Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

Event-based Media Processing and Analysis: A Survey of the Literature

Christos Tzelepis[a,b], Zhigang Ma[c], Vasileios Mezaris[a], Bogdan Ionescu[d], Ioannis Kompatsiaris[a], Giulia Boato[e], Nicu Sebe[e], Shuicheng Yan[f]

[a] *Information Technologies Institute (ITI), CERTH, Thermi, 57001, Thessaloniki, Greece*
[b] *Queen Mary, University of London, Mile End Campus, UK, E14NS*
[c] *School of Computer Science, Carnegie Mellon University, USA*
[d] *LAPI, University "Politehnica" Bucharest, 061071 Romania*
[e] *University of Trento, Italy*
[f] *Department of ECE, National University of Singapore, Singapore*

Abstract

Research on event-based processing and analysis of media is receiving an increasing attention from the scientific community due to its relevance for an abundance of applications, from consumer video management and video surveillance to lifelogging and social media. Events have the ability to semantically encode relationships of different informational modalities, such as visual-audio-text, time, involved agents and objects, with the spatio-temporal component of events being a key feature for contextual analysis. This unveils an enormous potential for exploiting new information sources and opening new research directions. In this paper, we survey the existing literature in this field. We extensively review the employed conceptualization of the notion of event in multimedia, the techniques for event representation and modeling, the feature representation and event inference approaches for the problems of event detection in audio, visual, and textual content. Furthermore, we review some key event-based multimedia applications, and various benchmarking activities that provide solid frameworks for measuring the performance of different event processing and analysis systems. We provide an in-depth discussion of the insights obtained from reviewing the literature and identify future directions and challenges.

Keywords: Event-based media processing and analysis, event conceptualization, event representation and modeling, multimedia event detection, event-based applications and benchmarking, survey of the literature

1. Introduction

In these times, people tend to collect dozens of photos and video clips every day using their smartphones, tablets, cameras, and such information is exchanged ceaselessly in a number of different ways (e.g., via social networks). The growing number of various types of sensors for capturing environmental conditions, in the moment of content creation, has led to multimedia content enriched with context-awareness that allows capturing experiences and events of interest from a very rich personal perspective. This unveils a growing demand, but also an enormous potential for event-centred data analysis. The core idea consists in using events as primary means for understanding, organizing and indexing content (e.g., audio, videos, news, social streams). Events have the distinctive ability to semantically encode relationships that come from different informational modalities. These modalities can include, but are not limited to: time, space, involved agents and objects, with the spatio-temporal component of events being a key feature for contextual analysis. A plethora of techniques have recently been presented to leverage contextual information for event-based analysis, covering audio-, visual-, and textual-based approaches.

Event-based media processing and analysis is currently a hot topic being used in a broad range of scientific and consumer domains, a sample of which include: (a) multimedia organization and consumer

Email addresses: tzelepis@iti.gr (Christos Tzelepis), kevinma@cs.cmu.edu (Zhigang Ma), bmezaris@iti.gr (Vasileios Mezaris), bionescu@alpha.imag.pub.ro (Bogdan Ionescu), ikom@iti.gr (Ioannis Kompatsiaris), boato@disi.unitn.it (Giulia Boato), sebe@disi.unitn.it (Nicu Sebe), eleyans@nus.edu.sg (Shuicheng Yan)

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

video management for digital preservation [46, 165, 228, 42, 34], (b) lifelogging [80, 200, 201, 153], (c) video surveillance [169, 39, 85, 101], (d) Journalism (News- and sport-related applications for community awareness) [203, 170, 146, 113, 166, 204, 51], (e) social media [151, 161, 174, 86], and (f) event visualization [48, 160, 38]. In this paper we provide an extensive overview of the literature in this growing domain.

1.1. Previous surveys and other resources

Event-based media processing and analysis is a vast research area to which the research community has contributed with many survey studies. Each of the existing studies focuses on a specific aspect of the problem. For instance, Scherp and Mezaris [173] surveyed existing event models along with commonly identified aspects of events, providing a detailed review focused on event representation approaches. Most of the existing surveys in this domain study the problem of event detection in video content. For instance, Jiang et al. [98] studied the problem particularly for unconstrained videos, such as those found in the Web. In [180], Snoek and Worring surveyed approaches to multimodal video indexing, focusing on methods for detecting various semantic concepts consisting of mainly objects and scenes. They also discussed video retrieval techniques exploring concept-based indexing, where the main application data domains were broadcast news and documentary videos. Brezeale and Cook [20] surveyed text, video, and audio features for classifying videos into a predefined set of genres, e.g., "sports" or "comedy", while Morsillo et al. [138] presented a brief review that focused on efficient and scalable methods for annotating Web videos at various levels including objects, scenes, actions, and high-level events. Further similar surveys on video event processing and analysis, covering topics such as visual feature extraction, event classification, and ontologies for knowledge representation and reasoning can be found in [11, 79, 215].

1.2. Taxonomy and scope of present survey

Different from the existing surveys, in this paper we have conducted a broader comprehensive analysis of the event-based media processing and analysis domain. Starting with the conceptualization of the notion of event for processing, where we reviewed several definitions depending on the complexity of events that are desired to be detected in multimedia content, going through the problem of event representation, where we surveyed approaches that aspire to model events in meaningful ways, we reached the issue of event detection in different media types, i.e., audio-, video-, and textual-based, and treated them individually in terms of feature representation and event inference. For the former, we discussed in detail various state-of-the-art feature representation schemes, such low-, intermediate-, and high-level ones, static- or motion-based, as well as audio and textual ones, in order to exploit every piece of information that is available for the specific detection problem in hand. We also looked into social event detection in multiple media collections. The study concludes by overviewing the current main applications of this particular field.

The remainder of the paper is organized as following. Section 2 presents a review on the different definitions of the notion of event in the multimedia field, while Section 3 gives the corresponding event models that attempt to represent such events. Section 4 surveys audio-visual event detection, i.e, the annotation of isolated audio-visual content items (e.g., a video with audio, or an audio-only file) with event labels, one content item at a time, while in Section 5 we discuss social event detection, which focuses on the processing of collections of media items, such as collections of images, videos, and/or text, and finding the associations between different content items. In Section 6, we present event-based applications and discuss various benchmarking activities for video annotation, surveillance, social event detection, as well as synchronization of multi-user event media. We conclude this survey in Section 7 with a discussion of promising directions for future research.

2. The notion of event in multimedia

The notion of event is ubiquitous in multimedia and shares different definitions in many different problem domains. In literature, the definition of an event can be heterogeneous even though it shares a common characteristic; events are in general said to occur, or happen, meaning that they are entities that unfold over time and/or space [203, 172]. As they can be seen as natural abstractions of *happenings*, or *observable occurrences* [173, 123], the research community adopts simple or more complex definitions of the event, depending on the specific problem under consideration.

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

Event Definition An event is defined as:	**Example(s)**	**Key References**
▶ a change of state in a multimedia entity ▶ a concept with a dynamic component	"ship stopping/moving"	Francois et al. [63] Ballan et al. [11]
▶ a collection of actions performed between agents ▶ a list of interactions between objects using any prior information concerning the context of a scene	"person stops moving left-hand"	Hakeem et al. [84] SanMiguel [169] Jiang et al. [94]
▶ a number of human actions, processes, and activities (loosely or tightly organized) having temporal and semantic relationships to the overarching activity ▶ a complex activity occurring at a specific place and time involving people interacting with other people or object(s)	"changing a vehicle tire", "making a cake", "attempting a bike trick"	Tong et al. [188] Jiang et al. [97] Over et al. [145]
▶ being any event (something happening at a specific time and place) of interest to the (news) media ▶ a story related to some news topic comprising of patterns that occurred at some specific time and space ▶ being planned and attended by people describing a social activity or a phenomenon that happened in real life	"demonstration", "speech", "concert"	Sayyadi et al. [170] Pahal et al. [146] Papadopoulos et al. [148]

Table 1: Overview of the event conceptualization scenarios (from simple definitions to complex formalization).

The less complex definition of an event introduces the notion of a *change of state* in a multimedia entity, such as an audio or visual stream; an event is typically triggered by a change of state captured [63]. Ballan et al. [11] refer to events as concepts with a dynamic component. This definition of the event includes simple movement events (such as "ship stopping" and "ship moving"), while these are used for defining more complex event patters (such as "trips", as a series of consecutive movements) [194].

Another common definition of an event, yet still simple, is that of a collection of *actions* performed between one or more agents [84]. In this case, agents are typically animates (people or machines) performing actions independently or dependently (for instance, "person stops moving left-hand"). In [169], SanMiguel et al. define event as a list of interactions between objects, which, along with any other prior information concerning the context of a scene (where the event evolves), are used for the problem of video surveillance. Similarly, in [94], an event is an activity-centered happening that involves people engaged in process-driven actions with other people and/or objects at a specific place and time. Consequently, the above definitions pose the notion of the event at the boundaries between video event detection and vision-based action recognition [154] problems.

Concerning the video understanding domain, the most dominant definition of an event introduces the notion of a complex activity occurring at a specific place and time which involves people interacting with other people and/or object(s) [188, 97]. Such events may include "changing a vehicle tire", "making a cake", or "attempting a bike trick", to name a few. In general, an event consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have temporal and semantic relationships to the overarching activity. This is also the definition of a (complex) event in the Multimedia Event Detection problem for the TRECVID benchmarking activity [145].

There are also higher-level definitions of events; for instance, a news event [170] is defined as being any event (something happening at a specific time and place) of interest to the (news) media. In [146], any news event encodes a story related to some news topic within itself. These event stories comprise of sequence of events or patterns that occurred at some specific time and space. Finally, one can also define *social events* [148], where the events are meant to be planned and attended by people, and the multimedia content immortalizing the event is also captured by people. A more specific definition of a social event describes a social activity or a phenomenon that happened in real life at some point in time and in specific place, either *planned* or *abrupt*.

In this section, we explored different definitions of the notion of an event in the field of multimedia. Despite their differences, multimedia events are fundamentally meant to unfold over space and time. They admit various definitions, simple or more complex, depending on the specific problem under consideration. These definitions predominantly determine both the model with which each category of events is modeled, and the approaches

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

followed for processing and analyzing multimedia content in order to build effective event processing and analysis tools. Table 1 summarizes the surveyed studies and gives indicative examples and key references for each class of approaches discussed in this section.

3. Event representation and modeling

Representing events with meaningful models depends heavily on their complexity. Simple events admit simple models, while more sophisticated ones demand more complex event models. Proposing and applying a meaningful event model is a tough process, to which the research community has contributed with several studies.

Simple event models have been proposed for representing elementary events. To this end, Event Calculus [104, 195, 25] has been considered for knowledge representation (without providing any spatial information) and the Situation Calculus [116] for representing changes in the real world. Moreover, Raimond and Abdallahas proposed Event Ontology [158] as a part of a music ontology framework, while Shaw et al. proposed the ontology on Linking Open Descriptions of Events (LODE) [175], which captures a minimal model of events. LODE permits modeling time and space using an absolute reference, but a relational reference is not allowed. The Simple Event Model (SEM) [194] supports absolute spatial information using the WGS841[1] vocabulary. Furthermore, Chen et al. proposed the Standard Ontology for Ubiquitous and Pervasive Computing (SOUPA) [31], which is the core of the Context Broker Architecture (CoBrA) [32]. Wang et al. presented the Context Ontology (CONON) [202] for modeling context in pervasive computing environments, while Yau et al. proposed the Situation Ontology [218] for an hierarchical modeling and sharing of situation knowledge. Matheus et al. proposed the Situation Awareness Assistant (SAWA) application [127, 126, 128], an event model that supports spatio-temporal composition.

Concerning some more complex event models, the following approaches permit the modeling of relative relations in space. The ISO-standard of the International Committee for Documentation on a Conceptual Reference Model (CIDOC/CRM) has been presented in [50, 178] for cultural heritage applications and supports hierarchical part-whole relationships. However, no further axiomatizations are provided for refining the mereological relationship by different criteria such as temporal and spatial constraints [178, 173]. Moreover, the XML-based OASIS standard of the Common Alerting Protocol (CAP), proposed in [182] for describing events in the domain of hazard emergency alerts and public warnings, allows for providing absolute information about space, but relative relations in space can only be provided by a textual description of the location. It does not provide support for modeling participants (objects and/or humans) in events due to the specific domain's restrictions; i.e., messages about upcoming hazardous events. Finally, IPTC defined for News events an XML-based event markup language called EventsML-G2 [88]. XML-based descriptions of events using EventsML-G2 can be embedded into a news item described in NewsML [87].

A more sophisticated set of event models proposed in literature, in an attempt to cover the design gaps in the above studies (such as the the capability for modeling sub-events), includes the Semantic-syntactic Video Model (SsVM) [56], the Networked multimedia event exploration (NMEE) model [5] by Appan and Sundaram, CASEˆE [84] providing an hierarchical event model for the analysis of videos, and the Video Event Representation Language (VERL) [56, 141] for video data. Events represented in VERL can be annotated using a companion mark-up language called the Video Event Markup Language (VEML) [141]. VEML is basically used to encode events described with VERL into video stream data [141]. CASEˆE allows for modeling temporal relations between events based on the Allen's time calculus [1] as well as sub-event relationships along the temporal dimension. NMEE presents the idea of different viewpoints to the same event. However, it is limited due to the fact that the definition of what is a viewpoint is underspecified. NMEE can just represent how many viewpoints an event has. Gkalelis et al. [74] proposed a graph model to represent events, where the nodes are events and the edges are relations between events. This model supports the notion mereology [173] by defining sub-events along the temporal dimension. Regarding interpretation of events, the authors used the properties *isInstantiatedBy* and *hasInstantiationTime*, while different interpretations of events are linked using a *sameAs* relation

[1] http:www.w3.org/2003/01/geo/, last visited: October 18, 2015

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

		Event Calculus	Situation Calculus	Event Ontology	LODE	SEM	SOUPA	CONON	Situation Ontology	SAWA	CIDOC CRM	CAP	EventsML-G2	SsVM	NMEE	CASEE	VERL&VEML	Gkalelis et al.	E	E*	Event-Model-F
Time	absolute	✓	✗	✓	✗	✓	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
	relative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Space	absolute	✗	✓	✗	✗	✗	✓	-	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	relative	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	-	✓	✓	✓	✗	✓	✓	✓	✓	✓
Relations	mereologic	✓	✓	-	✗	-	-	✗	-	✓	-	✗	✓	✓	✗	-	-	-	-	-	✓
	causal	✓	✓	-	✗	✗	-	✗	✗	-	-	✗	✗	-	✗	✓	-	✓	-	-	✓
	correlation	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

Table 2: Overview of event modeling approaches and their main characteristics, inspired from [173].

like in the Web Ontology Language (OWL)3 [3], which combines the different points of view and, thus, it is not possible to distinguish different interpretations from the same agent or person in different contextual situations.

