectors available. When relevant concepts are unavailable in the concept bank we started with, the results will not improve much, as can be seen for the events *Attempting bike trick*, *Marriage proposal*, and *Making sandwich*, but often better than using all. Figure 7(c)(d) shows the conceptlet for *Landing a fish* and *Flash mob gathering*. The conceptlet for the event *Landing a fish* consist of general concepts such as ‘Adult male human’, ‘Hand’, ‘3-or-more-people’, ‘Sea-Mammal’ and event-specific concepts such as ‘Hook’ and ‘Reel’. The conceptlet for *Flash mob gathering* shows several concepts that seem semantically relevant as well, such as ‘Walking-running’, ‘Crowd’, ‘Daytime-outdoor’. However, we also observe some concepts whose semantic connection is less apparent, such as ‘Water-bottle’ and ‘Ground-combat’. Note that the concepts are selected automatically from provided event examples only.

On the Columbia CV dataset (Figure 6(c)), we observe that conceptlets obtain 0.625 MAP, where using all 1,346 concepts results in 0.507 / 0.565 MAP. Conceptlets are always better and obtain a relative improvement of 23.2% / 10.6% with only 93 concepts per event on average. Conceptlets obtain a considerable relative improvement for events such as *Soccer*, *Biking* and *Graduation* using only 50, 50, and 200 of the available concept detectors. Interestingly, for the event *Birthday* the improvement compared to the linear SVM is as much as 87.2% (0.524 MAP against 0.280 MAP) using only 30 concepts. Figure 7(e)(f) highlights the conceptlets for *Biking* and *Birthday*. We observe that most of the selected concepts for event *Biking* in Figure 7(e), such as ‘Bicycling’, ‘Bicycle’, ‘Daytime outdoor’, ‘Road’, ‘Legs’, are semantically relevant to this event. In Figure 7(f), we show the conceptlet for *Birthday*. Beside semantically relevant concepts such as ‘Candle’ we observe several semantically irrelevant concepts such as ‘Abacus’. When we

(a) MediaMill 2012 MED

(b) Columbia CV

Figure 9: Correlation between the selected concepts and their training source per event in (a) MediaMill 2012 MED and (b) Columbia CV. Conceptlets automatically select the most informative concepts independent of their training source.

inspect the ImageNet images used for training of this conceptlet detector in Figure 8, we observe that the color beads of the abacus are visually similar to typical *Birthday* objects such as candles and balloons. When the quality of concept detectors further improves, we expect better selection in the conceptlets.

To explore the correlation between the selected concepts and the source of these concepts when using both the TRECVID and ImageNet annotations, we plot the fraction of selected concepts by their training source for all 25 events of MediaMill MED 2012 and the 15 events of Columbia CV in Figure 9. As can be observed, conceptlets automatically select the most informative concepts independent of their training source.

Since conceptlets need event video training examples we also investigated how many event videos are sufficient for conceptlet discovery. On the TRECVID MED 2010 dataset we train

Figure 10: Effect of number of positive event examples on classification with conceptlets (red) or an all-concept baseline (blue). Conceptlets are always better, even when examples are scarce.

event classifiers using 10, 20, 30, and 40 positive examples, and repeat the process of selecting examples 10 times. We compare event classifiers using all concepts with a linear SVM [81] versus our conceptlets with a linear SVM. The result in Figure 10 shows that training size affects the performance, more examples are beneficial for both the baseline and conceptlets, but conceptlets are always better, even when examples are scarce.

For future reference we also evaluate our conceptlets on the TRECVID 2013 MED test and Kindred sets [118]. While low-level feature representations are known to be more accurate on these benchmarks [118], we focus in our comparison on semantic representations and consider the state-of-the-art approach by Habibian *et al.* [28]. We observe similar behavior as before. Event classification using all concepts with a linear SVM (0.270, 0.289) is outperformed by all concepts with a non-linear SVM (0.296, 0.309), but conceptlets are best (0.309, 0.325).

The results of experiment 3 affirm that event classification using conceptlets outperforms a bank using all concepts and always contains significantly less semantic concepts, which often appear descriptive.

2.5.4 Conceptlets versus other selections

We plot the result of experiment 4 in Figure 11. On all three datasets conceptlets outperform selections based on Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression.

On the TRECVID 2010 MED dataset (Figure 11(b)), conceptlets score 0.483 MAP where Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression score 0.406 and 0.403. Where our algorithm selects 83, the others select 120 and 165. On this dataset we also compare with the initial version this work [76], with a MAP of 0.443. We improve our previous work by 10.0%, which we attribute to the fact that we optimize the conceptlet size, where we previously relied on fixed sizes only.

On the MediaMill 2012 MED dataset Minimum Redundancy Maximum Relevancy scores 0.304 MAP and L1-Regularized Logistic Regression scores 0.318 MAP, where conceptlets obtain

Figure 11: Experiment 4. Conceptlets versus other selections. For video event classification, conceptlets selected with our proposed cross-entropy algorithm outperform alternative concept selection algorithms.

0.329 MAP (Figure 11(b)). Again we reach better numbers with fewer selected concepts. Where conceptlets need on average 245 concepts to reach this results, the others need many more (458 vs 374). For the event *Flash mob gathering*, a conceptlet with 40 concepts obtains 0.574 AP where Minimum Redundancy Maximum Relevancy reaches 0.499 AP with 600 concepts and L1-Regularized Logistic Regression scores 0.516 AP with 298 concepts. The results on the

Table 4: Properties of conceptlets as a function of bank sizes considered during cross-entropy optimization. We consider sizes in the range 10 to 90, with a step size of 20. All results averaged over three events from the TRECVID 2010 MED dataset using the concept bank of Merler et al. [81]. In addition to L1-Regularized Logistic Regression (LRLR) and Minimum Redundancy Maximum Relevancy (MRMR) and the initial version of this work [76] we also report four other recent feature selections. Conceptlets are computationally more demanding than the other selections, but always result in the best event classification accuracy.

	Conceptlet					LRLR [89]	MRMR [96]	CIFE [16]	SFFS [1]	RELIEF [61]	DISR [82]	[76]
	[10]	[10,30]	[10,50]	[10,70]	[10,90]	–	–	–	–	–	–	–
Size	10	30	50	70	83	165	120	120	70	50	100	100
Time(m)	5	17	32	52	76	10	6	18	38	29	10	19
MAP	0.306	0.359	0.403	0.447	0.482	0.403	0.406	0.392	0.396	0.395	0.395	0.443

Columbia CV dataset in Figure 11(c) show similar behavior. Conceptlets obtain 0.625 MAP using on average 93 concepts per event where the others perform worse with more concepts selected (0.565 MAP/452 concepts vs 0.598 MAP/440 concepts).

When we inspect the selected concepts for the event *Landing fish* for Minimum Redundancy Maximum Relevancy (data not shown) we find several redundant concepts such as ‘Adult-Female-Human’, ‘Female-Human-Face-Closeup’, ‘Female-Person’, ‘Single-Person-Female’, and ‘Two-People’ which have a negative effect on the event classification result. Moreover, the redundancy makes the description less precise. Recall that Minimum Redundancy Maximum Relevancy ranks the concepts first based on the ratio of the relevance of the concept to the redundancy of the concepts in the set. Making this selection sensitive to the accuracy of the concept detectors. Moreover, some low ranked concept detectors may have poor average precision in isolation, but when combined with other concepts lead to a better event classification. Our algorithm is less sensitive to the performance of concept detectors. If the presence of a concept, either an accurate or inaccurate one, improves the accuracy of the event classifier it will try to maintain it in the conceptlet. Selection by L1-Regularized Logistic Regression is more competitive, but the selection always contains more concepts than our conceptlets. In addition to being less accurate than conceptlets, selection by L1-Regularized Logistic Regression is less descriptive.

In Table 4 we report the run time and mean average precision of conceptlets as a function of the number of bank sizes considered. We consider sizes in the range 10 to 100, with a step size of 10. In addition to L1-Regularized Logistic Regression, Minimum Redundancy Maximum Relevancy and the initial version of this work [76] we also report four other recent feature selections. As expected, the best result is obtained when we consider all sizes in the range, and the worst result is obtained when we only consider banks of 10 concepts. When we consider all ranges, without any preselection, conceptlets need 90 minutes per event on average. Using a non-linear instead of a linear SVM with conceptlets increases the MAP from 0.482 to 0.513, but needs 425 minutes per event on average. This is longer than the other selections, but can be sped up by parallelization when execution time is an issue. Conceptlets always result in better mean average precision.

We conclude that conceptlets, at the expense of a longer run time, are more effective than Minimum Redundancy Maximum Relevancy and L1-Regularized Logistic Regression by considering more discriminant and less redundant concepts.

2.6 CONCLUSIONS

We study event classification based on banks of concept detectors. Different from existing work, which simply includes in the bank all available detectors, we propose an algorithm that

learns to find from examples the most informative concept-subset per event, which we call the conceptlet. We formulate finding the conceptlet as an importance sampling problem which can be solved with a near-optimal cross-entropy optimization. We study the behavior of conceptlets by performing four experiments on three unconstrained web video collections from the 2010 and 2012 TRECVID Multimedia Event Detection task and the Columbia Consumer Video datasets using a total of 1,346 pre-trained concept detectors.

The results of experiment 1 show that there are individual concepts that are more discriminative and descriptive than others for representing events in internet video. The results of experiment 2 give an indication that large banks of concept detectors are important for covering a variety of complex events, as they may appear in unconstrained video. In general, the event classification accuracy increases with the number of concept detectors in the bank. However, we show that some concept-subsets are more informative than others for specific events, and this may result in improved event classification accuracy. The results of experiment 3 confirm that event classification using conceptlet outperform banks using all concepts, and always contains significantly less concept detectors. Finally, results of experiment 4 reveal that our conceptlets, at the expense of a longer run time, are more effective than selected concepts by state-of-the-art feature selection algorithms, by considering more discriminant and less redundant concepts. What is more, the conceptlets make sense for the events of interest, without being programmed to do so.

Further improvements of conceptlets with respect to the reduction of training examples, run time efficiency, classification accuracy and event descriptiveness can be envisioned, for example by the use of semantic query-by-video solutions [78]. We plan to elaborate on this direction in future work. For the moment we conclude that selective use of semantic concepts, by means of conceptlets, is beneficial for classifying video events and opens up the possibility to automatically describe and explain why a particular video was found.

ENCODING CONCEPT PROTOTYPES FOR VIDEO EVENT DETECTION AND SUMMARIZATION

3.1 INTRODUCTION

The goal of this paper is to detect an event from very few examples and to summarize a video containing the event in its most descriptive frames. The common tactic for few- and zero-example event detection is to rely on a video representation consisting of concept detector scores indicating presence of concepts like *cake*, *rock* or *tire* in the video [14, 15, 22, 28, 36, 62, 69, 75, 81]. Such a representation is known to outperform the traditional low-level representation when examples are scarce [11, 27, 68, 78]. In [14, 26, 39, 79], it is further shown that semantic representations are capable of event detection without examples, from just a textual description of an event (and the corresponding event-to-concept mapping). We observe that concept detectors are included in the representation whenever they are available and are learned from either image- or video-level annotations. In this paper we propose a new video representation for few- and zero-example event detection, which also uses concept detectors, but we place special emphasis on what concepts to include and from what video-frames they are learned.

In [81] Merler *et al.* used a mixture of 280 relevant and irrelevant concepts to the events which are trained from thousands of labeled web images. In [28] Habibian *et al.* used a collection of concepts consisting of 1,346 relevant and irrelevant concepts to the event trained from ImageNet [7] and TRECVID [95]. In [75] Mazloom *et al.* proposed a learning algorithm to find conceptlets, a set of concept detectors deemed relevant to an event, from the same 1,346 concepts used in [28]. In [14] Chen *et al.* use a collection of 2,000 concepts trained on Flickr images, deemed relevant to an event from analyzing the textual description. We note that in all these works [14, 28, 75, 81] concept detectors are trained from a collection of images, which have the appealing ability that they can be applied on each frame of a video. However the source of training data used in these methods is images from Flickr and ImageNet, applying them on frames extracted from web video results in an unreliable video representation due to the domain mismatch between images and videos. To counter the domain mismatch, recently several works propose to use web video for training concept detectors [8, 26, 27, 69, 114], which is not only much more efficient, but also suffers less from the domain mismatch problem. However, the drawback of these works is that they operate on a video-level, although concepts and events may be visible in a part of the video only. Hence, they include many irrelevant frames in their pooled representation leading to a sub-optimal representation.

In [48, 49, 56] methods are proposed that find relevant video parts by formulating the problem of video event detection as multiple instance learning. These works show the efficiency of training an event model only on relevant parts of video. Inspired by [48, 49, 56], in this paper we attempt to find an informative representation per concept by considering only the frames from web-video that are relevant to the concept, without using frame-level annotations. We call the informative

Figure 12: We propose an algorithm that learns from web data a set of relevant frames to the concept, which we call concept prototypes (y-axis). We encode videos as concept prototypes (x-axis) and show that for unseen videos it is able to detect events from few or zero examples. In addition, we demonstrate how the new representation can be exploited for unsupervised video event summarization (red border).

subset of frames per concept the concept prototypes and use them to represent video for few- and zero-example event detection.

In [18], Ding *et al.* propose a method to generate a text which summarizes the important information in a video of an event. The algorithm selects relevant concepts to the event using lots of video examples. We also aim for video event summarization, but we use concept prototypes by selecting the relevant frames of a video to the event without relying on any video examples. In [98], Potapov *et al.* proposed video event summarization by learning from lots of positive and negative video examples. Each event model is applied on various segments of a video, the segments with the highest scores are returned as summary. Since they train a model over the video-level representation of videos in the training set, including all irrelevant parts, applying their model on small video segments are unlikely to reflect an accurate score. Since our prototypes are a frame-level representation of concepts deemed relevant to the event, we have the ability of applying them on each video frame and computing a reliable score per frame without the need for any event model. We simply consider the highest scoring frames as the video summary, see Figure 12. To the best of our knowledge no method currently exists in the literature for summarizing a video event using only a textual event definition.

We make the following contributions:

- We model selecting the concept prototypes out of a large set of frames of positive videos per concept, as seeking the frames closest to the densest region in the feature space of positive and negative videos.

Figure 13: The flow chart for our proposal which consist of three modules. At first, we propose an algorithm which learns from a video dataset and corresponding captions a set of prototypes per concept. Secondly, we construct a prototype dictionary by collecting concept prototypes of all concepts and encode the video event in the dictionary. At the end we show the ability of our video representation for few- and zero-example event detection and unsupervised event summarization.

- We introduce concept prototypes as a new semantic video representation for few example event detection.
- We show the effectiveness of the video representation by concept prototypes in few- and zero-example event detection as well as unsupervised event summarization.

We organize the remainder of this paper as follows. We present our encoding using concept prototypes in Section 3.2. We introduce the experimental setup on the challenging TRECVID 2013 Multimedia Event Detection [91] and MED-summaries datasets [98] in Section 3.3. Results are presented in Section 3.4. We conclude in Section 3.5.

3.2 ENCODING CONCEPT PROTOTYPES

In this paper we aim to arrive at a semantic space with the ability for video event detection and summarization. We propose concept prototypes and leverage it for a new video representation. In this work, we define concept prototypes as a set of diversified video frames with variant visual appearance depicting the same concept in videos. For example, one can reasonably expect that for the concept *car*, prototypes are those frames related to various types and models of cars such as frames containing *city cars*, *sport car*, *truck*,. Because of their diversity, prototypes are beneficial over using a set of frames which describe only a specific type of car, e.g. *sport car*, or a set of frames which includes irrelevant frames to the concept *car*. In other words, concept prototypes for an arbitrary concept are those frames which appear in the most informative parts of the positive videos and not in the negative videos of the concept.

Since the concept prototypes are frame-level detectors, we have the ability of applying them in each frame of a video, which has several benefits. Firstly, by representing each frame of a video by concept prototypes, we can analyze each frame of the video individually. It gives us the opportunity of proposing a powerful video representation by aggregating all frames of a video in concept prototype space. Secondly, it allows us to do zero-example video event detection by considering the score of the frames in a video to each concept prototypes, and therefore help localize the key information in the video for zero shot event detection. It also offers us to do unsupervised video event summarization by selecting the relevant frames of video to the event.

