Scherp, Ansgar; Mezaris, Vasileios

Article

Survey on modeling and indexing events in multimedia

Multimedia Tools and Applications

Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Survey on Modeling and Indexing Events in Multimedia

Ansgar Scherp · Vasileios Mezaris

Submitted: March 5, 2013

Abstract Events have gained increasing interest in the area of multimedia in recent years. There have been many approaches published and research conducted on how to extract events from multimedia, represent it using appropriate models, and how to use events in end user applications. In this paper, we conduct an extensive analysis of existing event models along commonly identified aspects of events. In addition, we analyze how the different aspects of events relate to each other and how they can be applied together. Subsequently, we look into different approaches for how to index multimedia data. Finally, we elaborate on how to link the multimedia data with events in order to provide the basis for future event-based multimedia applications.

Keywords Event Models · Event Aspects · Event-based Indexing

1 Introduction

Events provide a natural abstraction of happenings in the real world. They are encoded in the multimedia content we all have created and shared or simply have encountered on the Web. For example, images of a concert we recently attended, an interesting location we visited during an overseas trip a long time ago, amateur video footage of a friend's wedding, official footage of an important game of our favorite football team, or eye- witness descriptions and infrared satellite images of a devastating tornado that made it to the headlines. All this content comes in different types of media (images, videos, text, etc.) using different modalities and was created by different devices under different conditions and with radically different usage in mind. What is common though for all these media items is that they all have captured and convey information about a real-life event. Efficiently and effectively understanding these underlying, heterogeneous

Ansgar Scherp (primary contact)
University of Mannheim
Institute of Computer Science and Business Informatics
B6, 26
68131 Mannheim, Germany
Phone: +49 621 181-2647
Fax: +49 621 181-2682
E-mail: ansgar@informatik.uni-mannheim.de

real-life events from the multimedia content and using them in order to better organize, share, retrieve and consume the content in any possible way, represents a significant challenge in today's connected world.

Events have been investigated since a long time in traditional sciences such as philosophy [12] and linguistics [35, 25]. Since some years, events have also drawn significant attention in computer science research. In summary, events can be characterized by six different aspects [86, 85]: time, space, participation, relations between events (in terms of mereology, causality, and correlation), documentation, and interpretation. In the subsequent section, we introduce these six aspects of events and analyze to which extent the existing event models support them. One can easily deduce from this analysis that human experience can only be fully captured and represented when capturing all these aspects satisfactorily. Thus, we argue in Section 3 for a high connectedness of the event aspects and provide a systematic analysis of the possible combinations of these aspects. We discuss existing approaches for indexing multimedia data in Section 4. Since the event-related information that can be reliably extracted from the multimedia content typically comprises somewhat lower-level concepts or activities (e.g. elementary human actions), we discuss the state of the art and challenges in the detection of events from multimedia content and their mapping to the different event aspects in Section 5, before we conclude the paper.

2 Analysis of Existing Event Models

Events have various aspects [85]. Until today, the most comprehensive list of event aspects can be found in the event model E [86, 85, 36], the Event-Model-F [68, 65] that is based on E, and the journalism interrogatives of the Eventory system [83]. The different aspects of events are illustrated in Figure 1.

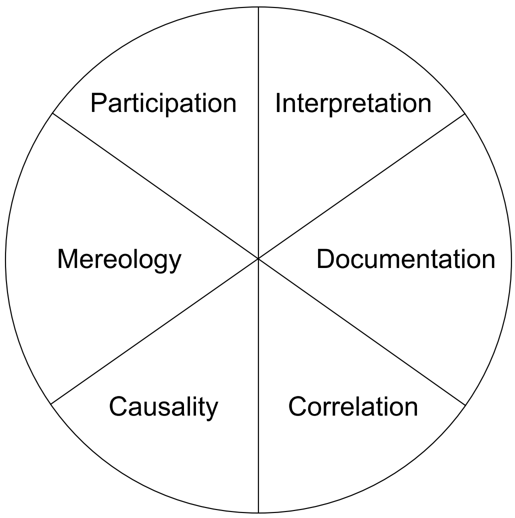

Fig. 1 The Six Aspects of Events (based on [85])

We describe the six aspects of events starting with the participation aspect, which is about the *participation of objects in events* [12]. Objects can be any living as well as non-living things and include people, buildings, and other even intangible objects like the roles a person plays in a specific situation. Events extend over time [14], while objects span in space [12]. The next three aspects are about relationships between events, namely *mereology*, *causality*, and *correlation*. The aspect of mereological relationship refers to the part-of relations of events [61], i. e., the fact that events can be composed of other events. For causal relationships, events are classified into those playing the role of a cause event and those that are an effect event [65, 43, 35]. Finally, correlation relationships are those where two (or more) events have a common cause, but this common cause cannot be explained or is not amenable (cf. [71]). Events can be documented using some media like photos or videos captured during the event. This is reflected by the *documentation aspect* (cf. experiential aspect [86]). Finally, the *interpretation aspect* allows to capture subjectivity that may exist on the other aspects of events [65, 85]. For example, in a soccer match the opponent teams might be of differing opinion about who has committed the foul.