Finally, in the direction of modeling a complex event using a sufficiently comprehensive representation, which differs from the above works in the number of event's aspects covered, we consider the event model E [91, 171, 206, 207] for event-based multimedia applications, its graph-based successor E* [79] with extended features such as for modeling time and space, and the Event-Model-F [171], which is an extension and formalization of the event model E. E* provides for elaborate features regarding expressing spatial relationships between events. Event-Model-F allows for describing complex mereological relationships (like the SsVM does), as well as supports spatio-temporal composition (similarly to the situation-awareness model of SAWA) [12, 173]. Only a limited number of event models allow for providing different interpretations of the same event. Inspired by the event model E, the Event-Model-F provides support for the event interpretation aspect [173], implemented in the form of so-called ontology design patterns [69]. Event-Model-F provides well formed, i.e., formally defined support for causal relationships between events [195]. Support for representing correlation relations between events can only be found with a specific extension of the event calculus [25] and the Event-Model-F.

As it is obvious, there is not a single, universal event representation that can model every event that can appear in a multimedia event processing problem. In the literature, there have been proposed several different models that mainly differ in their complexity and, thus, the categories of events that are capable of modeling, varying from elementary actions to complex high-level events where people interact with other people and/or objects in specific time and space. Table 2 illustrates the differences between many of event models presented in this paper.

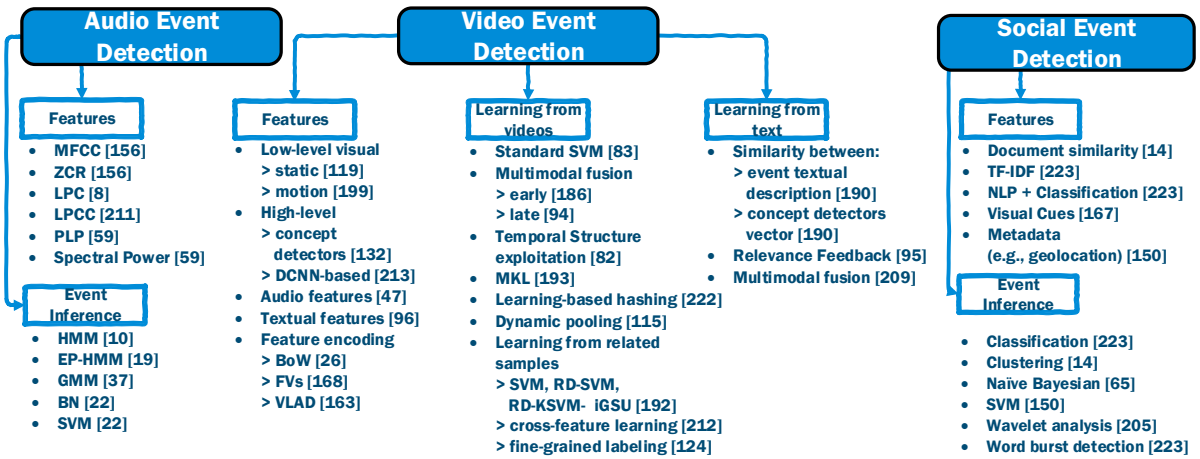
4. Audio-visual event detection

We define as audio-visual event detection the problem of processing isolated audio-visual content items (e.g., a video with audio, or an audio-only file), one at a time, so as to understand if this content item relates to one or more events from a specified event set. The resulting content-event associations are typically used for event-based retrieval of media items, within a large content collection. For solving this problem, features play a critical role: the way that the multimedia information is represented has been proven to be of utmost importance. The major categories of multimedia information exploited in an event detection system are audio-, visual-, and textual-based. Depending on the specific event detection problem in hand (audio-only, or video), different features are extracted, processed, and combined using various different approaches. Fig. 1a gives an outline of major research directions in audio-visual event detection.

4.1. Audio event detection

Audio event detection (AED), whose target is to recognize specific events in audio streams, has received considerable interest in recent years. Typically, audio events are less complex than the visual ones, while they are usually highly dependent

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005



(a) (b)

Figure 1: Outline of major media (a) audio-visual event detection and (b) social event detection directions in the literature and a few indicative references.

on the specific event detection application in hand. There exists a wide variety of audio events that could be of interest to detect in an audio stream. For instance, extracting audio highlights in a soccer game (e.g., cheering) could be advantageous to the consumer, who could automatically access indexed games recordings [211, 210]. On the other hand, detecting gunshots in noisy environments could be of particular value in a surveillance system [39, 85].

A typical AED framework includes a feature extraction and an audio event inference stage. The latter may make use of Hidden Markov Models (HMMs) for exploring the time structure of the event and/or model interconnections between key audio effects (e.g., an explosion being caused by a car accident), as well as classifiers, such as SVMs [196] for learning the audio event detector.

4.1.1. Features

The most frequently used audio representation is the Mel-Frequency Cepstral Coefficients (MFCCs) [156], which is widely used in the field of speech recognition and is designed to be robust to noise. MFCCs are used in many AED systems [10, 59, 39, 211, 210, 149]. Other audio features used in AED frameworks, both in the time and the frequency domain, include Zero-Crossing Rate (ZCR) [156], which was used in [8, 59, 211], Linear Prediction Coefficients (LPCs) and Linear Predictive Cepstral Coefficients (LPCCs) in [8, 211], Log-Frequency Cepstral Coefficients (LFCCs) in [8, 120] and Perceptual Linear Prediction (PLP) in [59, 120]. A number of other audio features that are less frequently used in this problem include Spectral power [59, 211], pitch frequency, sub-band power, brightness/bandwidth [59, 120], and short-time energy [39, 120]. In [24, 210], the well-studied MPEG-7 audio features are also used for the problem of audio event detection.

Many of the above audio features are usually combined for building an effective audio event detection system. This can lead to high-dimensional feature representations, which often result in lengthy training processes. In this case, it is common practice to use dimensionality reduction techniques, such as Principal Component Analysis (PCA) and/or Linear Discriminant Analysis (LDA) [59, 39], in order to transform high-dimensional audio features into lower-dimensional ones, leading to lower training times and possibly also to better detection results due to the potential denoising of the training data.

4.1.2. Event inference

Event inference is the stage of an audio event detection system responsible for deciding on whether an audio stream belongs to a specific audio event class or not. To this end, a frequently used approach is using Hidden Markov Model (HMM) [156], which is an effective tool for modeling time-evolving processes, widely used in speech recognition. Using HMMs raises two crucial issues

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

that have to be handled: a) the model selection process and b) the selection of both the optimal model size and number of (Gaussian) mixtures per state, where model size corresponds to the number of states. In [10], the authors use a continuous density HMM for modeling audio-based pattern classes. They tackle the above issues by carrying out a cross-validation process in order to identify a suitable number of states and mixtures per state. Similarly, Xiong et al. [210] used Entropic Prior HMM (EP-HMM) [19] for learning audio event detectors, following a preprocessing stage where background noise was detected based on energy and magnitude and removed.

Another popular classification technique used for building audio event detectors is the Gaussian Mixture Model (GMM) [35, 37]. For instance, Atrey et al. [8] proposed a multilevel classification approach for learning audio event detectors; i.e., a hierarchical classification approach to assign a label to an audio event in a given "audio frame" (a fixed-size audio segment which is extracted from the continuous audio stream). A GMM classifier was employed to classify an input, at the top level - into foreground or background, at second level - into vocal or nonvocal, and at third level - into excited events (e.g., shout/cry, door knock, running footsteps) or normal events (talk, walking footsteps). The authors considered four audio events ("talk", "shout", "knock", and "footsteps") and adopted a hierarchical (top-down) approach to model these events using a mixture of Gaussians (GMM). The top-down event modeling approach worked better compared to the single-level multi-class modeling approach.

Besides the above learning methods, there are also a few studies that use rule-based approaches [211], Bayesian Networks [22], standard SVM classifiers, and their combinations for building more elaborate audio event detectors. In their system's inference stage, Elo et al. [59] used various machine learning methods to provide a final classification of the audio events. These methods include rule-based approaches, GMMs, SVMs, and Bayesian Networks [22]. HMMs and SVMs were used for building one-against-all classifiers for each audio event, though the authors reported that better results could potentially be achieved by multiple-class classification. Clavel et al. [39] trained a GMM for each audio event class for their surveillance system. The appropriate number of Gaussians for each audio class was estimated based on the Bayesian Information Criterion [62]. The parameters of the models were estimated using the traditional Expectation-Maximization algorithm [136], initialized by a basic binary splitting vector quantization algorithm. Detection was made using the Maximum A Posteriori (MAP) decision rule: the mean a posteriori log-probability on a 0.5-second decision window was computed for each class model (by multiplying the probability obtained for each short-time analysis frame). The decision window was then classified according to the class that had the maximum a posteriori score. Xu et al. [211] designed a hierarchical SVM classifier for detecting three main audio event classes: "Whistling", "Commentator speech", and "Audience sound". Lu et al. [120] employed a sliding-window SVM classification module for which five audio event classes were defined: "speech", "music", "cheering", "applause", and others. In this work, a sliding window with a fixed length was used to pre-segment the input audio stream by moving from the beginning of the stream to the end. SVM was used as the classification method for classifying each segment into one of the five audio classes.

Another promising approach for building audio event detectors includes dictionary learning techniques together with Bayesian Networks. For instance, in [149], Penet et al. used a dictionary learning and segment quantization approach where they replaced the low-level audio features extracted for each segment with one or several symbols corresponding to audio words. The quantization dictionary learning phase was implemented with a k-means algorithm using product quantization [92]. After the quantization step, a classifier was learned on audio words. Bayesian Networks were used to define a probability distribution over the features. The structure of the BNs is a sensitive issue, but such a structure can be efficiently learned from the data [77]. Moreover, the huge advantage of BNs over the popular SVMs is to have a very low parameter learning cost, and no hyperparameters to tune, this yielding better generalization capabilities. Contrarily to SVMs, the number of parameters in BN is only dependent on the structure of the graph and, in the absence of latent variables, the parameters are learned by counting in the learning database. Nevertheless, the BN inference complexity grows very fast with the number of variables used.

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

4.2. Video event detection

Video event detection (VED) aims at deciding on the existence of complex (or high-level) events in a given set of videos. Typically, video events are usually more complex than audio ones, since the visual channel of a video provides a richer amount of information (appearance, motion, etc) than the audio channel. As discussed in Sect. 2, such events may include complex activities, occurring at specific places and times, which involve people interacting with other people and/or object(s) [188, 97]. Indicative examples of such event classes include "changing a vehicle tire", "making a cake", or "attempting a bike trick", etc.

A typical VED framework includes a feature extraction stage, which usually generates a plethora of low-, intermediate-, and high-level features from the different available modalities (audio, visual, and/or textual). Moreover, a classification stage is in most cases included, where an event detector is learnt by training one or more classifiers for each event class using available features, while a fusion approach is often followed in order to combine different modalities.

4.2.1. Features

Features matter. Not only generally in many image and video processing problems [71], but also specifically in the problem of video event detection (VED) [98]. Despite the fact that what makes a feature good is a set of multifactorial parameters, good features are primarily meant to be robust against variations, such that videos of the same event class can still be correctly recognized under different conditions.

Since videos typically consist in two main sources of information, i.e., a visual and an audio channel, the research community has provided an abundance of studies for exploiting them efficiently. Both channels convey useful information for the problem of event detection; that is, the visual channel depicts appearance information related to objects, scene settings, etc, and captures critical motion information related to the movements of the constituent objects. On the other hand, the audio channel may contain acoustic cues that could be useful in detecting specific complex events.

Static frame-based visual features

The majority of the video event detection systems use a set of static frame-based visual features, that is, features that capture appearance-based attributes and are computed from a single frame. The most known of them include Scale-Invariant Feature Transform (SIFT), proposed by Lowe [119] and used in several studies [186, 140, 221, 217, 144, 187, 214, 216, 219, 109], and their color extensions, e.g. colorSIFT [21] used in [186, 140, 221, 212, 214, 216, 94, 109]. Bay et al. [13] proposed Speeded Up Robust Features (SURF), as a faster alternative descriptor using 2D Haar wavelet responses. In [186, 140, 187], GIST[2], which is a very low-dimensional scene representation proposed by Oliva and Torralba [143] is also used for the problem video event detection. In [187], Tang et al. used the Local Binary Patterns (LBP) descriptor [142].

Motion (spatio-temporal) visual features

Since detecting complex, time-evolving events relies by definition heavily on the temporal dimension of the video, several spatio-temporal video features have been proposed in the literature and used in state-of-the-art event detection systems. First, in [186, 212, 216, 109], Motion SIFT (MoSIFT) [33, 135], proposed by Chen and Hauptmann as a 3D version of SIFT, is used, while spatio-temporal interest points (STIP) [110] have been recently applied in VED systems [186, 217, 214, 216, 219, 109]. Dense trajectories [197, 44, 198] and improved dense trajectories [199] use dense optical flow to track feature points. The feature points used for this are typically described by Histogram of Oriented Gradients (HoG) [43, 187], Histogram of optical Flow (HoF) [197], or Motion Boundary Histogram (MBH) [197] descriptors, and are frequently used as motion features [186, 212, 144, 214, 216, 109], due to their discriminative power concerning time-evolving events. Finally, in [187], Tang et al. experiments with 3D Histogram of Oriented Gradients (HoG3D) [43, 103] for complex event recognition.

