Visual Event Computing

Jonathan Weir; WeiQi Yan

bookboon
The eBook company

WeiQi Yan & Jonathan Weir

Visual Event Computing

Visual Event Computing

ISBN 978-87-7681-783-1

Contents

	Abstract	**7**
1	**Introduction**	**8**
1.1	Introduction to Visual Event Computing	8
2	**Events**	**10**
2.1	Towards Events	10
2.2	Events	11
2.2.1	Event Types	12
2.2.2	Event Aspects	13
2.2.3	Event Relationships	14
2.3	Defining `Events'	14
2.4	Experiential Data	16
2.5	Create a Web	17
2.6	Explicit Referential Links	18
2.7	Spatio-Temporal Links	19
2.8	IE: EventoScope	20
2.9	Causality	21
2.10	Casual Chains	22

2.11	A Scenario - Conferences	23
2.12	Crowdsourcing	26
2.13	Participatory Urban Sensing	27
3	**Visual Event Computing**	**28**
3.1	Content of Event Computing	28
3.1.1	Event Operations	29
3.1.2	Event Storage	29
3.1.3	Modelling Techniques to Facilitate Event Presentation	31
3.1.4	Event Mining and Reasoning	33
3.2	Applied Techniques in Event Computing	35
4	**Visual Event Capturing**	**41**
4.1	Live Data	41
4.2	Photo Event Detection	42
4.3	Video Event Detection	43
4.3.1	Event detection from sports video	52
4.3.2	Event detection from surveillance video	55
4.3.3	Meeting event detection from indoors scenes	58
5	**Multi-Modal Event Computing**	**60**
5.1	Multi-Modal Event Capturing in a Sentient Environment	60
5.1.1	Event	61

5.1.2	DocumentWeb and EventWeb	61
5.1.3	SEraja Structure	63
5.1.4	Event-based Analysis of Multi-Modal Data	64
5.2	Approaches	66
5.2.1	Capture Events Using Requirement-Specific Sensors	66
5.2.2	Store All Events From a Specific Environment in an Event Server using E	68
5.2.3	Develop Aggregation / Crawling Approach to Create a Situation Model for the Environment	68
5.2.4	Develop Event Interaction Environment	69
5.3	Event Detection Systems	70
5.3.1	Event Environments	70
5.3.2	Types of Events to be Detection	70
5.3.3	Sensors are Sources of Event Information	72
5.3.4	Environment Model	72
5.4	Detecting Events	73
5.4.1	Detecting Methods	73
5.4.2	Automatic or Semi-Automatic	73
5.5	Reports	73
5.5.1	Videos, Audio, and Sensory Data	74
5.5.2	Privacy Issues	74
	Bibliography	**75**

Abstract

In this book, the focus is on visual event computing. The events come from video and images. An introduction is presented on what an event is, then an introduction is presented on detecting events from videos and images. Furthermore, event search, retrieval, mining, and reasoning is examined and discussed. As an example, how to use visual event computing in a surveillance environment is proposed. The related software system is also presented.

The authors would like to thank Professor Ramesh Jain for his ideas and contribution towards improving this book on Visual Event Computing. This book would have been very difficult to produce without his help and expertise in the matter.

Chapter 1

Introduction

1.1 Introduction to Visual Event Computing

An event in the general sense is defined as something that happens at a given place and time (wordnet.princeton.edu/perl/webwn). The basic characteristics of an event to be considered in the sense of human centric computing systems are its ID, time, location and description. Events that take place at different times or positions are considered to be different events. People experience events associated with media either as a combination, e.g. video and audio, or individually. With the advent of new types of media, and the increasing diversity of media for capturing events, event-computing has emerged as a new research area. Previously, most work focused on event detection [1], however, recent work has started to address other functionalities such as storage, reasoning, interaction and exploration [2]. Consequently, new disciplines, such as data mining, have had to be embraced. The definition of an event in terms of visual computing is something that still has to be formally defined, throughout this book, a discussion is presented on the definition of an event in the hope of giving the reader a clear sense of what an event is seen as, even though this concept is still under discussion within visual computing.

In this book, visual event computing techniques are introduced along with their computational aspects. In Chapter 2, a description of what an event is and what characteristics are inherent in an event. This analysis describes how an event can be given a type classification. The composition of an event within a visual event computing system, including the different computational aspects and the relationships between events and their exploration. The chapter is concluded by reviewing the current modeling techniques used in event detection. In Chapter 3, events are classified according to various data types and acquisition methods. Chapter 4 focuses on capturing data for event computing. Current applications of the event paradigm

are introduced in Chapter 5. This book is aimed at those readers who expect to have a basic knowledge of events and how events are employed in visual computing. This book is limited to modern computing, however, the extension to general events is straight-forward.

An event to be captured for the purpose of further analysis includes detection, storage, search, retrieval, reasoning, mining, exploration and actions. Therefore, event computing systems must encapsulate each step. In detection, events are detected and analyzed semantically. Event detection is a procedure to match previously well defined patterns that have a high occurrence with new incoming patterns. Event detectors sense changes in the attributes of objects which participate in an event, ultimately causing a change in event status. The detectors analyze the event, it changes in status, and thus, a semantic evaluation can be provided. Following analysis, the detected events are stored in a database with their metadata, thereby allowing users to explore and retrieve particular events. Thus mining and reasoning of events can be facilitated by a database. As soon as events are changed, or presented in a new form, they are stored in a database as new events [3].

In this book, a taxonomy of knowledge is provided. While an effort to be as specific as possible in the classification of work has been made, the reader should be aware that there is some overlap where topics may be complementary.

Chapter 2

Events

2.1 Towards Events

Two things are distinct in emerging applications of information systems: they contain vast amount of multimedia (both live and archived) data and attention is moving away from examining isolated silos of data toward more holistic pictures of evolving situations. Multimedia systems, including text, video, images, and audio, provide both information and experience related to a dynamic situation. As information and communication technology evolved, multimedia has become increasingly ubiquitous; structured data is now a very small fraction of useful data in emerging applications.

Current information tools are reasonably good in dealing with entities, objects, and keywords. To address the needs of information management in dynamic multimedia environments, new concepts and techniques are needed. It is clear that the current concepts and tools are good for the text oriented and structured information systems. These tools are not good for dealing with images, video, audio, and other sensory information. An example of their limitation is the poor results that one sees on every major search engine for images and video. Since these search engines try to apply search tools effective for text to the text associated with images and video, but not processing images and video, their results, contrasting them to text search, are surprisingly bad.

Current information tools evolved before the wave of mobile phones, digital cameras, and broadband systems changed the landscape of information systems. With all these advances, experiences are becoming an integral part of information systems. In fact, signs indicating the dawn of experiential computing can already be seen. There is a very intimate relationship between events and experiences in experiential computing, events will play a central role.

It is believed that 'events' may be used as fundamental organizational concept in multimedia systems that are becoming ubiquitous. There are strong and deep conceptual, engineering, computational, and human centered design reasons to consider events as a primary structure for organizing and accessing dynamic multimedia systems. Multiple applications using event models are currently being developed to validate a hypothesis which states that events are effective in capturing multimedia semantics and building efficient systems to deal with multimedia information.

In the following sections, events are discussed, their role in organizing information and experiences, developing event-based approaches for emerging information and experience management systems, the EventWeb and its ubiquitous and fundamental role, and emerging role of the EventWeb as a parallel, but connected, structure to the current WWW which is predominantly a DocumentWeb.

2.2 Events

What is an event? Definitions of an event from various research fields are very diverse and tend to reflect the content of the assigned media. In text event detection and track, an event is something that happened somewhere at a certain time [4], whereas in pattern recognition, an event is defined as a pattern that can be matched with a certain class of pattern types. Meanwhile, from a signal processing viewpoint, an event is triggered by a status change in the signal. Thus, a uniform definition is required for all media.