Based on an earlier study [68, 65], we analyze and compare a comprehensive list of different existing event-based systems and event models. We have considered the event calculus [39, 55, 13] for knowledge representation and the situation calculus [42, 41] for representing changes in the real world. In addition, we have analyzed the Event Ontology [62] by Raimond and Abdallahas, which is part of a music ontology framework, the ontology on Linking Open Descriptions of Events (LODE) [70], capturing a minimal model of events, and the Simple Event Model (SEM) [81]. Further, we discuss the Standard Ontology for Ubiquituous and Pervasive Computing (SOUPA) [17], which is the core of the Context Broker Architecture (CoBrA) [18], the Context Ontology (CONON) [84] for modeling context in pervasive computing environments, the Situation Ontology [90] for an hierarchical modeling and sharing of situation knowledge, the Situation Awareness Assistant (SAWA) application [50, 48, 49, 47], and the Situation Theory Ontology (STO) [38]. We have looked into the ISO-standard of the International Committee for Documentation on a Conceptual Reference Model (CIDOC CRM) [20, 72] for cultural heritage, the XML-based OASIS standard of the Common Alerting Protocol (CAP) [58] for describing events in the domain of hazard emergency alerts and public warnings, and the XML-based IPTC standard of an event markup language called EventsML-G2 [33]. XML-based descriptions of events using EventsML-G2 can be embedded into a news item described in NewsML [34]. Another set of models analyzed in this study are the Semantic-syntactic Video Model (SsVM) [21], the Networked multimedia event exploration (NMEE) model [2] by Appan and Sundaram, CASE$^{\mathrm{E}}$ [32] providing an hierarchical event model for the analysis of videos, and the Video Event Representation Language (VERL) [22, 57] for video data. Events represented in VERL can be annotated using a companion mark-up language called the Video Event Markup Language (VEML) [57]. VEML is basically used to encode events described with VERL into video stream data [57]. Gkalelis et al. [26] propose a graph model to represent events, where the nodes are events and the edges are relations between events. The Eventory system [83] is an event-based application in the domain of journalism. Finally, we consider the event model E [64, 86, 85, 36] for event-based multimedia applications, its graph-based successor E* [30] with extended features such as for modeling time and space, and the Event-Model-F [64], which is an extension and formalization of the event model E. An overview of the analysis results and comparison to the six event aspects is depicted in Figure 2.

		Time		Space		Relations			Documen-tation	Interpre-tation
		Rel.	Abs.	Rel.	Abs.	Mereo-logic	Causal	Corre-lation		
Event Calculus	✗	✓	✓	✗	✗	✓	✓	✓	✗	✗
Situation Calculus	✓	✗	✓	✓	✗	~	✓	✗	✗	✗
Event Ontology	✓	✓	✓	✗	✓	~	~	✗	✗	✗
LODE	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
SEM	✓	✓	✓	✗	✓	~	✗	✗	✗	✗
SOUPA	✓	✓	✓	✓	✓	~	~	✗	✓	✗
CONON	✓	✗	✓	~	✓	✗	✗	✗	✓	✗
Situation Ontology	~	✓	✓	✗	✓	~	✗	✗	✗	✗
SAWA	✓	✗	✓	✓	✗	✓	~	✗	✗	✗
STO	✓	✗	✓	✓	✓	~	✗	✗	✗	✗
CIDOC CRM	✓	✓	✓	✓	✓	~	~	✗	✓	✗
CAP	✗	✗	✓	✓	~	✗	✗	✗	✓	✗
EventsML-G2	✓	✓	✓	✓	✓	✓	✗	✗	✓	~
SsVM	✓	✓	✓	✓	✓	✓	~	✗	✓	✗
NMEE	✓	✓	✓	✓	✓	✗	✗	✗	✓	~
CASE^E	✓	✓	✓	✓	✗	~	✓	✗	✗	✗
VERL and VEML	✓	✓	✓	✓	✓	~	~	✗	✓	✗
Gkalelis et al.	✓	✓	✓	✓	✓	~	✓	✗	✓	~
Eventory	✓	✓	✓	✓	✓	~	~	✗	✓	✗
E	✓	✓	✓	✓	✓	~	~	✗	✓	~
E*	✓	✓	✓	✓	✓	~	~	✗	✓	~
Event-Model-F	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Abbreviations: Rel.=relative, Abs.=absolute

Fig. 2 Comparison of Event Models and Event Aspects (extension of prior comparison [68])

Result of our analysis of the existing event-based systems and event models is that almost all of them provide support for the participation of objects in events and modeling the space aspect and time aspect. Many of the existing models for events also provide support for the documentation aspect. Only the Event Calculus and CAP do not provide support for modeling participants in events. Due to the domain considered, the CAP naturally does not support for capturing the participants of the events it describes, i. e., messages about upcoming hazardous events. For those future events, the participants are yet unknown. LODE does allow for modeling time and space by an absolute reference, but a relational reference is not possible. CAP allows for providing absolute information about space, but relative relations in space can only be provided by a textual description of the location. Only the Event Calculus does not provide for the spatial aspect. The SEM supports absolute spatial information using the WGS84[1] vocabulary.

Many of the considered models, however, are limited with respect to their support for representing relations between events such as hierarchical relations (mereology), causal relations, and correlation relations. Still, only a few models provide support for representing different points of view to the same events, namely providing different interpretations of events. Regarding mereological event relations, many event models allow for modeling simple, hierarchical part-of relationships based on decomposing the time period covered by the composite event. For example, the Event Ontology provides the `sub_event` property, SEM supports simple compositions along the time dimension through the `sem:hasSubEvent` relation and using the OWL Time ontology[2] for modeling time intervals, and the Eventory system has an informal notion of mereological relations between events. Again, LODE does not provide support for any relations between events including simple hierarchical sub-events. Hierarchical part-whole relationships are supported in CIDOC CRM [72]. However, no further axiomatizations are provided for refining the mereological relationship by different criteria such as temporal and spatial constraints. CASE$^{\mathrm{E}}$ 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. Gkalelis et al. support mereology by defining sub-events along the temporal dimension. Whether sub-events along other dimensions like space are also supported is not explicitly specified. E* provides for elaborate features regarding expressing spatial relationships between events. SsVM allows for describing complex mereological relationships and the situation-awareness model of SAWA supports spatio-temporal composition [8] like the Event-Model-F does.