High-level visual features

Besides the above low-level visual features for video event detection, the research community has extensively studied higher-level representation

[2]The "gist" is an abstract representation of the scene that spontaneously activates memory representations of scene categories (a city, a mountain, etc). See Friedman [64].

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

schemes that aim to bridge the semantic gap between human and machine interpretation of multimedia. Merler et al. [132] and others [139, 75] proposed semantic model vectors, an intermediate-level representation as a basis for modeling and detecting complex events in Internet videos. The semantic model vectors were extracted using a set of discriminative semantic classifiers, each being an ensemble of SVM models trained from thousands of labeled web images, for a total of 280 generic concepts. The authors reported that the proposed representation approach outperforms -and is complementary to- other low-level visual descriptors for video event modeling.

In general, such intermediate- or high-level features are represented by a set of confidence scores estimating the probability of observing specific concepts in a video. This is more consistent with human's understanding and reasoning about the task, where an event is characterized by the presence/absence of certain concepts rather than interest points [94, 181, 114]. For instance, a video would be understood as belonging to the event class "birthday party" if visual concepts such as "birthday cake", "faces", or "cheering" are present. There are many studies dealing with video event detection that use intermediate- or high-level video representations, frequently along with a set of low-level features, e.g. [78, 40, 217, 224, 191, 117, 129, 109, 108, 83, 82, 179, 220, 118, 162].

Besides the above types of high-level features, which are extracted from images (i.e., keyframes of videos), there are also many studies that exploit the time-evolving nature of videos for representing them meaningfully in order to detect complex events. For instance, in [184], the authors proposed ACTIVE, a video event detection framework for exploiting activity concept transitions in video events. A video is treated as a sequence of short clips, all of which are observations corresponding to latent activity concept variables in a Hidden Markov Model (HMM). Fisher Kernel techniques are applied so that the concept transitions over time can be encoded into a compact and fixed length feature vector very efficiently. Moreover, Habibian et al. [83] studied on how to create an effective vocabulary of concepts for the problem of video event detection by examining the roles of the number, the type, the specificity, and the quality of the detectors in the concept vocabulary. Furthermore, in [82], Habibian et al. proposed VideoStory, a video representation scheme particularly suitable for learning event detectors from few training video samples. VideoStory, that achieves state-of-the-art results in detecting events in videos, learns from Web data a multimedia embedding, optimizing the visual projection for event recognition. Finally, in [125] Ma et al. proposed to leverage attributes at video level (video attributes), i.e., by using the semantic labels of external videos. Compared to complex event videos, these external videos contain simple contents such as objects, scenes and actions which are the basic elements of complex events. Specifically, building upon a correlation vector which correlates the attributes and the complex event, the authors incorporated video attributes latently as extra informative cues into the event detector learnt from complex event videos.

Selecting meaningful concepts for building a vocabulary for representing videos is an important challenge, since irrelevant concepts in a video representation scheme may introduce noise, resulting in worse detection results. For this, there are studies that choose to prune specific concepts from the concept vocabulary, depending on the coherence of those concepts with the event class that is desired to be detected. For instance, in [179], Singh et al. proposed an event detection algorithm that constructs pairs of automatically discovered concepts and then prunes those concepts that are unlikely to be helpful for retrieval. Pruning depends both on the query and on the specific video instance being evaluated. Similarly, Chang et al. [29] proposed a framework for pruning irrelevant noisy concepts towards improving event detection in Web videos.

Deep learning for video representation

In the last few years, the use of deep convolutional neural networks (CNN) for feature extraction has shown excellent results in image and video understanding and indexing problems [68]. Deep learning tries to model a high-level abstraction of data by using model architectures composed of multiple nonlinear transformations. Specifically, CNNs [111, 15] correspond to a biologically-inspired class of deep learning models that have demonstrated excellent abilities for high-level vision tasks, such as image classification [105, 52], object detection [71], and scene labeling [61]. Moreover, the features learned by large networks trained on the ImageNet dataset [49] show great generalization ability that yields state-of-the-art performance beyond standard image classification tasks, e.g., on several action recognition datasets [102, 177]. Be-

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

sides, the problem of understanding and visualizing deep CNNs [226, 60] has also attracted some attention. Very recently, [176, 226] proposed to localize the objects in images in a weakly supervised manner without relying on bounding box annotations. Compared to still image data and shot action videos, there is relatively little work on applying CNNs to multimedia event detection and recounting tasks.

In [213], Xu et al. chose to utilize the deep learning approach, especially Convolutional Neural Networks (CNNs), for the problem of video event detection, given their overwhelming accuracy in image analysis and high processing speed, which is achieved by leveraging the massive parallel processing power of GPUs [105]. They effectively leveraged deep CNNs to advance event detection, where only frame-level static descriptors can be extracted by the existing CNN toolkits. In [213], the authors make two contributions to the inference of CNN video representation. First, while average pooling and max pooling have long been the standard approaches to aggregating frame level static features, they show that performance can be significantly improved by taking advantage of an appropriate encoding method. Second, they proposed using a set of latent concept descriptors as the frame descriptor, which enriches visual information while keeping it computationally affordable. The integration of the two contributions results in state-of-the-art performance in event detection over the largest video datasets. Moreover, Zha et al. [227] conducted an in-depth exploration of different strategies for doing event detection in videos using CNNs trained for image classification. They studied different ways of performing spatial and temporal pooling, feature normalization, choice of CNN layers as well as choice of classifiers. Making judicious choices along these dimensions led to a significant increase in performance over previous, simpler approaches. Finally, Ye et al. [220] proposed EventNet, a large scale event-specific concept library that covers as many real-world events and their concepts as possible. This approach includes the training of a Convolutional Neural Network (CNN) model on a large set of videos over the 500 events, which subsequently is used to extract deep features from video content. With the learned deep learning features, a set of 4490 binary SVM classifiers are trained as the event-specific concept library. The concepts and events are further organized in a hierarchical structure (EventNet). The EventNet concept library is used to generate a concept-based representation of event videos.

Audio features

Acoustic information can be valuable for video event detection, particularly in the case where videos are captured under realistic conditions. Mel-Frequency Cepstral Coefficients (MFCC) [156] is one of the most popular audio features used to this end [47, 209, 214, 219, 109]. Other frequently used audio features are those discussed in Sect. 4.1.1.

Textual features

Textual information is often useful for detecting complex events in videos. The most useful techniques employed in event detection systems use Automatic Speech Recognition (ASR) [134] and Optical Character Recognition (OCR) [137]. ASR provides complementary information for events that are characterized by acoustic evidence. It is especially effective for close-to-camera and narrative events such as "town hall meeting" and "asking for directions" [96]. OCR captures characters in videos; the recognized characters are often not meaningful words but sometimes can be a clue for fine-grained detection, e.g., distinguishing between wording such as "baby shower" and "wedding shower". [96].

Feature encoding

Local features vary in number across different frames of videos, due to the different complexity, content, duration, etc. of the different pieces of visual information. This causes difficulties in measuring video/frame similarities, as most measurements require fixed-dimensional vectors. To this end, once the local low-level features are extracted, encodings such as the Fisher Vector (FV) [168] or VLAD [93, 163] representations, which were found to be the most effective ones in a recent evaluation study of feature pooling techniques for object recognition [30], are used to construct a signature characterizing the video. The FV extends the Bag-of-Words (BoW) representation [26, 100], which until recently was widely used for video classification [216, 94, 89]. The BoW approach relies on the quantization of the local descriptor space using off-line k-means clustering on a large collection of local descriptors.

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

4.2.2. Learning from training video examples

Learning from training video samples is the dominant approach in the video event detection community. The most common approach for building efficient video event detectors includes a) feature extraction from different modalities (i.e., static and motion visual features, audio features, and text-based attributes), b) learning of one classifier for each feature that is extracted (a standard kernel SVM is among the state-of-the-art ones), and c) an appropriate fusion technique that combines effectively the above. There are many video event detection systems that roughly follow this approach [219, 94, 2, 83, 179, 78, 224, 129, 109, 6, 73].

Concerning the combination of feature modalities, as discussed in Sect. 4.2.1, there are two popular fusion strategies [221] used in video event detection systems; i.e. early and late. Early fusion, also known as feature-level fusion has been widely used in computer vision and multimedia applications [9, 70]. Late fusion, on the other hand, aims at combining the confidence scores of the models constructed using different feature modalities, in which each confidence score measures the probability of classifying a test video sample into the positive event class by one specific model. There are also other, more sophisticated fusion methods that employ additional learning stage(s) for optimizing the weights with which each feature modality contributes to the final fused event detector. Also, many works build event detectors by applying combinations of early, late, and other fusion techniques [78, 219, 94, 2, 129, 109, 186, 221, 140, 214, 187, 83, 179].

Several video analysis works explicitly study the temporal structure of the video in order to effectively detect time-evolving complex events. In [16], the authors first propose a video representation that captures the temporal dynamics of mid-level concepts by expressing each video as an ordered vector time-series. Then, they plug their new features into linear SVMs for building event detectors. In [36], a video is represented by a sequence of visual words learnt from the video, and the Sequence Memoizer [208] is applied in order to capture long-range dependencies in a temporal context in the visual sequence. Then, an event detector is learnt (e.g., using a standard SVM). In [107], Lai et al. proposed an instance-based video event detection approach where each video is represented as multiple "instances", defined as video segments of different temporal intervals. Then, an instance-level event detector based only on video-level labels is learnt by applying a large-margin classifier that treats the instance labels as hidden latent variables and simultaneously infers the instance labels and the instance-level classification model. In a similar fashion, Vahdat et al. [193] proposed a compositional model for video event detection where a video is modeled using a collection of both global and segment-level features (which are treated as a latent variable). Then, a multiple kernel learning (MKL) latent SVM is defined and used to combine and re-weight multiple feature types while simultaneously operating within the latent variable framework (see also [185]). In [222], Ye et al. proposed a learning-based hashing method for video event detection, where a minimization problem over a structure-regularized empirical loss is efficiently solved by an Accelerated Proximal Gradient (APG) method. In particular, the structure regularization exploits the common local visual patterns occurring in video frames that are associated with the same semantic class, and simultaneously preserves the temporal consistency over successive frames from the same video. Assari et al. [7] proposed a contextual approach to video classification based on Generalized Maximum Clique Problem (GMCP), which uses the co-occurrence of concepts as the context model. More specifically, an event class is represented based on the co-occurrence of its concepts and a video is classified based on matching its semantic co-occurrence pattern to each class representation. The authors argued that, in principle, the co-occurrence of concepts yields a richer representation of a video compared to other approaches. Additionally, they proposed a novel optimal solution to GMCP based on Mixed Binary Integer Programming (MBIP). Finally, Ramanathan et al. [159] proposed to learn temporal embeddings of video frames for complex video analysis. The authors used unlabeled Web video data, which possess the implicit weak label that they are sequences of temporally and semantically coherent images. The authors leveraged this information in order to learn temporal embeddings for video frames by associating frames with the temporal context that they appear in. To do this, they proposed a scheme for incorporating temporal context based on past and future frames in videos, and compared this to other contextual representations. In addition, they showed how data augmentation using multi-resolution samples and hard negatives helps to sig-

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

nificantly improve the quality of the learned embeddings.

Selecting pooling regions for video event detection problem is characteristic of another set of literature approaches. In [115], Li et al. defined a dynamic pooling operator in order to enable a unified solution to the problems of event specific video segmentation, temporal structure modeling, and event detection. Video is first decomposed into segments, and the segments most informative for detecting a given event are identified, so as to dynamically determine the pooling operator most suited for each sequence. This dynamic pooling is implemented by treating the locations of characteristic segments as hidden information, which is inferred, on a sequence-by-sequence basis, via a large-margin classification rule with latent variables. Moreover, for the problem of video event retrieval, Douze et al. [53] proposed various hyper-pooling strategies that encode the frame descriptors into a representation of the video sequence in a stable manner and introduced a query expansion technique to improve the ranking in retrieval. In [23], Cao et al. proposed a visual representation, namely scene aligned pooling, for the task of event recognition in complex videos. Based on the observation that a video clip is often composed of shots corresponding to different scenes, the key idea of scene aligned pooling is to decompose any video features into concurrent scene components, and to construct classification models adaptive to different scenes.

While building an efficient and effective learning method for video event detection is a challenge in its own right, finding a sufficient number of videos that depict the event so as to use them as positive training samples for training any machine learning method is also not an easy feat. In fact, video event detection is even more challenging when the available positive training samples are limited. Ma et al. [121, 122] deal with the problem of learning from a few positive video samples by employing knowledge adaptation [54, 99] to facilitate event detection. Another way of addressing the scarcity of positive samples is to take advantage of any available videos that do not exactly fulfill the requirements to be characterized as true positive examples of an event class, but nevertheless are closely related to it. In [191, 190], the authors proposed Relevance Degree SVM (RD-SVM) in order to take advantage of related videos by exploiting them as weighted positives or weighted negatives in conjunction with an automatic weighting selection scheme. Also, in [192], an extension of kernel SVM that models and takes into consideration the uncertainty of each video sample, called kernel SVM with Isotropic Gaussian Sample Uncertainty (KSVM-iGSU), along with RD-KSVM and its RD-KSVM-iGSU extension are used for the problem of learning event detectors from a few positive and a few related video samples. Differently from the above, in [212] Xu et al. dealt with problem of utilizing related examples for complex event detection only when multiple features are available for training. The authors proposed an algorithm which adaptively utilizes the related examples by cross-feature learning. Ordinal labels were used to represent the multiple relevance levels of the related videos. Label candidates of related examples were generated by exploring the possible relevance levels of each related example via a cross-feature voting strategy. The maximum margin criterion was then applied in order to discriminate the positive and negative examples, as well as the related examples exhibiting different relevance levels. Finally, in [124], the authors proposed a framework for treating negative video samples differently than pure negatives; that is, they assigned fine-grained labels to negative examples for more effective exploitation based on the assumption that many negative videos may resemble the positive videos in different degrees. Since the resulting fine-grained labels may not be accurate enough to characterize the negative videos, the authors proposed to jointly optimize the fine-grained labels with the knowledge from the visual features and the attributes representations, which brings mutual reciprocality.