An event is symbolic abstraction for the semantic segmentation of happenings in a specific spatio-temporal volume of the real world. Happenings include the presence of a discernable entity or entities present throughout the temporal space of a piece of media (for the case of visual computing); however, it does not mean that all entities have to be included in a specific event. State changes in objects, or their movement in the real world, can trigger meaningful events [5][6][7]. To find the correct relationships between events and their related data, the relationships between sub-events should be realized. This step requires the object attributes involved with the related events and sub-events. For a better understanding, let's look back at theories about the ultimate entities which compose the real world [8]. The real world is a spatio-temporal framework that requires entities to both its dimensions: time and space. All definitions about the constituents of the world refer to their spatial, temporal or spatio-temporal aspects. From this viewpoint, events, objects, their properties, and the existing relationships between them, are the main subjects in terms of the spatio-temporal aspect of the real world.

An event is understood as a fundamental semantic concept in multimedia systems which are becoming ubiquitous. There are strong and deep conceptual, engineering, computational, and human centered design reasons to consider events as a primary structure for organizing and accessing dynamic multimedia systems. There is an intimate relationship between events and experiences in experiential computing. Events play a central role in that they are related to multiple information sources such that changes in multimedia lead to the triggering of events. Event detection requires examining objects, actions, and their inter-relationships automatically [9]. Events are considered as abstractive entities of the actual world's composing elements in terms of their spatial aspect. Another attribute of events is that they are the source of causal relationships; they cause other events or objects to be created, continued, degraded or terminated. They consist of time intervals, with a beginning and end point. For higher level events, there might be some blank time intervals because nothing has happened related to an event. Events should have occurrences, patterns and categories. The term *event* is very wide in its own sense; hence, it has different classifications according to different categories. In the next section, events are defined as either telic or atelic, based on their temporal properties, and as either atomic or composite, based on their composition.

2.2.1 Event Types

- *Telic and Atelic*

  Telic events have patterns and types which make them purposeful, while atelic events do not have such attributes. The time interval in which a telic event occurs has end points, whilst atelic events do not.

- *Atomic and composite*

  An event is thought as a physical reality [10]. Modeling events need knowledge about both atomic and composite events. An atomic event is an elementary one which cannot be divided into any other events and is the simplest type of event inferred from the observables in the visual data. It is the event in which exactly one object having one or more attributes is involved in exactly one activity.

  Composite events are defined by composition of two or more atomic events [5]. A composite event indicates the part-whole hierarchical relationships in events. If composite events contain simultaneously multiple single events, the events are called multithreaded. Gehani et al [11] composed composite events from primitive or atomic ones in 1992. Atomic events are basic ones optionally qualified by a mask, which is used to hide or mask the occurrence of an event. Composite events are specified as event expressions, which are formed using event operators. The event actions in a database include 'insert', 'delete' and 'update'. An event occurrence is a 2-tuple of the form <*atomic event*, *event identifier*>, in which event identifiers are used to define a total ordering.

  An event history is a finite set of event occurrences in which two event occurrences have not the same event identifier. Obviously, composite events are created by considering structural and causal combinations that are meaningful in a given context for specific people.

  A multi-object event is a composite event that involves multiple objects. An atomic event is regarded as a consistent motion state of an object, and are often inferred directly from motion trajectories [12]. A multi-object event is then learned based on a coarse-to-fine strategy. An event class is detected by a bottom-up/top-down search algorithm, where some distinctive local event properties are propagated to infer a likely global event configuration. In an attempt to make the concept of an event useable in the visual computing paradigm, a closer look at its aspects should be considered [13].

2.2.2 Event Aspects

An event, defined in human centric computing systems, has six aspects: *temporal*, *causal*, *spatial*, *experiential*, *informational* and *structural*.

The temporal aspect contains the *Physical time*, indicating the time stamp (event-starting time and event-time-duration); *Logical time*, the temporal domain concept; and *Relative time*, which is the temporal relationship of the specific event to other events.

The causal aspect refers to a chain of events [14] which cause other events to be produced, live longer, be degraded or be terminated. This causal relationship also applies between an event as the cause and the object as the effect.

The spatial aspect indicates the relative location, the spatial relationships to other events, the logical location, the spatial domain concept, physical location, GPS position, geographic and frame region.

The informational aspect of an event contains the information about the media content related to the captured event, which will be used in its context information. It also includes the tags and comments which users add to the related event.

The experiential aspect includes all the media files stored in a separate database with paths related to specific events. These play the role of objects occupying events.

The structural aspect refers to the structure of composite events containing sub-events. These sub-events may contain other composite events or atomic events. The structural aspect of an event can be displayed with a tree structure.

2.2.3 Event Relationships

There are four kinds of relationship between events: *referential*, *structural*, *causal*, and *similar*. Hopkins et al [15] conducted research on determining the causes of events and strove to find an effective algorithm that could determine whether one event causes another. Brute-force and intervention-proving approaches were used to prune the search space and project a casual world onto a reduced set of variables.

2.3 Defining 'Events'

Objects in common language remain vague and are understood more clearly and precisely in a specific context. For computers, objects must be defined more precisely. In computer science one defines objects as having two important components: data associated with the object and methods that operate or access data in predefined ways. The data is usually accessed only through the methods. In programming, one first defines classes of objects where all data associated with objects is defined, along with its type, and methods that will operate on all these data fields are defined. Each object in a program is an instance of a class. Each class may have subclasses and these subclasses may inherit some data types and methods from their parent class.

An event, defined in computing environments, should be a mechanism to define three important aspects of an event clearly and explicitly. These three aspects are:

information about the event, experiences related to the event, and structural and causal relationships with other events.

An event in computational form should represent data associated with the above aspects and processes to acquire and present these as may be needed. By providing flexible and expressive mechanisms to define these three components and associated methods, one could define events effectively. The event environment should provide tools to define any event of interest from many disparate application domains. One may first define event classes and then each event in the system may be an instance of a class.

The basic characteristics of an event will be the ID of the event, its time and location. A similar event may take place at different time and space and will be considered a different event. In this sense, an event is a defined in spatio-temporal space. There are good reasons to consider events as either point events or interval events. Point events are just points in the spatio-temporal space; interval events are regions in spatio-temporal space.

The informational characteristics of an event will be similar to data elements defined for objects. All informational attributes will be data of specific types and methods may be defined to access these attributes and operate on those. The information components may consider fields like participants, objects, and similar data fields commonly defined as attributes of objects. Many of these fields will have similar data types and methods.

2.4 Experiential Data

The experiential attributes of an event are fundamentally different from informational attributes. Each experiential attribute represents a data (stream) that is experienced using a specific natural human sensor. Thus the types of information obtained are visual data, audio, tactile, olfactory, and taste related data. Currently good sensing and reproduction techniques are available for visual (image and video) and audio data. Tactile is improving fast and others are slowly getting developed. Thus, experiential data will be defined as the data of a particular sensory type rather than an integer or real or character type as commonly used in informational data.

Experiential data is usually much larger in volume than other data types. For historical reasons, in computing most representations evolved to represent simple data like numbers and characters. A collection of numbers representing an image is thus represented using an array of data, an intensity value at different pixels forming an image. A video will be an array of such arrays. In databases, when designers faced such data, they usually lumped it all and called it a "binary large object", or a 'blob'. Search engines analyze a text file and identify words in it by analyzing arrangement of characters, but usually don't open an image to analyze it. In general, except few people specializing in particular experiential data analysis, people have avoided dealing with experiential data. Interestingly, slowly experiential data started becoming popular and now photos, video, and audio are becoming the central data elements in computing.

Another distinguishing feature of the experiential data is that it is always grounded in space and time. A sensor captures data at a point in space and in many cases over a time period. Thus the data captures a physical phenomenon at a given point in space and over a particular time interval.

The operators and methods to be applied to experiential data are significantly different than the methods used for processing alpha-numeric data that was commonly used in many traditional computing fields. Of course due to the nature of digital computers, the most basic operations must be reduced to the basic processing operation in computing. For human abstraction and use, however, these operations are fundamentally different. The computational techniques for experiential data are emerging and clearly are not as well developed as techniques for computing and managing alpha-numeric data.

Rapid progress in sensing, storage, processing and display (reproduction) technology is making experiential data rapidly popular. It is rapidly becoming not the secondary source of information, but a primary source of experiences and communication. It may be noticed that computer as well as mobile phone manufacturers usually advertise their devices based on their experiential characteristics? They make it

clear how good the camera or video processing capability of the device is. They know that experiential data appeals to humans much more than the abstract numbers.