Modeling support for causal relationships is mostly limited, if it can be found in the existing event models at all. Examples of models that support causal relations are CIDOC CRM, SsVM, E, and E*. However, it is limited to modeling cause-effect relationships along simple properties like `factor` and `product` in the Event Ontology and the "resulted in" property in CIDOC CRM. Thus, the causal relationships cannot be further annotated with, e. g., a justification for that relationship as is the case for the Event-Model-F. Only the event calculus, situation calculus, and Event-Model-F provide well formed, i. e., formally defined support for causal relationships between events [42]. Support for representing correlation relations between events can only be found with a specific extension of the event calculus [13] and the Event-Model-F. Also, only very few 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 implemented in the form of so-called ontology design patterns [23]. NMEE also has the idea of different viewpoints to the same event. However, it is limited because the definition of what is a viewpoint is underspecified. NMEE can just represent how many viewpoints an event has. Regarding interpretation of events,

[1] `http://www.w3.org/2003/01/geo/`, last visited: December 14, 2012

[2] `http://www.w3.org/TR/owl-time/`

Gkalelis et al. use the properties `isInstantiatedBy` and `hasInstantiationTime`. The different interpretations of events are linked using a `sameAs` relation like in the Web Ontology Language (OWL)[3], which mixes up the different points of views. Thus, it is not possible to distinguish different interpretations from the same agent or person in different contextual situations. Finally, EventsML-G2 provides relationships between events that are unique among the considered models such as `broader` and `narrower` like in the Simple Knowledge Organization System (SKOS)[4] and `sameAs` relations between events.

3 Combining the Aspects of Events

We have discussed six aspects of events in Section 2, namely participation of objects in events, mereologic event relationships (event compositions), causal event relationships, correlative event relationships, documentation of events, and interpretation of events. Unlike other surveys on events [30, 89, 7, 37], we here investigate the combinability of the different aspects of events, i. e., we describe which aspects of events can be combined and how they can be combined. The matrix depicted in Figure 3 gives an overview of the different possible combinations of the event aspects. The x- and y-axis show the different aspects along the order they have been introduced. A tick means that the patterns are combinable. A tilde indicates that the two patterns can be combined under some certain condition. Finally, a cross stands for non-combineable patterns.

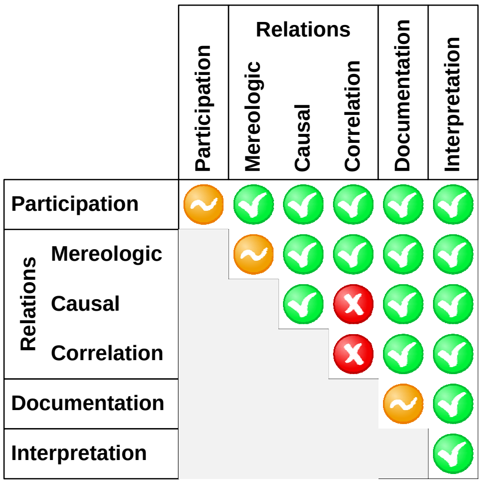

Fig. 3 Combining the Aspects of Events

In the following, we discuss the possible combinations of the event aspects along the matrix depicted in Figure 3. We start with the first row and go through the columns from left to right. We only need to consider the combinations of the diagonal and above

[3] `http://www.w3.org/TR/owl-features/`
[4] `http://www.w3.org/2004/02/skos/`

the diagonal, as it makes in principle no difference if we first consider the participation aspect of a specific event and then the mereologic aspect or vice versa.

The participation aspect expresses which objects participate in an event such as persons and non-living things, where the event happened, and when it happened. It can be considered from different point of views on the same event. This means that the participation aspect can be interpreted differently, depending on one's source of information. In an event of a bar fight, one witness might say that Jim, Bob, and Tom were involved, whereas someone else is saying that Bob and John were fighting. The mereologic aspect decomposes an event into a composite event and multiple component events [61]. Both the composite event as well as the component events might be individually described using the participation aspect. For example, a composite event of a soccer game might be decomposed into two component events of the first and second half. Players that have participated in the first half and second half are then described along the participation aspect. Two events with some participants involved can be part of a causal relationship and thus be either a cause event or effect event [35, 43]. For example, in the event of a heavy storm the participant is the amount of rain that falls. This event is the cause for the dam to burst, which has as participants the objects located around the dam when the burst happened. In a different scenario, a patient is undergoing a medical examination. Here, several correlate events can be identified such as measuring a high blood pressure, finding out about an increasing irritation of the skin, and the patient complaining of some pain. Each of these events happen at the same time. Thus, they are correlate events. However, there is no causal relationship between them; they are all effects of some other (yet unknown) causal event. The patient might have eaten something wrong, has too much stress, or something else. Each of the correlating events can have further information regarding the participation aspect. In the event of observing a high blood pressure, the participating objects might be the specific medicine the patient has to take. Regarding the increasing skin irritation, the participating objects considered are the certain kind of bacteria involved. The persons involved in the dam burst are represented using the participation aspect. The event can be documented by attaching some assets to the event, e.g., some photos taken during the burst. Finally, there might be different interpretations of the same event involving one or multiple other aspects like the participation aspect or the causal aspect. For example, in the event of the bar fight above there might be Jim saying that Bob started and Tom participated. Whereas Tom is saying that Jim has started and Bob has not participated.