4.2.3. Learning from an event's textual description

Learning video event detectors from zero positive video examples draws motivation from the image classification domain. That is, due to the rapidly increasing number of images on the Web, extensive research efforts have been devoted in multi-label, zero-example (or few-example) classification in images [131, 147]. In [58], Elhoseiny et al. proposed a method for predicting unseen image classes from a textual description, using knowledge transfer from textual to visual features.

In the video domain, learning from zero positive examples is investigated primarily in the context of video event detection or video activity recognition [66]. In [82] this problem is addressed by transforming both the event's textual description and the visual content of un-classified videos in a

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

high-dimensional concept-based representation, using a large pool of concept detectors; then relevant videos are retrieved by computing the similarities between these representations. In [209], multiple low-level representations using both the visual and the audio content of the videos are extracted, along with higher-level semantic features coming from ASR transcripts, OCR, and off-the-shelf video concept detectors. This way, both audio-visual and textual features are expressed in a common high-dimensional concept space, where the computation of similarity is possible. In [81], logical operators are used to discover different types of composite concepts, which leads to better event detection performance.

Employing user's feedback on the detection results (relevance feedback) has shown to improve detection performance in the absence of training examples. Differently from the above works that do not use any feedback approach, Jiang et al. [95] proposed a relevance feedback approach in detecting complex events in videos using zero positive video examples achieving promising results. In this study, the authors used a relevance feedback approach in order to improve event detection results in the zero-example problem using features computed from several modalities. The main idea is to use the textual information that describes the event class in order to create queries for each modality. Then, the system results in ranked video lists, one per each modality. The top videos from these lists are used as a "pseudo label" video set on which a joint model is trained, and a new ranked list is produced and used for creating a new "pseudo label" set; this process is iterated a few times.

Moreover, using visual concepts in the zero-example video event detection problem achieved state-of-the-art results in [96], where the authors proposed E-Lamp. E-Lamp is a zero-example event detection system made of four subsystems. The first one is an off-line indexing component, while the rest of them compose the on-line event search module. In the off-line module, each video is represented with 4043 visual concepts along with ASR and OCR high-level features. Then, in the on-line search module, the user-specified event description is translated into a set of relevant concepts, called *system query*. This system query is used to retrieve the videos that are most relevant to the event. Finally, a pseudo-relevance feedback approach is exploited in order to improve the results.

In [90], differently from traditional zero-shot approaches, Jain et al. did not demand the design and specification of attribute classifiers and class-to-attribute mappings to allow for transfer from seen classes to unseen classes. Instead, in their proposed approach, called objects2action, they used a semantic word embedding that is spanned by a skip-gram model of thousands of object categories, where action labels are assigned to an object encoding of unseen video based on a convex combination of action and object affinities. The proposed embedding has three main characteristics to accommodate for the specifics of actions. First, the authors proposed a mechanism in order to exploit multiple-word descriptions of actions and objects. Second, they incorporated the automated selection of the most responsive objects per action. And finally, they demonstrated how to extend our zero-shot approach to the spatio-temporal localization of actions in video. Moreover, Gan et al. [67] addressed the problem of action recognition when no positive exemplars of that class are provided. For this, the authors, differently from other zero-shot learning approaches, which exploit attributes as the intermediate layer for the knowledge transfer, proposed the use of inter-class relationships (SIR), which directly leverages the semantic inter-class relationships between the known and unknown actions followed by label transfer learning. The inter-class semantic relationships are automatically measured by continuous word vectors, which learned by the skip-gram model using the large-scale text corpus.

In [57], Elhoseiny et al. proposed a novel zero-shot video event detection method by multi-modal distributional semantic embedding of videos. The proposed method embed object and action concepts, as well as other available modalities from videos, into a distributional semantic space. This is the first zero-shot event detection model that is built on top of distributional semantics and extends it in the following directions: (a) semantic embedding of multimodal information in videos (with focus on the visual modalities), (b) automatically determining relevance of concepts/attributes to a free text query, and (c) retrieving videos by free text event query (e.g., changing a vehicle tire) based on their content. The authors embedded videos into a distributional semantic space and then measured the similarity between videos and the event query in a free text form.

In [28], Chang et al. dealt with the problem of complex event detection in long Internet videos. A major challenge in this setting is that only a few

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

shots in a long video are relevant to the event of interest while others are irrelevant or even misleading. Instead of indifferently pooling the shots, the authors first defined a novel notion of semantic saliency that assesses the relevance of each shot with the event of interest; then, they prioritized the shots according to their saliency scores since shots that are semantically more salient are expected to contribute more to the final event detector. Next, they proposed a new isotonic regularizer that is able to exploit the semantic ordering information. The resulting nearly-isotonic SVM classifier exhibited higher discriminative power. In [27], Chang et al. first pre-trained a bundle of concept classifiers using data from other sources. Then, they evaluated the semantic correlation of each concept w.r.t. the event of interest and picked up the relevant concept classifiers, which were applied on all test videos to get multiple prediction score vectors. While most existing systems combine the predictions of the concept classifiers with fixed weights, they proposed to learn the optimal weights of the concept classifiers for each testing video by exploring a set of online available videos with free-form text descriptions of their content. Moreover, Mazloom et al. [130] proposed a new semantic video representation, called TagBook, that is based on freely available social tagged videos only, without the need for training any intermediate concept detectors. They introduced a simple algorithm that propagates tags from a videos nearest neighbors, similar in spirit to the ones used for image retrieval, but redesigned it for video event detection by including video source set refinement and varying the video tag assignment.

In an attempt to evaluate the impact of various different design choices, some of which discussed above, for building zero-example video event detectors, Tzelepis et al. [190] first identified a general learning framework from the textual information of an event class and then studied the impact of different design choices for various stages of this framework. This study goes beyond the classic semantic similarity comparison between a given event title (or other user-specified event cues) and each concept title from a concept pool, and tries to enrich each concept by automatically searching in Google or Wikipedia in order to find more information for it; this enables finding semantic similarities between events and concepts more effectively.

4.2.4. Video event recounting

Another challenging problem closely related to video event detection is video event recounting, which –given a video and its event-level annotation– aims to describe in a human-comprehensible way the key semantic entities that are depicted in this video and support the premise that the video belongs to the said event class [76, 224, 133, 189]. In other words, video event recounting refers to the task of providing comprehensible evidences to justify a detection result, e.g., why is this video classified as a "birthday party" event? This problem is a problem that was highlighted by the TRECVID Multimedia Event Recounting tasks [145, 72] in 2012–2014.

Most works in the literature focus on the temporal localization of an event's key evidences. In [155, 224, 183], the authors applied object and action detectors or low-level visual features, in order to localize temporal key evidences. They trained a video-level classifier and then used it to rank the keyframes or shots. These approaches are based on the assumption that the video-level classifiers that can distinguish positive and negative exemplars can also be used to distinguish the informative shots. These approaches equally treat the shots or key frames within the video, and thus the classifier may be confused by the ubiquitous but noninformative shots in videos. To overcome these limitations, in [106, 107], the authors formulated the problem as a multiple instance learning problem, aiming at learning an instance-level event detection and recounting model by selecting the informative shots or keyfames during the training process. In [29], Chang et al. proposed a joint framework to simultaneously classify high-level events and locate semantic evidences for each complex event. After extracting a semantic, albeit noisy, video representation, they introduced a recounting model that can localize key evidences both concept-wise and temporal-wise, and a detection model based on the infinite push support vector machine [164] that significantly enhanced the discriminative power.

In contrast to the above approaches that can only localize temporal key evidences, in [68], the authors proposed a flexible deep CNN infrastructure, namely Deep Event Network (DevNet), that simultaneously detects pre-defined events and provides key spatial-temporal evidences. Taking key frames of videos as input, they first detect the event of interest at the video level by aggregating the

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

CNN features of the key frames. The pieces of evidences which recount the detection results are also automatically localized, both temporally and spatially. Based on the intrinsic property of CNNs, the authors first generated a spatial-temporal saliency map by back passing through DevNet, which then could be used to find the key frames which are most indicative to the event, as well as to localize the specific spatial position, usually an object, in the frame of the highly indicative area.

5. Social event detection

Social Event Detection (SED) is about identifying events organized and attended by people, and captured in multiple media by them. Detecting social events typically involves processing collections of media items, such as collections of images, videos, and/or text, and finding the associations between different content items; this distinguishes the problem of social event detection from audio-visual event detection discussed in section 4, where processing and detection is performed on isolated content items (e.g., a single video at each time). Fig. 1b gives an outline of major research directions in social event detection. Social events are dominant in the content available in on-line social networks like Facebook[3], Twitter[4], Google plus[5], etc. As discussed above, social events are meant to happen in a certain point in time and at a specific place. In the literature, a more specific definition of a social event describes a social activity or a phenomenon that happened in real life at some point in time and in specific place, either *planned* or *abrupt*. For example, users in social networks tend to post updates about their daily activities and news that include social events such as athletic events, concerts, and exhibitions, but also disastrous natural phenomena like earthquakes, floods, and fires. However, since a vast amount of messages (e.g., tweets or Facebook posts) appear in social streams every minute, identifying interesting social events, free of irrelevant information, is a challenging task. Some of the challenges arising in building a useful social event detection system associate with the amount of data in social streams, the heterogeneity of possible social events, and the presence of fake or misleading content in them. The latter emerges as a particularly crucial parameter, especially when the events that need to be detected include dangerous or critical situations (e.g., in the case of an earthquake or a fire). Social event detection systems are designed to overcome the above challenges and discover real-time event instances separated from the rest of the noisy and often dominant in the social streams content.

A fundamental categorization of SED systems relies on the media type used for extracting features. Social event detection methods belong to one or more of the following: a) methods based solely on the textual information of social streams, b) methods that rely on visual information, such as images or videos, and c) methods that use meta-data information such as tags geo-location. Text-based detection systems, in the majority of related works, rely on Natural Language Processing (NLP) techniques, followed by some learning stage in order to generate features like Latent Dirichlet Allocation (LDA) [17] or Latent Semantic Indexing (LSI) [55]. On the other hand, visual-based detection systems apply state-of-the-art techniques derived from the field of computer vision and machine learning (see Sect. 4.2).

Another dimension along which SED systems can be categorized has to do with whether they use classification or clustering techniques. This depends, to some extent, on the range of social events that need to be detected in social streams. That is, clustering-based approaches aim to detect meaningful sets of media items in the streams in order to cluster all content into a (non-fixed) number of classes. In contrast, classification-based event detection approaches try to decide on whether or not a social event takes place. Furthermore, when the number of events is known a priori, these techniques can be applied to classify media into pre-specified social event classes.

Finally, another categorization criterion for SED systems is the nature of social events whose detection is desired. That is, planned and abrupt social events admit different detection algorithms, since planned events have been scheduled before the time they occur (e.g., athletic events, elections, exhibitions), while abrupt events are defined as non-scheduled events with no prior knowledge about the time or place they occur (e.g., earthquakes, accidents, fires). Due to this fact, abrupt social event detection systems typically use the temporal information in social streams by monitoring for abnor-

[3]https://www.facebook.com/
[4]https://twitter.com/
[5]https://plus.google.com/

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

mal topics or sudden bursts of a topic in them. On the contrary, planned social event detection systems often use scheduling information and features of the event that are known beforehand, such as the scheduled time and venue/place of an event, the description provided by the organizers and/or users about the event, etc.

The majority of the studies dealing with social event detection use the textual information of social streams and employ text-based analysis for building event detectors. Weng and Lee in [205] proposed EDCoW (Event Detection with Clustering of Wavelet-based Signals), a social event detection system that generates a signal for each word in a Twitter stream corpus and then applies wavelet analysis in order to detect the signal's bursts. After filtering recurring bursts (using their autocorrelation), the signals are cross-correlated and clustered using a graph partitioning of the resulting cross-correlation matrix. Finally, a measurement of the importance of each event is computed in order to distinguish between big events and trivial ones. Becker et al. [14] proposed a clustering approach for detecting social events in Flickr. They aim to identify documents that refer to a specific event given an amount of social media data. For this, they proposed a discriminative representation scheme that applied document similarity metrics in order to cluster and detect events. For every input document, they used the name of the user that created the document, the title and the name of the document, a short textual description that summarizes the document content, a set of tags describing the document content, and time and the location of the document publication. Subsequently, they transformed their textual features into a TF-IDF [112] weight vector and used cosine similarity as a similarity metric. As a typical additional step, they removed stop-words and applied stemming to their textual features. Yin et al. [223] developed a system for extracting situation awareness information based on Twitter data. The proposed framework detects bursts of words in the textual data, by modeling the number of tweets using a binomial distribution in order to estimate the number of tweets that contain a specific word. If the actual number of word occurrences becomes higher than the estimated number, then the word is characterized as burst. Events of interest include destructions in infrastructure, such as roads, bridges, railways, etc. To detect such events, the authors trained SVMs and Naive Bayesian classifiers [65]. In order to detect important and emerging topics, an online incremental clustering algorithm [14] was also applied. In contrast to [14], the authors in [223] used solely the Term Frequency-Inverse Document Frequency (TF-IDF) vector from the tweet (unimodal approach).

Finally, Sakaki et al. [167] considered Twitter users as "sensors" and tweets as "sensor information". They assumed that a user (sensor) detects a target event and reports it in Twitter. Events of interest in this work include earthquakes and typhoons. The proposed model is constructed in three steps: i) an SVM classifier decides on whether a tweet is related to a specific social event or not, ii) a temporal analysis of the tweets is performed to estimate a waiting time for raising an alarm, and iii) the location information of each tweet is used to calculate an estimate of the earthquake center or the trajectory of the typhoon. From the above studies, Becker et al. [14], Weng and Lee [205] used clustering-based approaches, while Sakaki et al. [167], Yin et al. [223] employed classification techniques. Additionally, in [14, 205, 167, 223], events of interest include solely abrupt social events, while in [14] detection of planned events is also supported.