2.5 Create a Web

A fundamental insight brought into the creation of WWW by Sir Tim Barners-Lee was that documents could be linked to each other by creating links explicitly among them. Before that each document on Internet was an independent document. By creating tools and environment so documents could be linked and could be created and accessed easily, he created the Web. The tradition of linking documents explicitly has existed for long time through footnotes, references at the end of articles or books, and by explicitly mentioning other documents in the text. The tradition of creating a link between an article and another one started with concept of hyperlinks. And this was taken to a very different utility level in the Web. These links are created to refer to another document explicitly that is considered relevant in that context. These are known as referential links. Ultimately the Web is the Web due to links among documents.

EventWeb will be created by creating such explicit links among different events. The links among events are much stronger in many senses, as discussed in the following, than they are in documents. Links among events are also much more natural than they are in documents. There are implicit relationships among documents of different kinds and techniques for discovering and presenting such links are emerging slowly. The same will happen in the case of events. In fact, as discussed later, insight is the result of discovering such links among the myriad events that surround us in all aspects of our life.

Referential links are similar to the links in documents. An event may refer to other event by creating explicit links from the first to the second. These links are usually one directional, the pointed node does not necessarily point to the pointing node. This is a big feature in the Document Web that has been utilized by many systems, most notably Google in their PageRank algorithm.

An event (node) can also refer to another event (node) using any of the creative ways that the author wants to use. This could range creating a link from the title, description, or reports. The referential node, like referential nodes in DocumentWeb, will be one way node and will be created by the publisher of the pointing (or referring) event. The referenced node may not be aware of this link.

2.6 Explicit Referential Links

Referential nodes are behind most of the current Web. All of these nodes are created by authors of documents, usually at the time of the creation of the document. Of course, the author of the document, commonly called a page, is free to edit the document at a later time and create new links. Increasingly tools are becoming available to create relevant links. Such tools help an author to find appropriate links to be used.

Recently efforts are being made to discover other types of links among documents by utilizing some automatic techniques to find implicit relationships and similarity among documents. Text mining techniques are busy discovering such links by developing techniques for topic detection and clustering. This has become an active area of research in the last few years because with increasing number of pages on the Web it is important to find relevant pages and basic search techniques give too many results. Techniques based only on PageRanking must utilize other filtering to provide results that will be meaningful and useful to users.

However, it is important to realize that such implicit link discovery techniques are in their early infancy. Most of the Web is the result of explicit referential links created by authors of documents.

In case of events also, initially most links may be created explicitly by people who will create events and these may be referential. Event creators will think of related events in the past as well as upcoming related events and may explicitly link those to their event. Also, events have close references to documents and experiential data. These links may also be created. Unlike documents, however, many links to documents and experiential data may be created after the event is not only created but has passed. These links will be related to the reports. Also, it is expected that in the case of events, many explicit referential links may be created by other people who participated in the event in some form. In this aspect, links in events may be different majority links for events may be owned by other people than the creator of the event.

In the early stages of the EventWeb, most links will be explicit manually created links to other events and other data sources. As technology will evolve, implicit links will be discovered and used for events also. The EventWeb will emerge in the next few years without such implicit links, however.

Events are, inherently more closely linked to other events and documents. Also, in case of events, there are other referential links that play an equally important role.

2.7 Spatio-Temporal Links

For events that are considered interval events, there is a dominant structure that is commonly used. These events have strong spatio-temporal taxonomic structure. This structure is present in both organized events as well as natural events.

An event may contain many other events that took place during this event. The world cup soccer tournament has many subevents first there are league games, each league game has many games in which two teams belonging to the same league play games. Of course each game is itself an event. And in each game, there are two halfs. Again each half can be divided into many possessions, penalties, scoring a goal, and whatever else is considered a relevant event. After the league games there are play-off rounds leading to finals. This is the temporal structure. There is spatial structure also. Each game is played in a different location in a different city and even the events in a game take place at the different part of the play-field. Similar structures are used by people to represent different natural event. And the after effects also came at different times. The rescue efforts and the rebuilding efforts due to the Tsunami could also be considered part of the event and will have different subevents that will be spatially and temporally different but structurally part of the major event.

The above spatio-temporal structure of event could be considered the human desire to organize related events in a structure that could be easily and efficiently represented. Since this is usually imposed by humans, and humans found taxonomy to be an efficient mechanism to represent such relationships, the spatio-temporal structure is naturally represent as a taxonomy. As per the above examples, the first one, the world-cup, is designed by humans and has strong taxonomic structure. The second, the Tsunami, is effort by humans based on their observations to bring them under a structure and hence uses an implied spatio-temporal structure.

Organized events will usually have strong taxonomy structure. Natural events will have taxonomy structure imposed by human analysts. In these cases, events and subevents will be defined using some attributes of these events and will be usually fit in this structure. Like all taxonomy structures imposed on natural observations, like the periodic table and animal species, one may find exceptions and situations where an event may clearly belong to two parents leading to "weak taxonomy", rather than a pure tree like taxonomy structure.

Another important difference in these two types of events is that in organized events the structure is given and the events are generated from that structure. In natural events, multiple events will take place. Powerful event mining techniques based on space time and attributes of events will be required to group them into meaningful classes that will fit into the taxonomy structure. In simpler cases, one may use a known structure and fit given events into that structure, while in other cases the structure will evolve from given events based on general characteristics of events.

In both cases this structure plays very important role in our understanding of the relationship among events. The relationships among these events can be represented by spatial and temporal subevent and superevent links. And these links will be important part of the EventWeb.

2.8 IE: EventoScope

When there are lots of events that are all interrelated and are captured using multiple modality, how should one browse/search them and how should they be presented? Currently, the Web is strongly influenced by the sequential and static nature of text a legacy of paper rather than the screen on which it is usually presented. Due to the legacy of paper, it presents most information in sequential manner just see how the results of search are presented. In fact two major limitations of the current Search Engines used to navigate the Web are the Keyword Box and the list of results. One has to articulate the query using a few keywords. These search engines have trained us so well that when they don't work. This is great for search engines, but a terrible way to deal with people who want to use them. The second limitation is the presentation of the list format. The reason one needs ranking algorithms and measures performance using precision is the result of the list that must be presented on the limited screen.

When there are billions, if not gogolian, of events how should they be navigated? How should they be searched? How should the details be obtained? How can the kind of events be requested? There are many of these questions that come to mind. It is clear that the tedium of the keyword box will become unbearable and is totally inadequate. There has to be something different.

One thing comes to mind that if defined and developed correctly, maybe an interesting approach. One should present lots of events that are somehow interesting. One could use one of many different criteria one approach could be to combine the approaches adopted in current News sites with the culture of the Web. Editorial selection of events could be combined with the random presentation along with the search based selection. These events should be presented not as a list but in some other presentation format. This format should have the ability to focus on one or more of those and explore more details. Like a microscope or a telescope used to explore either micro or remote environments, this system should give the ability to start with a wider field of view but then focus to the one that is of specific interest. This environment, like the microscope and telescopes, also should give the ability to steer the device so as to move to the area of interest and then focus.

2.9 Causality

Wikipedia says:

> "Causality is the centerpiece of the universe and so the main subject of human knowledge; for comprehending the nature, meaning, kinds, varieties, and ordering of cause and effect amounts to knowing the beginnings and endings of things, to uncovering the implicit mechanisms of world dynamics, or to having the fundamental scientific knowledge."

Given two events W and X it is possible that W is the cause of X; or that X is the effect or the result of W. This causal relationship among events plays a key role in understanding why something happened or why some event took place. Most of the analysis is concerned with finding this cause and effect relationship. Of course, many times one finds that people get confused between cause and effect, but that is not the issue here.

The important point here is that events could be related using causality or causal links. A causal link will exist between W and X if W is the cause and X is the effect. If there is another causal link between X and Y then there is a causal chain between W and Y. And this chain could be extended further. Most present events are the result of some earlier events and they result in some future events. This is a continuous process. Most tools for dynamic analysis try to capture and utilize causality using powerful mathematical tools. Dynamic systems use powerful approaches to deal with causality. At any given time, the state of a system is defined as the result of the summation of the effects of all inputs (causes) coming from various sources.

Of course in nature and in most other systems, causality links are not explicitly specified. One of the main objectives of analysis of systems is to uncover causal links between different components of the system and the events taking place there. Much of science is concerned with discovering all these causal relationships. Once these relationships are discovered, they are formalized become part of the established knowledge.