At the second row of combining the aspects of events in Figure 3, we look at the combined application of the mereologic aspect to an event. The mereologic aspect can describe an event either on the same level or hierarchically. Describing the mereologic event relationships on the same level means that there are different compositions of the same composite event, e. g., under different interpretations as described above. Providing an event hierarchy means that an event plays the role of a component event while others play the role of composite events. For example, the event of a soccer game can be divided into the first half and second half. The first half can then be divided again into several sub-events such as the foul that led to a red card and so on. Composite events and the component events can be in some causal relationships. For example, the composite event of an important soccer game during the FIFA World Cup is the cause event for the effect events of the streets being empty in the two participating countries and people watching television. The component event of a foul in the soccer game is the cause for the component event of a red card in the game. Thus, the red card is an effect event of the cause event of a foul. In the case of a patient suffering from some disease, different correlate events are observed like measuring higher blood pressure, observing the increasing skin irritation over time, and the patient complaining about some pain. These correlate events can also be understood as some composite events that happen during the event of the patient examination. A single correlation event can also be in a mereologic event relationship. For example, the increase of the patient's skin irritation is observed through several measurements (events where the medical doctor has looked at the skin irritation) that have been conducted over a period of time. Composite events and component events can be documented like any other events with artifacts. For example, the composite events in the soccer game and the patient examination for the increasing skin irritation can be documented using some photographs. In addition, composite events and component events can be interpreted differently. For example, the composite event of the soccer game above might be interpreted differently by the two media coverages in the participating countries. This applies in particular with respect to the component events of the foul and red card.

There can be chains of causal relations. For example, in the soccer game the foul leads to the red card which might further lead to a quarrel among the players and coaches of the two teams. Correlate events are defined as events that are effects of some common causal relationship, although this cause is (initially) unknown [65]. If this causal relationship is discovered, e.g., the reason why the patient is suffering from the blood pressure, skin irritation, and pain, it can be explicitly modeled as cause-effect-relationship. However, when the common cause of the correlate events is not known such a cause-effect-relationship cannot be established (see Section 2). Cause events and effect events can be documented as usual using some artifacts. For example, the cause event of the foul as well the effect event of the red card is documented by the photographers and journalists attending the game. Cause events and effect events can also be interpreted differently. For example, to explain the effect event of the dam burst different theories of the involved objects and mereologic relationships to other events might exist.

As said, correlate events are defined as events that are effects of some common cause [71]. However, there is no causal relationship among the different correlate events. As such, it does not make sense to provide multiple correlation relationships between the same set of events. Correlate events can be documented using some assets. For example, the examination of the patient's skin irritation can be documented using some photographs (see above). There can be different interpretations of correlate events like the skin irritations of the patient. These different interpretations are, e. g., bacteria, chemical reactions, etc. happening in the skin of the patient. An event that is documented can be interpreted in different ways. For example, different images of the same event can be taken and used for arguing that the event is about different things. Finally, there can be several interpretations of the same event with the same information about the participation aspect and other aspects. For example, in the case of a power outage during the event of a flooding, different emergency control officers might have different interpretations about what has caused the event of the power outage.

4 Approaches for Indexing Multimedia Data

The primary application of event modeling and event-based processing, in relation to multimedia, is the event-based organization, indexing and retrieval of the content [26].

In order to map events detected and extracted from multimedia data to the event aspects implemented by the event models discussed above, we give in this section a short overview of the different broad categories of multimedia indexing methods, metadata standards and models.

4.1 Content-independent vs. Content-based Multimedia Indexing

One broad categorization of multimedia indexing approaches can be made on whether they rely on pre-existing content metadata alone, or also use the content itself (e.g., the image, video, audio content) as a source of indexing information. In the former case, we refer to content-independent indexing. Under this category of approaches, administrative metadata such as those provided by the Exchangeable Image File Format (EXIF) standard [76], constitute the simplest way to index the content. These are designed to hold content-independent information usually supplied by the creator of the multimedia data, such as when and how an image was created, digital camera specifications, and other technical information such as location and lighting conditions [77]. Although useful, it is evident that administrative metadata by themselves can offer only limited search and retrieval capabilities and can hardly reveal the event-content associations, as they ignore the actual multimedia content and the perceptual and conceptual information therein.

4.2 Categories of Content-based Multimedia Indexing Approaches

A piece of multimedia content may consist of multiple sources of information, such as text, images, audio, video. Moreover, access to it at different granularity levels may be required, e.g., to search or retrieve the entire video or just a specific video shot or scene. In such cases, the content-independent schemes discussed in the previous paragraph are inadequate as they treat each piece of multimedia data as a whole, ignoring its content and structure. To address these issues, a large number of content-based indexing techniques have been proposed.

- Indexing using perceptual information: A large fraction of content-based indexing approaches use media segmentation algorithms together with objective measurements at the perceptual level, i.e., derive features by processing the low-level visual or audio information within each content segment. These features are then used for indexing the data [15], e.g., MPEG-7 color and texture features [46], local neighborhood features expressed with the use of the scale-invariant feature transform (SIFT) [44] or similar local descriptors, e.g., speeded-up robust features (SURF) [9], color-variants of SIFT [80]), mel-frequency cepstral coefficients (MFCC) for audio, and others. Such approaches, although they are a necessary part of any multimedia indexing scheme, when used in isolation present several limitations. The most important is that they fail to capture the conceptual and contextual information conveyed by the multimedia content.
- Concept-based indexing: Many works have appeared on combining the aforementioned low-level features with machine learning algorithms in order to achieve the association of content with semantic concepts such as “person”, “outdoors”, etc., or different actions [60, 52]. The content segments can then be retrieved with the