Our framework contains three parts, schematically illustrated in Figure 13. At first, we learn from a video dataset harvested from the web a set of concept prototypes per concept. Secondly, we construct a prototype dictionary by collecting concept prototypes of all concepts and encode each video from a train and test video event set in the prototype dictionary. At the end we show the ability of the concept prototype encoding for event video detection and summarization. Each part is detailed next.

3.2.1 Concept prototype learning

Our algorithm for learning concept prototypes of an arbitrary concept consists of two steps. i) Scoring positive frames which aims to compute a score for each frame of a positive videos containing the concept. ii) Selecting a subset of positive frames and consider them as concept prototypes.

Scoring positive frames: Suppose D is a dataset of web videos which are annotated for d concepts in a collection $C = \{c_1, c_2, \dots, c_d\}$. Let consider $X_c \subset D$ is a set of n video/label pairs for arbitrary concept $c \in C$. We denote positive videos as V_i^+ which consist of n_i^+ frames v_{ij}^+ , $j = 1, \dots, n_i^+$. l^+ shown the number of positive videos in X_c , $X_c = \{\{V_i^+\}_{i=1}^{l^+}, \{V_i^-\}_{i=1}^{l^-}\}$. Similarly, V_i^- , n_i^- , v_{ij}^- , and l^- represent negative videos, the number of frames in the video, the j^{th} frames in the video, and number of negative video in X_c respectively. We suppose all frames v_{ij} belong to the feature space $\mathbb{R}^m$. We collect all frames of positive videos V_i^+ of concept c in P_c , $P_c = \{v_{ij}\}(i = 1, \dots, l^+, j = 1, \dots, n_i^+)$.

To compute a score per frame $v \in P_c$ we need a function to precisely return a value for v indicating its confidence of being applied as a prototype of the concept. It has to describe how relevant the frame v is w.r.t. concept c in the positive videos, V_i^+ , and while departing away from the negative videos, V_i^- . We use a diverse density function which is well known and frequently used in machine learning [73, 74]. The Diverse Density function (*DD*) at frame $v \in P_c$ returns a real value which expresses the probability that frame v agrees with the densest regions within the

frames of both positive and negative videos. It reflects the coherence with the visual appearance of the concept and is defined as:

$$DD(v) = Pr(v|V_1^+, \dots, V_l^+, V_1^-, \dots, V_l^-) \quad (3.1)$$

where $Pr()$ represents the probability function. By applying Bayes' rule to Eq. 3.1 and assuming that all videos are conditionally independent given v , we write Eq. 3.1 as:

$$DD(v) = \prod_{i=1}^{l^+} Pr(v|V_i^+) \prod_{i=1}^{l^-} Pr(v|V_i^-) \quad (3.2)$$

To estimate $Pr(v|V_i^+)$ and $Pr(v|V_i^-)$ we use the proposed way in [73]:

$$Pr(v|V_i^+) = \max_j \exp(-\|v_{ij}^+ - v\|) \quad (3.3)$$

$$Pr(v|V_i^-) = 1 - \max_j \exp(-\|v_{ij}^+ - v\|) \quad (3.4)$$

Eq. 3.3 uses the maximum similarity between all frames in a positive video and the given frame as the probability of choosing v as a prototype of video V_i^+ . On the contrary, Eq. 3.4 measures how dissimilar a negative video should depart from the candidate prototype frame. After considering Eq. 3.3 and Eq. 3.4 in Eq. 3.2 the DD function computes a real value between 0 and 1 for frame v . A larger value of DD at frame v indicates a higher probability that the frame v fits better with the frames from all positive videos than with those from all negative videos, and therefore have higher chance to be chosen as a concept prototype.

Selecting concept prototypes: After computing a score per each frame of P_c , we select the concept prototypes for concept c by considering two constraints: i) They need to be distinctive from each other so that the prototype collection has no redundancy. ii) They need to have large DD values. Applying the first constraint, we remove some of the frames that are similar to each other. In our algorithm we control the repetition of frames by considering coefficient α . It is a parameter which defines the degree of closeness of frames to concept prototypes. Its value is between 0 and 1. Closer to 1 causes our algorithm to generate a small number of prototypes per concept. The second constraint limits concept prototypes to those that are most informative in terms of co-occurrence in different positive videos. In our algorithm, we consider it by picking frames with DD value greater than a threshold T .

Our learning algorithm for obtaining concept prototypes for concept category c is summarized in Table 5. The inputs are a set of positive and negative videos for concept c , threshold T and coefficient α which are parameters to define the number of concept prototypes for concept c . The pseudo code learns a set of concept prototypes for concept c which is represented by CP_c . In the pseudo code in Table 5, lines 1-6 compute the value of the DD function for all frames of positive videos in P_c . Line 7-12 describe an iterative process to select a collection of distinct concept prototypes. In each iteration, a frame of P_c with the maximum value of DD function is selected as a concept prototype (line 9). The frames of P_c that are close to the selected concept prototype or that have DD values lower than a threshold are removed from P_c (line 10). A new iteration starts when there is no any frame in P_c . In Figure 14 we visualize several concept prototypes.

3.2.2 Video encoding

In this section we first describe the process of constructing a prototype dictionary using concept prototypes, then we explain how to encode video events of train and test set into the prototype dictionary.

Table 5: Pseudocode of learning concept prototypes for concept category c .

INPUT: Positive videos for concept c (V_i^+),
 Negative videos for concept c (V_i^-),
 Threshold (T), and coefficient α

OUTPUT: Concept prototypes for concept c (CP_c)

1. Set P_c be the set of all frames of videos in V_i^+
2. $H = []$ an array for keeping DD values
3. $M = \text{length}(P_c)$
4. for $q = 1, \dots, M$
5. $H(q) = \text{compute } DD(P_c(q))$ using Eq. 3.2
6. *end*
7. $i = 1$
8. Repeat
9.   Select concept prototype, a frame of P_c corresponding to the maximum value of DD
        $CP_c(i) = P_c(\text{find}(H = \max(H)))$
10.   Remove from H all elements a satisfying
        $\|CP_c(i) - P_c(a)\| < \alpha \|CP_c(i)\|$ OR $H(a) < T$
11. $i = i + 1$
12. While ($P_c \neq []$)
13. Output: CP_c , concept prototypes for concept c

Constructing prototype dictionary: To make a prototype dictionary we first learn m concept prototypes for each of d concept $c_i \in C$ using the method which we explained in section 3.2.1. We consider $CP_{c_i} = \{cp_{i1}, cp_{i2}, \dots, cp_{im}\}$ as concept prototypes for concept c_i . Then we construct a prototype dictionary PD by collecting all concept prototypes of concepts in C , $PD = \{CP_{c_1}, CP_{c_2}, \dots, CP_{c_d}\}$ which element cp_{jk} is k^{th} concept prototype of j^{th} concept.

Encoding video in prototype dictionary: We consider each element cp_{jk} as a feature in new feature space $\mathbb{F}_{PD}$. We define a function $g_{jk}(\cdot)$ to compute the value of the feature cp_{jk} on video event V , which consist of n frames $V = \{v_1, v_2, \dots, v_n\}$ as:

$$g_{jk}(V) = Pr(cp_{jk}|V) = \max_{i=1, \dots, n} \text{similarity}(cp_{jk}, v_i). \quad (3.5)$$

We can interpret $g_{jk}(V)$ as a measure of similarity between the concept prototype cp_{jk} and the video event V . It is determined by the value of similarity between concept prototype cp_{jk} and the closest frame in the video V .

As a result, each video event V can be encoded in $\mathbb{F}_{PD}$ using all concept prototypes in PD . It can be defined as a point $\varphi(V)$, in feature space $\mathbb{F}_{PD}$ which define as:

$$\varphi(V) = [g_{11}(V), g_{12}(V), \dots, g_{jk}(V), \dots, g_{dm}(V)]^T \quad (3.6)$$

where each element of $\varphi(V)$ is defined by one concept prototype and one frame from the video event V , the frame that is *closest* to the concept prototype. Also it can be viewed as a measure of the degree that a concept prototype is visible in the video event V .

Figure 14: Learned prototypes for five arbitrary concepts. Prototypes ordered from left (best) to right.

3.2.3 Event detection and summarization

After representing video events in concept prototypes space, we use it for the problem of few- and zero-example event detection, and unsupervised video event summarization.

Few-example event detection: Suppose $TR = \{V_1^+, \dots, V_{n^+}^+, V_1^-, \dots, V_{n^-}^-\}$ is a train set consist of n^+ positive videos and n^- negative videos, and $TE = \{W_1, \dots, W_n\}$ is a test sets for an event category e . By mapping each video of train and test set to concept prototype space using Eq. 3.6 we define TR and TE as two matrix representation of all videos in $\mathbb{F}_{PD}$:

$$TR = [\varphi(V_1^+), \dots, \varphi(V_{n^+}^+), \varphi(V_1^-), \dots, \varphi(V_{n^-}^-)]^T \quad (3.7)$$

$$TE = [\varphi(W_1), \dots, \varphi(W_n)]^T \quad (3.8)$$

Each row of TR and TE represents a video, k^{th} column of TR and TE represent k^{th} feature in $\mathbb{F}_{PD}$. We train a model on TR and report the result of event detection accuracy on TE .

Zero-example event detection: In this scenario there is only a textual query explaining the event. Suppose T is an explanation for event e which consist of n words, $T = \{w_1, w_2, \dots, w_n\}$, $C = \{c_1, c_2, \dots, c_d\}$ is a set of concepts, and PD is a prototype dictionary which consists of $m \cdot d$ elements (m concept prototypes for each of d concept in C). Let's consider $W_i = \{w_{i1}, w_{i2}, \dots, w_{il}\}$, which consist of l frames, be a video of event e in test set TE . We use following steps for computing a score per each video W_i in TE . i) At first we define a binary representation for event

e as a query, $\mathbf{Q} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d) \in \mathbb{F}_{PD}$, by comparing the concepts in C with the words in T . The value of a_i computes as:

$$a_i = \begin{cases} \vec{1} \in \mathbb{N}^m & \text{if } c_i \in T \\ \vec{0} \in \mathbb{N}^m & \text{if } c_i \notin T \end{cases}$$

ii) We map each frame of W_i to prototype dictionary PD by computing similarity between each frame of W_i and each of concept prototypes in PD . As a result, each frame of W_i is represented as a vector of concept prototype score, $W_i = \{x_{i1}, x_{i2}, \dots, x_{il}\}, x_{ij} \in \mathbb{F}_{PD}$. iii) We compute a score per each frame of W_i by multiplying query $\mathbf{Q}$ with $x_{ij}, j = 1, \dots, l$ and define vector $Y = [Q \cdot x_{i1}, Q \cdot x_{i2}, \dots, Q \cdot x_{il}]$. The value of j^{th} element of Y describes how much the content of the j^{th} frame of W_i matches to event e . At the end iv) We score video W_i with maximum value of vector Y , $S(W_i) = \max(Y)$.

After computing a score S per each video $W_i, i = 1, \dots, n$ of TE , we sort the videos with their scores and compute the accuracy of zero-example event detection for event e .

Unsupervised video event summarization: We demonstrate the feasibility of using our concept prototypes, for summarizing the content of a video event by selecting relevant frames to an event category. Suppose W_i is a positive video of event category e consist of l frame, $W_i = \{w_{i1}, w_{i2}, \dots, w_{il}\}$. We follow the first three steps of zero-example event detection, to compute vector Y which the value of j^{th} element of Y show how much the content of the j^{th} frame of W_i is match to event e . We sort the vector Y from the best frame, that completely matches the content of the event, to the worst frame. We select the best frames based on how many percent of frames of video W_i we want to report as summarization of W_i .

3.3 EXPERIMENTAL SETUP

3.3.1 Datasets

In our experiments we report on two large corpora of challenging real-world web video: the TRECVID 2013 Multimedia Event Detection dataset [91], and MED-summaries provided by Potapov *et al.* [98].

Event detection: TRECVID 2013 MED [91]. This corpus contains 56k user-generated web videos with a large variation in quality, length and content of 20 real-world events. This corpus consists of several partitions, we consider the Event Kit training, Background training, and MED test set, containing about 200, 5K, and 27K, videos. We follow the *10Ex evaluation procedure* outlined by the NIST TRECVID event detection task [95]. It means that for each event the training data consist of 10 positive videos from the Event Kit training data along with about 5K negative videos from the Background training data. We also consider the *0Ex evaluation procedure* as specified by NIST, where we don't have any positive video example available. We rely only on a provided brief textual definition of the event containing some evidences of concepts which can be expected to be visible in positive video examples of a particular event.

Event summary: MED-summaries [98]. This set consists of 160 videos selected from the TRECVID 2011 Multimedia Event Detection task: the validation set consist of 60 videos and the test set consist of 100 videos. Every segment of all these 160 videos is annotated with a category-specific importance value (0, 0.333, 0.666, and 1). All segments of all videos are annotated by different users to obtain the ground truth video summary for evaluation. High scores indicate the human annotators consider the segment is close to the event of interest. We report all results on the test set containing 10 videos for a total of 10 events. We adapt the provided annotation format from segment level to frame level by simply propagating the segment annotation to each of its frames.

3.3.2 Implementation details

Concept prototypes training data. To train our concept prototypes, we use a video dataset that is disjunct from both the TRECVID 2013 MED and MED summaries collections. We adapt the VideoStory46K from Habibian *et al.* [27] which contains a collection of 46k video from YouTube. Every video has a short title caption provided by the user who uploaded the video. From these captions we generate individual terms which serves as positive concept labels for the corresponding video. In total there are 19,159 unique terms in the captions. We filter the terms by: removing stop words, noisy terms, terms which are not visually detectable such as *God* (we used the visualness filter from [27]), terms with frequency less than 100, and keep those terms which match with the words in the definition text of the events. After this step we reach to a list of 479 different concept labels.

Features. For representing the frames in all of our datasets we use deep learning features. At first we extract the frames of videos uniformly every 2 seconds. Then, similar to [47], we use a pre-trained convolutional neural network for mapping every frame of videos to a 4,096-dimensional feature vector which is the output of the second fully connected layer of the convolutional neural network. The network is pre-trained on all the 15,0254 categories in the ImageNet dataset, for which at least 50 positive examples are available.

Concept prototypes learning. For learning concept prototypes for each of the 479 concepts we repeat the following steps. i) We construct a set of videos for each concept label by considering all its positive videos and supplement it with negatives up to ten times the number of positives by random sampling from other concept labels. ii) We apply our concept prototypes learning algorithm (Section 3.2.1). We set parameter T as the average of the maximum and minimum of the DD function in H (Table 5), and we empirically found 0.05 to be a good value for α . We map the videos from TRECVID 2013 MED and MED-summaries to concept prototypes space using cosine similarity.

Event detection. We train for each event a one-versus-all linear support vector machine [107] and fix the value of its regularization parameter C to 1.

3.3.3 Experiments

In order to establish the effectiveness of our concept prototypes for event detection and summarization we perform three experiments.

Experiment 1: Few-example event detection In this experiment we compare the event detection accuracy of our proposed concept prototype encoding versus five baselines. *Baseline 1:* DeepNet. Video representation using deep learning features. *Baseline 2:* 15k-Concepts. A video event representation consisting of 15k concept detectors trained on ImageNet (inspired by [28]). *Baseline 3:* Conceptlets [75]. Video representation consisting of the most informative ImageNet 15K concepts per event. *Baseline 4:* 479-Concepts. Video representation using 479 video-level concept detectors trained on VideoStory46k (inspired by [36]). *Baseline 5:* VideoStory [27]. Video event representation which strives to encode the caption of a video rather than a single word. All six methods are evaluated by their event detection performance on the TRECVID 2013 MED test set.