Further studies additionally exploit the visual information extracted from social streams for building SED systems. In [167], Sakaki et al. used visual cues (images and videos) for deciding on whether a tweet belongs to a specific social event, while Petkos et al. [150] proposed a framework for the event-based clustering of multimedia content from social networks such as Flickr. The authors used a multimodal approach (metadata information along with visual descriptors) and defined the "same cluster" relationships between samples of the dataset by computing the similarities between all available modalities. A classification stage is then introduced in order to determine if a pair of images belongs to the same event. Thus, the matrix of pairwise distances between items is transformed to a pairwise similarity indicator matrix. k-means clustering is then applied on this indicator matrix in order to assign every image to an event. Both Sakaki et al. [167] and Petkos et al. [150] used classification techniques, while in [150] additionally used clustering methods in their system. The detection of both planned and abrupt social events is addressed in [150], in contrast to [167] that is designed for detecting solely abrupt events (e.g. earthquakes).

Finally, there are many works that use metadata

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

information, such as geo-location tags, for facilitating event detection in social streams. Becker et al. [14] used the names of the users that created documents in Flickr, a set of tags describing the document content, etc. Sakaki et al. [167] used location information in Twitter, while Petkos et al. [150] combined metadata information along with visual descriptors. Finally, Rafailidis et al. [157] presented a data-driven technique for social event detection. In the latter work, the collected social multimedia contained noisy metadata, with missing and possibly erroneous values. To counter these data implementations, they built initial clusters from content that contains spatial metadata, and they created singleton events for content with missing spatial information. Subsequently, a single-pass procedure was followed for clustering based on temporal information, and the anchored clusters (i.e., sets of data with fixed spatio-temporal information) were created. Then, the inter-correlations between anchored and singletons, or among singleton clusters, are computed to merge them into clusters. The inter-correlation between clusters is computed as the aggregated similarity from the various available modalities.

6. Event applications and evaluation activities

Event-based processing and analysis finds numerous applications in the domain of multimedia. Video surveillance, multimedia organization and video management are a few application domains where Event-based processing and analysis have been extensively used heretofore. News and sports applications rely significantly these days in event detection techniques, while Lifelogging and healthcare-related applications have also attempted to support and facilitate everyday life. Event visualization applications have also been implemented in order to allow navigation to a complex event, or just alleviate information overload in human analysis systems. Finally, as social media become more pervasive in on-line everyday life, social event detection applications draw increasing attention in the multimedia community. Thus, since Event-based processing and analysis systems are ubiquitous in multimedia, their performance needs to be evaluated in an open and fair way. To this end, a few well-known benchmarking activities have been organized in recent years.

6.1. Event applications

Video surveillance is a challenging and time-critical problem since it typically involves subtleties that are readily understood by humans, but difficult to encode for machine learning approaches, and can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc. Research efforts on multimedia event detection approaches for surveillance applications have recently shown a significant increase. SanMiguel et al. [169] used video event detection techniques using a list of interactions between objects, which, along with any other prior information concerning the context of a scene where the event evolves are used for the problem of video surveillance. In [39], Clavel et al. proposed a surveillance system that uses audio event detection. In the same direction, surveillance system in [39, 85] used audio event detection techniques for detecting gunshots in noisy environments in order to facilitate their surveillance system. Joo and Chellappa [101] proposed an event detection application in order to recognize atomic events in parking lot surveillance.

Event detection has recently gained significant attention in multimedia organization and consumer multimedia management. Dao et al. [46] addressed the problem of associating personal image collections with events by analyzing the photo collection of an event as a whole, rather than looking at individual images. The proposed method aims at detecting event types such as graduation, wedding, or different types of vacations and sports events, which describe the collection. Moreover, in [165], Ruocco and Ramampiaro dealt with the problem of event-based organization of images that are available in online photo-sharing applications such as Flickr. They proposed a clustering approach, which takes into account textual annotations as well time and geo-location metadata of the images. Finally, Zigkolis et al. [228] presented CrEve, a semi-automatic collaborative event annotation framework for the event-based organization of online images, which facilitates the annotation process and increases the coverage of the generated ground truth.

In the video domain, event-based techniques have also been proposed for consumer video management. Cricri et al. [42] exploited the readings of auxiliary sensors (such as accelerometers and GPS receivers, which are typically included in camera-enabled devices), for detecting interesting events in

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

user generated videos for the event-based organization of user-captured video content. They extracted high-level contextual information about the recording activity, while they also exploited multiple audio-visual recordings of a common event (e.g., music concerts), when available, to extract additional event-related information such as regions of interest in the videos. Furthermore, a healthcare event-based approach for detecting events in the daily activities of seniors within their home has been proposed in [34], where Cheng et al. addressed the problem by introducing a subspace Naive-Bayesian Mutual Information Maximization (sNBMIM) algorithm. The presented senior home activity recognition system was evaluated for eight categories of everyday home events, such as sleep, eat, wash.

Lifelogging is a user-controlled form of gathering personal multimedia information using wearable sensors in order to capture everyday activities [80]. The goal of lifelogging is to analyze a person's everyday behavior and experiences in terms of events, states, and relationships [200] in order to support and facilitate everyday life. Events, as discrete and repetitive activities, are defined as the unit of interaction in a lifelogging system [200, 201]. A challenging problem in a lifelogging system is to detect specific events after identifying first the boundaries of them. In [200], the authors tackle this challenge by semantically enriching lifelog events and creating semantic links between those descriptors and other external knowledge. Wang et al. [201] proposed an interestingness-based semantic aggregation and representation algorithm, to tackle the problem of event management and representation in visual lifelogging. Semantic concept interestingness is calculated by fusing image-level concepts which are then exploited to select a representation for the semantic event correlated to various event topics. Finally, in [153], a lifelogging approach is employed for therapeutic support to people suffering from dementia in order to help them maintain or regain cognition of their identity.

In the domain of journalism, Wang et al. [203] proposed Eventory, an event driven media sharing repository to facilitate community awareness. Sayyadi et al. [170] studied event detection algorithm for news-related retrieval systems. Moreover, in [146], Pahal et al. proposed an ontology-driven approach where any news event encodes a story related to some news topic within itself. These event stories comprise of sequence of events or patterns that occurred at some specific time and space. Special attention have drawn event-based applications concerning sports [113, 166, 204, 51].

As social media applications proliferate, an ever-increasing amount of multimedia content available on the Web is being created and, thus, effective social event detection systems are in critical need. Such applications relate to methods that can detect event-related media and group them by the events they illustrate or refer to. From the end user's perspective, finding digital content related to social events is challenging, requiring to search large volumes of data, possibly at different sources and sites [151, 161, 174, 45]. To this end, Iliakopoulou et al. [86] proposed a multi-step multimedia retrieval framework that collects relevant and diverse multimedia content from multiple social media sources given an input news story or event of interest. This framework utilizes a query formulation method in combination with relevance prediction. The query formulation method relies on the construction of a graph of keywords for generating refined queries about the event/news story of interest based on the results of a first-step high precision query. Relevance prediction is based on supervised learning using 12 features computed from the content (text, visual) and social context (popularity, publication time) of posted items.

There are also studies for visualizing multimedia events in order to permit navigation (through space and/or time) to a complex event, or just alleviate information overload in human analysis systems. For instance, in [48], Deligiannidis et al. proposed Semantic Event Tracker (SET), an interactive visualization tool for analyzing events in a three-dimensional environment. The authors modeled an event as an object that describes an action, its location, time, and relations to other objects. SET is capable of visualizing as well as navigating through the event data in all three aspects of space, time and theme. Temporal data is illustrated as a 3D multi-line in the 3D environment that connects consecutive events. The line is marked with user-selectable objects that represent the events being visualized. Upon an events selection, SET informs the user about semantically associated information via voice commands. Then, upon the users verbal command, SET can display semantically associated media such as digital images, audio and/or video clips. The system provides access to multi-source, heterogeneous, multimedia data, and is capable of visualizing events that contain geographic and time

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

Event visualization application	**Dataset(s)**	**Event types**
Deligiannidis et al. [48]	►Location and date/time information about hundreds of terrorist events in 600km proximity of Zaragoza (2000-2002)	Terrorism-related events
Reinders et al. [160]	Tracking of features of the following types: ►Synthetic data (motion paths with hundreds of nodes) ►Computational fluid dynamics simulation with turbulent vortex structures (100 frames) ►NASA Ames Research Center data for the application of flow past a tapered cylinder (400 frames)	In the context of a feature: (i) Continuation (ii) Birth and death (iii) Entry and exit (iv) Split and merge (v) Unresolved event
Chung et al. [38]	►Tucson Police Department's (TPD) databases (1.4 million incident records)	Crime incidents events

Table 3: Event visualization applications with datasets and event types descriptions

.

information. In a different direction, Reinders et al. [160] proposed a method to analyze and visualize time-dependent evolution of features in videos. The task of the visualization method is to extract the features from all frames, to determine the correspondences between features in successive frames, to detect significant events or stages in the evolution of the features, and, finally, to visualize the results. Finally, in [38], Chung et al. proposed a taxonomy of event visualization and presented COPLINK Spatio-Temporal Visualizer (STV), an event visualization tool integrates spatial, temporal, and aggregated data in order to support coordinated visualization of crime events. The tool can be used to summarize crime data, identify crime trends and reveal criminals' behavioral patterns. Table 3 gives an overview of the above event visualization applications with respect to the utilized datasets and the event types they support.

6.2. Evaluation activities

In the context of video event detection, the most popular benchmarking activity is organized yearly by TREC Video Retrieval Evaluation (TRECVID), whose goal is to promote progress in content-based exploitation of digital video via open, metrics-based evaluation [145, 225, 18, 72]. Specifically, in TRECVID Multimedia Event Detection (MED) task, the goal is to detect high-level pre-specified or ad-hoc events in a given set of videos. Such high-level events include "Attempting a bike trick", "Tuning a musical instrument", or "Horse riding competition", to name a few. Typically, a MED dataset consists in hundreds of hours of videos. Moreover, these videos belong to 20 pre-specified (PS) events in years 2012-2015, while there are also the so-called ad-hoc (AH) event classes (5 in 2012, 20 in 2013, and 10 in 2014-2015). Typical evaluation metrics for measuring the performance of a MED system are Mean Average Precision (MAP) and Inferred MAP. Moreover, TRECVID benchmarking activity includes Surveillance Event Detection (SED) [145] as a task, which aims at automatically detection of observable events of interest in surveillance videos. Such events in SED 2015 belong to two categories: (a) events that require understanding of the articulated body motion of a single person, such as "CellToEar", "ObjectPut", and "Pointing", and (b) events that can be revealed by the moving trajectories of a single person, such as "OpposingFlow" and "ElevatorNoEntry". The dataset used in SED 2015 task included approximately 100 hours of video. Normalized Detection Cost Rate (NDCR), Minimum NDCR, and NDCR at Target Operating Error Ratio (NDCR@TOER) were used as evaluate metrics for assessing a SED system's performance.

Another major benchmarking activity in the field of multimedia is MediaEval, which includes the Synchronization of Multi-User Event Media (SEM) [4, 41] task, which aims at aligning and presenting of media galleries of different users in a consistent way, so as to preserve the temporal evolution of the event. This is a challenging problem considering that the time information attached to some of the captured media may be wrong and geolocation information may be missing. Datasets used in this task include Tour De France 2014 (TDF14), NAMM Show 2015 (NAMM15), Salford Test Shoot (SAL), and Spring Parti Salesiani 2015 (SPS15) [41], which consist in thousands of images, hundreds of audio files and videos. The number of events supported in this benchmarking activity are in the magnitude of ten (e.g., 10 for SAL, 89 for TDF14 datasets) for 2015 [41]. For evaluating the performance of a SEM system, Jaccard index (JI)

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

and the clustering F1 score were used.

Especially for detecting social events in multimedia, the Social Event Detection (SED) task [161] of MediaEval, which has been organized between 2011 and 2014, has drawn the attention of the research community recently [151, 152, 174]. The datasets and event classes used in SED task in the period between 2011 and 2014, are as follows. In SED 2011, events were related to two categories: (a) soccer matches in Barcelona and Rome, (b) concerts in Paradiso and Parc del Forum; 73645 Flickr photos from five cities were used. In SED 2012, events were related to three categories: (a) technical events (e.g., exhibitions) in Germany, (b) soccer events in Hamburg and Madrid, (c) Indignados movement events in Madrid; 167332 Flickr photos from five cities were used. In SED 2013, the task required (a) to cluster the given photo collection, consisting in 57165 Instagram photos, into eight event types or non-event, and (b) to attach YouTube videos to the discovered events. In SED 2014, two datasets (362578 and 110541 images, respectively, collected from Flickr) were used. For both datasets, the actual image files and their metadata were available. Images were associated to distinct events in Last.fm[6] and Upcoming[7]. Finally, typical evaluation metrics for assessing the performance of event detectors were the harmonic mean of Precision and Recall (F-score), and the Normalized Mutual Information (NMI).

7. Conclusions and future challenges

In this paper, we have conducted a comprehensive analysis on an extensive set of event media processing and analysis techniques. We have reviewed several event definitions, depending on the complexity of events that are desired to be detected in multimedia content, as well as various event representation approaches that aspire to model events in meaningful ways. We also discussed event detection approaches in different media types, i.e., audio-, video-, and textual-based, and treated them individually in terms of feature representation and event inference. For the former, we discussed in detail a plethora of various state-of-the-art feature representation schemes, such as low-, intermediate-, and high-level ones, exploiting audio, visual (static- or motion-based), or textual information, in order to exploit every piece of information that is available for the specific detection problem in hand. Subsequently, we looked into event inference approaches that, given media representations, lead to building event detectors. More particularly, in this topic we discussed some typical state-of-the-art event detection learning approaches, but also more sophisticated ones that take thoroughly into consideration the time-evolving nature of the problem; for instance, by employing concept vocabularies and video attributes using external videos from the Web, to name a few. Finally, we discussed several issues that emerged because of particular event detection application requirements, such as video surveillance, multimedia organization, lifelogging, etc, while we further discussed about popular benchmarking activities for different problem of event detection in multimedia content, i.e., video event detection, social event detection, etc.

Events are everywhere and almost any media can be formalized as an event-based concept, e.g., images and audios are recorded in time, videos are naturally space-time representations, objects are enduring entities that extend across time just as they do in space, etc. Therefore, media processing and analysis can be seen as to be inherently event-driven. With the increase of the multi-modality awareness of the current new age of data (Big Multimedia Data), where data is now considered to be rather a spatio-temporal representation of information (visual-audio-textual-temporal) than seen independently as a set of different modalities that are merged together, event-based modeling became naturally part of the processing chain. Although significant progress has been made in this field and connected areas, we can identify some specific problems that are still open issues.