use of the detected concepts [74, 16]. These methods represent a significant improvement over the methods of the previous category. However, they still do not fully capture the meaning that the content has to a human observer, who typically "sees" in it more than just a few depicted objects or elementary actions. For this to happen, the automatically detected concepts need to be seen in combination and be used for deriving higher-level interpretations of the content.
- Semantics-based indexing: A large number of multimedia indexing techniques based on technologies related to the vision of the Semantic Web [69] have recently emerged in various application domains [19, 11, 10, 53] to support the higher-level interpretation of the content. These include attempts to develop an MPEG-7 multimedia ontology using RDF or OWL. However, in these attempts there is an one to one translation of MPEG-7 types to RDF/OWL concepts and properties. Thus, MPEG-7 related problems may arise such as the lack of syntactic and semantic interoperability between multimedia applications using these models [79, 3, 56]. In [11], the content of a video is first described using MPEG-7 and then a domain level ontology in OWL is used to annotate the content segments. However, domain ontologies developed in this work lack formal semantics. To fill this gap, the core ontology for multimedia (COMM) [3] builds upon the descriptive ontology for linguistic and cognitive engineering (DOLCE) [24] and is a manual mapping of the MPEG-7 features to an ontology. DOLCE is a foundational ontology, providing a domain independent vocabulary that explicitly includes formal definitions of foundational concepts like objects, events, and others. In addition, DOLCE's descriptions and situations (DnS) ontology design pattern offers formal representation of context. This allows COMM to, e.g., conduct different simultaneous descriptions (viewpoints) of the same multimedia content segments. COMM suffers from some redundancies and inherent limitations. For example, it provides three ontology design patterns to conduct annotations [63, 66]. In addition, annotations of technical details like resolutions of the images could not be directly attached to the image data itself but rather to an abstraction of the data, called the information object [67, 63]. These limitations have been removed by the Multimedia Metadata Ontology (M3O) [63]. It abstracts from the features of today's metadata formats and metadata standards and provides a generic modeling framework for integrating existing metadata models such EXIF and MPEG-7 [67]. Annotations using the COMM or M3O are not centered around events, in contrast to the fact that human memory organizes experiences in events. However, the M3O can be used in combination with event models like the Event-Model-F [68]. Here, the M3O serves as implementation of the event documentation aspect.

Addressing the above drawbacks requires the development of multimedia indexing methods that are explicitly designed to organize multimedia data in terms of events. These indexing methods have to take into account the different event aspects that can be captured and represented by the event models, and further link them to the actual multimedia content. Performing this linking in an automated way forms the basis of *event-based multimedia indexing* approaches, as discussed in the following section.

5 Linking Multimedia Data to Event Model Aspects

Following the modelling of events, an important and difficult problem is how to automatically (or semi-automatically) link the different multimedia content such as images, videos, audio, text files with the different aspects of event representation. Whilst this is a rather wide area, some hints and references that can serve as a starting point for further study are given here. The short review below focuses on the most recent techniques for event detection in audio-visual content items (i.e., video that may include audio). This choice is based on the fact that such items convey a high amount of information (typically higher that still images or audio alone) and are far more challenging that text in their automatic analysis, for the purpose of detecting events.

During the past few years significant research has been devoted to the detection and recognition of events in broad domains (e.g. the detection of LSCOM[5] or TRECVID events in large and rather unconstrained video collections). In [87], a Bag-of-Words (BoW) algorithm is combined with a multilevel sub-clip pyramid method to represent a video clip in the temporal domain; the Earth Mover's Distance (EMD) is then applied for recognizing LSCOM events in the TRECVID 2005 dataset. In [82], motion relativity and visual relatedness are exploited, and relative motion histograms of BoWs are computed to represent video clips. Then, similarly to [87], the EMD and Support Vector Machines (SVMs) are used for recognizing video events. In [6], knowledge embedded into ontologies and concept detectors based on SVMs are employed. In [88], three types of features, namely, spatiotemporal interest points, SIFT features, and a bag of MFCC audio words, are used to train SVM-based classifiers for recognizing the three events of the TRECVID 2010 Multimedia Event Detection Task (TRECVID MED 2010). For the same events, in [45] a wide range of static and dynamic features are extracted such as 13 different visual descriptors on eight granularities, histograms of gradients and flow, MFCC audio features, and others. These features are then used for training 272 SVM-based semantic detectors. Thus, they represent videos with model vector sequences. Subsequently, hierarchical HMMs are applied for event recognition. In [27], the temporal evolution of specific visual concept patterns is examined. A model vector-based approach is adopted again, where visual concept detectors are used for automatically describing a video sequence in a concept space. The study in [51] reveals that the semantic model vector representation outperforms other low-level visual descriptors for video event detection, and is complementary to other such descriptors. In later editions of the TRECVID MED task (2011 and 2012 [59]), most approaches adopted a model vector representation; they usually combined this with a multitude of other features, coming either from the visual modality (e.g. low-level features) or also from non-speech audio, analysis of speech recognition transcripts, and optical character recognition that was applied on the video frames (see, for instance, [40] and [31]). Typically, considering multiple features resulted in higher performance, at the expense of similarly higher computational cost and system complexity.

In all the above and similar model vector approaches, the use of concept detectors trained on pre-existing datasets (e.g., MediaMill [75], TRECVID Semantic Indexing (SIN) Task [73]) for forming the model vectors significantly reduces the required training time, since no additional time is introduced for training event and/or concept detectors specifically for the dataset of interest. At the same time, the model vector approach allows for a more natural linking of media items and events, using concepts (i.e., natural objects, elementary actions, etc.) as an intermediate link. In this way, it facilitates some form of event recounting, i.e., the process of understanding why a video segment was identified as depicting a specific event. In [27], the link between concepts and events is implemented with the use of a novel discriminant analysis technique, the

[5] Large Scale Concept Ontology for Multimedia, http://www.lscom.org/

Mixture Subclass Discriminant Analysis (MSDA) [28]. MSDA (or further extensions of it, as in [29]) is invoked for identifying the semantic concepts that best describe each event, thus defining a discriminant concept subspace. In the latter, the nearest neighbor classifier (NN) along with the median Hausdorff distance are used for recognizing the events. This approach was also evaluated in the TRECVID MED 2010 task. Further to this, after the detection of an event, one can follow the inverse path in order to find out with the help of MSDA which concepts contributed the most to this detection, similarly to [54].