Experiment 2: Zero-example event detection We compare the accuracy of our concept prototypes for zero-example event detection versus four baselines. *Baseline 1:* Flickr concepts. As a first baseline we report the result in [14] which used 2,000 relevant concepts per event as trained from flickr images. *Baseline 2:* 15k-Concepts. Same as above. *Baseline 3:* Composite concepts [26]. Video representation using 138 video-level concept detectors which are the result of combining several concepts using logical connectors. *Baseline 4:* 479-Concepts. Video

Table 6: Experiment 1: Few-example event detection on TRECVID 2013 MED test set. Best AP result per event in bold.

Event	DeepNet	15k-Concepts [28]	Conceptlet [75]	479-Concepts [36]	VideoStory [27]	This paper
Birthday party	0.137	0.114	0.145	0.146	0.118	0.188
Changing a vehicle tire	0.391	0.388	0.401	0.444	0.103	0.464
Flash mob gathering	0.405	0.347	0.377	0.425	0.535	0.439
Getting a vehicle unstuck	0.334	0.323	0.390	0.377	0.319	0.418
Grooming an animal	0.084	0.108	0.110	0.126	0.151	0.154
Making a sandwich	0.031	0.074	0.075	0.100	0.074	0.131
Parade	0.171	0.109	0.145	0.246	0.452	0.303
Parkour	0.320	0.309	0.319	0.317	0.721	0.326
Repairing an appliance	0.169	0.127	0.133	0.320	0.184	0.244
Working on a sewing project	0.058	0.071	0.103	0.072	0.151	0.109
Attempting a bike trick	0.054	0.030	0.047	0.106	0.061	0.144
Cleaning an appliance	0.021	0.019	0.020	0.031	0.078	0.055
Dog show	0.232	0.134	0.165	0.379	0.354	0.313
Giving directions to a location	0.012	0.005	0.008	0.020	0.004	0.022
Marriage proposal	0.002	0.002	0.003	0.004	0.004	0.004
Renovating a home	0.019	0.024	0.032	0.056	0.051	0.033
Rock climbing	0.070	0.063	0.083	0.080	0.100	0.110
Town hall meeting	0.268	0.201	0.229	0.158	0.118	0.290
Winning a race without a vehicle	0.150	0.126	0.164	0.116	0.217	0.182
Working on a metal crafts project	0.054	0.068	0.078	0.116	0.118	0.144
MAP	0.150	0.135	0.156	0.181	0.196	0.202

representation using 479 video-level concept detectors trained on VideoStory46k. For all methods we build a query Q per event as a binary vector indicating presence or absence of the concept in the event description, as detailed in Section 3.2.3. We compute a score for a video in the test set by multiplying Q with a histogram containing the scores of concept detectors when applied on the video. From the resulting ranking, we compute the zero-example accuracy. We again evaluate the accuracy of all methods on the TRECVID 2013 MED test set.

Experiment 3: Unsupervised event summarization We compare the accuracy of our concept prototypes in video event summarization versus three baselines. *Baseline 1*: Random. Video event summarization with relevant frames selected randomly. To cancel out the accidental effects of randomness, we repeat the random frame selection 10 times and report the mean. *Baseline 2*: 15k-Concepts. A video event summarization with relevant frames selected using 15k ImageNet concepts. *Baseline 3*: 479-Concepts. A video event summarization with relevant frames selected using concept detectors. *Oracle* Video event summarization using the ground truth user annotation value from [98], to create an upper bound for video event summarization. In *Baseline 2* and *3*, similar to our method in section 3.2.3, we make a vector Y with the value of each of its elements being the result of multiplying Q , a binary representation of event, and a histogram of applying concept detectors on each frame of video. Then we follow the steps which we explained in Section 3.2.3 for summarizing the video. We evaluated the accuracy of all methods in video event summarization on the MED-summaries dataset [98] by keeping 10, 20, 30, ..., 100 percent of the frames per video. For those videos annotated by more than one user, we report the mean accuracy. At the end we report the mean accuracy for all 10 videos of each of the 10 events.

Evaluation criteria For event detection performance we consider the average precision (AP) [110]. We also report the average performance over all events as the mean average precision (MAP). Similar to [98], for evaluating event summarization performance we propose a new metric, important ratio. We defined important ratio (IR) as average of user annotation value of selected p frames of the video. The value of IR shows the quality of the video event summarization. We consider the average of the p highest user annotation value as a ground truth. The closeness of the value of IR to ground truth value shows the better summarization.

Table 7: Experiment 2: Zero-example event detection on TRECVID 2013 MED test set. Best AP result per event in bold.

Event	Flickr concepts [14]	15k-Concepts [28]	Composite Concepts [26]	479-Concepts [36]	This paper
Birthday party	0.032	0.022	0.076	0.096	0.154
Changing a vehicle tire	0.038	0.099	0.018	0.238	0.320
Flash mob gathering	0.058	0.104	0.319	0.266	0.271
Getting a vehicle unstuck	0.010	0.107	0.055	0.356	0.406
Grooming an animal	0.011	0.019	0.009	0.079	0.095
Making a sandwich	0.028	0.021	0.079	0.082	0.164
Parade	0.130	0.094	0.223	0.221	0.240
Parkour	0.027	0.020	0.021	0.077	0.112
Repairing an appliance	0.048	0.078	0.025	0.221	0.213
Working on a sewing project	0.012	0.016	0.014	0.087	0.089
Attempting a bike trick	0.003	0.017	0.021	0.056	0.061
Cleaning an appliance	0.013	0.006	0.006	0.022	0.026
Dog show	0.003	0.003	0.001	0.011	0.011
Giving directions to a location	0.002	0.004	0.025	0.004	0.008
Marriage proposal	0.009	0.004	0.002	0.003	0.005
Renovating a home	0.002	0.017	0.023	0.029	0.026
Rock climbing	0.014	0.003	0.146	0.015	0.036
Town hall meeting	0.020	0.008	0.014	0.023	0.035
Winning a race without a vehicle	0.028	0.012	0.110	0.095	0.101
Working on a metal crafts project	0.002	0.002	0.006	0.010	0.014
MAP	0.024	0.032	0.063	0.100	0.119

3.4 RESULTS

3.4.1 Few-example event detection

We show the result of experiment 1 in Table 6. Results demonstrate that the video event representation based on our concept prototypes performs better than all baselines. Our concept prototypes reach 0.202 MAP in event detection, where using DeepNet feature, 15k-Concepts, Conceptlet, 479-Concepts, and VideoStory results in 0.150, 0.135, 0.156, 0.181, and 0.196 respectively.

We explain the good results of our concept prototypes from several observations. First, most of the concepts in 15k-concepts are irrelevant to the event categories. Moreover, as these detectors originate from ImageNet, they are trained on images and not on video leading to a domain-mismatch. Using Conceptlets [75] to find the most informative concepts, out of the 15k, per event improves the event detection accuracy, but it cannot compensate for the domain mismatch. Interestingly, learning the events directly from DeepNet features is advantageous over using 15k-concepts. However, when we learn the detectors from video annotations, as suggested in [36], result improve considerably (from 0.150 to 0.181). VideoStory further improves upon this result, but our concept prototypes are more accurate. There are two reasons. First, VideoStory trains detectors using all frames from both positive and negative training videos. By adding lots of irrelevant frames it leads to a noisy video concept representations. In contrast, our concept prototypes identify the relevant frames of video only, making the representation more clean. Secondly, since the concept detectors from VideoStory are trained over a video-level representation, they have to apply on a video-level representation for event detection as well. Since most of the events happen in a part of the video only, aggregating all frames is sub-optimal. In contrast, our concept prototypes have the ability to apply the concepts on individual frames, which decreases the effect of irrelevant frames.

The results of experiment 1 confirm that few-example event detection accuracy profits from using concepts that are relevant to the event and which are learned from a video, rather than the image dataset. Moreover, it is beneficial to represent video using concept prototypes as the influence of irrelevant frames is reduced.

Figure 15: Experiment 3. Video event summarization on MED-summaries dataset. The result shows the effectiveness of our concept prototypes.

3.4.2 Zero-example event detection

The results of this experiment are shown in Table 7. It indicates our concept prototypes outperform all four baselines in zero-example event detection. We reach to 0.119 MAP where the alternatives reach to 0.024, 0.032, 0.063, and 0.100 MAP. Since our concept prototypes are frame-based we have the ability to apply them on each frame. As a result we describe how well the frame is related to an event. Irrelevant frames are simply ignored.

Flickr concepts are relevant to the event but they are trained on images of Flickr leading to a domain mismatch. The same holds for the 15k-Concepts trained on ImageNet. Moreover, most of its concepts are irrelevant to the event of interest. Both the composite concepts and 479-Concepts are trained on video-level labels. It means they suffer less from the domain mismatch leading to improved zero-example event detection results. However, both of these methods suffer from the fact that they treat all frames equally, which adds lots of irrelevant frames to the representation. Our concept prototypes contain concepts that are relevant to the event, are trained on video, and consider only the most relevant frames.

3.4.3 Unsupervised event summarization

We show the result of experiment 3 on MED-summaries [98] in Figure 15. The result demonstrates the effectiveness of our concept prototypes for unsupervised video event summarization against random frame selection, 15k-Concepts, and 479-Concepts for all summarization settings. When

Figure 16: Experiment 3. Two examples of video event summarization using concept prototypes and 479-Concepts. We denote the ground truth annotation value of each frame with four colors. Green indicates 1, yellow 0.666, orange 0.333, and red 0 (higher is better). The color coding shows the effectiveness of concept prototypes.

we request a summary containing 30% of the video we reach to 0.600 accuracy in *IR*, where random frame selection, 15k-Concepts, and 479-Concepts reach to 0.221, 0.274, and 0.384 accuracy in *IR*. The accuracy of the oracle upper bond is 0.923 *IR*. Since our concept prototypes are frame-based, we can apply them on each frame of a video and reach to an accurate video representation. Where the others are video-level by nature which makes them less suited for frame-level classification, resulting in a less reliable summary.

We visualize the result of video event summarization for two videos of the events *birthday party* and *Changing a vehicle tire* using our concept prototypes and the 479-concept detectors, as the best baseline, in Figure 16. In both examples we see the effectiveness of using our concept prototypes against concept detectors in video event summarization.

3.5 CONCLUSION

In this chapter we propose *concept prototypes* a new semantic video representation for few and zero example event detection and unsupervised video summarization. Different from existing works, which obtain a semantic representation by training concepts over images or an entire video, we propose an algorithm that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations.

We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames containing both positive and negative training videos. The prototypes represent a class of frames that are more likely to appear in positive videos than in negative videos. Since the concept prototypes are a frame-level representation of concepts, we have the ability of mapping each frame of a video in concept prototype space which has several benefits. First, few-example event detection accuracy increases when using relevant concept prototypes. Second, the accuracy of zero-example event detection increases by analyzing each frame of a video individually in concept prototype space. Finally, unsupervised video event summarization using concept prototypes is more accurate than recent alternatives.

We conclude that selective use of relevant frames of a video to the concept, by means of concept prototypes, is beneficial for detecting and summarizing video events.

QUERYING FOR VIDEO EVENTS BY SEMANTIC SIGNATURES FROM FEW EXAMPLES

4.1 INTRODUCTION

The goal of this chapter is to retrieve complex events from web video using only a handful of video query examples, like the ones in Figure 17. So far, the common approach to event retrieval is to represent a video in terms of fused audiovisual features and to learn a ranker from hundreds of positives and negative labeled examples *e.g.*, [50, 85]. We differ from this supervised classification solution to event retrieval in three ways. 1) We focus on the more realistic scenario where only a few positive examples for an event are available. In contrast to [67, 70] deliberately ignoring the negative examples as these are not necessarily available at query time. 2) We prefer a high-level over a low-level video representation as events are often characterized by similarity in semantics rather than appearance. 3) We query directly by the example videos rather than a textual label.

We draw inspiration from progress in query-by-image retrieval, where it is known to be advantageous to fuse multiple queries for retrieving concepts [87] or specific objects [2], but only for a low-level feature representation. In this chapter we study the influence of combining multiple queries as well, but for the problem of retrieving complex events in video. In addition, we consider a *semantic signature* representation, consisting of off-the-shelf concept detectors [7, 110], which we deem more suited for capturing the variation in semantic appearance of events. We extend upon existing semantic image search approaches [99] by explicitly addressing fusion of multiple queries and pooling of the semantic representation for video. While approaches for video event classification using semantic representations have recently started to appear in the literature [29, 81], to the best of our knowledge this is the first work on query-by-video retrieval using semantic representations with multiple examples.

Since it is unknown what similarity metric and query fusion to use in such an event retrieval setting, as well as the influence of having only a handful of video query examples available, this chapter conducts an experimental study to shed light on the matter. We also demonstrate the utility and interpretability of semantic signatures for event retrieval in a video search engine. Before we discuss our experimental setup, we first motivate our semantic query fusion methodology.

4.2 SEMANTIC QUERY FUSION

4.2.1 *Semantic Signatures*

A video representation based on semantic signatures has the advantage of evaluating video similarity at a higher level of abstraction, and therefore potentially better semantic generalization

---

Published in the *Proceeding of ACM Multimedia*, 2013 [78].

Published in the *Proceeding of the ACM International Conference on Multimedia Retrieval*, 2014 [25].

Figure 17: Each row shows a video query for the event Attempting a board trick. Due to the variance in both audiovisual and semantic appearance of events in video, a multiple query approach to retrieval, as we propose and evaluate, seems mandatory.

than what is possible with a low level bag-of-words representation (see Figure 18). To arrive at a semantic representation for video retrieval, we consider a lexicon consisting of n concept detectors that include *scenes*, *objects*, *people*, and *activities*. Thanks to efforts like the TRECVID Semantic Indexing task [110] and the ImageNet Large-Scale Visual Recognition Challenge [7] these detectors are commonly available these days, *e.g.*, [29, 81]. We compute concept detector scores per video frame, which are extracted once every s seconds. By concatenating and normalizing the detector outputs, each frame is represented by a concept score histogram of n elements. Finally the concept score histograms are aggregated into a video-level representation by average-pooling, which is known to be a stable choice for video classification [81]. We call the final histogram of concept detector scores a semantic signature.

4.2.2 Similarity Metrics

There are many well known methods for measuring the similarity distance between two histograms, see Table 8. However, it is not clear what metric is the most suited for the purpose of querying video by semantic signatures. In a semantic signature the presence of a concept is much more descriptive than absence of the concept. Hence, we expect that a similarity metric which only relies on the difference between elements while ignoring their values, such as the chi-square and histogram intersection, perform worse for retrieval. To illustrate, consider the cases where two videos contain a certain concept with the same probability. For the first case suppose this probability is 1 for both videos, and for the second case they are 0. From the chi-square point of view for both cases the similarity is identical. However, we know that for the second case the video content is different. A different semantic signature that results in the same similarity is problematic for fusing multiple queries. Since normalized correlation explicitly considers the value of the histogram elements, we expect this metric to be more suited. The same holds for metrics inspired by information theoretic divergence such as Kullback-Leibler and Jensen-Shannon.

Figure 18: Semantic signature for a web video showing the event Attempting a board trick. Since the semantic signature is both expressive and compact, it can be exploited for on-the-fly event detection. What is more, it provides a crude form of recounting.

4.2.3 Query Fusion

Since a single video query cannot cover all possible semantic variations of an event (see Figure 17), we consider a retrieval scenario where a limited set of video example queries is available. In such a multi query scenario, it is important to select the appropriate query fusion. We consider early and late query fusion. In early query fusion we simply combine the semantic signatures of each query into a single signature, which we call *signature pooling*. To arrive at a single signature, we consider both the average and max operator. In early query fusion, we query videos in the test set based on a single semantic signature, making it a very efficient query method. In contrast, for late query fusion, we search the test set based on each query individually making it slightly more demanding than early query fusion. For late query fusion we combine the retrieval results afterwards by *score pooling*. Again, we consider both the average and max operator. We anticipate that the average operator is preferred for both early and late query fusion, as it decreases the influence of noisy or irrelevant concept scores. Continuing the example from Figure 17, the average operator decreases the effect of the concepts *monkey* and *animal* which appear in the query at the bottom.

In order to establish the effectiveness of using query fusion with semantic signatures for video event retrieval, we perform three experiments on a large corpus of challenging real-world web video.