The first challenge is the model gap. There are many ways in conceptualizing the notion of event in order to facilitate its processing with the existing machine tools. The research community has contributed with many and of varying complexity definitions of the notion of an event in the field of multimedia. These definitions predominantly determine both the model with which each category of events is modeled, and the approaches followed for processing and analyzing multimedia content in order to build effective event processing and analysis tools. Finding an optimal trade-off between an event model's complexity and its detection performance remains a challenging open issue [173].

[6] http://www.last.fm/
[7] http://en.wikipedia.org/wiki/Upcoming

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

The second challenge is the data dimensionality and synchronization gap. As for the multimedia processing, there are currently countless ways of extracting meaningful information from the data, such as those feature representation schemes studied in this survey for exploiting audio-visual and textual information. Events are even more complex, being often multi-dimensional entities, each of the dimension being multi-modal itself. This raises, apart from the problem of the curse of dimensionality traditionally encountered for classification tasks, the problem of synchronization between different sources of information. This is still an open problem for which the research community continues to direct its efforts towards [4, 41].

Finally, another problem, which is also the problem of any multimedia system, is the semantic gap, i.e., the difference between the meaning that can be inferred from the information automatically extracted from data and its actual, human-based, understanding. As for the previous gap, this is inherently amplified by the much higher complexity of event data. To address this issue, the research community has recently shown promising results by employing Deep Convolutional Neural Networks (DCNNs). For instance, Karpathy et al. [102] studied on how to extend the connectivity of a DCNN in time domain in order to take advantage of local spatio-temporal information.

8. Acknowledgment

This work was supported by the European Union's Horizon 2020 and FP7 research and innovation programmes under grant agreements H2020-687786 InVID, H2020-693092 MOVING, FP7-611346 xLiMe and FP7-600826 ForgetIT.

References

[1] J. F. Allen, Maintaining knowledge about temporal intervals, Communications of the ACM 26 (11) (1983) 832–843.

[2] T. Althoff, H. O. Song, T. Darrell, Detection bank: an object detection based video representation for multimedia event recognition, in: Proceedings of the 20th ACM Int. Conf. on Multimedia, ACM, 2012, pp. 1065–1068.

[3] G. Antoniou, F. van Harmelen, Web ontology language: OWL, in: Handbook on ontologies, Springer, 2009, pp. 91–110.

[4] K. Apostolidis, C. Papagiannopoulou, V. Mezaris, CERTH at MediaEval 2014 Synchronization of Multi-User Event Media task, Proc. of MediaEval Workshop 1263.

[5] P. Appan, H. Sundaram, Networked multimedia event exploration, in: Proceedings of the 12th annual ACM Int. Conf. on Multimedia, ACM, 2004, pp. 40–47.

[6] S. Arestis-Chartampilas, N. Gkalelis, V. Mezaris, GPU accelerated generalised subclass discriminant analysis for event and concept detection in video, in: Proc. of the 23rd Annual ACM Conf. on Multimedia, ACM, 2015, pp. 1219–1222.

[7] S. M. Assari, A. R. Zamir, M. Shah, Video classification using semantic concept co-occurrences, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2529–2536.

[8] P. K. Atrey, N. C. Maddage, M. S. Kankanhalli, Audio based event detection for multimedia surveillance, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE Int. Conf. on, vol. 5, IEEE, 2006, pp. V–V.

[9] F. R. Bach, G. R. Lanckriet, M. I. Jordan, Multiple kernel learning, conic duality, and the smo algorithm, in: Proceedings of the twenty-first Int. Conf. on Machine Learning, ACM, 2004, p. 6.

[10] M. Baillie, J. M. Jose, Audio-based event detection for sports video, in: Image and Video Retrieval, Springer, 2003, pp. 300–309.

[11] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra, Event detection and recognition for semantic annotation of video, Multimedia Tools and Applications 51 (1) (2011) 279–302.

[12] N. Baumgartner, W. Retschitzegger, A survey of upper ontologies for situation awareness, in: Proc. of the 4th IASTED Int. Conf. on Knowledge Sharing and Collaborative Engineering, St. Thomas, US VI, 2006, pp. 1–9.

[13] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer vision and image understanding 110 (3) (2008) 346–359.

[14] H. Becker, M. Naaman, L. Gravano, Learning similarity metrics for event identification in social media, in: Proceedings of the third ACM Int. Conf. on Web search and data mining, ACM, 2010, pp. 291–300.

[15] Y. Bengio, Learning deep architectures for AI, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127.

[16] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, M. Shah, Recognition of complex events: Exploiting temporal dynamics between underlying concepts, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2243–2250.

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[18] R. Bolles, B. Burns, J. Herson, et al., The 2014 SESAME multimedia event detection and recounting system, in: Proc. TRECVID Workshop, 2014.

[19] M. Brand, Structure learning in conditional probability models via an entropic prior and parameter extinction, Neural Computation 11 (5) (1999) 1155–1182.

[20] D. Brezeale, D. J. Cook, Automatic video classification: A survey of the literature, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Trans. on 38 (3) (2008) 416–430.

[21] G. J. Burghouts, J.-M. Geusebroek, Performance evaluation of local colour invariants, Computer Vision and Image Understanding 113 (1) (2009) 48–62.

[22] L.-H. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, L.-H. Cai,

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

A flexible framework for key audio effects detection and auditory context inference, Audio, Speech, and Language Processing, IEEE Trans. on 14 (3) (2006) 1026–1039.

[23] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, J. R. Smith, Scene aligned pooling for complex video recognition, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 688–701.

[24] M. Casey, Mpeg-7 sound-recognition tools, IEEE Trans. on circuits and Systems for video Technology 11 (6) (2001) 737–747.

[25] I. Cervesato, M. Franceschet, A. Montanari, A guided tour through some extensions of the event calculus, Computational Intelligence 16 (2) (2000) 307–347.

[26] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, J. Luo, Large-scale multimodal semantic concept detection for consumer video, in: Proceedings of the Int. Workshop on Workshop on Multimedia Information Retrieval, ACM, 2007, pp. 255–264.

[27] X. Chang, Y. Yang, G. Long, C. Zhang, A. G. Hauptmann, Dynamic concept composition for zero-example event detection, in: Proc. of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 2016, pp. 3464–3470.

[28] X. Chang, Y. Yang, E. Xing, Y. Yu, Complex event detection using semantic saliency and nearly-isotonic svm, in: Proc. of the 32nd int. conf. on machine learning (ICML), 2015, pp. 1348–1357.

[29] X. Chang, Y.-L. Yu, Y. Yang, A. G. Hauptmann, Searching persuasively: Joint event detection and evidence recounting with limited supervision, in: Proceedings of the 23rd Annual ACM Conf. on Multimedia Conf., ACM, 2015, pp. 581–590.

[30] K. Chatfield, V. S. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods., in: BMVC, vol. 2, 2011, p. 8.

[31] H. Chen, T. Finin, A. Joshi, The SOUPA ontology for pervasive computing, in: Ontologies for agents: Theory and experiences, Springer, 2005, pp. 233–258.

[32] H. Chen, T. Finin, A. Joshi, Using owl in a pervasive computing broker, Tech. rep., DTIC Document (2005).

[33] M.-y. Chen, A. Hauptmann, MoSIFT: Recognizing human actions in surveillance videos, Technical Report CMU-CS, 2009.

[34] H. Cheng, Z. Liu, Y. Zhao, G. Ye, X. Sun, Real world activity summary for senior home monitoring, Multimedia Tools and Applications 70 (1) (2014) 177–197.

[35] W.-H. Cheng, W.-T. Chu, J.-L. Wu, Semantic context detection based on hierarchical audio models, in: Proceedings of the 5th ACM SIGMM Int. Workshop on Multimedia information retrieval, ACM, 2003, pp. 109–115.

[36] Y. Cheng, Q. Fan, S. Pankanti, A. Choudhary, Temporal sequence modeling for video event detection, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2235–2242.

[37] W.-T. Chu, W.-H. Cheng, J.-L. Wu, J. Y.-j. Hsu, A study of semantic context detection by using svm and gmm approaches, in: Multimedia and Expo, 2004. ICME'04. 2004 IEEE Int. Conf. on, vol. 3, IEEE, 2004, pp. 1591–1594.

[38] W. Chung, H. Chen, L. G. Chaboya, C. D. OToole, H. Atabakhsh, Evaluating event visualization: a usability study of COPLINK spatio-temporal visualizer, Int. Journal of Human-Computer Studies 62 (1) (2005) 127–157.

[39] C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system, in: Multimedia and Expo, 2005. ICME 2005. IEEE Int. Conf. on, IEEE, 2005, pp. 1306–1309.

[40] N. C. Codella, A. Natsev, G. Hua, M. Hill, L. Cao, L. Gong, J. R. Smith, Video event detection using temporal pyramids of visual semantics with kernel optimization and model subspace boosting, in: Multimedia and Expo (ICME), 2012 IEEE Int. Conf. on, IEEE, 2012, pp. 747–752.

[41] N. Conci, F. De Natale, V. Mezaris, M. Matton, Synchronization of multi-user event media at mediaeval 2015: Task description, datasets, and evaluation, in: MediaEval 2015 Workshop, Wurzen, Germany, 2015.

[42] F. Cricri, K. Dabov, I. D. Curcio, S. Mate, M. Gabbouj, Multimodal extraction of events and of information about the recording activity in user generated videos, Multimedia Tools and Applications 70 (1) (2014) 119–158.

[43] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conf. on, vol. 1, IEEE, 2005, pp. 886–893.

[44] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 428–441.

[45] M.-S. Dao, G. Boato, F. De Natale, T.-V. Nguyen, Jointly exploiting visual and non-visual information for event-related social media retrieval, in: Proc. of the 3rd ACM conf. on Int. conference on multimedia retrieval, ACM, 2013, pp. 159–166.

[46] M.-S. Dao, D.-T. Dang-Nguyen, F. De Natale, Robust event discovery from photo collections using signature image bases (sibs), Multimedia Tools and Applications 70 (1) (2014) 25–53.

[47] S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, Acoustics, Speech and Signal Processing, IEEE Trans. on 28 (4) (1980) 357–366.

[48] L. Deligiannidis, F. Hakimpour, A. P. Sheth, Event visualization in a 3d environment, in: Human System Interactions, 2008 Conf. on, IEEE, 2008, pp. 158–164.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conf. on, IEEE, 2009, pp. 248–255.

[50] M. Doerr, C.-E. Ore, S. Stead, The CIDOC conceptual reference model: a new standard for knowledge sharing, in: Tutorials, posters, panels and industrial contributions at the 26th Int. Conf. on Conceptual modeling-Volume 83, Australian Computer Society, Inc., 2007, pp. 51–56.

[51] K. Doman, T. Tomita, I. Ide, D. Deguchi, H. Murase, Event detection based on twitter enthusiasm degree for generating a sports highlight video, in: Proceedings of the ACM Int. Conf. on Multimedia, ACM, 2014, pp. 949–952.

[52] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, arXiv

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

preprint arXiv:1310.1531.

[53] M. Douze, J. Revaud, C. Schmid, H. Jégou, Stable hyper-pooling and query expansion for event detection, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 1825–1832.

[54] L. Duan, D. Xu, I.-H. Tsang, J. Luo, Visual event recognition in videos by learning from web data, Pattern Analysis and Machine Intelligence, IEEE Trans. on 34 (9) (2012) 1667–1680.

[55] S. T. Dumais, Latent semantic analysis, Annual review of information science and technology 38 (1) (2004) 188–230.

[56] A. Ekin, R. Mehrotra, et al., Integrated semantic-syntactic video modeling for search and browsing, Multimedia, IEEE Trans. on 6 (6) (2004) 839–851.

[57] M. Elhoseiny, J. Liu, H. Cheng, H. Sawhney, A. Elgammal, Zero-shot event detection by multimodal distributional semantic embedding of videos, arXiv preprint arXiv:1512.00818.

[58] M. Elhoseiny, B. Saleh, A. Elgammal, Write a classifier: Zero-shot learning using purely textual descriptions, in: Computer Vision (ICCV), IEEE Int. Conf. on, IEEE, 2013, pp. 2584–2591.

[59] J. P. Elo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, A. Serralheiro, Non-speech audio event detection, in: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE Int. Conf. on, IEEE, 2009, pp. 1973–1976.

[60] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network, University of Montreal 1341.

[61] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (8) (2013) 1915–1929.

[62] C. Fraley, A. E. Raftery, How many clusters? which clustering method? answers via model-based cluster analysis, The computer journal 41 (8) (1998) 578–588.

[63] A. R. Francois, R. Nevatia, J. Hobbs, R. C. Bolles, J. R. Smith, VERL: an ontology framework for representing and annotating video events, MultiMedia, IEEE 12 (4) (2005) 76–86.

[64] A. Friedman, Framing pictures: the role of knowledge in automatized encoding and memory for gist., Journal of Experimental Psychology: General 108 (3) (1979) 316.

[65] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine learning 29 (2-3) (1997) 131–163.

[66] C. Gan, M. Lin, Y. Yang, G. de Melo, A. G. Hauptmann, Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition, in: Proceedings of the 30th AAAI Conf. on Artificial Intelligence (AAAI 2016), AAAI Press, 2016.

[67] C. Gan, M. Lin, Y. Yang, Y. Zhuang, A. G. Hauptmann, Exploring semantic inter-class relationships (sir) for zero-shot action recognition, in: Twenty-Ninth AAAI Conf. on Artificial Intelligence, 2015.

[68] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, A. G. Hauptmann, DevNet: A deep event network for multimedia event detection and evidence recounting, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 2568–2577.

[69] A. Gangemi, V. Presutti, Ontology design patterns, in: Handbook on ontologies, Springer, 2009, pp. 221–243.

[70] P. Gehler, S. Nowozin, On feature combination for multiclass object classification, in: Computer Vision, 2009 IEEE 12th Int. Conf. on, IEEE, 2009, pp. 221–228.