The detection of video events in more constrained application domains such as surveillance or sport has also received significant attention. In [4], a method for capturing surveillance events such as humans standing, objects being abandoned, or humans shouting is presented. This method exploits visual and audio information and is used as part of a complete event-based multi-camera surveillance system. In [78], a method for detecting and classifying sports events (e.g. goal, foul) is proposed, and is applied to soccer and basketball video. This method combines logical rule-based models and statistical information coming from annotated data sets for detecting the sport events. In [5], visual clustering and ontology-based reasoning are used for sport event detection, focusing on soccer video. In [60], the detection of events or, more generally, the detection of human actions and other semantic classes of interest in sports and news video is addressed. To this end, a new representation of the motion information contained in the video is combined with well-studied classification techniques like Hidden Markov Models.

6 Conclusions

In this paper, we have conducted a detailed analysis on an extensive set of existing event-based systems and event models with respect to the different aspects of events. Subsequently, we have looked into how these aspects of events relate to each other and how they can be applied together. Then, based on the fact that it is multimedia data that typically documents an event, we looked into different approaches for indexing multimedia and linking it with the event information. This forms the basis for future event-based multimedia applications.

In our analysis, we have concentrated on the central aspects of events, namely time, space, participation, relations between events (in terms of mereology, causality, and correlation), documentation, and interpretation. A discussion of additional characteristics such as if the model is formally defined, if uncertainty information can be added to events, and if media assets documenting an event can be decomposed into parts can be found for some of the event models in a work by Gkalelis et al. [26]. Further detailed surveys on event detection and recognition in video, covering topics such as visual feature extraction, event classification, and ontologies for knowledge representation and reasoning can be found in [30, 89, 7, 37].

References

1. J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26 (11):832–843, Nov. 1983. ISSN 0001-0782. doi: 10.1145/182.358434. URL `http://doi.acm.org/10.1145/182.358434`.

2. P. Appan and H. Sundaram. Networked multimedia event exploration. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 40–47, New York, NY, USA, 2004. ACM. ISBN 1-58113-893-8. doi: 10.1145/1027527.1027536. URL `http://doi.acm.org/10.1145/1027527.1027536`.
3. R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. COMM: designing a well-founded multimedia ontology for the web. In *The Semantic Web: ISWC 2007 + ASWC 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 30–43, Berlin, 2008. Springer.
4. P. K. Atrey, A. E. Saddik, and M. S. Kankanhalli. Effective multimedia surveillance using a human-centric approach. *Multimedia Tools and Applications*, 51(2):697–721, January 2011.
5. L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. *Multimedia Tools and Applications*, 48(2):313–337, 2010.
6. L. Ballan, M. Bertini, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE Multimedia*, 17(4):80–88, Oct. 2010.
7. L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, January 2011.
8. N. Baumgartner and W. Retschitzegger. A survey of upper ontologies for situation awareness. In *Knowledge Sharing and Collaborative Engineering; St. Thomas, VI, USA*, pages 1–9. ACTA Press, 2006.
9. H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.
10. M. Bertini, A. D. Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, and R. Vezzani. Dynamic pictorially enriched ontologies for digital video libraries. *IEEE MultiMedia*, 16:42–51, Apr/Jun 2009.
11. A. Carbonaro. Ontology-based video retrieval in a semantic-based learning environment. *Journal of e-Learning and Knowledge Society*, 4(3):203–212, Sept. 2008.
12. R. Casati and A. Varzi. Events. Stanford Encyclopedia of Philosophy, 2006. `http://plato.stanford.edu/entries/events`.
13. I. Cervesato, M. Franceschet, and A. Montanari. A guided tour through some extensions of the event calculus. *Computational Intelligence*, 16:200–0, 1999.
14. K. M. Chandy, M. Charpentier, and A. Capponi. Towards a theory of events. In *Proceedings of the 2007 inaugural international conference on Distributed event-based systems*, DEBS '07, pages 180–187, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-665-3. doi: 10.1145/1266894.1266929. URL `http://doi.acm.org/10.1145/1266894.1266929`.
15. S.-F. Chang, J. He, Y.-G. Jiang, E. E. Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *Proc. TRECVID 2008 Workshop*, Gaithersburg, MD, USA, Nov. 2008.
16. G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon. Large-scale content-based audio retrieval from text queries. In *Proc. 1st ACM Int. Conf. on Multimedia Information Retrieval (MIR '08)*, pages 105–112, Vancouver, BC, Canada, Oct. 2008.
17. H. Chen and A. Joshi. *The SOUPA Ontology for Pervasive Computing.* Birkhauser Publishing Ltd., April 2004.