4.3 EXPERIMENTAL SETUP

4.3.1 Data Sets

Video data For the event retrieval experiments, we rely on the web video training corpus from the TRECVID 2012 Multimedia Event Detection task [110]. It comes with ground truth annotations at video level for 25 real-world events, including life events, instructional events, sport events, etc.. Following the protocol for event classification outlined in [29], we split the corpus into

Table 8: Similarity metrics that we consider when querying for video events by a semantic signature containing n concept detector scores, where q and d are two semantic signatures for the query and a retrieved video, respectively.

Similarity Metric	Formula
Normalized Correlation	$\frac{\sum_{i=1}^n q_i * d_i}{\sqrt{\sum_{i=1}^n q_i^2} * \sqrt{\sum_{i=1}^n d_i^2}}$
Kullback-Leibler	$\sum_{i=1}^n q_i \log\left(\frac{q_i}{d_i}\right)$
Histogram Intersection	$\frac{\sum_{i=1}^n \min(q_i, d_i)}{\min(\sum_{i=1}^n q_i, \sum_{i=1}^n d_i)}$
Chi-Square	$\frac{1}{2} \sum_{i=1}^n \frac{(q_i - d_i)^2}{(q_i + d_i)}$
Jensen-Shannon	$\frac{1}{2} \sum_{i=1}^n q_i \log\left(\frac{2q_i}{q_i + d_i}\right) + \frac{1}{2} \sum_{i=1}^n d_i \log\left(\frac{2d_i}{d_i + q_i}\right)$

two partitions: consisting of 1,736 and 4,434 videos respectively. In this paper we use the first partition as the *query set* that consists of several queries for each of 25 event classes (and an additional set of 7,104 negatives which we use for comparison against event classification approaches). We report all results on the second partition, the independent *test set*.

Concept detectors For ease of comparison we adopt the publicly available concept detector scores for this dataset provided by [29]. It consists of a lexicon of 1,346 concept detectors. The 1,346 concept detectors are trained using the training data for 346 concepts from the TRECVID 2012 Semantic Indexing task [110] and for 1,000 objects from the ImageNet Large-Scale Visual Recognition Challenge 2011 [7]. The detectors are trained using a linear SVM atop a standard bag-of-words of densely sampled color SIFT with Fisher vector coding and spatial pyramids, for implementation details see [29]. The concept scores are computed for 1 frame every 2 seconds. We convert them into semantic signatures using the procedure outline in Section 4.2.1. We experimented with the length of the signature and observed only a small retrieval performance difference when increasing length from 500 to 1,346 concepts. For all the reported experiments we use the full, length as it provided the best retrieval result overall. Since we use a compact semantic signature for representing the videos, the process of retrieving the videos on unseen video collections can be done in real-time.

4.3.2 Experiments

Experiment 1: Event retrieval by single query In this experiment we consider the problem of video event retrieval using a single query example. We evaluate the event retrieval accuracy and compare the time and memory efficiency of semantic signatures versus a standard bag-of-words using densely sampled SIFT descriptors with VLAD difference coding [38] using a 1024 words codebook. We consider each of available positive examples of event as a query once and compute the similarity score between query and each video of the test set. We rank videos in the test set by their similarity scores and compute the event retrieval accuracy for each query. At the end we report the average of the accuracy of using each query as a final event retrieval accuracy. As a similarity metric between query and unseen video of test set, we consider all five similarity metrics from Table 8.

Table 9: *Experiment 1: Single Query Baseline. The results are averaged over 25 events and repeated 500 times. Best MAP result in bold. We observe only a minimal variance for all events, therefore not shown.*

Video representation	Random	Similarity Metric				
		Normalized Correlation	Kullback-Leibler	Histogram Intersection	Chi-Square	Jensen-Shannon
Semantic signature	0.011	0.059	0.054	0.033	0.030	0.053
Bag-of-words	0.011	0.046	0.040	0.032	0.024	0.038

Experiment 2: Event retrieval by multiple query To assess the effect of using multiple queries instead of using only one query in video event retrieval, we compare the early and late query fusion schemes described in section 4.2.3. We perform the experiment using semantic signatures with the best similarity metric from experiment 1. We follow [2, 87, 99] and use eight queries per event, which we select randomly from all available queries. We compute the retrieval accuracy using early and late fusion. To compensate for the random effect we repeat this process 500 times. The event retrieval accuracy are reported in terms of average results of all random cases.

Experiment 3: Event retrieval vs event classification In this experiment we compare event retrieval with event classification when only a limited number of positive examples are available. We vary the number of video queries (or positive examples) from 1 to 20 by randomly sampling from our pool of positive query examples. For event retrieval we use the best performing fusion scheme from experiment 2. For event classification we employ a linear Support Vector Machine on top of each semantic signature, similar to [29]. We consider the video query examples per event jointly as positive examples and the other videos from the query set as negative examples. We also report results using an exemplar-SVM [70] trained on each individual positive query and all available negative examples, and then fused over all queries using average pooling. We measure event retrieval and event classification performance on the test set. We repeat this process 50 times for both event retrieval and event classification with different number of positive examples, and a fixed number of negatives for the event classification scenario. Note that we do not rely on negative examples for the event retrieval scenario.

Experiment 4: Interpretability of semantic signature We showcase three scenarios of the capabilities of semantic signatures in event retrieval:

1. **Video to video:** In this scenario the goal is to retrieve a video using semantic signatures where the query is a video example of an event. We show the ability of semantic signatures for providing a crude interpretation on why a certain video has been retrieved.
2. **Video to text:** In this scenario the goal is to show the ability of semantic signatures for generating a text for a video query. We assume that each video in the test set is associated to a text which describes the visual contents of the video, for example obtained from a YouTube video title caption. We suggest a text for the video query consisting of the most frequent terms of the text associated to the k visual nearest neighbor videos of the query.
3. **Text to video:** We demonstrate the capability of semantic signatures in video retrieval when the query is a text by comparing it's terms with the concepts in semantic signatures. We simply represent each video with semantic signatures and rank the video of the test set by computing the similarity score with the query.

Evaluation criteria The retrieval performance is measured in terms of the well known average precision (AP), which combines precision and recall into a single metric [110]. We also report the average retrieval performance over all events as the mean average precision (MAP).

Table 10: Experiment 2: Early vs Late Query Fusion. For the fusion we sample eight queries per iteration. The results are averaged over 25 events and repeated 500 times. Best MAP result in bold. We observe only a minimal variance for all events, therefore not shown.

Event category	Available Queries	Single Query	Early Query Fusion		Late Query Fusion	
			AVG	MAX	AVG	MAX
Attempting a board trick	98	0.062	0.123	0.058	0.135	0.117
Feeding an animal	95	0.023	0.056	0.052	0.069	0.030
Landing a fish	71	0.086	0.120	0.102	0.128	0.061
Wedding ceremony	69	0.078	0.232	0.089	0.294	0.037
Working on a woodworking project	79	0.026	0.092	0.086	0.087	0.037
Birthday party	121	0.055	0.151	0.184	0.159	0.054
Changing a vehicle tire	75	0.045	0.106	0.078	0.118	0.027
Flash mob gathering	115	0.175	0.254	0.224	0.305	0.115
Getting a vehicle unstuck	85	0.083	0.147	0.101	0.159	0.055
Grooming an animal	91	0.068	0.195	0.071	0.214	0.016
Making a sandwich	83	0.047	0.141	0.087	0.138	0.020
Parade	105	0.120	0.221	0.202	0.221	0.055
Parkour	75	0.082	0.231	0.121	0.219	0.037
Repairing an appliance	85	0.082	0.213	0.146	0.221	0.028
Working on a sewing project	86	0.046	0.094	0.052	0.101	0.042
Attempting a bike trick	43	0.057	0.108	0.101	0.110	0.046
Cleaning an appliance	43	0.024	0.069	0.070	0.072	0.051
Dog show	43	0.059	0.116	0.067	0.130	0.083
Giving directions to a location	43	0.022	0.102	0.010	0.114	0.006
Marriage proposal	43	0.012	0.019	0.009	0.025	0.005
Renovating a home	43	0.054	0.200	0.130	0.207	0.023
Rock climbing	43	0.042	0.071	0.103	0.065	0.019
Town hall meeting	43	0.036	0.060	0.113	0.055	0.076
Winning a race without a vehicle	43	0.065	0.083	0.074	0.085	0.121
Working on a metal crafts project	43	0.021	0.029	0.022	0.030	0.015
Mean average precision		0.059	0.129	0.095	0.138	0.047

4.4 RESULTS

Experiment 1: Single query baseline We present the results of experiment 1 in Table 9. While the overall MAP using a single query is modest, but always much better than random, the video representation using semantic signatures outperforms bag-of-words for all metrics. As expected, the semantic signatures can generalize better than the bag-of-words. Table 9 also confirms that the normalized correlation metric and two information theory based metrics, are better than the histogram intersection and chi-square metrics for computing the similarity between events, as predicted in Section 4.2.2. For the remaining experiments we consider the semantic signatures with normalized correlation.

Experiment 2: Early vs Late Query Fusion Table 10 presents the results of experiment 2. The results show considerable improvement in event retrieval performance when we use multiple queries instead of a single query (0.138 vs 0.059). For both fusion methods using the average operator is better than the maximum. We explain this by the fact that the average operator reduces the effect of noisy and irrelevant concepts that may occur accidentally in one or two event queries, as noted in Section 4.2.3. By contrast, the max operator is sensitive to irrelevant and noisy concepts scores with a high value. This results in a big drop in retrieval performance (from 0.138 to 0.047), even lower than the single query baseline. However, for some events such as *Rock climbing*, *Town hall meeting* and *Birthday party* early query fusion with the max operator results in the best retrieval performance. We attribute this to the presence of reliable relevant concepts in one of the queries. Indicating, that much is to be expected from adaptive semantic

Figure 19: Experiment 3: Event Retrieval vs Event Classification. When only a handful of semantic signature queries are available per event, classification is outperformed by retrieval.

Figure 20: Video to video: User interface of our video event retrieval based on semantic signatures, with a user marked, green border, example as a video query (left). The results of our event retrieval on unseen web video (right). Note that with only one positive event example, we can obtain proper event retrieval result. In the right side we also prepare a video interpreter which shows the semantic signature of retrieved video. The user may inspect why the video is retrieved by inspecting the signatures.

signatures that determine concept relevance at query time. The results of experiment 2 show that using multiple queries improves the performance of event video retrieval in comparison to a single query, especially for late query fusion. Moreover, we observe that in multiple query video retrieval the average operator is preferred for the fusion.

Experiment 3: Event Retrieval vs Event Classification We plot the result of experiment 3 in Figure 19. As expected the accuracy of both event retrieval and classification increases when more and more queries or positive event example are available. However, when we use only a limited number of queries, *i.e.*, from 1 to 8, we observe that the accuracy of event retrieval is higher than event classification. After increasing the number of queries, *i.e.*, from 8 to 20, we see

Figure 21: Video to text. The interface of our event video search engine for explaining video content. The upper part provides a summary of the video by plotting the histogram of its dominant concepts. This histogram can be used to explain why the video has been retrieved. The lower part of the figure, demonstrates video translation in terms of textual descriptions. Moreover, the tag-cloud indicates the most frequent terms extracted from the translations.

the difference in accuracy increasing in favor of event classification. Using the available positive examples jointly is a better choice than bagging them individually with an exemplar-SVM. The exemplar-SVM needs up to 16 positive examples (and a bunch of negative examples) before it outperforms event retrieval. We conclude that when more than eight positive examples are available per event, as well as a set of negative examples, it pays off to learn a classifier. However, for more realistic event video retrieval scenarios where only a handful of positive examples are available, it is advantageous to rely on late query fusion retrieval with an average operator.

Experiment 4: Interpretability of semantic signature

Video to video: In Figure 20 we show the user interface of video search engine using semantic signature when the query is a video example. The visual similarity between each video of the test set and the video query is computed by selecting the most dominant concepts within semantic signatures for the video query. To some extent, we can explain why a video has been retrieved for a specific video query by exploring it’s semantic signature.

Video to text: We show in Figure 21 the process of generating a text for a video query using semantic signatures. After finding visual nearest neighbor videos to the query, we propagate the sentences associated to them to the query and consider more frequent terms to generate a tag cloud which translate the query video. Figure 21 shows some examples of the retrieved descriptions for a video.

Text to video: In Figure 22 we demonstrate the capability of semantic signature for video event retrieval when the query is a text. The result shows the semantic signature are promising even in the absence of visual example of an event. Especially in this challenging scenario low-level video representations will fail.

Figure 22: Text to video. We automatically project a text query into a semantic signature, which is also used for representing video. Then we retrieved the videos using the similarity of their semantic signature with the binary signature of the text query and rank the video.

4.5 CONCLUSIONS

This chapter studies the behavior of multiple video queries for event retrieval, by performing three experiments on an unconstrained web video collection. The result of experiment 1 provides an indication that querying for event video using semantic signatures generalizes better than a low-level bag-of-words alternative. In addition, we find that normalized correlation is a suitable similarity metric when considering retrieval using semantic signatures. The results of experiment 2 shows that video retrieval using multiple event queries in combination with late query fusion and an average operator outperforms event retrieval using only a single query. With only a handful of examples the results increase from 0.059 to 0.138. A considerable improvement. The result in experiment 3 reveals that it is advantageous to rely on event retrieval rather than event classification, when the number of available video examples is less than eight. At the end we demonstrate a semantic video search engine which leverages semantic signatures to allow users to search video by video example, by event description and even offers a crude form of video translation.

TAGBOOK: A SEMANTIC VIDEO REPRESENTATION WITHOUT SUPERVISION FOR EVENT DETECTION

5.1 INTRODUCTION

The goal of this chapter is to detect events such as *dog show*, *felling a tree*, and *wedding dance* in arbitrary video content (Fig. 23). The topic of event detection has a long tradition in the discipline of multimedia, see [5, 42, 52] for recent surveys. Early works considered knowledge-intensive approaches using relatively little video data, *e.g.* [3, 9, 30, 54]. The state-of-the-art is to exploit big video data sets, such as the Columbia Consumer Video collection [44] and the TRECVID Multimedia event detection corpus [91], and to learn an event classifier from dozens of carefully labeled examples, *e.g.* [33, 85, 94, 127]. However, as events become more specific, the harder it will be to find sufficient relevant examples for learning, even on the socially-tagged web [14, 27, 124]. Different from the dominant strategy in the event detection literature, we consider in this paper event detection scenarios where video examples of the event are scarce, and even completely absent.

The key to event detection is to have a discriminative video representation. Traditional video representations for event detection rely on low-level audiovisual features. Often combining bag-of-words derived from SIFT descriptors, MFCC audio features and space-time interest points [41, 45, 83, 85, 92, 115], or by localizing temporal evidence by formulating the problem of video event detection as multiple instance learning in a low-level feature space [48, 49, 56]. Oneata *et al.* demonstrate the effectiveness and robustness of the improved dense trajectories with a Fisher vector encoding [93, 94]. Based on the great success in image recognition [47], learned representations derived from convolutional neural network layers are becoming popular for video event detection as well, *e.g.* [77, 127, 129]. Since most low-level representations have a high feature vector dimensionality, they reach good performance in the presence of sufficient positive training examples per event. However, the applicability of the low-level representation is limited when only a few positive event examples are available [68, 78]. Especially in event detection scenarios where only a description of an event is available, *i.e.* without any video training example, the low-level representations by themselves are more or less useless. To tackle the scarcity of positive examples, video samples which do not precisely describe an event but are still relevant to help detect the event are exploited in [128]. A novel joint training protocol is developed in [13] to simultaneously conduct event detection and recounting, where the recounting model assists detection by filtering out noisy irrelevant information. We take a direction orthogonal to these works, aiming to find a semantic video representation capable of detecting events in the presence of few examples and even zero examples.

Others have also studied semantic video representations in the context of few-example [26, 27, 68, 78] and zero-example [12, 14, 26, 39, 124] event detection. All these works build a semantic

---

Published in the *IEEE Transaction on Multimedia*, 2016. [80].

Published in the *Proceeding of the ACM International Conference on Multimedia Retrieval*, 2014. [79].