When all these events are taking place, if causal relationships are known and can be represented explicitly then one can easily answer questions related to why a particular event took place by reverse tracking the causal events. This assumes deterministic situations in which there is only one way in which an event takes place, as a result of some other events. Deterministic situations have a single causal chain and hence are easy to create and maintain. Even in those cases, when causal links are not known, people can identify and easily create and maintain those. Such links are easy to traverse and explore. In fact, except for the label that these links are causal, in all aspects these are similar to referential links that are the common links in the current WWW.

2.10 Casual Chains

In many cases there may be many causes for an event and these could even be different possible combinations. Thus an event X may result because one of the events $W_1, W_2, \ldots, W_n$. Or it could result because W_1 and W_2 occurred or W_2, and W_3, and W_4 occurred. There are many potential combinations of events that could be the cause of the event X. In some cases it may be known what are the potential causes for X to occur; while in equal or more number of cases, it may not even be known what are the potential cause (or causes) for X.

There is an interesting causal chain. So let us say for X to occur events W_1 and W_2 should occur. Now for W_1 to occur, V_{11} and V_{12} should occur and for W_2 to occur, V_{21}, V_{22}, V_{23} should occur. Now let's extend this one more level. For V_{11} to occur, U_{111}, U_{112}, U_{113}, U_{114} should occur. And for V_{23} to occur, U_{231} and U_{232} should occur. This chain can keep expanding further in the past. There is exponential growth that takes place in the causal chain as time progresses further into the past. The rate of growth clearly depends on how many events could be the cause of events on this chain.

And how does this chain stop? Or does it ever stop? Are there any well established areas where they faced similar situations and they developed an approach for similar situation?

The above are some interesting questions that need explorations. One thing is

clear that most intellectual efforts and even some of our routine analysis involves a reasoning similar to the one presented above. This will be examined in more detail.

What about an area that comes the closest to this? I think much of the systems theory, particularly the control system theory, is based on causality. It is assumed that the response of a system at any time is the sum of all responses to inputs that may have been applied to the system in any form. Thus, the output or response of the system is not necessarily the result of the input applied to it only at that time instant. The response at a time is the result of inputs, over an extended time period in the past. The time period depends on the characteristics of the system.

2.11 A Scenario - Conferences

Examples usually help in illustrating concepts. So here is an example from a familiar area of conferences. Obviously, this could be easily extended to other things like sports, meeting, seminars, weddings, etc. In this and some other posts, however, it will help to focus on this application.

Conferences are major events to build strong communities. Conferences are planned and organized well in advance. People who are part of the community start planning to attend conferences, they contribute to different mail lists, now 'feeds', to receive information. They contribute to the organization of conferences in many ways ranging from organizers to authors to participants. People who want to be in a particular area, say Multimedia, they find out important conference events in that area and plan to attend.

Then conference takes place. There are many events at the conference. Papers are presented in many sessions. Demonstrations and posters are presented in others. With some conferences there are exhibits. Reception, business meetings, banquets, and social programs are as much a part of the conference as the main conference. Most conferences usually publish a 'proceedings' of the conference.

Now suppose that each paper is created, each demo, each social function, and any other event at the conference as an event. Text-based information such as papers, presentations (say powerpoints) photos and videos of what happens in that event is collected. All these things are posted as an Event (the term 'Event' is used to denote computer representation of all the material related to an event). After the event, people can post photos and videos that they may have taken related to the event. People can also comment on the event and have discussions going.

A paper presented at a conference may have relationship with other paper at that conference. It may also have relationship with other papers presented at other conferences and published in journals. It may be related to some product launches or other events taking place at other place. Multiple papers may have relationships among them based on authors, topics, organizations, or any other aspect. A paper may result in spawning new fields, killing an existing field, development of a product, tenure for a researcher, or even friendship between two researchers.

Conferences are related to each other in many ways also. ACM as a professional organization is involved in making sure that all aspects of computing progress. It encourages organization of many conferences and workshops. These events are related to each other in many different ways. There are workshops and tutorials that take place in conjunction to a conference. As research fields evolve, common topics are addressed. The relationships among such events (conferences or workshops) can be judged by sub-events (papers and their topics), and also by participants. If the papers at two conferences are from similar topic areas and the people involved are similar then one knows that the two events are similar.

One judges the prominence of people in a field by analyzing their role in these events. Similarly the strength of a field is judged by how these events are changing over time and space. For example, there are many conferences starting in multimedia related area in Asian countries reflecting the growth of this field in Asia more than Europe.

As may be clear, there are many different relationships among these events and can be captured by creating Events on them.

Events have a life cycle that continues far beyond the time when the event takes place. One could consider three important stages in this lifecycle. These could be easily considered in three clear stages: Pre-event, Live, and Post-event.

Of course some events just happen or their Pre phase is so fast that it is not considered. This case will be discussed later.

Most events are planned. First the time, duration, and location of the event are fixed. This then blocks a region in spatio-temporal space for the event to take place. In most events there are people who are invited and people who play role as organizers. In many cases, invitations could be open, meaning anybody who is interested may attend the event participate in it in some way. There is some information that must be shared with different groups. All this information is attached to this event. In some cases, to attract people to this event, examples of similar events are presented, or information about the participants in the planned event is used for marketing. All this material also becomes part of the information related to the event.

It is common to form one or more mailing lists related to the event. These mailing lists are used to disseminate appropriate information. It is now common to send all this information easily using e-mail. Of course there are problems related to spam, but this is beyond the scope of this discussion. These mails sometime are used to discuss logistics related to the organization and attending of events.

Obviously, people attend events because they are interested in it. Surprisingly, however, in many cases people attend events because they want to either meet other people, want to be seen at the event, or just don't want to feel that they missed something. Thus, events are a social occasion for a special group of people. Events are organized many times to facilitate meeting of people with common interest. Events are used to build communities. No wonder that events are advertised to a particular group of people by using mail lists, publication of details in newspapers, magazines or even radio and TV. In the last year many sites on the Web have started providing facilities to help people find events that they may be interested in. These sites provide calendar of events. Some sites have started providing environment so people can publish their events and share with others. This is clearly very important and is increasing in popularity. There are sites that help people to organize events by providing environment to invite people to events and helping in managing attendee list, RSVPs, and thank-you notes.

What is interesting is that in many cases the anticipation of the event is built by talking about similar events in the past and giving details of those events. Obviously, past events are always used to predict the future events.

An important part of the build-up leading to an event is discussions among people related to the event. These discussions take place at many different levels and in many different forms. It is important to provide an environment to encourage these discussions for creating more anticipation. Results of a poll about who is going to win a match published with much demographics of votes provide interesting example of how much people want to talk about events they are interested. People do want to meet people who are interested in similar events.

2.12 Crowdsourcing

Events are unique in the sense of attracting people. People are attracted to attend events and know about past events. More importantly, people are interested in reports on events. And people really want to get reports on events from many different people in the hope that some of them will be similar to them in socio-economic-political orientations. No wonder that when an important event takes place (such as a local team winning a championship) TV people interview many people asking silly questions like "how do you feel about our team winning championship?".

In this context, it is clear that media is slowly going towards crowdsourcing. Wikipedia defines crowdsourcing as:

> 'Crowdsourcing' is a term coined by Wired magazine writer Jeff Howe and editor Mark Robinson in June 2006. Like outsourcing, crowdsourcing is a model that depends on work being done outside the traditional company walls, but while outsourcing is typically performed by lower paid professionals, crowdsourcing relies on a combination of volunteers and low-paid amateurs who use their spare time to create content, solve problems, or even do corporate Research and Development.

Crowds targeted for crowdsourcing include garage scientists, amateur videographers, freelancers, photo enthusiasts, data companies, writers, smart mobs and the electronic herd.

Lately people have talked a lot about crowdsourcing news commonly called citizen journalism. A citizen journalist is a common person who reports through blog posts, photographs, and videos, is an event. It is becoming increasingly common that when an unexpected event takes place, since there are no professional journalists to prepare a report (text, audio, or visual), most early reports are 'published' by people who just happened to be there. These people have their camera, also called mobile phone, and can capture the event using photographs, audio, and video. Since there are many people with their camera, it is possible that many people capture the event and are willing to share their 'report' with others.