[71] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 580–587.

[72] N. Gkalelis, F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, K. Avgerinakis, N. Pittaras, S. Vrochidis, V. Mezaris, I. Kompatsiaris, I. Patras, ITI-CERTH participation to TRECVID 2014, in: Proc. TRECVID Workshop, 2014.

[73] N. Gkalelis, V. Mezaris, Video event detection using generalized subclass discriminant analysis and linear support vector machines, in: Proc. of int. conf. on multimedia retrieval, ACM, 2014, p. 25.

[74] N. Gkalelis, V. Mezaris, I. Kompatsiaris, A joint content-event model for event-centric multimedia indexing, in: Semantic Computing (ICSC), 2010 IEEE Fourth Int. Conf. on, IEEE, 2010, pp. 79–84.

[75] N. Gkalelis, V. Mezaris, I. Kompatsiaris, High-level event detection in video exploiting discriminant concepts, in: Content-Based Multimedia Indexing (CBMI), 2011 9th Int. Workshop on, IEEE, 2011, pp. 85–90.

[76] N. Gkalelis, V. Mezaris, I. Kompatsiaris, T. Stathaki, Video event recounting using mixture subclass discriminant analysis, in: Image Processing (ICIP), 2013 20th IEEE Int. Conf. on, IEEE, 2013, pp. 4372–4376.

[77] G. Gravier, C.-H. Demarty, S. Baghdadi, P. Gros, Classification-oriented structure learning in bayesian networks for multimodal event detection in videos, Multimedia tools and applications 70 (3) (2014) 1421–1437.

[78] J. Guo, D. Scott, F. Hopfgartner, C. Gurrin, Detecting complex events in user-generated video using concept classifiers, in: Content-Based Multimedia Indexing (CBMI), 2012 10th Int. Workshop on, IEEE, 2012, pp. 1–6.

[79] A. Gupta, R. Jain, Managing event information: Modeling, retrieval, and applications, Synthesis Lectures on Data Management 3 (4) (2011) 1–141.

[80] C. Gurrin, A. F. Smeaton, A. R. Doherty, Lifelogging: Personal big data, Foundations and Trends in Information Retrieval 8 (1) (2014) 1–125.

[81] A. Habibian, T. Mensink, C. G. Snoek, Composite concept discovery for zero-shot video event detection, in: Proceedings of Int. Conf. on Multimedia Retrieval, ACM, 2014, p. 17.

[82] A. Habibian, T. Mensink, C. G. Snoek, VideoStory: A new multimedia embedding for few-example recognition and translation of events, in: Proceedings of the ACM Int. Conf. on Multimedia, ACM, 2014, pp. 17–26.

[83] A. Habibian, K. E. van de Sande, C. G. Snoek, Recommendations for video event recognition using concept vocabularies, in: Proceedings of the 3rd ACM Conf. on Int. Conf. on Multimedia Retrieval, ACM, 2013, pp. 89–96.

[84] A. Hakeem, Y. Sheikh, M. Shah, CASEˆ E: A hierarchical event representation for the analysis of videos, in: AAAI, 2004, pp. 263–268.

[85] A. Härmä, M. F. McKinney, J. Skowronek, Automatic

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

surveillance of the acoustic activity in our living environment, in: Multimedia and Expo, 2005. ICME 2005. IEEE Int. Conf. on, IEEE, 2005, pp. 4–pp.

[86] K. Iliakopoulou, S. Papadopoulos, Y. Kompatsiaris, News-oriented multimedia search over multiple social networks, in: Content-Based Multimedia Indexing (CBMI), 2015 13th Int. Workshop on, IEEE, 2015, pp. 1–6.

[87] U. IPTC Int. Press Telecommunications Council, London, NewsML, Tech. rep., https://iptc.org/standards/newsml-g2/.

[88] U. IPTC Int. Press Telecommunications Council, London, EventML, 2012., Tech. rep., https://iptc.org/standards/eventsml-g2/ (2012).

[89] H. Izadinia, M. Shah, Recognizing complex events using large margin joint low-level event model, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 430–444.

[90] M. Jain, J. C. van Gemert, T. Mensink, C. G. Snoek, Objects2action: Classifying and localizing actions without any video example, in: Proceedings of the IEEE Int. Conf. on Computer Vision, 2015, pp. 4588–4596.

[91] R. Jain, EventWeb: Developing a human-centered computing system, Computer (2) (2008) 42–50.

[92] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, Pattern Analysis and Machine Intelligence, IEEE Trans. on 33 (1) (2011) 117–128.

[93] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (9) (2012) 1704–1716.

[94] L. Jiang, A. G. Hauptmann, G. Xiang, Leveraging high-level and low-level features for multimedia event detection, in: Proceedings of the 20th ACM Int. Conf. on Multimedia, ACM, 2012, pp. 449–458.

[95] L. Jiang, T. Mitamura, S.-I. Yu, A. G. Hauptmann, Zero-example event search using multimodal pseudo relevance feedback, in: Proceedings of Int. Conf. on Multimedia Retrieval, ACM, 2014, p. 297.

[96] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, A. G. Hauptmann, Bridging the ultimate semantic gap: A semantic search engine for internet videos, in: Int. Conf. on Multimedia Retrieval, 2015.

[97] Y. Jiang, Q. Dai, T. Mei, Y. Rui, S. Chang, Super fast event recognition in internet videos, IEEE Trans. on Multimedia 17 (8) (2015) 1174–1186.

[98] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, M. Shah, High-level event recognition in unconstrained videos, Int. Journal of Multimedia Information Retrieval 2 (2) (2013) 73–101.

[99] Y.-G. Jiang, C.-W. Ngo, S.-F. Chang, Semantic context transfer across heterogeneous sources for domain adaptive video search, in: Proceedings of the 17th ACM Int. Conf. on Multimedia, ACM, 2009, pp. 155–164.

[100] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. G. Hauptmann, Representations of keypoint-based semantic concept detection: A comprehensive study, Multimedia, IEEE Trans. on 12 (1) (2010) 42–53.

[101] S.-W. Joo, R. Chellappa, Attribute grammar-based event recognition and anomaly detection, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conf. on, IEEE, 2006, pp. 107–107.

[102] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 1725–1732.

[103] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conf., British Machine Vision Association, 2008, pp. 275–1.

[104] R. Kowalski, M. Sergot, A logic-based calculus of events, in: Foundations of knowledge base management, Springer, 1989, pp. 23–55.

[105] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[106] K.-T. Lai, D. Liu, M.-S. Chen, S.-F. Chang, Recognizing complex events in videos by learning key static-dynamic evidences, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 675–688.

[107] K.-T. Lai, F. X. Yu, M.-S. Chen, S.-F. Chang, Video event detection by inferring temporal instance labels, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2251–2258.

[108] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, A. G. Hauptmann, Multimedia classification and event detection using double fusion, Multimedia Tools and Applications 71 (1) (2014) 333–347.

[109] Z.-Z. Lan, Y. Yang, N. Ballas, S.-I. Yu, A. Haputmann, Resource constrained multimedia event detection, in: Multimedia Modeling, Springer, 2014, pp. 388–399.

[110] I. Laptev, On space-time interest points, Int. Journal of Computer Vision 64 (2-3) (2005) 107–123.

[111] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[112] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of massive datasets, Cambridge University Press, 2014.

[113] B. Li, M. I. Sezan, Event detection and summarization in sports video, in: Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001). IEEE Workshop on, IEEE, 2001, pp. 132–138.

[114] L.-J. Li, H. Su, L. Fei-Fei, E. P. Xing, Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.

[115] W. Li, Q. Yu, A. Divakaran, N. Vasconcelos, Dynamic pooling for complex event recognition, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 2728–2735.

[116] F. Lin, Embracing causality in specifying the indeterminate effects of actions, in: Proceedings of the thirteenth National Conf. on Artificial Intelligence-Volume 1, AAAI Press, 1996, pp. 670–676.

[117] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, H. Sawhney, Video event recognition using concept attributes, in: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE, 2013, pp. 339–346.

[118] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, A. Yanagawa, Kodak's consumer video benchmark data set: concept definition

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

and annotation, in: Proceedings of the Int. Workshop on Workshop on Multimedia Information Retrieval, ACM, 2007, pp. 245–254.

[119] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. journal of computer vision 60 (2) (2004) 91–110.

[120] L. Lu, F. Ge, Q. Zhao, Y. Yan, A svm-based audio event detection system, in: Electrical and Control Engineering (ICECE), 2010 Int. Conf. on, IEEE, 2010, pp. 292–295.

[121] Z. Ma, Y. Yang, Y. Cai, N. Sebe, A. G. Hauptmann, Knowledge adaptation for ad hoc multimedia event detection with few exemplars, in: Proceedings of the 20th ACM Int. Conf. on Multimedia, ACM, 2012, pp. 469–478.

[122] Z. Ma, Y. Yang, N. Sebe, A. G. Hauptmann, Knowledge adaptation with partiallyshared features for event detectionusing few exemplars, Pattern Analysis and Machine Intelligence, IEEE Trans. on 36 (9) (2014) 1789–1802.

[123] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. G. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, Multimedia, IEEE Trans. on 15 (7) (2013) 1628–1637.

[124] Z. Ma, Y. Yang, Z. Xu, N. Sebe, A. G. Hauptmann, We are not equally negative: fine-grained labeling for multimedia event detection, in: Proceedings of the 21st ACM Int. Conf. on Multimedia, ACM, 2013, pp. 293–302.

[125] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, A. G. Hauptmann, Complex event detection via multi-source video attributes, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conf. on, IEEE, 2013, pp. 2627–2633.

[126] C. J. Matheus, M. M. Kokar, K. Baclawski, A core ontology for situation awareness, in: Proceedings of the Sixth Int. Conf. on Information Fusion, vol. 1, 2003, pp. 545–552.

[127] C. J. Matheus, M. M. Kokar, K. Baclawski, J. A. Letkowski, C. Call, M. L. Hinman, J. J. Salerno, D. M. Boulware, SAWA: An assistant for higher-level fusion and situation awareness, in: Defense and Security, Int. Society for Optics and Photonics, 2005, pp. 75–85.

[128] C. J. Matheus, M. M. Kokar, K. Baclawski, J. J. Letkowski, An application of semantic web technologies to situation awareness, in: The Semantic Web–ISWC 2005, Springer, 2005, pp. 944–958.

[129] M. Mazloom, E. Gavves, K. van de Sande, C. Snoek, Searching informative concept banks for video event detection, in: Proceedings of the 3rd ACM Conf. on Int. Conf. on Multimedia Retrieval, ACM, 2013, pp. 255–262.

[130] M. Mazloom, X. Li, C. G. Snoek, TagBook: A semantic video representation without supervision for event detection, arXiv preprint arXiv:1510.02899.

[131] T. Mensink, E. Gavves, C. G. Snoek, COSTA: Co-occurrence statistics for zero-shot classification, in: Computer Vision and Pattern Recognition (CVPR), IEEE Conf. on, IEEE, 2014, pp. 2441–2448.

[132] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, Multimedia, IEEE Trans. on 14 (1) (2012) 88–101.

[133] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, C. G. Snoek, Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting, in: Proceedings of the 5th ACM on Int. Conf. on Multimedia Retrieval, ACM, 2015, pp. 427–434.

[134] Y. Miao, L. Jiang, H. Zhang, F. Metze, Improvements to speaker adaptive training of deep neural networks, in: Spoken Language Technology Workshop (SLT), 2014 IEEE, IEEE, 2014, pp. 165–170.

[135] Y. Ming, Human activity recognition based on 3d mesh mosift feature descriptor, in: Int. Conf. on Social Computing, SocialCom 2013, Washington, DC, USA, 8-14 September, 2013, 2013, pp. 959–962.

[136] T. K. Moon, The expectation-maximization algorithm, Signal processing magazine, IEEE 13 (6) (1996) 47–60.

[137] S. Mori, H. Nishida, H. Yamada, Optical character recognition, John Wiley & Sons, Inc., 1999.

[138] N. Morsillo, G. Mann, C. Pal, Youtube scale, large vocabulary video annotation, in: Video Search and Mining, Springer, 2010, pp. 357–386.

[139] A. Moumtzidou, A. Dimou, N. Gkalelis, S. Vrochidis, V. Mezaris, I. Kompatsiaris, ITI-CERTH participation to TRECVID 2010., in: TRECVID, 2010.

[140] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, P. Natarajan, Multimodal feature fusion for robust event detection in web videos, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf. on, IEEE, 2012, pp. 1298–1305.

[141] R. Nevatia, J. Hobbs, B. Bolles, An ontology for video event representation, in: Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conf. on, IEEE, 2004, pp. 119–119.

[142] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Trans. on 24 (7) (2002) 971–987.

[143] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. journal of computer vision 42 (3) (2001) 145–175.

[144] D. Oneata, J. Verbeek, C. Schmid, Action and event recognition with fisher vectors on a compact feature set, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 1817–1824.

[145] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, G. Quenot, An overview of the goals, tasks, data, evaluation mechanisms and metrics, in: Proc. of TRECVID 2014, NIST, USA, 2014.

[146] N. Pahal, S. Chaudhury, V. Gaur, B. Lall, A. Mallik, Detecting and correlating video-based event patterns: An ontology driven approach, in: Proceedings of the 2014 IEEE/WIC/ACM Int. Joint Conf.s on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01, IEEE Computer Society, 2014, pp. 438–445.

[147] M. Palatucci, D. Pomerleau, G. E. Hinton, T. M. Mitchell, Zero-shot learning with semantic output codes, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta (eds.), Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009, pp. 1410–1418.

[148] S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, I. Kompatsiaris, Social event detection at mediaeval 2011: Challenges, dataset and evaluation., in: Medi-

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

aEval, 2011.

[149] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, Audio event detection in movies using multiple audio words and contextual bayesian networks, in: Content-Based Multimedia Indexing (CBMI), 2013 11th Int. Workshop on, IEEE, 2013, pp. 17–22.

[150] G. Petkos, S. Papadopoulos, Y. Kompatsiaris, Social event detection using multimodal clustering and integrating supervisory signals, in: Proceedings of the 2nd ACM Int. Conf. on Multimedia Retrieval, ACM, 2012, p. 23.