Figure 23: Example videos for the events dog show, felling a tree, and wedding dance. Despite the challenging diversity in visual appearance, each event maintains specific semantics in a consistent fashion. This paper studies whether an event representation based on tags assigned from the video’s nearest neighbors can be an effective semantic representation for few- and zero-example event detection.

representation on top of concept detectors such as ‘dog’, ‘tree’ and ‘groom’. Such an approach has become feasible to some extent thanks to the availability of thousand of concept annotations as part of the TRECVID benchmark [84, 110] and the ImageNet challenge [102], as well as social-tagged image and video resources [14, 27, 46, 124]. While this allows for few-example event detection indeed, the need for event examples has effectively been substituted for the even bigger problem of acquiring appropriate concept examples. Not to mention the computational demand for training the individual detectors. In contrast, we propose a new semantic video representation that is based only on freely available social tagged videos, without the need for training *any* concept detectors. Before detailing our approach on how to arrive at the new video representation, we first discuss in more detail related work on semantic video representations.

5.2 RELATED WORK

5.2.1 *Representations from supervised concepts*

There are good efforts for achieving a semantic representation by automatically recognizing concepts in a video’s audiovisual content. The standard approaches attempt to train a classifier per concept and use the corresponding classifier confidence values as the building block for a video representation, which in turn is leveraged for event detection, e.g. [8, 20, 22, 28, 36, 66, 75, 81, 114]. In [20] for example, Ebadollahi *et al.* employ 39 pre-defined concepts from the large scale concept ontology [84] for detecting events in broadcast news footage. Mazloom *et al.* [75] introduce a feature selection algorithm that learns the best concept representation for an event from a large bank of more than thousand concept detectors trained on ImageNet [102] and TRECVID [95]. Bhattacharya *et al.* [8] leverage the temporal dynamics of concept detector scores in their representation using linear dynamical system models. Naturally, a semantic representation can be mixed with a low-level one, as successfully shown by Ma *et al.* [69]. All these works rely on carefully annotated images or video fragments to arrive at their concept detectors. Since it is hard to determine a priori what concepts will be needed, we prefer a more flexible video representation that builds its representation by learning from many weakly annotated web videos, e.g., YouTube videos with social tags.

5.2.2 *Representations from weakly-supervised concepts*

Weakly supervised web resources have been explored by others as well. In [27], Habibian *et al.* harvest YouTube videos as a resource on which they base their representation. To accommodate for the ambiguity of the video descriptions they define a set of initial filters on the video collection, covering grammar and visualness of the descriptions, to assure the most reliable descriptions remain. To further alleviate the ambiguity an algorithm is proposed that learns an embedding of the joint video-description space. The embedding essentially groups several terms into topics to allow for a robust visual predictor, while maintaining descriptiveness. Rather than obtaining a semantic representation by training concepts over web video examples, Mazloom *et al.* [77] propose an algorithm that learns a set of relevant frames as the concept prototypes, without the need for frame-level annotations. Since the concept prototypes are a frame-level representation of concepts, they offer the ability of mapping each frame of a video into the concept prototype space, which can be leveraged for both few-example and zero-example event detection. Wu *et al.* [124] leverage off-the-shelf detectors as well as various video and image collections that come with textual descriptions to learn a large set of concept detectors using various multimedia features. To allow for zero-example detection, both the event description and concept detectors are mapped into the same textual space, in which their similarity is computed using the cosine distance. Chen *et al.* [14] also start from a set of events and their textual descriptions. They first extract tags deemed relevant for the events. After verifying that the tags are meaningful and visually detectable, each tag is used as query on a photo sharing website. By doing so the authors harvest 400,000 image examples to build a representation containing a total of 2,000 concept detectors. Similar to Chen *et al.* [14] we rely on social tagged media, be it that we focus on tagged videos as also used by Habibian *et al.* [27] and Mazloom *et al.* [77]. However, rather than building concept detectors from the tagged videos [14, 27, 77, 124], we prefer to use the tags directly for video representation.

5.2.3 Representations from tags

We are inspired by recent progress in socially tagged image retrieval [59], where many have demonstrated the value of tags for image retrieval. While it is well known that tags are often ambiguous, faulty, and incomplete, these limitations can be overcome to some extent by clever algorithms. Two representative and good performing [59] algorithms are neighbor voting by Li *et al.* [58] and TagProp by Guillaumin *et al.* [23]. Given an image, the neighbor voting algorithm first retrieves its nearest neighbors from a source set in terms of low-level visual similarity. To determine the relevance of each tag of the input image, the algorithm then simply counts the tag’s occurrence in annotations of the top- k most similar images. Apart from tag refinement, the algorithm can also be leveraged for tag assignment. In this scenario the tags from the neighbors are sorted in descending order in terms of their occurrence frequency, and the top ranked tags are propagated to the input image. Different from the neighbor voting algorithm which considers the neighbors equally important, TagProp assigns rank-based or distance-based weights to the individual neighbors such that tags from neighbors closer to the input image will be enhanced in the tag propagation process. Our solution is grounded on tag propagation similar to [23, 58], but takes two steps further to make it more suited for video event detection. One, instead of frame-level tag propagation as a straightforward application of [23, 58] to the video domain, we conduct video-level tag propagation. Since the number of videos is much smaller than the number of video frames, this design ensures good scalability of our solution to deal with large-scale video sets. Two, we conduct tag refinement on the weakly labeled training video set before using it as a resource for tag propagation. This resolves to some extent the inaccuracy and the incompleteness of social tags assigned to the source videos. As a consequence, more relevant tags will be propagated to the input video.

Propagating tags between videos has been studied in the context of tag recommendation [6, 109]. There, tags are meant to be used by end users, mainly for video browsing and retrieval. In contrast, we propagate tags for the purpose of using them as video representation for computing (cross-media) relevance between an unlabeled video and a specific event.

5.2.4 Contributions

Our work makes the following contributions. First of all, we propose a new semantic video representation for event detection using social tags that can be associated to videos. To the best of our knowledge, no method currently exists in the literature able to represent a video for event detection using just its tags, other than our previous conference paper [79]. It should be noted that [79] proposes a language model on top of the representation for video retrieval using query by zero, one or multiple positive examples. Here we prefer the parameter-free cosine distance for zero-shot event detection and exploit a support vector machine for the scenario where a few positive *and* many negative video examples are available to learn an event classifier. In addition, we introduce source set refinement, which differentiates between the tags of neighbor videos in advance to tag propagation. Consequently, we obtain an improved bag of tags per video by considering source set refinement and multiple tag assignment functions. We show the merit of our proposal by performing several experiments on more than 1,000 hours of arbitrary Internet videos from the TRECVID Multimedia Event Detection task 2013, 2014 and the Columbia Consumer Video dataset. We call our approach *TagBook*, and detail its construction for few-example and zero-example event detection next.

Table 11: Main notations defined in this work

Notation	Definition
v	a video
e	a video event
t	a tag
$\mathbf{b}(v)$	a tag vector of a given video
$\mathbf{b}(e)$	a tag vector of a given event
$\mathbf{b}_{few}(e)$	few-example version of $\mathbf{b}(e)$
$\mathbf{b}_{zero}(e)$	zero-example version of $\mathbf{b}(e)$
$f(v, e)$	a relevance function computed as $\cosine(\mathbf{b}(v), \mathbf{b}(e))$
$\mathcal{T}$	a vocabulary of m tags
$\mathcal{V}$	a set of unlabeled test videos
$\mathcal{V}_l$	a set of labeled video samples of a given event
$\mathcal{V}_s$	a set of socially tagged videos for tag propagation
$s(v, v')$	visual similarity between two videos
$\llbracket v_s, t \rrbracket$	a binary function indicating if $v_s \in \mathcal{V}_s$ is labeled with t
$r(v_s, t)$	the relevance score between $v_s \in \mathcal{V}_s$ and tag t

5.3 TAGBOOK BASED VIDEO EVENT DETECTION

5.3.1 Problem Formalization

Given a user specified event, video event detection is to retrieve videos showing the event from a large set of unlabeled videos. For the ease of consistent description, we use e to indicate the given event, v be a video, and $\mathcal{V} = \{v_1, \dots, v_n\}$ a test set of n videos. We aim to construct a real-valued function $f(v, e)$ which produces the relevance score between the video and the event. By sorting $\mathcal{V}$ according to $f(v, e)$ in descending order, videos most relevant with respect to the event will be obtained.

Let $\mathcal{V}_l = \{(v_{l,1}, y_1), \dots, (v_{l,p}, y_p)\}$ be a set of p labeled video samples available for a specific event, where $y_i = 1$ means positive samples and $y_i = -1$ for negative samples. The difficulty in constructing $f(v, e)$ largely depends on the size of $\mathcal{V}_l$. Here the amount of positive samples is our concern, as the occurrence of a specific event in a video collection tends to be rare, making the acquisition of positive samples much more expensive than obtaining negative samples. In practice, even finding a single sample could be tricky, and one has no other choice than to express the event in words.

We now describe more formally the two scenarios of video event detection, in an order of increasing difficulty:

1. *Few-example* video event detection: Finding videos relevant to a specific event e from $\mathcal{V}$, given $|\mathcal{V}_l| \geq 1$. Typically $\mathcal{V}_l$ has a handful of positive examples.
2. *Zero-example* video event detection: Finding videos relevant to a specific event e from $\mathcal{V}$, given $\mathcal{V}_l = \emptyset$. In this case, the event is described by a natural language sentence q .

The scarcity of video samples combined with the high dimensionality of low-level visual features makes it nontrivial to construct $f(v, e)$ effectively. Moreover, in the zero-example scenario, the visual features are inapplicable to compute cross-media similarity between a video and a description. To resolve these difficulties, we present TagBook, a compact and semantic representation of an entire video, which works for both scenarios.

Figure 24: A conceptual framework for generating a TagBook representation for an unlabeled video. From a source set of web videos annotated by online users we propagate tags from the visually similar video neighbors of the input video. Depending on how the neighbors are weighted and whether the tags of the source set are refined, we derive three variants of TagBook, i.e., TagBook-hard (equal neighbors and raw tags), TagBook-soft (weighted neighbors and raw tags), and TagBook-refine (weighted neighbors and refined tags).

The key idea of TagBook is to represent an unlabeled video v by a fixed-length tag vector, denoted as $\mathbf{b}(v)$. Let $\mathcal{T} = \{t_1, \dots, t_m\}$ be a vocabulary of m distinct tags used in the TagBook. Each dimension of the tag vector uniquely corresponds to a specific tag, where $\mathbf{b}(v, i)$ is the relevance score between the tag t_i and the video v . Hence, TagBook essentially embeds a video into an m -dimensional tag space.

Next, we show in Section 5.3.2 how to tackle video event detection using TagBook, followed by a solution to implement this representation in Section 5.3.3. For the ease of reference, Table 11 lists the main notation used throughout this work.

5.3.2 Two-Scenarios for Video Event Detection using TagBook

We explain how a specific event e can be represented as a TagBook. Let $\mathbf{b}(e)$ be the tag vector of an event. The relevance between this event and a video boils down to computing the cosine similarity between the two tag vectors. That is,

$$f(v, e) := \text{cosine}(\mathbf{b}(v), \mathbf{b}(e)). \quad (5.1)$$

Notice that we have also investigated other similarity metrics including the Euclidean distance, the Spearman rank correlation, the Jensen–Shannon divergence, the χ^2 distance, histogram intersection, and the Earth Mover’s Distance. Among them, the cosine similarity strikes the

best balance between effectiveness and efficiency. We use $\mathbf{b}_{few}(e)$ and $\mathbf{b}_{zero}(e)$ to indicate two variants corresponding to the few-example and zero-example scenarios, respectively.

In the few-example scenario, the event e is expressed in terms of p labeled video samples. Some of these samples could be more important than others for modeling the event. Hence, we consider the tag vector of the event as a weighted combination of its samples. In particular, we define:

$$\mathbf{b}_{few}(e) := \sum_{i=1}^p \alpha_i y_i \mathbf{b}(v_{l,i}), \quad (5.2)$$

where $\{\alpha_i\}$ are weight parameters. Notice that Eq. (5.2) bears high resemblance to the decision function of a linear Support Vector Machine (SVM). Hence, we optimize the weights by a linear SVM solver [107].

In the zero-example scenario, a textual description q of the event is provided. Using the classical bag-of-words model, q is converted to a tag vector. Accordingly, $\mathbf{b}_{zero}(e, i)$ is 1 if t_i is in q , and 0 otherwise.

With $\mathbf{b}(e)$ in Eq. (5.1) replaced by $\mathbf{b}_{few}(e)$ and $\mathbf{b}_{zero}(e)$ separately, we have the relevance functions $f_{few}(x, e)$ and $f_{zero}(x, e)$ for each of the two scenarios.

5.3.3 TagBook Construction by Content-Based Tag Propagation

We propose to construct the TagBook representation of an unlabeled video by propagating tags from a large set of N socially tagged videos, denoted by $\mathcal{V}_s = \{v_{s,1}, \dots, v_{s,N}\}$. Each video $v_s \in \mathcal{V}_s$ is assigned with a limited number of social tags. For each tag $t \in \mathcal{T}$, we use a binary labeling function $\llbracket v_s, t \rrbracket$, which outputs 1 if v_s is labeled with t , and 0 otherwise. Due to the subjective nature of social tagging, some of the assigned tags could be irrelevant with respect to the visual content of v_s .

With the hypothesis that visually similar images shall have similar tags, content-based tag propagation has been exploited in the context of image auto-tagging [23, 58]. Tags are propagated from neighbor images which are visually close to a test image, where the neighbors are treated either equally [58] or weighted in terms of their visual distance to the test image [23]. In our context, let $\{\hat{v}_{s,1}, \dots, \hat{v}_{s,k}\}$ be the k nearest neighbor videos retrieved from $\mathcal{V}_s$ by a predefined video similarity $s(v, v')$. A general formula of tag propagation can be expressed as

$$\mathbf{b}(v, i) = \frac{1}{k} \sum_{j=1}^k h(s(v, \hat{v}_{s,j})) \cdot r(\hat{v}_{s,j}, t_i), \quad (5.3)$$

where $r(v_s, t)$ measures the relevance of a specific tag t with respect to a specific video $v_s \in \mathcal{V}_s$. To simplify our notation, we abuse $s(v, v')$ to let it also indicate the contribution of a neighbor video in the tag propagation process. For instance, in a hard assignment mode, the output of $s(v, v')$ will be binary, producing 1 if the rank of the neighbor is within k , and 0 otherwise. Tags of higher occurrence in $\mathcal{V}_s$ are more likely to be propagated. In order to reduce such an effect, we subtract $\mathbf{b}(v, i)$ by a term related to tag occurrence, i.e.,

$$\begin{aligned} \mathbf{b}(v, i) &= \frac{1}{k} \sum_{j=1}^k h(s(v, \hat{v}_{s,j})) \cdot r(\hat{v}_{s,j}, t_i) \\ &\quad - \frac{1}{N} \sum_{j=1}^N h(s(v, v_{s,j})) \cdot r(v_{s,j}, t_i) \end{aligned} \quad (5.4)$$

Concerning $r(v_s, t)$, a straightforward choice is to instantiate it using the labeling function $\llbracket v_s, t \rrbracket$. As aforementioned, this choice is questionable due to the inaccuracy and sparseness of

social tags. We therefore conduct tag refinement on the source set $\mathcal{V}_s$ *before* using it for TagBook construction. Again, tag propagation is employed, computing $r(v_s, t)$ by

$$r(v_s, t) = \frac{1}{k_r} \sum_{j=1}^{k_r} s(v_s, \tilde{v}_{s,j}) \cdot \llbracket \tilde{v}_{s,j}, t \rrbracket - \frac{1}{N} \sum_{j=1}^N s(v_s, v_{s,j}) \cdot \llbracket v_{s,j}, t \rrbracket \quad (5.5)$$

where $\{\tilde{v}_{s,1}, \dots, \tilde{v}_{s,k}\}$ are the k_r nearest neighbors of v_s retrieved from the source set. Both k and k_r are empirically set to be 500.