Let's now consider a slightly different scenario. Suppose that an event takes place. There is a place where citizen journalists can send their reports in any form, photos, video, audio, or text and these reports somehow are attached to that event. If people knew that this is possible, then they will start sending reports on all important events and not-so-important-events like an accident in a neighbourhood or a wonderful ballet performance at an elementary school to this location. This results in a place for people to publish their reports. But more importantly, a place will exist to find reports on all events and from all sources. Thus not limited to a perspective only

of a major TV station or a newspaper, but of anybody, including those sources, making available a kaleidoscope of reports on the event of interest. Of course, there should be methods to organize these reports in some forms so they can be searched for particular types of perspectives only.

This scenario is not very different from what is happening now a days in the blogosphere or even on social networking oriented sites. This is just a way to bring together power of new media to facilitate what people want to do in order to find out about events and how different people react to events.

2.13 Participatory Urban Sensing

Two billion people carry mobile phones. These ubiquitous devices can act as sensor nodes: they are increasingly capable of capturing, classifying and transmitting image, acoustic, location and other data, interactively or autonomously. Though there is much interest and research in distributed sensing for the sciences, industry and defense, much less is known about its function and utility in the public sphere, when the components are owned and operated by everyday users.

These sensors will form infrastructure needed for creating an EventWeb for some interesting as yet unimagined applications. It will be truly interesting to see how this direction of research evolves.

Chapter 3

Visual Event Computing

3.1 Content of Event Computing

In this chapter, event modelling, detection, storage, exploration, mining, reasoning and operations, are examined and reviewed. Event modelling is essentially a discovery problem and involves the use of pattern recognition techniques, such as cluster analysis if prior information is unknown [16].

As explained earlier, events can be related to multiple information sources with changes in those multimedia leading to the triggering of events. Using heterogeneous data types in an information system leads one to think about an event-centric model [17]. Many media-centric applications have been developed which have used events based on their specific requirements, however, no generic event model has been implemented to capture events from different applications. A generic event-model [13] would offer the opportunities for reusable components and techniques for event visualization, exploration and event query languages. There are some new experiential applications such as chronicles, life logs, and event-centric media managers, but so far, none of these applications shares a common event model. They use specific event models based on their application requirements.

A common approach to representing events helps to reduce the different specialized event models, developed each time for a new event-based application, to one reusable model which can be used in different applications, irrespective of the media. Such a model can incorporate other event-based applications because it is generic. A common event management structure can provide reusable implementation platforms for lots of applications. It should also be extensible and adaptable in order to promote the applicability of the event model. Furthermore, it should be capable of integrating events from heterogeneous applications. Extensibility and adaptability

are general concepts, so, for a better understanding, the following explanations are provided.

- *Extensibility*: To perform retrieval easily, event types should be assigned to events in order to reduce the exploration time. However, there are various event types that might not be considered in the first place while designing an event-based system, or some event types may be created in future. Therefore, the best idea would be to design a system that lets users add event types, properties, and associations.
- *Adaptability*: As events have different description aspects, it is desirable that the system be adaptable enough to let users choose the representation for those descriptions compatible with the applications needs.

The prototype consists of three components: (1) event monitoring, (2) tagging, (3) search and retrieval, in order to test the fundamental event-base information storage framework and the relationship network.

3.1.1 Event Operations

Event operations usually refer to two kinds of operations: unary and binary. The former includes *projection*, *selection*, and *renaming*, whilst the latter consists of *union*, *concatenation*, *conditional sequence*, *iteration* and *aggregates*. A prototype system, called Cayuga [18], has been implemented using the event operations of algebra theory. It adds built-in support for parameterization, aggregates and selection over infinite domains, and support for arbitrary streams of events and events with non-trivial duration.

3.1.2 Event Storage

Databases are employed to save events. Sunrise [19] is an industrial-strength database system for real-time event processing and aggregation for telecommunication applications that has been developed in Bell Labs since 1998. It is a main-memory database platform that supports scalability and parallel processing with the service authoring environment.

The instantly indexed multimedia database system [20], as the name suggests, performs real-time indexing of real world events as they take place, called Lucentvision, it has a rich set of indices derived from disparate sources and allows domain-specific retrieval and visualization of multimedia data. Lucentvision exemplifies an emerging

paradigm of instantly indexed multimedia databases that convert's real world events in real-time into a form that enables a new multimedia experience for remote users: 1) immersion in a virtual environment where a viewer can choose to view any part of the event from any desired viewpoint along with any desired speed; 2) the ability to visualize statistics and implicit information hidden in media data; 3) the ability to search, retrieve, compare and analyze content including video sequences, virtual replays and a variety of new visualizations; 4) the ability to access this information in real-time over diverse networks.

Based on an event conceptual model, Pack et al [21] identify a set of design requirements that guide the development of a storage and processing system architecture. The system allows for multiple methods of event detection (manual detection, web crawling, video and audio processing) that can be used to create an event summary, thereby facilitating an easy search method for heterogeneous media. The work identified the criteria of event extensibility, event persistence, search and update efficiency, and consistency. The advantages of this system include flexibility and explicitness.

Supported by a Video Event Representation Language (VERL), events can be represented by MPEG-7 semantic description schemes (DSs) as described in [22]. MPEG-7 DSs are designed primarily to describe higher-level audio-visual (AV) features such as regions, segments, objects, events and other immutable metadata related to creation and production, usage, etc. The DSs produce more complex descriptions by integrating together multiple descriptors and DSs, and by declaring relationships among the description components. In MPEG-7, the DSs are categorized as pertaining to the multimedia, audio, or visual domain. Typically, the multimedia DSs describe content consisting of a combination of audio, visual, and possibly textual data, whereas, the audio or visual DSs refer specifically to features unique to the audio or visual domain, respectively.

3.1.3 Modelling Techniques to Facilitate Event Presentation

Franois et al [23] introduce the VERL and the Video Event Mark-up Language (VEML). VERL is used to represent video events and works with VEML as a companion annotation framework [23]. VEML [24] is a language for recording the observation of concept instances defined in VERL's video event and object ontology. VEML consists of a set of structures, compatible with the VERL definition, that allows links to physical evidence. VERL is designed to encode six items (ontology, data streams, context, objects, events and others) for events that have been automatically extracted, or interactively annotated, in a set of streaming data. The functions of VERL include: *process*, *primitive*, *single-thread*, *multiple-thread*, *subtype*, *rule*, and *sequence*. The language provides: *repeat-until*, *while-do*, *conditions*, etc. There are six possible basic relationships that can exist between two events: *before*, *meets*, *overlaps*, *begins*, *contains* and *ends* [5]. Distinguishing features of VEML are the underlying set of high-level data structure encoding and the relationships between the event ontology, scene-centric and stream-centric representations.

An ontology of events requires a means of describing the structure and function of events. The structure indicates how an event is composed of lower-level states and sub-events, the function introduces the roles an event plays in its environment and how it in turn participates in larger-scale events. Nevatia et al [5] represent video events using VERL to annotate instances of the events described in VERL. This paper provides a summary of VERL and VEML, as well as the considerations associated with the specific design choices. They also advocate [5] use of hierarchical decomposition and single or multiple threads to naturally represent complex spatio-temporal events, common in the physical world, by a composition of simple events. The events are abstracted into three hierarchies: *primitive events*, *single-thread composite events*, and *multiple-thread composite events*. This leads to a language, the Event Recognition Language (ERL), which allows users to conveniently define events of interest without interacting with the low-level processing in the program. The data types in this language include object, location, interval, and numerical value. Hongeng et al [25] point out that a single-thread action is represented by a stochastic finite automation of event states, which are recognized from the characteristics of the trajectory and shape of moving blobs associated with an actor using Bayesian methods. Scenario events are modeled from shape and trajectory features using a hierarchical activity representation, where events are organized into several layers of abstraction, providing flexibility and modularity in the modeling scheme. Multi-agent events are recognized by propagating the constraints and the likelihood of event threads in the event graph. Events in the scenario library are modeled using a hierarchical event representation, in which a hierarchy of entities is defined to bridge the gap between a high-level event description and the pixel level information. Several layers of more abstract mobile object properties and scenarios are constructed

explicitly by users to describe a more complex and abstract activity shown at the highest layers. The links between a mobile object property at a higher layer, to a set of properties at the lower layers, represents the relationship between them. Scenarios are defined from a set of properties or sub-scenarios, and the structure of a scenario is hierarchical. Event representation of the scenario level maps closely correspond to how humans would describe events - little expertise is expected from users.