[151] G. Petkos, S. Papadopoulos, V. Mezaris, Y. Kompatsiaris, Social event detection at mediaeval 2014: Challenges, datasets, and evaluation, in: MediaEval 2014 Workshop, Barcelona, Spain, 2014.

[152] G. Petkos, S. Papadopoulos, V. Mezaris, R. Troncy, P. Cimiano, T. Reuter, Y. Kompatsiaris, Social event detection at mediaeval: a three-year retrospect of tasks and results, in: Proc. ACM ICMR 2014 Workshop on Social Events in Web Multimedia (SEWM), 2014.

[153] P. Piasek, A. F. Smeaton, et al., Using lifelogging to help construct the identity of people with dementia, in: Irish Human Computer Interaction Conf. 2014, DCU, Dublin, Ireland, 2014.

[154] R. Poppe, A survey on vision-based human action recognition, Image and vision computing 28 (6) (2010) 976–990.

[155] D. Potapov, M. Douze, Z. Harchaoui, C. Schmid, Category-specific video summarization, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 540–555.

[156] L. Rabiner, B.-H. Juang, Fundamentals of speech recognition.

[157] D. Rafailidis, T. Semertzidis, M. Lazaridis, M. G. Strintzis, P. Daras, A data-driven approach for social event detection., in: MediaEval, 2013.

[158] Y. Raimond, S. Abdallah, The event ontology, Tech. rep., Technical report, 2007. http://motools. sourceforge. net/event (2007).

[159] V. Ramanathan, K. Tang, G. Mori, L. Fei-Fei, Learning temporal embeddings for complex video analysis, in: Proceedings of the IEEE Int. Conf. on Computer Vision, 2015, pp. 4471–4479.

[160] F. Reinders, F. H. Post, H. J. Spoelder, Visualization of time-dependent data with feature tracking and event detection, The Visual Computer 17 (1) (2001) 55–71.

[161] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, S. Geva, Social event detection at mediaeval 2013: Challenges, datasets, and evaluation, in: Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013, 2013.

[162] A. Rosani, G. Boato, F. De Natale, A game-based framework for event-saliency identification in images, IEEE Trans. on Multimedia 17 (8) (2015) 1359–1371.

[163] N. Rostamzadeh, J. Uijlings, I. Mironica, M. K. Abadi, B. Ionescu, N. Sebe, Cluster encoding for modelling temporal variation in video, in: Image Processing (ICIP), 2015 IEEE Int. Conf. on, IEEE, 2015.

[164] C. Rudin, The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list, The Journal of Machine Learning Research 10 (2009) 2233–2271.

[165] M. Ruocco, H. Ramampiaro, A scalable algorithm for extraction and clustering of event-related pictures, Multimedia Tools and Applications 70 (1) (2014) 55–88.

[166] D. Sadlier, N. E. O'Connor, et al., Event detection in field sports video using audio-visual features and a support vector machine, Circuits and Systems for Video Technology, IEEE Trans. on 15 (10) (2005) 1225–1233.

[167] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th Int. Conf. on World wide web, ACM, 2010, pp. 851–860.

[168] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, Int. journal of computer vision 105 (3) (2013) 222–245.

[169] J. C. SanMiguel, J. M. Martínez, Á. García, An ontology for event detection and its application in surveillance video, in: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE Int. Conf. on, IEEE, 2009, pp. 220–225.

[170] H. Sayyadi, M. Hurst, A. Maykov, Event detection and tracking in social streams., in: ICWSM, 2009.

[171] A. Scherp, S. Agaram, R. Jain, Event-centric media management, in: Electronic Imaging 2008, Int. Society for Optics and Photonics, 2008, pp. 68200C–68200C.

[172] A. Scherp, T. Franz, C. Saathoff, S. Staab, F–a model of events based on the foundational ontology DOLCE+DnS Ultralight, in: Proceedings of the fifth Int. Conf. on Knowledge capture, ACM, 2009, pp. 137–144.

[173] A. Scherp, V. Mezaris, Survey on modeling and indexing events in multimedia, Multimedia Tools and Applications 70 (1) (2014) 7–23.

[174] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, P. A. Mitkas, Multimodal graph-based event detection and summarization in social media streams, in: Proceedings of the 23rd Annual ACM Conf. on Multimedia Conf., ACM, 2015, pp. 189–192.

[175] R. Shaw, R. Troncy, L. Hardman, LODE: Linking open descriptions of events, in: The Semantic Web, Springer, 2009, pp. 153–167.

[176] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034.

[177] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[178] P. Sinclair, M. Addis, F. Choi, M. Doerr, P. Lewis, K. Martinez, The use of CRM core in multimedia annotation, in: In, First Int. Workshop on Semantic Web Annotations for Multimedia (SWAMM), Edinburgh, Scotland,, 2006.

[179] B. Singh, X. Han, Z. Wu, V. I. Morariu, L. S. Davis, Selecting relevant web trained concepts for automated event retrieval, in: Computer Vision (ICCV), 2015 IEEE Int. Conf. on, IEEE, 2015.

[180] C. G. Snoek, M. Worring, Concept-based video retrieval, Foundations and Trends in Information Retrieval 2 (4) (2008) 215–322.

[181] C. G. Snoek, M. Worring, A. W. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th annual ACM Int. Conf. on Multimedia, ACM, 2005, pp. 399–402.

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

[182] O. Standard, Common alerting protocol version 1.2 (2010).

[183] C. Sun, B. Burns, R. Nevatia, C. Snoek, B. Bolles, G. Myers, W. Wang, E. Yeh, ISOMER: Informative segment observations for multimedia event recounting, in: Proceedings of Int. Conf. on Multimedia Retrieval, ACM, 2014, p. 241.

[184] C. Sun, R. Nevatia, ACTIVE: Activity concept transitions in video event classification, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 913–920.

[185] C. Sun, R. Nevatia, DISCOVER: Discovering important segments for classification of video events and recounting, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2569–2576.

[186] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, H. Sawhney, Evaluation of low-level features and their combinations for complex event detection in open source videos, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf. on, IEEE, 2012, pp. 3681–3688.

[187] K. Tang, B. Yao, L. Fei-Fei, D. Koller, Combining the right features for complex event recognition, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 2696–2703.

[188] W. Tong, Y. Yang, L. Jiang, S.-I. Yu, Z. Lan, Z. Ma, W. Sze, E. Younessian, A. G. Hauptmann, E-lamp: integration of innovative ideas for multimedia event detection, Machine vision and applications 25 (1) (2014) 5–15.

[189] C.-Y. Tsai, M. L. Alexander, N. Okwara, J. R. Kender, Highly efficient multimedia event recounting from user semantic preferences, in: Proceedings of Int. Conf. on Multimedia Retrieval, ACM, 2014, p. 419.

[190] C. Tzelepis, D. Galanopoulos, V. Mezaris, I. Patras, Learning to detect video events from zero or very few video examples, Image and vision Computing, 2015.

[191] C. Tzelepis, N. Gkalelis, V. Mezaris, I. Kompatsiaris, Improving event detection using related videos and relevance degree support vector machines, in: Proceedings of the 21st ACM Int. Conf. on Multimedia, ACM, 2013, pp. 673–676.

[192] C. Tzelepis, V. Mezaris, I. Patras, Video event detection using kernel support vector machine with isotropic gaussian sample uncertainty (KSVM-iGSU), in: MultiMedia Modeling - 22nd Int. Conf., MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I, Springer, 2016, pp. 3–15.

[193] A. Vahdat, K. Cannons, G. Mori, S. Oh, I. Kim, Compositional models for video event detection: A multiple kernel learning latent variable approach, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 1185–1192.

[194] W. R. Van Hage, V. Malaisé, G. K. de Vries, G. Schreiber, M. W. van Someren, Abstracting and reasoning over ship trajectories and web data with the simple event model (sem), Multimedia Tools and Applications 57 (1) (2012) 175–197.

[195] F. Van Harmelen, V. Lifschitz, B. Porter, Handbook of knowledge representation, vol. 1, Elsevier, 2008.

[196] V. N. Vapnik, V. Vapnik, Statistical learning theory, vol. 1, Wiley New York, 1998.

[197] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conf. on, IEEE, 2011, pp. 3169–3176.

[198] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. journal of computer vision 103 (1) (2013) 60–79.

[199] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE Int. Conf. on Computer Vision, Sydney, Australia, 2013.
URL `http://hal.inria.fr/hal-00873267`

[200] P. Wang, A. Smeaton, A. Mileo, Semantically enhancing multimedia lifelog events, in: Advances in Multimedia Information Processing–PCM 2014, Springer, 2014, pp. 163–172.

[201] P. Wang, A. F. Smeaton, Aggregating semantic concepts for event representation in lifelogging, in: Proceedings of the Int. Workshop on Semantic Web Information Management, ACM, 2011, p. 8.

[202] X. H. Wang, D. Q. Zhang, T. Gu, H. K. Pung, Ontology based context modeling and reasoning using owl, in: Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conf. on, IEEE, 2004, pp. 18–22.

[203] X.-J. Wang, S. Mamadgi, A. Thekdi, A. Kelliher, H. Sundaram, Eventory–an event based media repository, in: Semantic Computing, 2007. ICSC 2007. Int. Conf. on, IEEE, 2007, pp. 95–104.

[204] H.-K. Wen, W.-C. Chang, C.-H. Chang, Y.-T. Lin, J.-L. Wu, Event detection in broadcasting video for halfpipe sports, in: Proceedings of the ACM Int. Conf. on Multimedia, ACM, 2014, pp. 727–728.

[205] J. Weng, B.-S. Lee, Event detection in twitter., ICWSM 11 (2011) 401–408.

[206] U. Westermann, R. Jain, {rm E}-a generic event model for event-centric multimedia data management in echronicle applications, in: Data Engineering Workshops, 2006. Proceedings. 22nd Int. Conf. on, IEEE, 2006, pp. x106–x106.

[207] U. Westermann, R. Jain, Toward a common event model for multimedia applications, IEEE MultiMedia (1) (2007) 19–29.

[208] F. Wood, C. Archambeau, J. Gasthaus, L. James, Y. W. Teh, A stochastic memoizer for sequence data, in: Proceedings of the 26th Annual Int. Conf. on Machine Learning, ACM, 2009, pp. 1129–1136.

[209] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, P. Natarajan, Zero-shot event detection using multi-modal fusion of weakly supervised concepts, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on, IEEE, 2014, pp. 2665–2672.

[210] Z. Xiong, R. Radhakrishnan, A. Divakaran, T. S. Huang, Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework, in: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE Int. Conf. on, vol. 5, IEEE, 2003, pp. V–632.

[211] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, Q. Tian, Creating audio keywords for event detection in soccer video, in: Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 Int. Conf. on, vol. 2, IEEE, 2003, pp. II–281.

[212] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, A. G. Hauptmann, Event detection using multi-level relevance labels and multiple features, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on,

Image and Vision Computing Journal, Elsevier, 2016, accepted for publication.
The published paper can be found at http://dx.doi.org/10.1016/j.imavis.2016.05.005

IEEE, 2014, pp. 97–104.

[213] Z. Xu, Y. Yang, A. G. Hauptmann, A discriminative cnn video representation for event detection, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 1798–1807.

[214] Z. Xu, Y. Yang, I. Tsang, N. Sebe, A. G. Hauptmann, Feature weighting via optimal thresholding for video analysis, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 3440–3447.

[215] W. Yan, D. F. Kieran, S. Rafatirad, R. Jain, A comprehensive study of visual event computing, Multimedia Tools and Applications 55 (3) (2011) 443–481.

[216] Y. Yang, Z. Ma, Z. Xu, S. Yan, A. G. Hauptmann, How related exemplars help complex event detection in web videos?, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 2104–2111.

[217] Y. Yang, M. Shah, Complex events detection using data-driven concepts, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 722–735.

[218] S. S. Yau, J. Liu, Hierarchical situation modeling and reasoning for pervasive computing, in: Software Technologies for Future Embedded and Ubiquitous Systems, 2006 and the 2006 Second Int. Workshop on Collaborative Computing, Integration, and Assurance. SEUS 2006/WCCIA 2006. The Fourth IEEE Workshop on, IEEE, 2006, pp. 6–pp.

[219] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. Lee, S.-F. Chang, et al., Joint audio-visual bi-modal codewords for video event detection, in: Proceedings of the 2nd ACM Int. Conf. on Multimedia Retrieval, ACM, 2012, p. 39.

[220] G. Ye, Y. Li, H. Xu, D. Liu, S.-F. Chang, EventNet: A large scale structured concept library for complex event detection in video, in: Proceedings of the 23rd Annual ACM Conf. on Multimedia Conf., ACM, 2015, pp. 471–480.

[221] G. Ye, D. Liu, I. Jhuo, S. Chang, Robust late fusion with multi-task low rank minimization, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf. on, IEEE, 2012.

[222] G. Ye, D. Liu, J. Wang, S.-F. Chang, Large-scale video hashing via structure learning, in: Computer Vision (ICCV), 2013 IEEE Int. Conf. on, IEEE, 2013, pp. 2272–2279.

[223] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power, Using social media to enhance emergency situation awareness, IEEE Intelligent Systems (6) (2012) 52–59.

[224] Q. Yu, J. Liu, H. Cheng, A. Divakaran, H. Sawhney, Multimedia event recounting with concept based representation, in: Proceedings of the 20th ACM Int. Conf. on Multimedia, ACM, 2012, pp. 1073–1076.

[225] S.-I. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, et al., Informedia at TRECVID 2014 MED and MER, in: NIST TRECVID Video Retrieval Evaluation Workshop, 2014.

[226] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer vision–ECCV 2014, Springer, 2014, pp. 818–833.

[227] S. Zha, F. Luisier, W. Andrews, N. Srivastava, R. Salakhutdinov, Exploiting image-trained CNN architectures for unconstrained video classification, 26th British Machine Vision Conference (BMVC) (2015) 60.1–60.13.

[228] C. Zigkolis, S. Papadopoulos, G. Filippou, Y. Kompatsiaris, A. Vakali, Collaborative event annotation in tagged photo collections, Multimedia Tools and Applications 70 (1) (2014) 89–118.