Concerning the content-based similarity between two videos $s(v, v')$, we use convolutional neural network (CNN) features for their well recognized performance. In particular, we train the AlexNet [47] for over 15k ImageNet classes, each having at least 50 positive examples. Given a video, we extract its frames uniformly with a time interval of two seconds. The second fully connected layer (FC2) is used, representing each frame with a 4,096-dimensional feature vector. The video-level feature vector is obtained by average pooling over all the frame-level vectors. The video similarity $s(v, v')$ is computed as the cosine similarity between the corresponding CNN feature vectors.

Depending on how the weights of the neighbors and $r(v_s, t)$ are implemented, we present three variants of TagBook’s tag assignment, namely

1. TagBook-hard: Neighbor videos are assigned with binary weights, i.e., 1 if the rank of the neighbor is within k , and 0 otherwise, and $r(v_s, t)$ as $\llbracket v_s, t \rrbracket$.
2. TagBook-soft: Neighbor videos are weighted in terms of their similarity scores, and $r(v_s, t)$ as $\llbracket v_s, t \rrbracket$. TagBook-soft corresponds to the representation used in [79].
3. TagBook-refine: Neighbor videos are weighted in terms of their similarity scores, and $r(v_s, t)$ as Eq. (5.5).

Fig. 24 illustrates the TagBook generation process. Next, we evaluate TagBook-hard, TagBook-soft, and TagBook-refine for video event detection on three benchmark datasets.

5.4 EXPERIMENTS

5.4.1 Datasets

Source set. As our social-tagged video collection $\mathcal{V}_s$, we adopt the VideoStory46K dataset from Habibiian *et al.* [27] which contains 46k videos from YouTube. Every video has a short caption provided by the person who uploaded the video. From the captions we remove stop words and words not visually detectable such as God (we used the visualness filter from [27]) and finally obtain a vocabulary $\mathcal{T}$ of 19,159 unique tags.

Test datasets 1 and 2: TRECVID MED 2013 and MED 2014 [91]. The MED corpus contains user-generated web videos with a large variation in quality, length and content of real-world events including life events, instructional events, sport events, etc. Both the 2013 and 2014 corpus consist of several partitions with ground truth annotation at video-level for 30 event categories, with 10 of those events overlapping in both 2013 and 2014. For the few-example scenario, we follow the TRECVID 10Ex evaluation procedure [95]. That is, for each event its training data $\mathcal{V}_l$ contains 10 positive video samples from the Event Kit training data, and 5K negative video samples from the Background training data. In the zero-example scenario, we rely

Table 12: Comparing three variants of TagBook on TRECVID MED 2013. For each scenario, top performers per event are highlighted in bold font.

Event	Zero-example			Few-example		
	TagBook-hard	TagBook-soft	TagBook-refine	TagBook-hard	TagBook-soft	TagBook-refine
Birthday party	0.028	0.051	0.065	0.099	0.136	0.149
Changing a vehicle tire	0.081	0.108	0.125	0.278	0.402	0.466
Flash mob gathering	0.145	0.194	0.221	0.294	0.372	0.399
Getting a vehicle unstuck	0.198	0.211	0.235	0.499	0.547	0.587
Grooming an animal	0.046	0.066	0.095	0.101	0.165	0.201
Making a sandwich	0.019	0.021	0.036	0.038	0.040	0.076
Parade	0.192	0.201	0.204	0.210	0.228	0.230
Parkour	0.094	0.100	0.109	0.229	0.308	0.334
Repairing an appliance	0.200	0.277	0.298	0.256	0.376	0.381
Working on a sewing project	0.034	0.031	0.027	0.087	0.066	0.072
Attempting a bike trick	0.029	0.067	0.087	0.083	0.146	0.199
Cleaning an appliance	0.008	0.004	0.019	0.016	0.009	0.028
Dog show	0.125	0.084	0.091	0.232	0.121	0.143
Giving directions to a location	0.003	0.006	0.006	0.006	0.008	0.009
Marriage proposal	0.002	0.002	0.003	0.005	0.007	0.009
Renovating a home	0.008	0.010	0.013	0.023	0.028	0.044
Rock climbing	0.043	0.021	0.026	0.124	0.084	0.098
Town hall meeting	0.094	0.071	0.077	0.182	0.165	0.177
Winning a race without a vehicle	0.011	0.055	0.071	0.171	0.232	0.275
Working on a metal crafts project	0.003	0.006	0.007	0.014	0.050	0.067
MAP	0.068	0.079	0.091	0.148	0.174	0.198

Figure 25: TagBook-refine vs Other TagBooks for video description generation, tested on TRECVID MED 2013. TagBook-refined generates more accurate descriptions.

only on the TRECVID provided textual definition of a test event. For both scenarios we report results on the MED 2013 test set and the MED 2014 test set, each containing 27K videos.

Test dataset 3: Columbia CV [44]. This corpus consists of 9,317 YouTube videos, and crowd-sourced ground truth with respect to 20 visual concepts. Fifteen of the concepts correspond

Figure 26: Video examples along with their expert-provided description for the events birthday party (top) and grooming an animal. Tags predicted by TagBook-refine (left), TagBook-soft (middle), and TagBook-hard (right) are summarized as tag clouds. Tags generated by TagBook-refine tend to result in the best overlap with the ground truth (see Fig. 25).

to specific video events such as *Ice skating*, *Birthday*, and *Music performance*, so only these event-related concepts are considered in our experiments. We start from the official data partition, i.e., a training set of 4,625 videos and a test set of 4,637 videos. For few-example event detection, similar to Habibian et al. [27] we down-sample the training set to have at most 10 positive training examples per event, which are obtained based on the alphabetical order of the video names. Different from the TRECVID datasets, the Columbia CV dataset does not provide textual definition of events. So we do not perform zero-example video event detection on this dataset.

In what follows, we first use the MED 2013 dataset to find a good implementation of TagBook, achieved by evaluating varied choices including refining annotations of the source set, visual neighbor re-weighting, and the TagBook size. To study whether a more complex model with higher non-linear capability would help improve the accuracy of event detection, we compare in the few-example setting the linear model and a non-linear variant, reporting both speed and accuracy. To assess how the learned implementation generalizes to new test data, we evaluate it using the other two test sets, with a comparison to several state-of-the-art video representations.

As performance metrics, Average Precision (AP) per event and Mean Average Precision (MAP) per dataset are reported.

5.4.2 Experiment 1: Finding a Good TagBook

Table 12 gives the performance of TagBook-hard, TagBook-soft, and TagBook-refine on the MED 2013 test set. TagBook-soft performs better than TagBook-hard, with 0.079 *versus* 0.068

Figure 27: The influence of the TagBook size on (a) few-example and (b) zero-example video event detection. Compared to the full-sized TagBook, TagBooks consisting of around 2,000 most frequent tags yield the best performance on the TRECVID MED 2013 test set. Since Conceptlet and PCA require visual examples, they are inapplicable in the zero-example scenario.

for zero-example video event detection and 0.174 *versus* 0.148 in the few-example scenario. TagBook-refine performs the best, scoring MAP of 0.091 and 0.198 in zero-example and few-example, respectively. Recall that the only difference between TagBook-hard and TagBook-soft is that the latter re-weights neighbor videos in terms of their visual similarity to a test video, and the only difference between TagBook-soft and TagBook-refine is that the latter uses enriched annotations of the source set. The result shows the joint use of source set refinement and neighbor re-weighting is beneficial for extracting a better TagBook representation from unlabeled videos.

We make a further comparison between the three variants of TagBook to see how well they describe a video. Given a test video, its TagBook based description is automatically generated by sorting tags in terms of their $\mathbf{b}(v, i)$ and keeping the top κ ranked tags. We report the result of video description generation on the positive videos of each event class for which expert-provided descriptions are available. Following the protocol of [27], we use ROUGE-1, a performance metric computing the recall of the ground truth words in the generated description, thus increasing along with κ . The performance curve is shown in Fig. 25, with real examples in Fig. 26. Both figures demonstrate that TagBook-refine generates more accurate video descriptions.

To assess the effect of the TagBook size on video event detection performance, we investigate three dimension reduction methods. The first and the most straightforward method is to preserve the top frequent tags in the source set. We term it *Frequent tags*. The second is the classical Principal Component Analysis (PCA). The last is Conceptlet [75], a state-of-the-art concept selection algorithm, aiming for the best subset of concepts per event by considering correlations between concepts. Notice that PCA and Conceptlet require positive video examples, making them inapplicable in the zero-example scenario. As shown in Figure 27, for all the three methods, size-reduced TagBooks score higher MAP than the full-sized TagBook. In particular, peak performance is reached at the size of 2,000 for the few-example case, and 2,500 for the one-example case. In the remaining part of the evaluation, we use TagBook-refine reduced by the *Frequent tags* method, for its good performance, simplicity, and applicability for both scenarios.

Finally, we also assessed the scenario where TagBook-refine relies on a non-linear χ^2 feature embedding [120] rather than a linear kernel. As Table 13 shows, few-example event detection with TagBook-refine profits from a non-linear kernel at the expense of an increased computation

Table 13: Linear vs non-linear kernel, using TagBook-refine with varying sizes, for few-example event detection on TRECVID MED 2013. We report the time needed for training models and testing them for all events, measured in seconds on Intel Xeon Processor E5-2690. The non-linear kernel is more effective at the expense of an almost ten-fold increase in computation time on average.

TagBook size	Linear			Nonlinear		
	MAP	Training time	Test time	MAP	Training time	Test time
1,000	0.209	9.01	6.00	0.224	55.11	39.89
1,500	0.219	9.88	7.12	0.235	81.00	69.00
2,000	0.225	14.02	8.98	0.244	114.03	85.98
2,500	0.218	16.01	10.99	0.247	141.86	116.14

Table 14: TagBook versus others on TRECVID MED 2014.

Event	Few-example				Zero-example		
	CNN-FC2	ConceptVec-15k	ConceptVec-2k	TagBook	ConceptVec-15k	ConceptVec-2k	TagBook
Attempting a bike trick	0.057	0.127	0.134	0.139	0.016	0.042	0.075
Cleaning an appliance	0.022	0.062	0.072	0.119	0.014	0.071	0.080
Dog show	0.215	0.361	0.271	0.312	0.016	0.162	0.157
Giving directions to a location	0.013	0.051	0.030	0.032	0.003	0.004	0.006
Marriage proposal	0.003	0.005	0.008	0.008	0.008	0.005	0.005
Renovating a home	0.022	0.050	0.050	0.083	0.016	0.046	0.047
Rock climbing	0.066	0.061	0.101	0.089	0.005	0.008	0.020
Town hall meeting	0.268	0.212	0.204	0.228	0.008	0.102	0.120
Winning a race without a vehicle	0.126	0.121	0.130	0.175	0.017	0.087	0.063
Working on a metal crafts project	0.038	0.037	0.082	0.072	0.003	0.006	0.005
Beekeeping	0.410	0.525	0.461	0.502	0.062	0.003	0.009
Wedding shower	0.074	0.044	0.072	0.117	0.021	0.012	0.035
Non-motorized vehicle repair	0.228	0.407	0.415	0.398	0.003	0.074	0.265
Fixing musical instrument	0.077	0.085	0.097	0.114	0.008	0.003	0.009
Horse riding competition	0.390	0.280	0.344	0.392	0.124	0.075	0.118
Felling a tree	0.030	0.100	0.086	0.118	0.006	0.021	0.072
Parking a vehicle	0.111	0.231	0.088	0.119	0.198	0.011	0.035
Playing fetch	0.007	0.033	0.031	0.055	0.005	0.020	0.035
Tailgating	0.110	0.149	0.136	0.176	0.005	0.003	0.006
Tuning musical instrument	0.040	0.079	0.055	0.079	0.022	0.006	0.009
MAP	0.115	0.151	0.141	0.166	0.028	0.038	0.059

time. The mean average precision tends to be about 10% higher for the non-linear kernel, but computation is also about ten times as much. In the remaining few-example event detection experiments we rely on the linear kernel for its good accuracy and efficiency tradeoff.

5.4.3 Experiment 2: TagBook versus Others

We compare TagBook with several state-of-the-art video representations for event detection.

1. *CNN-FC2*. This representation has been described in Section 5.3 for finding similar videos.

2. *ConceptVec-15k*. For each sampled frame of a specific video, instead of the CNN-FC2 layer we adopt the output of the AlexNet’s softmax layer. The output is a nonnegative vector, where each dimension corresponds to one of the 15k ImageNet concepts and its value is a probabilistic estimation of the concept present in the frame. Average pooling is used to obtain the video-level representation.

3. *ConceptVec-2k*. As aforementioned, the TagBook is essentially constructed by neighbor voting based on tag propagation. One might consider using more advanced mode-based techniques such as SVMs. To address this concern, for each of the top 2k most frequent tags in our source set, we learn a separate linear SVMs classifier with CNN-FC2 as the underlying feature. A video

in the source set is taken as positive training examples if its caption contains the tag, and used as negatives otherwise. By applying the classifiers, each video is represented by a 2k vector of concept detector outputs.

4. *VideoStory* [27]. This video event representation strives to embed the caption of a video and its visual features in a joint space by grouping tags. We follow the author suggested implementation [27], which encodes each video as a Fisher vector over MBH descriptors along the motion trajectories. We learn the joint embedding from the source set with an optimal target dimensionality of 2,048.

5. *Concept Prototypes* [77]. Video event representation that learns a set of relevant frames as the concept prototypes and uses the prototypes for representing a video. We follow the author suggested implementation [77], which first encodes each video frame as CNN-FC2, and maps it to a concept prototype space learned for 479 concepts.

Except for CNN-FC2, all video representations are semantic and can therefore be used in both few-example and zero-example scenarios. For fair comparisons, the same event modeling technique, i.e., Eq. (5.2), is used, making the choice of video representation the only variable. This setting allows us to precisely identify which representation is the best.

The performance of video event detectors built on the varied representations is summarized in Table 17, Table 14, and Table 15, corresponding to TRECVID MED 2013, TRECVID MED 2014, and CCV, respectively. We directly cite AP scores from the original papers whenever applicable. Consequently, the results of VideoStory and Concept Prototypes are only partially available.

TagBook outperforms its competitors on all the three test datasets. In particular, as TagBook is built on top of CNN-FC2, its superior performance shows that TagBook is a more compact yet more semantic enriched video representation than the CNN feature.

For the model-based video representations, we observe that ConceptVec-2k is better than ConceptVec-15k in general. The main reason is that the ImageNet classes emphasize image objects, many of which are fine-grained classes of animals and plants. They are not meant for describing video events. By contrast, the source set from [27] was collected from YouTube using event-like descriptions as queries. Learned from such data, ConceptVec-2k is more suited than ConceptVec-15k for video event detection.

TagBook is better than ConceptVec-2k, although they use the same source set and the same visual feature as their starting point. The main technical difference between TagBook and ConceptVec-2k is that the former is built in a model-free manner while the latter is model-based. User tags are known to be subjective and ambiguous, meaning large divergence in their imagery. Model-free approaches as neighbor voting can figure out a decision boundary much more complex than linear classifiers, making it more suited for addressing subjective tags. Besides, model-based approaches are more sensitive to noise. Because of these reasons, model-free approaches like neighbor voting are more effective for learning from user-tagged video data.

TagBook also compares favorably against VideoStory [27] and Concept Prototypes [77]. Recall that they all use the VideoStory46K dataset as their source set. Both VideoStory and Concept Prototypes trust the (weak) annotations and learn their representation directly on top of the dataset. TagBook, in contrast, enriches the source set first by suppressing noise and generating more relevant tags per video. Moreover, TagBook considers a weight per tag, rather than a binary presence or absence value.

We also compared against our previous work [79], which relies on TagBook-soft and a language model for retrieval. For fair comparison we use the same CNN features and the same source set. On TRECVID MED 2013, TagBook-refine improves over [79] for few-example event detection from 0.221 to 0.225 and for zero-example from 0.113 to 0.129

Finally, we make a system level comparison between the proposed TagBook based system and several state-of-the-art alternatives for zero-example video event detection. The results

Table 15: TagBook versus others on Columbia Consumer Video.