Ontological semantics aims at building resources which would be maximally applicable for reproducing the results of human language processing ability. Malaia [26] proposed using an ontological description of the semantics of lexical entries to describe a real-world event. The typical event taxonomy deals with four types: *state*, *process*, *accomplishment* and *achievement*. Accomplishments and achievements are more complex and share several important traits: telic complete (incomplete) or complete.

In event interactions [27], highly intuitive graphical operations are used to perform event-level manipulations such as merging, altering and creating new events. Event interactions allow reasoning about the semantically hierarchical nature of events. The creation of capabilities is required for performing drill-down and roll-up operations on event hierarchies and visualizing their spatial and temporal characteristics. A collection of specialized interfaces allows users to visualize and interact with various semantically relevant event characteristics.

CASE^E [24] bridges the gap between low- and high-level events. Based on the CASE^E representation of natural language, an event is regarded as a collection of actions performed by one or more agents, and an event detection involves matching a sub-tree pattern. The detected events are represented hierarchically in terms of sub-events, case-list and temporal logic based on interval algebra.

A conceptual representation for the complex spatial arrangement of image features in large multimedia datasets is introduced in [28]. Spatial Event Cubes (SEC) are a scalable approach to mining spatial events in large image datasets based on spatial occurrence of perceptually classified image features. It not only visualizes the dominant spatial arrangements of feature classes but also discovers non-obvious configurations.

Stemming from natural language processing, a representation of activities as bags of event n-grams is introduced in [29], where the global structural information of activities using their local event statistics are analyzed. Based on these discovered sub-classes, a definition of anomalous activities is given and a way to detect the activities is provided. Making use of this representation, the work shows how activity subclasses can be discovered by exploiting the notion of maximal cliques in an edge-weighted graph. Finally, an incremental information-theoretic method of a new

activity-instance detection and classification, without re-analyzing the entire activity data-set, is proposed.

Events are represented [30] by using event descriptors, each of which has a physical quantity expression that reflects an interpretation of an event and composition rules. The event representation is a middle language between sensor reading values and natural language phrase descriptions. Observable events are represented using physical quantities of an object state such as position, velocity and temperature. This approach brings an ontological structure to a set of event descriptors.

In [31], events in video sequences are presented using reversible context-free grammars. By using the classification entropy as a heuristic cost function, the grammars are iteratively learnt using a search method. Context-free grammars, with their flexible representation, provide more expressive power with a straightforward design. A search-based iterative algorithm for learning the grammar structure and parameters for each class of motion is employed in a semi-supervised learning strategy.

In syntactic pattern recognition [32], the given data is represented by a string of discrete (terminal) symbols from an alphabet (a finite set of terminal symbols). For event recognition, the terminal symbols correspond to what are called primitive events extracted from the video. Abnormal events are detected when the input does not follow the grammar syntax, or the attributes do not satisfy the constraints in the attribute grammar to some degree.

Discriminative actions can be used to describe the fundamental units in distinguishing between events. Actions are captured first, these actions are modelled, and their usefulness in discriminating between events is estimated as a score. The score highlights the important parts (or actions) of the event from the recognition aspect [33].

3.1.4 Event Mining and Reasoning

Data mining [34][35] is the principle of sorting through large amounts of data and picking out relevant information. It has been described as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and “the science of extracting useful information from large data sets or databases” [36]. Event mining delivers a whole variety of information by searching for patterns in data.

Event mining was firstly addressed within the context of a database. Tesic et al [28] mine spatial events to discover interesting spatial patterns in an extended image database. Their SEC data structure supports the extension of the general association rule approach to multimedia databases so as to identify frequently occurring item sets.

A propositional language, called HOT [37], is proposed using two sets of symbols for temporal reasoning, whose inference engine is based on qualitative and quantitative temporal constraints, and is defined over holds, occurs and temporal propositions. The language has a tractable core which allows others to make weak inferences. An alternative propositional semantics to HOT, that decouples propositional constraints from temporal constraints, is the Conditional Temporal Network. Such a network is consistent i.f.f. there exists a minimal model of the propositional constraints so that the set of temporal constraint propagation techniques is applicable to reasoning about events.

Events are reasoned about using k-AMA [38], a sublanguage of event logic, and develop a specific-to-general algorithm for learning event definitions in k-AMA. The events are recognized from video input using temporal, relational and force-dynamic representations. An event-recognition component determines which events from a library of event definitions occurred in the model, and recognizes events in the video. Lower and upper bounds algorithms of the subsumption and generalization problems for two expressively powerful subsets of this logic are proposed, and a positive-examples-only specific-to-general learning method based on the resulting algorithms is used [39].

3.2 Applied Techniques in Event Computing

Currently, events are detected by Dynamic Bayesian Network (DBN) [40] [41][42][43], TemporalBoosting [43], Bootstrapping [44], Continuous State Machine (CSM) [45] and Finite State Machines (FSM) [46], Kalman Filtering, Radial Reach Filter (RRF) [47], Maximum Entropy [48], etc. These algorithms are mostly taken from statistical pattern classification and recognition, machine learning and artificial intelligence, etc.

- *Event detection using Dynamic Bayesian Network*

  A Bayesian network is a graphical model for representing conditional independencies between a set of random variables [49]. A Bayesian approach starts with a priori knowledge about the model structure and model parameters. The initial knowledge is updated using the data to obtain a posterior probability distribution over both models and parameters that usually peaks around the likelihood maxima. The Expectation-Maximization (EM) algorithm is used to estimate the likelihood maxima [50] with hidden variables. A DBN is used to represent sequences of variables. These are often a time-series or a sequence of symbols. The hidden Markov model (HMM) can be considered as the simplest type of DBN. DBNs generalize two well-known signal modelling tools: Kalman filters for continuous state linear dynamic systems, and HMMs for classification of discrete state sequences [50].

  HMMs [51] have been widely used in visual event detection [52][53][25] [24][40][54] [55][56][57][58] [59]. The main reason for this is that events can be regarded as continuous and having temporal coherence, which can be well modelled by HMMs.

  A HMM model is usually formulated as a triple $< A, B, \pi >$, where $A = (a_{ij})_{n\times n}$ is the state transition matrix, $a_{ij} = \sum \mathbf{1}_{(i\to j)}/F$ is the probability of the j-th state given the i-th state, $\mathbf{1}_{(i\to j)}(\cdot) = 1$ solely inference i from j. $B = (b_i)_{n\times 1}$ is the matrix of overall observation symbol probability, $b_i = \sum \delta_{si}/F$ is distribution of the i-th state, $\delta_{si}(\cdot)$ is the Kronecker function. π is the primal state sequence, it is generated from local states in the detecting procedure.

  A HMM is a statistical model in which the system being modelled is assumed to be a Markov process with unknown parameters, the challenge being to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis. In a HMM, the state is not directly visible, but variables are influenced by the state. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives information about the sequence of states. In event detection, a HMM is regarded as a generative model

within the maximum-likelihood framework. Each event class is described by several state models. These states are used to represent different sub-events of different event types. The HMM parameters are trained by the Baum-Welch algorithm, whereas the Viterbi algorithm is used for classification [50].

A semi-supervised approach [60] for event recognition is provided in situations where there is not enough labelled training data, and the high dimensionality of the observation space requires a large amount of labelled data to capture the event characteristics. The framework is general and can be easily applied to many cases in which collecting labelled data is difficult, but collecting a large amount of unlabeled data is easy. The corresponding sequence of events is obtained by applying the Viterbi decoding algorithm.