18. H. Chen, T. W. Finin, and A. Joshi. Using OWL in a pervasive computing broker. In *Ontologies in Agent Systems; Melbourne, Australia*, volume 73 of *CEUR Workshop Proceedings*, pages 9–16. CEUR-WS.org, 2003.
19. S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. Papastathis, and M. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 15(10):1210–1224, Oct. 2005.
20. M. Doerr, C.-E. Ore, and S. Stead. The CIDOC conceptual reference model: a new standard for knowledge sharing. In *Conceptual modeling*, pages 51–56. Australian Computer Society, Inc., 2007. ISBN 978-1-920682-64-4.
21. A. Ekin, A. M. Tekalp, and R. Mehrotra. Integrated semantic-syntactic video modeling for search and browsing. *IEEE Transactions on Multimedia*, 6(6):839–851, 2004.
22. A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles. VERL: An ontology framework for representing and annotating video events. *IEEE MultiMedia*, 12(4), 2005.
23. A. Gangemi and V. Presutti. Ontology design patterns. In S. Staab and R. Studer, editors, *Handbook of Ontologies*, International Handbooks on Information Systems. Springer, 2nd edition, 2009.
24. A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. In *Proc. of the 13th Int. Conf. on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW '02)*, pages 166–181, London, UK, Oct. 2002.
25. A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening Ontologies with DOLCE. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181, London, UK, 2002. Springer. ISBN 3-540-44268-5.
26. N. Gkalelis, V. Mezaris, and I. Kompatsiaris. A joint content-event model for event-centric multimedia indexing. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC 2010), September 22-24, 2010, Carnegie Mellon University, Pittsburgh, PA, USA*, pages 79–84. IEEE, 2010.
27. N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011)*, pages 85–90, Madrid, Spain, June 2011.
28. N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Mixture subclass discriminant analysis. *IEEE Signal Processing Letters*, 18(5):319–322, May 2011.
29. N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations. *IEEE Transactions on Neural Networks and Learning Systems*, 24(1):8–21, January 2013.
30. A. Gupta and R. Jain. *Managing Event Information: Modeling, Retrieval, and Applications.* Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.
31. H. Cheng, J. Liu, S. Ali et. al. Sri-sarnoff aurora system at trecvid 2012 multimedia event detection and recounting. In *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.
32. A. Hakeem, Y. Sheikh, and M. Shah. Casee: A hierarchical event representation for the analysis of videos. In D. L. McGuinness and G. Ferguson, editors, *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose,*

California, USA, pages 263–268. AAAI Press / The MIT Press, 2004. ISBN 0-262-51183-5.
33. IPTC International Press Telecommunications Council, London, UK. EventML, 2012. `http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/Specification/`.
34. IPTC International Press Telecommunications Council, London, UK. NewsML, 2012. `http://www.iptc.org/site/News_Exchange_Formats/NewsML-G2/`.
35. E. Itkonen. *Causality in Linguistic Theory.* Indiana Univ. Press, Bloomington, Indiana, USA, 1983.
36. R. Jain. EventWeb: Developing a human-centered computing system. *Computer*, 41(2):42–50, 2008. ISSN 0018-9162. doi: http://dx.doi.org/10.1109/MC.2008.49.
37. Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2012, accepted for publication.
38. M. M. Kokar, C. J. Matheus, and K. Baclawski. Ontology-based situation awareness. *Inf. Fusion*, 10(1):83–98, 2009. ISSN 1566-2535. doi: http://dx.doi.org/10.1016/j.inffus.2007.01.004.
39. R. Kowalski and M. Sergot. A logic-based calculus of events. *New Gen. Comput.*, 4(1):67–95, Jan. 1986. ISSN 0288-3635. doi: 10.1007/BF03037383. URL `http://dx.doi.org/10.1007/BF03037383`.
40. L. Cao, N. Codella, L. Gong et. al. Ibm research and columbia university trecvid-2012 multimedia event detection (med), multimedia event recounting (mer), and semantic indexing (sin) systems. In *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.
41. F. Lin. Embracing causality in specifying the indeterminate effects of actions. In *AAAI/IAAI, Vol. 1*, pages 670–676, 1996.
42. F. Lin. *Handbook of Knowledge Representation*, chapter Situtation Calculus. Elsevier, 2008.
43. L. Lombard. *Events: A metaphysical study.* Routledge & Kegan Paul, 1986.
44. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
45. M. Hill, G. Hua, A. Natsev et. al. IBM research TRECVID 2010 video copy detection and multimedia event detection system. In *Proc. TRECVID 2010 Workshop*, Gaithersburg, MD, USA, Nov. 2010.
46. B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):703–715, June 2001.
47. C. Matheus, M. Kokar, K. Baclawski, J. Letkowski, C. Call, M. Hinman, J. Salerno, and D. Boulware. Sawa: An assistant for higher-level fusion and situation awareness. In *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications; Orlando, USA*, pages 75–85. SPIE, March 2005.
48. C. J. Matheus, M. M. Kokar, and K. Baclawski. A core ontology for situation awareness; Cairns, Australia. In *Information Fusion*, pages 545–552, July 2003.
49. C. J. Matheus, K. Baclawski, M. M. Kokar, and J. Letkowski. Using SWRL and OWL to capture domain knowledge for a situation awareness application applied to a supply logistics scenario. In *Rules and Rule Markup Languages for the Semantic Web*, volume 3791 of *LNCS*, pages 130–144. Springer, 2005.
50. C. J. Matheus, M. M. Kokar, K. Baclawski, and J. Letkowski. An application of semantic web technologies to situation awareness. In *International Semantic Web*

Conference, volume 3729 of *LNCS*, pages 944–958. Springer, 2005.