Event	CNN-FC2	ConceptVec-15k	ConceptVec-2k	VideoStory	TagBook
Basketball	0.466	0.515	0.547	0.553	0.633
Baseball	0.551	0.608	0.563	0.299	0.594
Soccer	0.507	0.504	0.546	0.505	0.574
Ice skating	0.580	0.700	0.769	0.675	0.722
Skiing	0.745	0.794	0.796	0.671	0.796
Swimming	0.719	0.665	0.755	0.764	0.762
Biking	0.435	0.435	0.507	0.561	0.621
Graduation	0.261	0.295	0.278	0.121	0.290
Birthday	0.330	0.292	0.502	0.257	0.492
Wedding reception	0.214	0.174	0.161	0.117	0.196
Wedding ceremony	0.463	0.412	0.439	0.324	0.454
Wedding dance	0.399	0.296	0.423	0.521	0.503
Music performance	0.291	0.317	0.289	0.201	0.385
Non-music performance	0.188	0.240	0.226	0.282	0.289
Parade	0.487	0.354	0.512	0.634	0.521
MAP	0.442	0.440	0.487	0.432	0.522

Table 16: A system level comparison to the state-of-art for zero-example video event detection on TRECVID MED 2013.

System	MAP
Chen <i>et al.</i> [14]	0.024
Habibian <i>et al.</i> [26]	0.063
Ye <i>et al.</i> [129]	0.089
Chang <i>et al.</i> [12]	0.096
Jiang <i>et al.</i> [39]	0.101
Mazloom <i>et al.</i> [77]	0.119
This chapter	0.129

shown in Table 16 again confirms the effectiveness of the TagBook as a new video representation for event detection. In the most recent TRECVID MED evaluations from 2014 and 2015, other approaches have proven effective for few-example and zero-example event detection as well [111, 131]. In [131] the Informedia team from Carnegie Mellon University showed how a mixture of multimodal features, concepts and fusion schemes leads to state-of-the-art few-example event detection results. In addition, they have repeatedly demonstrated that pseudo-relevance feedback improves zero-example event detection [40]. In [111] the MediaMill team from the University of Amsterdam proposed a better CNN video feature by enriched pretraining, leading to state-of-the-art results for few-example event detection. The PROGRESS set used in the TRECVID MED benchmark is for blind testing by NIST only, so we cannot compare directly, but we note that TagBook will profit from more discriminative and multimodal representations as well and is orthogonal to pseudo-relevance feedback.

5.5 CONCLUSIONS

This chapter proposes TagBook, a new semantic video representation for video event detection. TagBook is based on freely available socially tagged videos, without the need for training any intermediate concept detectors. We introduce an algorithm that propagates tags to unlabeled videos from many socially tagged videos. The algorithm is inspired by image neighbor voting, but is improved by refining the source set, i.e., removing existing noisy tags and generating new tags, before tag propagation. Experiments on the TRECVID 2013 and 2014 multimedia event detection datasets and the Columbia Consumer Video dataset show that TagBook outperforms the current state-of-the-art semantic video representations for both zero- and few-example video event detection.

Table 17: TagBook versus others on TRECVID MED 2013.

Event	Few-example					Zero-example				
	CNN-FC2	ConceptVec-15k	ConceptVec-2k	VideoStory [27]	Concept Prototypes [77]	TagBook	ConceptVec-15k	ConceptVec-2k	Concept Prototypes [77]	TagBook
Birthday party	0.137	0.114	0.156	0.118	0.188	0.182	0.022	0.075	0.154	0.155
Changing a vehicle tire	0.391	0.388	0.411	0.103	0.464	0.560	0.099	0.181	0.320	0.337
Flash mob gathering	0.405	0.347	0.421	0.535	0.439	0.317	0.104	0.178	0.271	0.174
Getting a vehicle unstuck	0.334	0.323	0.436	0.319	0.418	0.602	0.107	0.201	0.406	0.312
Grooming an animal	0.084	0.108	0.149	0.151	0.154	0.247	0.019	0.101	0.095	0.201
Making a sandwich	0.031	0.074	0.087	0.074	0.131	0.108	0.021	0.031	0.164	0.099
Parade	0.171	0.109	0.271	0.452	0.303	0.279	0.094	0.135	0.240	0.185
ParLOUR	0.330	0.309	0.378	0.721	0.326	0.467	0.020	0.131	0.112	0.215
Repairing an appliance	0.169	0.127	0.261	0.184	0.244	0.395	0.078	0.157	0.213	0.211
Working on a sewing project	0.058	0.071	0.107	0.151	0.109	0.126	0.016	0.036	0.089	0.098
Attempting a bike trick	0.054	0.030	0.123	0.061	0.144	0.200	0.017	0.067	0.061	0.066
Cleaning an appliance	0.021	0.019	0.035	0.078	0.055	0.038	0.006	0.019	0.026	0.023
Dog show	0.232	0.134	0.254	0.354	0.313	0.243	0.003	0.155	0.011	0.200
Giving directions to a location	0.012	0.005	0.011	0.004	0.022	0.013	0.004	0.004	0.008	0.005
Marriage proposal	0.002	0.002	0.009	0.004	0.004	0.007	0.004	0.002	0.005	0.003
Renovating a home	0.019	0.024	0.046	0.051	0.033	0.053	0.017	0.011	0.026	0.018
Rock climbing	0.070	0.063	0.127	0.100	0.110	0.097	0.003	0.020	0.036	0.026
Town hall meeting	0.268	0.201	0.200	0.118	0.290	0.236	0.008	0.087	0.148	0.099
Winning a race without a vehicle	0.150	0.126	0.153	0.217	0.182	0.245	0.012	0.045	0.101	0.099
Working on a metal crafts project	0.054	0.068	0.099	0.118	0.144	0.079	0.002	0.005	0.014	0.002
MAP	0.150	0.132	0.188	0.196	0.204	0.225	0.032	0.081	0.119	0.129

SUMMARY AND CONCLUSION

6.1 SUMMARY OF IN SEARCH OF VIDEO EVENT SEMANTICS

In this thesis we aim to represent an event in a video using semantic features. We start from a bank of concept detectors for representing events in video. At first we considered the relevance of concepts to the event inside the video representation. Then, we concentrated on the accuracy of concept detectors. Finally, we consider the problem of searching video events with and without semantic concepts. Before reviewing the questions considered in the introduction, we will give a brief overview of the chapters.

In Chapter 2, we address the problem of video event classification using a bank of concept detector scores. Different from existing work, which simply relies on a bank containing all available detectors, we propose an algorithm that learns from examples what concepts in a bank are most informative per event, which we call the conceptlets. We model finding the conceptlets out of a large set of concept detectors as an importance sampling problem. Our proposed approximate algorithm finds the optimal conceptlets using a cross-entropy optimization. We study the behavior of video event classification based on conceptlets by performing four experiments on challenging internet video from the 2010 and 2012 TRECVID multimedia event detection tasks and Columbia’s consumer video dataset. Starting from a concept bank of more than thousand pre-computed detectors, our experiments establish there are (sets of) individual concept detectors that are more discriminative and appear to be more descriptive for a particular event than others, event classification using an automatically obtained conceptlet is more robust than using all available concepts, and conceptlets obtained with our cross-entropy algorithm are better than conceptlets from state-of-the-art feature selection algorithms. What is more, the conceptlets make sense for the events of interest, without being programmed to do so.

In Chapter 3 we propose a new semantic video representation for few and zero example event detection and unsupervised video event summarization. Different from existing works, which obtain a semantic representation by training concepts over images or entire video clips, we propose an algorithm that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations, and use them for representing an event video. We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames from both positive and negative training videos of a target concept. We study the behavior of our video event representation based on concept prototypes by performing three experiments on challenging web videos from the TRECVID 2013 multimedia event detection task and the MED-summaries dataset. Our experiments establish that i) Event detection accuracy increases when mapping each video into concept prototype space. ii) Zero example event detection increases by analyzing each frame of a video individually in concept prototype space, rather than considering the holistic videos. iii) Unsupervised video event summarization using concept prototypes is more accurate than using video-level concept detectors.

In Chapter 4 we aim at querying web videos for complex events using only a handful of video query examples, where the standard approach learns a ranker from hundreds of examples. We consider a semantic signature representation, consisting of off-the-shelf concept detectors, to capture the variance in semantic appearance of events. Since it is unknown what similarity metric and query fusion to use in such an event retrieval setting, we perform three experiments on unconstrained web videos from the TRECVID 2012 event detection task dataset. It reveals that retrieval with semantic signatures using normalized correlation as similarity metric is more accurate and faster than a low-level bag-of-words alternative, multiple queries are best combined using late fusion with an average operator, and event retrieval is preferred over event classification when less than eight positive video examples are available. We demonstrate a capability to yield semantic signature in a video search engine. We show how the semantic signatures provide a crude interpretation on why a certain video has been retrieved.

The aim of the Chapter 5 is event detection in video for scenarios where only few, or even zero example is available for training. For this challenging setting, the prevailing solutions in literature rely on a semantic video representation obtained from thousands of pre-trained concept detectors. Different from existing work, we propose a new semantic video representation that is based on freely available social tagged videos only, without the need for training any intermediate concept detectors. We introduce a simple algorithm that propagates tags from a video's nearest neighbors, similar in spirit to the ones used for image retrieval, but redesign it for video event detection by including video source set refinement and varying the video tag assignment. We call our approach *TagBook* and study its construction, descriptiveness and detection performance on the TRECVID 2013 and 2014 multimedia event detection datasets and on the Columbia Consumer Video dataset. Despite its simple nature, the proposed TagBook video representation is remarkably effective for few-example and zero-example event detection, even outperforming very recent state-of-the-art alternatives building on supervised representations.

6.2 GENERAL CONCLUSION

The aim of this thesis was to study semantic representations of video that could lead to a better event search. Several aspects of this problem have been explored, each of them contribute to answer the fundamental question of the thesis: *How to represent events for video search?*

We began with the question: *what concepts matter for an event?* Our findings support the conclusion that using all concepts in a bank of concepts for representing an event leads to a sub-optimal representation. By treating concept selection from the importance sampling perspective, we are able to discover few concepts the ones that are particularly relevant for representing the event. By emphasizing the concepts relevant to the event, we reach to a more precise event representation.

The relevancy of a concept to an event is a necessary but not a sufficient condition. Hence, we posed our second question: *what concept frames matter for an event?* We expressed the problem in terms of multiple instance learning. By learning a set of relevant frames of videos to the concepts, we reach to a more precise video event representation. The method is suited for video event summarization. Encouraged by the result of video event detection using semantic concepts, we then investigated the question *how to search events with concepts?* Using a semantic video representation, retrieval is performed at a much higher level of abstraction. This allows for better event search, especially when the events are semantically diverse as shown in Figure 2 in chapter 1. Last, we questioned *How to search events without concepts?* We propose a representation by propagating tags from nearest neighbor videos. The results showed that this is possible, even outperforming representations building upon concept detectors. Search of video event semantics

has ended by bringing us further to the ideal of automated solutions that process and memorize events as humans do.

BIBLIOGRAPHY

- [1] D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In *AISTATS*, 1995.
- [2] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.
- [3] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *TMM*, 4(1):68–75, 2002.
- [4] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Video event classification using string kernels. *MTAP*, 48(1), 2010.
- [5] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *MTAP*, 51, 2011.
- [6] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Enriching and localizing semantic tags in internet videos. In *MM*, 2011.
- [7] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li. ImageNet large scale visual recognition challenge 2011. <http://www.image-net.org/challenges/LSVRC/2011>.
- [8] S. Bhattacharya, M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events exploiting temporal dynamics between underlying concepts. In *CVPR*, 2014.
- [9] A. Bonzanini, R. Leonardi, and P. Migliorati. Event recognition in sport programs using low-level motion indices. In *ICME*, 2001.
- [10] J. A. Bucklew. *Introduction to rare event simulation*. Springer, 2004.
- [11] E. F. Can and R. Manmatha. Modeling concept dependencies for event detection. In *ICMR*, 2014.
- [12] 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 *Proc. of IJCAI*, 2015.
- [13] X. Chang, Y.-L. Yu, Y. Yang, and A. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *Proc. of ACM MM*, 2015.
- [14] 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.
- [15] Y. Cui, D. Liu, J. Chen, and S.-F. Chang. Building a large concept bank for representing events in video. *arXiv preprint arXiv:1403.7591*, 2014.
- [16] X. T. Dahua Lin. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *ECCV*, 2006.
- [17] 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.
- [18] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. Beyond audio and video retrieval: Towards multimedia summarization. In *ICMR*, 2012.
- [19] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *PAMI*, 34(9), 2012.
- [20] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *ICME*, 2006.
- [21] S. Finkelstein. *Understanding the gift of endless memory*. 60 Minutes, CBS Interactive, 2011.
- [22] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *CBMI*, 2011.
- [23] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.
- [24] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3, 2003.
- [25] A. Habibian, M. Mazloom, and C. G. M. Snoek. On-the-fly video event search by semantic signatures. In *ICMR*, 2014.
- [26] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

- [27] A. Habibian, T. Mensink, and C. G. M. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *MM*, 2014.
- [28] A. Habibian and C. G. M. Snoek. Recommendations for recognizing video events by concept vocabularies. *CVIU*, 124:110–122, 2014.
- [29] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.
- [30] N. Haering, R. Qian, and I. Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *CSVT*, 2000.
- [31] 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. *TMM*, 9(5), 2007.
- [32] B. Huurink, K. Hofmann, and M. de Rijke. Assessing concept selection for video retrieval. In *MIR*, 2008.
- [33] N. Inoue et al. TokyoTech+Canon at TRECVID 2011. In *NIST TRECVID Workshop*, 2011.
- [34] I. Inza, P. Larranaga, and B. S. R. Etxeberria. Feature subset selection by bayesian network-based optimization. *AI*, 2000.
- [35] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *PAMI*, 22(8), 2000.
- [36] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.
- [37] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8), 1996.
- [38] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 2012.
- [39] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.
- [40] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. Hauptmann. Fast and accurate content-based semantic search in 100m Internet videos. In *Proc. of ACM MM*, 2015.
- [41] Y. G. Jiang. Super: towards real-time event recognition in internet videos. In *ICMR*, 2012.
- [42] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMR*, 2013.
- [43] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *TMM*, 12(1), 2010.
- [44] 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.
- [45] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.
- [46] S. Kordumova, X. Li, and C. G. M. Snoek. Best practices for learning video concept detectors from social media examples. *MTAP*, 74(4):1291–1315, 2015.
- [47] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [48] K.-T. Lai, D. Liu, M.-S. Chen, and S.-F. Chang. Recognizing complex events in videos by learning key static-dynamic evidences. In *ECCV*, 2014.
- [49] K.-T. Lai, F. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.
- [50] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *MMM*, 2012.
- [51] I. Laptev. On space-time interest points. *IJCV*, 64(2-3), 2005.
- [52] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in videos. *TSMC*, 39(5), 2009.
- [53] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [54] B. Li and M. I. Sezan. Event detection and summarization in sports video. In *IEEE W'shop on Content-Based Access to Video and Image Libraries*, 2001.