However, whilst HMMs provide a good method of modelling temporal sequences, they suffer from overfitting when faced with a large number of parameters, long and complex temporal sequences, and relatively small amounts of training data [58]. HMMs also have difficulty modelling long term temporal relationships in data. This is due to the state transition distribution which obeys the Markov assumption where the current state only depends on the previous state. To overcome this drawback, layered HMMs were proposed in [59][61]. In [59], the first layer, namely the feature HMM, is used to produce a posterior probability for each of the midlevel clusters at each time t in the sequence. This layer is built by using unsupervised clustering and segmentation of the training data. The second layer is trained using the output of the first layer. This is supervised training using higher level events. So the higher level event HMM produces a probability of a higher level event at time t. In [61], individual action features are extracted as vectors, which are subsequently fed forward to a continuous Gaussian mixture HMM that provides a segmentation and recognition of events via the Viterbi algorithm.

A Bayesian network is one of the main tools in video event detection. Hongeng et al [25] recognize events by computing the probabilities of simple events at each frame based on Bayesian analysis. Simple events are represented by a Bayesian network and are inferred from the properties of moving objects. Complex events are defined as a temporal sequence of sub-events, and are represented by a finite-state automation. The recognition process consists of decoding the pattern of sub-events and segmenting them into their corresponding state. Multiple-thread events are recognized by combining the probabilities of complex event threads whose temporal segmentations satisfy the logical and time constraints.

Multiple agents have been applied to detect and represent events embedded in a long video. Hakeem et al [24] first classify the multiple agent based event detection algorithms into three categories: 1) manual encoding or con-

straints (grammar, rules); 2) modelling single person activities, or requiring prior knowledge and data variation (HMM, Bayesian network, etc); 3) graph partitioning of the weight matrix. Most recently, detection of unusual and surprising events has become a promising research direction. Hereafter, unusual and/or surprising events refer to events that seldom occur. In [62], events are detected in an event graph. For training events, they adopt a directed acyclic graph approach for representing temporal relationships, and use an event correlation graph to represent temporal conditional dependencies between sub-events. Based on a video event graph, an event correlation graph, and the weight matrix, events are detected using normalized cuts, which is an unbiased method of partitioning a graph into two segments. Using the learned event models, event detection in video is preceded by estimating the weight matrix for each event model. Normalized cuts are then applied to obtain event clusters. These event representations capture temporal variations of only those sub-events, via applying normal cuts to the training video and sub-event alignment of the segmented events.

In [40], events are observed as a sequence of binarized distance relationships amongst objects. The goal of event detection is to find the occurrence of interesting events in a large corpus of video footages [53]. Events are modelled using semantic spatial primitives that enable generalization well beyond the training data. Semantic primitives are defined as Bayesian networks learned from the training data. A HMM is used to model spatial-temporal relationships of objects participating in an event of interest, with observables consisting of a sequence of semantic primitives derived from the binarized distance relationships. Semantic observables outperform direct continuous observables in terms of generalizing to unseen data with little training data. This enables the detection of rare events in video footage [40].

Another approach to event detection is to exploit different kinds of information by segmenting videos into two groups and to progressively select interesting video segments using a maximum likelihood criteria. Piriou et al [63] use probabilistic image motion (camera and image motion) models to detect events in tennis videos. The motion models associated to pre-identified classes of meaningful events are learned from a training set of video samples. Three kinds of events are taken into consideration: *rally*, *serve*, and *change of side*.

For events embedded in news videos, Boykin et al [54] compare an FSM segmentation system with an automatically induced one using HMMs. The Viterbi algorithm is employed to predict the segmentation of a video shot from assigned states such as *start*, *stop*, and *advertisement*.

- *Event detection using Finite / Continuous State Machine*

A FSM [64] contains a finite number of states, and produces outputs on state transitions after receiving inputs. There are two types of FSM: Mealy machines and Moore machines. The machines have the attributes of equivalence, isomorphism and minimization. A FSM is represented by a state transition diagram, a directed graph whose vertices correspond to the states of the machine and whose edges correspond to the state transitions. Each edge is labelled with the input and output associated with the transition. Events are regarded as a state change. The FSM approach has been proven to be robust in modelling temporal transition patterns and has the advantage of not requiring a training process [65]. FSMs [3] can easily represent temporal or logical relationships between simple events and have the power to store extra information about each state, such as how long an object has been in that particular state. Another useful property is the ease with which state transitions can be based on multiple events.

A CSM [45] has a state transition rule in a continuous state space and classifies time-varying patterns from different types of single sources. CSMs support dynamic time warping and robustness against noise. They integrate sequential optimization, such as a Kalman filter, with class discrimination methods based on recurrent neural networks. The state transition rule is derived as the minimization of a time-varying energy function that integrates external inputs and predicted states. Time-varying patterns are embedded in a state space as their corresponding trajectories in the learning phase. Since a basic HMM represents just one class, each class is modeled as an independent state space and a state transition rule. On the other hand, CSMs embed the classes in a single state space as trajectory attractors. When a time-varying pattern is observed, the CSM changes its state along one of the learned trajectories. As a result, the time-varying pattern is classified by the trajectory to which the state is attracted. Kawashima et al [45] detect and recognize events from multiple sources using CSMs that act as simplified Kalman filters. The interaction enables the system to dynamically focus over the multiple sources, and improves reliability and accuracy of event classification in dynamically changing situations. To design a system that recognizes events robustly in an unconstrained environment, the system integrates information from distributed multiple sensors. The dynamics of individual CSMs and the interaction among them are controlled by a certainty distribution during the recognition phase.

- *Event detection using Filtering Algorithms*

  Kalman filtering [66] has been regarded as the optimal solution too many tracking and data prediction tasks. The filter is constructed as a mean squared error minimizer, and related to maximum likelihood statistics. It is optimal in the mean-squared error sense, but it is limited from a practical viewpoint by

the quality and accuracy of the embedded model. Kalman filtering [67] uses a moving object's center position and size as tokens $t(k)$at time k. In order to estimate the token change $\triangle t(k)$, Kalman filtering is applied to the system state $x(k)$, which is defined as a 4-D vector of the positional change per unit time interval of the target object and its change in size. Once the system and measurement models are defined, a recursive Kalman filtering operation can be applied to obtain optimal linear minimum variance estimates of the motion parameters.

The RRF [47] has been used to determine, on a pixel-by-pixel basis, similar and dissimilar areas between a background image and a current image of a scene. It evaluates the local texture at pixel-level resolution while reducing the effects of lighting variations. Satoh et al[47] use this approach to detect events from real world image sequences using grayscale information, which is much more stably acquired compared with color information. The RRF is firstly calculated with reached points, these are then used to calculate evaluation points, and finally the similarity of the RRF value is obtained.

- *Event detection using Support Vector Machine*

  A Support Vector Machine (SVM) [68][69] defines basic functions that are composed of a subset of the training data, which is selected during training. The advantage of an SVM is that although the training involves nonlinear optimization, the objective function is convex and the solution of the optimization problem is relatively straightforward. SVM training always finds a global minimum, and their simple geometric interpretation provides fertile ground for further investigation. An SVM is characterized by the choice of its kernel. In [70], a generic framework for event detection in field-sports broadcast video using an SVM is outlined. Given a shot, the corresponding feature data is aggregated into a Shot Feature Vector. An event model is inferred from evidence derived in turn from feature detectors, which are chosen such that they are recyclable across multiple sports within the field-sport domain. The five feature detectors are *crowd image detection*, *speech-band audio activity*,*on-screen graphics tracking*, *motion activity measure* and *field line orientation*.

- *Event detection Using Conditional Random Fields*

  The Sequential Monte Carlo (SMC) method provides a finite dimensional approximation to a posterior probability given past observations [71]. Conditional Random Fields (CRFs) are undirected probabilistic models designed for segmenting and labelling sequence data. Compared with the traditional approaches of SVM and HMMs, CRF based event detection offers several unique advantages. To detect an event, a three-level framework, based on multi-modality fusion and mid-level keywords, is adopted. The first level extracts audiovisual features, the mid-level detects semantic keywords, and the high-level infers semantic events from multiple keyword sequences.

Chapter 4

Visual Event Capturing

Currently, visual events can be considered from within the context of photographs, video streams, and archived video. Photo events are well studied in recent years since photographs are easily acquired and are ubiquitous. Furthermore, detected events help us in photograph organization, search and retrieval. In live streams, surveillance and broadcasting videos are the subject of much research. In particular, research into event detection in sports video constitutes a large portion of the research effort.