51. M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic Model Vectors for Complex Video Event Recognition. *IEEE Transactions on Multimedia*, 14(1): 88–101, February 2012.
52. V. Mezaris, A. Dimou, and I. Kompatsiaris. On the use of feature tracks for dynamic concept detection in video. In *Proc. IEEE International Conference on Image Processing (ICIP 2010)*, pages 4697–4700, Hong Kong, China, September 2010.
53. V. Mezaris, S. Gidaros, G. Papadopoulos, W. Kasper, J. Steffen, R. Ordelman, M. Huijbregts, F. de Jong, I. Kompatsiaris, and M. Strintzis. A system for the semantic multi-modal analysis of news audio-visual content. *EURASIP J. on Advances in Signal Processing*, 2010.
54. A. Moumtzidou, N. Gkalelis, P. Sidiropoulos, M. Dimopoulos, S. Nikolopoulos, S. Vrochidis, V. Mezaris, and I. Kompatsiaris. Iti-certh participation to trecvid 2012. In *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.
55. E. T. Mueller. *Handbook of Knowledge Representation*, chapter Event Calculus. Elsevier, 2008.
56. F. Nack, J. Ossenbruggen, and L. Hardman. That obscure object of desire: Multimedia metadata on the Web, Part 2. *IEEE Multimedia*, 12(1):54–63, Jan./Mar. 2005.
57. R. Nevatia, J. Hobbs, and B. Bolles. An ontology for video event representation. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 7 - Volume 07*, CVPRW '04, pages 119–, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2158-4. URL `http://dl.acm.org/citation.cfm?id=1032638.1033010`.
58. OASIS Emergency Management TC. Common alerting protocol version 1.2 (oasis standard), July 2010. `http://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2.doc`.
59. P. Over, J. Fiscus, G. Sanders, B. Shaw, G. Awad, M. Michel, A. Smeaton, W. Kraaij, and G. .Quenot. Trecvid 2012 - goals, tasks, data, evaluation mechanisms and metrics. In *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.
60. G. Papadopoulos, A. Briassouli, V. Mezaris, I. Kompatsiaris, and M. Strintzis. Statistical motion information extraction and representation for semantic video analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, 19(10): 1513–1528, October 2009.
61. A. Quinton. Objects and events. *Mind*, 88(350):197–214, April 1979.
62. Y. Raimond and S. Abdallah. The event ontology, October 2007. `http://motools.sf.net/event`.
63. C. Saathoff and A. Scherp. Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In *World Wide Web Conference; Raleigh, NC, USA*, pages 831–840. ACM, 2010.
64. A. Scherp, S. Agaram, and R. Jain. Event-centric media management. In *SPIE*, volume 6820, 2008.
65. A. Scherp, T. Franz, C. Saathoff, and S. Staab. F–a model of events based on the foundational ontology dolce+dns ultralight. In *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), September 1-4, 2009, Redondo Beach, California, USA*, pages 137–144. ACM, 2009. ISBN 978-1-60558-658-8.

66. A. Scherp, C. Saathoff, T. Franz, and S. Staab. Designing core ontologies. *Applied Ontology*, 6(3):177–221, 2011.
67. A. Scherp, D. Eißing, and C. Saathoff. A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology. *Int. J. Semantic Computing*, 6(1):25–50, 2012.
68. A. Scherp, T. Franz, C. Saathoff, and S. Staab. A core ontology on events for representing occurrences in the real world. *Multimedia Tools Appl.*, 58(2):293–331, 2012.
69. N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE Intell. Syst.*, 21(3):96–101, May 2006.
70. R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *The Semantic Web, Fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. Proceedings*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009. ISBN 978-3-642-10870-9.
71. B. Shipley. *Cause and Correlation in Biology.* Cambridge Univ. Press, 2002.
72. P. Sinclair, M. Addis, F. Choi, M. Doerr, P. Lewis, and K. Martinez. The use of CRM core in multimedia annotation. In *Semantic Web Annotations for Multimedia*, May 2006.
73. A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.
74. C. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.
75. C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. ACM Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.
76. Technical Standardization Committee on AV & IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif Version 2.2. Technical report, April 2002.
77. J. Tesic. Metadata practices for consumer photos. *IEEE Multimedia*, 12(3):86–92, July 2005.
78. D. W. Tjondronegoro and Y.-P. Chen. Knowledge-discounted event detection in sports video. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 40(5):1009–1024, September 2010.
79. R. Troncy, O. Celma, S. Little, R. Garcia, and C. Tsinaraki. MPEG-7 based multimedia ontologies: Interoperability support or interoperability issue ? In *Proc. 1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies*, Genova, Italy, Dec. 2007.
80. K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
81. W. R. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Abstracting and reasoning over ship trajectories and web data with the simple event model (sem). *Multimedia Tools Appl.*, 57(1):175–197, 2012.
82. F. Wang, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and visual relatedness. In *Proc. 16th ACM international conference on Multimedia*, pages 239–248, Vancouver, British Columbia, Canada, Oct. 2008.

83. X. Wang, S. Mamadgi, A. Thekdi, A. Kelliher, and H. Sundaram. Eventory – an event based media repository. In *Semantic Computing*, pages 95–104, Washington, DC, USA, 2007. IEEE. ISBN 0-7695-2997-6.
84. X. H. Wang, D. Q. Zhang, T. Gu, and H. K. Pung. Ontology based context modeling and reasoning using OWL. In *Pervasive Computing and Communications Workshops*, page 18, Washington, DC, USA, 2004. IEEE. ISBN 0-7695-2106-1.
85. U. Westermann and R. Jain. E - a generic event model for event-centric multimedia data management in echronicle applications. In *DataEngineering Workshops*, page 106, Washington, DC, USA, 2006. IEEE. ISBN 0-7695-2571-7. doi: http://dx.doi.org/10.1109/ICDEW.2006.1.
86. U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14(1):19–29, 2007.
87. D. Xu and S.-F. Chang. Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11): 1985–1997, Nov. 2008.
88. Y. Jiang, X. Zeng, G. Ye et. al. Columbia-UCF TRECVID 2010 multimedia event detection: combining multiple modalities, contextual concepts, and temporal matching. In *Proc. TRECVID 2010 Workshop*. Gaithersburg, MD, USA, 2010.
89. W. Yan, D. F. Kieran, S. Rafatirad, and R. Jain. A comprehensive study of visual event computing. *Multimedia Tools Appl.*, 55(3):443–481, 2011.
90. S. S. Yau and J. Liu. Hierarchical situation modeling and reasoning for pervasive computing. In *Software Technologies for Future Embedded and Ubiquitous Systems*, pages 5–10, Washington, DC, USA, 2006. IEEE. ISBN 0-7695-2560-1.