- [55] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.
- [56] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013.
- [57] X. Li, E. Gavves, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Personalizing automated image annotation using cross-entropy. In *MM*, 2011.
- [58] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 2009.
- [59] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *CoRR*, abs/1503.08248, 2015.
- [60] X. Li, D. Wang, J. Li, and B. Zhang. Video search in concept subspace: a text-like paradigm. In *CIVR*, 2007.
- [61] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007.
- [62] 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.
- [63] M. H. Lloyd. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, 1999.
- [64] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.
- [65] A. R. Luria. *The mind of a mnemonics: a little book about a vast memory*. Harvard University, 1987.
- [66] Z. Ma. *From Concepts to Events: A Progressive Process for Multimedia Content Analysis*. PhD thesis, University of Trento, Trento, Italy, 2013.
- [67] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *MM*, 2012.
- [68] Z. Ma, Y. Yang, Z. Xu, N. Sebe, and A. G. Hauptmann. We are not equally negative: fine-grained labeling for multimedia event detection. In *MM*, 2013.
- [69] Z. Ma, Y. Yang, Z. Xu, N. Sebe, S. Yan, and A. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.
- [70] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.
- [71] R. Manmatha and S. C. Ravela. Syntactic characterization of appearance and its application to image retrieval. *SPIE*, 3016, 1997.
- [72] S. Mannor, D. Peleg, and R. Rubinstein. The cross entropy method for classification. In *ICML*, 2005.
- [73] O. Maron. *Learning from Ambiguity*. PhD thesis, Massachusetts Institute of Technology, 1998.
- [74] O. Maron and T. Lozano-Pérez. A framework for multiple instance learning. In *NIPS*, 1998.
- [75] M. Mazloom, E. Gavves, and C. G. M. Snoek. Conceptlets: Selective semantics for classifying video events. *TMM*, 16(8):2214–2228, 2014.
- [76] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. G. M. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.
- [77] M. Mazloom, A. Habibian, D. Liu, C. G. M. Snoek, and S.-F. Chang. Encoding concept prototypes for video event detection and summarization. In *ICMR*, 2015.
- [78] M. Mazloom, A. Habibian, and C. G. M. Snoek. Querying for video events by semantic signatures from few examples. In *MM*, 2013.
- [79] M. Mazloom, X. Li, and C. G. M. Snoek. Few-example video event retrieval using tag propagation. In *ICMR*, 2014.
- [80] M. Mazloom, X. Li, and C. G. M. Snoek. Tagbook: A semantic video representation without supervision for event detection. *TMM*, 18(7):1378–1388, 2016.
- [81] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *TMM*, 14(1), 2012.
- [82] P. E. Meyer, C. Schretter, and G. Bontempi. Information-theoretic feature selection in microarray data using variable complementarity. *J-STSP*, 2(3):261–274, 2008.
- [83] 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. *MVA*, 25(1), 2014.

Bibliography

- [84] M. R. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *TMM*, 13(3), 2006.
- [85] P. Natarajan, S. Wu, S. N. P. 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.
- [86] A. Natsev, A. Haubold, J. Tesci, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *MM*, 2007.
- [87] A. Natsev, M. R. Naphade, and J. Tesci. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM MM*, 2005.
- [88] 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.
- [89] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *ICML*, 2004.
- [90] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. Qbic project: querying images by content, using color, texture, and shape. *Proc. SPIE*, 1908, 1993.
- [91] NIST TRECVID Multimedia Event Detection (MED) Evaluation Track. <http://www.nist.gov/itl/iad/mig/med.cfm>.
- [92] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. G. Amitha Perera, M. Pandey, and J. I. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *MVA*, 25(1), 2014.
- [93] D. Oneata, M. Douze, J. Revaud, S. Jochen, 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. NIST TRECVID Workshop, 2012.
- [94] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Proc. of ICCV*, 2013.
- [95] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2013.
- [96] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 2005.
- [97] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *IJCV*, 18(3), 1996.
- [98] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.
- [99] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *TMM*, 9(5), 2007.
- [100] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, 2004.
- [101] S. Rudinac, M. Larson, and A. Hanjalic. Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval. *IJMIR*, 2012.
- [102] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. In press.
- [103] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.
- [104] Y. Saeyns, I. naki Inza, and pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23, 2007.
- [105] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11), 1975.
- [106] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.
- [107] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Math. Program.*, 127(1), 2011.
- [108] W. Siedlecki and J. Sklansky. On automatic feature selection. *IJPR*, 1988.
- [109] S. Siersdorfer, J. S. Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *SIGIR*, 2009.
- [110] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *MIR*, 2006.

- [111] C. Snoek, S. Cappallo, D. Fontijne, D. Julian, D. Koelma, P. Mettes, K. van de Sande, A. Sarah, H. Stokman, and R. Towal. Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing Concepts, Objects, and Events in Video. In *NIST TRECVID Workshop*, 2015.
- [112] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *TMM*, 9, 2007.
- [113] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *FnTIR*, 2(4), 2009.
- [114] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, 2014.
- [115] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. S. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.
- [116] R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSSB*, 58, 1994.
- [117] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.
- [118] TRECVID Multimedia Event Detection Evaluation Track, 2012. <http://www.nist.gov/itl/iad/mig/med.cfm>.
- [119] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. PAMI*, 32(9), 2010.
- [120] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34(3), 2011.
- [121] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.
- [122] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang. Selection of concept detectors for video search by ontology-enriched semantic spaces. *TMM*, 10(6), 2008.
- [123] H. Williams, M. Conway, and G. Cohen. *Autobiographical memory*. Psychology Press, 2008.
- [124] 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.
- [125] L. Xie, H. Sundaram, and M. Campbell. Event mining in multimedia streams. *Proceedings of the IEEE*, 2008.
- [126] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *PAMI*, 30(11), 2008.
- [127] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *CVPR*, pages 1798–1807, 2015.
- [128] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. Hauptmann. How related exemplars help complex event detection in web videos? In *Proc. of ICCV*, 2013.
- [129] 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 *MM*, 2015.
- [130] E. Younessian, T. Mitamura, and A. G. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ICMR*, 2012.
- [131] S.-I. Yu, L. Jiang, Z. Xu, Z. Lan, S. Xu, X. Chang, X. Li, Z. Mao, C. Gan, Y. Miao, X. Du, Y. Cai, L. Martin, N. Wolfe, A. Kumar, H. Li, M. Lin, Z. Ma, Y. Yang, D. Meng, S. Shan, P. Sahin, S. Burger, F. Metze, R. Singh, B. Raj, T. Mitamura, R. Stern, and A. Hauptmann. Informedia@ TRECVID 2014 MED and MER. In *NIST TRECVID Workshop*, 2014.

Op zoek naar de semantiek van evenementen in video

In dit proefschrift hebben we als doel het representeren van een evenement in een video met gebruik van semantische features. We beginnen met een set van concept detectoren om evenementen in videos te representeren. In het begin hebben we de relevantie van concepten met betrekking tot het evenement binnen in de video representatie beschouwd. Daarna hebben we ons geconcentreerd op de accuraatheid van concept detectoren. Uiteindelijk beschouwen we het probleem van het zoeken in video evenementen met en zonder semantische concepten. Voordat de onderzoeksvragen van de introductie worden besproken, geven we een kort overzicht van de hoofdstukken.

In Hoofdstuk 2 adresseren we het probleem van het classificeren van video evenementen met gebruik van een set van scores van concept detectoren. Anders dan bestaande werken, die simpelweg afhankelijk zijn van een set die alle beschikbare detectoren bevat, stellen we een algoritme voor die leert van voorbeelden welke concepten in een set het meest informatief zijn per evenement, genaamd conceptlets. We modelleren het vinden van conceptlets uit een grote collectie concept detectoren als een importance sampling probleem. Ons voorgestelde algoritme vindt de optimale conceptlets door middel van een cross-entropy optimalisatie. We studeren het gedrag van het classificeren van video evenementen op basis van conceptlets door vier experimenten uit te voeren op uitdagende internet videos van de 2010 en 2013 TRECVID multimedia event detection taken en Columbia's consumenten evenementen dataset. Beginnend met een set van meer dan duizend voorberekende concept detectoren, laten onze experimenten zien dat er (collecties van) individuele concept detectoren zijn die discriminatiever en discriptiever voor een specifiek evenement lijken te zijn dan anderen, dat het classificeren van evenementen met gebruik van automatisch verkregen conceptlets meer robuust is dan gebruik te maken van alle beschikbare concepten, en dat conceptlets verkregen met onze cross-entropy algoritme beter zijn dan concepten van state-of-the-art feature selectie algoritmes. Daar bovenop zijn de conceptlets logischer voor het evenement waarin men is geïnteresseerd is, zonder daarvoor geprogrammeerd te zijn.

In Hoofdstuk 3 stellen we een nieuwe semantische video representatie voor voor het detecteren en samenvatten van evenementen met weinig of geen voorbeelden. Anders dan bestaande werken, die een semantische representatie verkrijgen door concepten te trainen van plaatjes en gehele video clips, stellen we een algoritme voor die een set van relevante frames leert als de concept prototypes van voorbeelden van web videos, zonder frame-annotaties nodig te hebben, en gebruiken die om videos met evenementen te representeren. We formuleren het probleem van het leren van concept prototypes als het zoeken naar de frames die het dichtst in de buurt van het drukste gebied van de feature ruimte van video frames zijn van zowel positieve als negatieve training voorbeelden voor een specifiek concept. We studeren het gedrag van onze video evenementen representatie gebaseerd op concept prototypes door drie experimenten te doen op uitdagende video van het internet uit de TRECVID 2013 multimedia event detection taak en de MED-summaries dataset. Onze experimenten laten zien dat i) het detecteren van evenementen beter werkt wanneer elke video wordt geprojecteerd in de concept prototype ruimte, ii) het detecteren van evenementen zonder trainings voorbeelden werkt beter doordat elke frame van een video apart wordt bekeken in de concept prototype ruimte, in plaats van alleen de gehele video te bekijken, iii) het samenvatten van evenement videos met gebruik van concept prototypes zonder trainings voorbeelden is accurater dan gebruik maken van concept detectoren van de gehele video.

In Hoofdstuk 4 hebben we tot doel om videos van het internet op te halen voor complexe evenementen met gebruik van slechts een handvol voorbeelden, daar waar de standaard aanpak algoritmes leert van honderden voorbeelden. We overwegen een semantische representatie die bestaat uit concept detectoren van de plank, om de variatie in het semantische uiterlijk vast te leggen. Omdat het niet bekend is welke vergelijkingsfunctie en fusie van queries moeten worden gebruikt in zo'n evenementen setting, doen we drie experimenten op videos van het internet van de TRECVID 2012 multimedia event detection taak. Die laten zien dat het zoeken met semantische representaties met normalized correlation als vergelijkingsfunctie accurater en sneller is dan gebruik te maken van een niet-semantische bag-of-words alternatief, dat meerdere queries het best gecombineerd kunnen worden door late fusion met een middellings operatie,

en dat het vinden van evenementen geprefereerd wordt aan het classificeren van evenementen wanneer er minder dan acht positieve voorbeelden beschikbaar zijn. We laten de mogelijkheid om semantische representaties te genereren zien in een video zoekmachine. We laten ook zien hoe de semantische representaties een grove interpretatie geven over waarom een specifieke video gevonden is.

Het doel van Hoofdstuk 5 is het detecteren van evenementen wanneer er slechts enkele, of zelfs geen, voorbeelden beschikbaar zijn tijdens het trainen. Voor deze uitdagende setting berust de huidige literatuur op een semantische video representatie verkregen van duizenden voorgetrainde concept detectoren. Anders dan de huidige werken, stellen wij een nieuwe semantische representatie voor dat is gebaseerd op verkrijgbare videos die enkel zijn getagged in sociale media, zonder een trainingsstap nodig te hebben om concept detectoren te verkrijgen. We introduceren een simpel algoritme die tags van de dichtste burens van een video verspreid, zoals ook wordt gebruikt bij het vinden van plaatjes, maar opnieuw ontworpen voor het detecteren van evenement videos door een verbetering toe te voegen met gebruik van een bron van videos en door de toewijzing van tags te variëren. We noemen onze aanpak TagBook en bestuderen de constructie, mate van descriptie, en het functioneren van de detectie op de TRECVID 2013 en 2014 multimedia event detection datasets en op de Columbia Consumer Video dataset. Ondanks zijn simpelheid, is de voorgestelde TagBook evenementen representatie opmerkelijk effectief voor detecteren van evenementen met weinig of geen voorbeelden, en werkt zelfs beter dan zeer recente alternatieven die bouwen op representaties van veel voorbeelden.

ACKNOWLEDGEMENTS

I learnt from scientific writing in computer science to avoid using the singular, first person pronoun “I”. I always took it as an implicit acknowledgment that no research is ever completed without the help of others specially in case of a dissertation. Many people have inspired and helped me along the course of my PhD journey and, I am deeply grateful to all of them. Here, I would like to thank people who practically or spiritually helped me accomplish the journey.

First, I would like to express gratitude to my daily supervisor and my great friend Cees G. M. Snoek and Arnold W. M. Smeulders, my supervisor, my promotor. Without them this thesis would not have been, and I would not have enjoyed working on it as much as I did. Special thanks go to Cees for the opportunity he gave me to come to The Netherlands and work on challenging and complex research project, collaborations and papers with smart colleagues in Intelligent Systems Lab Amsterdam. I have so many memories with Cees from our first meeting in University of Berkeley till now. We had countless discussions about crazy new ideas, which he asked me to put in freezer where many of them remained, or the problems, scientific and personal. Cees, was always happy to listen, to discuss, to give smart advice and suggestion on how to go straight from Amsterdam, initial state, to Tehran, goal state. Cees, thanks for the trust and support, and for listening to my rants and raves throughout the past few years.

Arnold deserves a special mention as mentor as well; always a very critical supervisor whose help was paramount, particularly towards the end of my thesis. Many thanks Arnold for letting me have a special experience as a visiting scholar in one of the best multimedia labs at Colombia University. Arnold, thanks for the help and all your support.

I would also like to thank Dennis Koelma, for preparing materials which greatly facilitated my experiments in all chapters of this thesis. With him I had many fruitful research discussions and his advise helped me at various stages of my PhD.

I would like to acknowledge all kinds of help I have got from people during my PhD at the University of Amsterdam: Xirong Li, with whom we wrote two papers, Efstratios Gavves, with whom our extensive collaboration led to my first conference and first journal paper. A great thanks go to Amirhossein Habibian for our very interesting discussions and investigations that lead to publish three papers, and hopefully many more in the future. Also I would like to thanks Svetlana Kordomova, Ran Tao, and Zhenyang Li for helping me in so many different ways specially during 3 month in DVMM lab at Colombia University as visiting scholar. You were always kind, and supportive.

Furthermore, I would like to thank the people from ISLA. Marcel, Theo, Dennis, Xirong, Thomas, Jan, Jasper, Koen, Gosia, Ivo, Efstratios, Hamdi, Dung, Amirhossein, Zhenyang, Ran, Svetlana, Sezar, Morris, Silvia, Kandan, Mihir, Spencer, Pascal, Honza, Stevan, Jörn, and Zhaochun, who forged a great environment not only for research but also for having fun, playing volleyball, football, and ping pong, or having BBQs, ISLA dinners, and group outings and social activities. My time with all of them was great, with lots of memories. Special thanks go to Morris and Svetlana for designing the cover, as well as Bob, Pascal, and Dennis for translating my thesis summary into Dutch. Special thanks go to Virginie for helping with all the problems in the past years.

Moreover, I would like to thank my Iranian friends at Amsterdam, where I met so many great friend in particular those from Science Park: Amirhoseein and Irene, Yaser and Fereshteh, Hosein, Keyvan, Hadi, Mostafa and Samira, Ali and Azadeh, Mehran and Nahid, Mohammad and Fahimeh, Mehdy and Mahdiah, Naser and Sareh, Naser and Aylar, Mehdy and Samira, Hodjat and Marzieh, Narges and Shayan, Behnaz and Rori, Parsa, Parisa, Rojman and Elaheh, Mehdy and Marzieh, Hamid and Marzieh, Mehdy and Maryam, Farzin and Mahsa, Amin and Fatemeh, Sara and Vahid, Amin and Parisa, Mojdeh, Mojtaba, Abbas, and Mehran. Thanks to them, five years spent in Amsterdam were not purely scientific.

I would like to thank my committee, Prof. dr. H. Afsarmanesh, Prof. dr. A. Hanjalic, Prof. dr. A. G. Hauptmann, Prof. dr. M. Worrington, and dr. T. E. J. Mensink for honoring me with reading this thesis and participating in my defence.

Last, I would like to express my gratitude to my family. First, I want to express my gratitude to my parents, my father, Abolghasem, and my mother, Fatemeh. Because of you I possess the perpetual self-confidence needed for accomplishing this work. Your love has been always unconditional and unlimited.

Acknowledgements

I would like to thank my brothers, Reza, Mohsen, and Saeed, and my sister, Maryam, for your love and always believing me. My loving thanks go to Hoda Alemi, my best friend, for being my biggest inspiration and support through the years and for making me a better person. Finally, my son, my hero, Sina, who I dedicate this work to. Because of you, it has become simple to make choices and focus on what is important in life. Love you son!

Masoud