4.1 Live Data

A video event can be looked as a kind of semantic unit for expressing a story. The ultimate purpose of most tracking systems is the extraction of symbolic descriptions of scene activity. In [72], each event consists of a light-weight data structure that contains the identifier and parameters of the objects involved, as well as further information such as frame number and time stamp. Object trackers are combined into a distributed infrastructure for visual surveillance applications. The tracker generates application independent events on the basis of generic incidents and target interactions detected in the video stream. These events can then be received and interpreted.

An approach for event detection in live sports, based on the analysis and alignment of web-casting text and broadcast sports video [73]. The contributions are: (1) detecting live events using only partial content captured from the web and TV; (2) extracting detailed event semantics and detecting exact event boundaries; (3) creating a personalized summary related to a certain event, player, or team, according to a user's preference. Three modules are provided: live video capturing, live video analysis, and live video alignment.

A patent, filed by Engle and Odutola, concerned a method and system for identifying commercial segments of a video signal [74]. This method mainly involved monitoring a digital bit stream comprising the video signal, detecting a change in a control field of the digital bit stream, and then selectively generating a commercial event notification in response to the detection step. In addition, the method may also involve detecting a change in an informational parameter of the video signal, exclusive of the audio-visual content.

4.2 Photo Event Detection

Photographs associated with an event often exhibit little coherence in terms of both low-level image features and visual similarity [75]. The purpose of photo event detection is to integrate metadata and content-based information for automatic photographical organization. The basic framework is to quantitatively assess structure in the collection at multiple scales and feed this data into several different classifiers. Three event detection algorithms are provided: scale-space analysis of the raw timestamp data, time-based similarity analysis, and time and content-based analysis. For event clustering, confidence scores, combined from the average similarity between photographs within a cluster (average intra-cluster similarity) and the similarity between adjacent clusters (average inter-cluster similarity), are calculated.

Digital photograph collections are automatically organized into event-based clusters in [76]. An automatic unsupervised algorithm for partitioning a collection of digital photographs based on either temporal or content-based similarity. A learning vector quantization codebook is employed to discriminate between "event boundary" and "event interior" classes. The codebook vectors for each class are used for Nearest-Neighbor classification of the novelty features for each photograph in the test set. Dynamic programming and the Bayesian information criterion are introduced to select clustering boundaries.

From a user study and survey, Lim et al. [77] divide events, based on the visual content of photographs, into gatherings, family activities and a place visit. It is possible to model visual events in order to automatically extract relevant semantic tokens using visual event graphs and a visual vocabulary. The event models are generated from labelled photographs, and semantics extracted from photographs annotated with events. Loui et al. [78] provide event segmentation based on date / time metadata information, as well as the color content of pictures. Two photographic events are found to be popular: chronicle and subject order. An event clustering algorithm organizes pictures into events and sub-events based on date and time of picture capture, and content similarity between pictures. Loui et al [79] refined previous work and provided event clustering and screening of low-quality images to deal with

problematic photos.

An event can be thought as a series of consecutive photographs that were taken in the same context. Naaman et al [80] leverage time and location context to resolve identity in their PhotoCompas system. As many users annotate their identities, patterns and events, the system uses these patterns to generate label suggestions for identities that were not annotated. Naaman et al [81] describe the contextual metadata automatically assembled for a photograph, as well as a browser interface that utilizes the metadata. PhotoCompas adopts time and location information to automatically group photos into hierarchies of location and time-based events. The categories subsume *outdoor/indoor*, *who*, *day/night*, *camera*, *mood*, *captions*, *camera settings*.

An event is considered to be a semantically meaningful human activity, taking place within a selected environment, and containing a number of necessary objects [82]. Events are classified into static images by integrating scene and object categorizations. This is achieved by integrative and holistic recognition through a generative graphical model. Similar to object and scene recognition, event classification is both an intriguing scientific question as well as a highly useful engineering application. Event classification is part of the ongoing effort of providing effective tools to retrieve and search semantically meaningful visual data. Event classification is also particularly useful for the automatic annotation of images.

A hierarchical clustering of photographs, based on a similarity matrix of color histograms, and summarization of photographs, based on a novel contrast context histogram technique, is employed in [83]. In [69], Principal Component Analysis (PCA) is employed to reduce the dimensionality of a histogram based on color descriptors. An SVM is then used to classify the images into various high level categories corresponding to histogram subspaces.

Photographic events are characterized by the coherence of multimodality including time, content and camera settings [84]. An event is taken as a latent semantic concept, and discovered by fitting a generative model using Expectation-Maximization (EM). This approach is general and unsupervised, without any training procedure or predefined threshold. The multimodal metadata used in [84] includes contextual information about the time and camera parameters, and perceptual image content such as color, texture and face number. The parameters for photo clustering are iteratively estimated using an EM algorithm.

4.3 Video Event Detection

Video encapsulates spatio-temporal features, as well as content and contextual information. Object trajectories and motion are the salient features for video event

detection.

- *Event detection using trajectory and action cylinder*

  A trajectory is a set of time-indexed locations combined into a single hypothesized entity. It is at a higher level of abstraction than time-ordered locations used by other correspondence algorithms. The use of trajectories for event detection has a relatively long history [1][85][86]. In the latter, a statistical model of object trajectories is learned from image sequences. Both simple and complex events are recognized by attaching meaning to prototypes representing instantaneous movements and complete trajectories. Trajectory prediction can be achieved in a similar way by labeling nodes whose prototypes represent complete trajectories - with information acquired automatically in a further learning phase. Partial trajectories can then activate the node representing the most similar complete trajectory.

  More recently [87], dynamic time warping was employed to match trajectories using a view invariant similarity measure. It works in an unsupervised manner for motion capturing, action representation and learning. In earlier work [88], the spatio-temporal curvature of a trajectory was represented by a sequence of dynamic instants and intervals. Video is automatically segmented into individual actions, and a view invariant representation for each action is calculated. The proposed model also learns parameters from different actions, and discovers different instances of the same actions performed by different people.

Syeda-Mahmood et al [89] recognize action events based on a trajectory cylinder obtained from multiple views. The shape formed from successive perspective projections of an object is visualized as a generalized cylinder, called action cylinder. This is a spatio-temporal solid formed by combining successive cross-sections obtained by the intersection of the 3D body with a plane. The action cylinder can be viewed as the perspective projection of the object as it undergoes motion.

An effective representation of sports tactics, called aggregate trajectory, is constructed of multiple trajectories obtained using a novel analysis of the spatio-temporal interaction among players and the ball in [90]. The interactive relationship between the playing region information and hypothesis testing for the spatio-temporal distribution of trajectories is exploited to analyze tactical patterns in a hierarchical coarse-to-fine framework.

A trajectory that records an object's position from entering to exiting a scene is one of the most useful information types for embedding a moving object's behaviour [91]. A generic rule induction framework, based on trajectory series analysis, is proposed to learn event rules. The trajectories acquired by a tracking system are mapped into a set of primitive events that represent basic motion patterns of moving objects. A grammar induction algorithm, based on the Minimum Description Length (MDL) principle, is adopted to infer meaningful rules from the primitive event series. Grammar induction from artificial intelligence and natural language processing aims at identifying a set of grammar rules from a set of training sentences. PCA and Euclidean distance are adopted to compute the similarity between two trajectory segments. A spectral clustering algorithm is then used to partition the segments into several motion patterns. This allows trajectory classes, corresponding to different driving lanes, to be separated correctly. A HMM that takes into account the uncertainty of the low level processing is trained on each cluster. These HMMs are used as the detectors of primitive events. For a given trajectory segment, the HMM yields the maximum likelihood.

A standard approach to detect events is to break the model into parts, allowing the parts to move independently, and to measure the joint appearance and geometric matching score of the parts. Allowing parts to move makes the template more robust to the spatial and temporal variability inherent in actions. Examples of this approach are given in [92][93]. Specifically, integral video and box features of a sequence are extracted. A detected volume over all locations in space and time is then scanned at different spatial and temporal scales and windows. The detector is trained and tested on real videos with the actions *sit-down*, *stand-up*, *close-laptop* and *grab-cup* actions having different camera view, scale variations, and changing speeds at which actions are performed.

The efficient matching of event models is via over-segmented spatio-temporal volumes. The models are derived from a single example and are manually constructed. Automatic generation of event models from weakly-labeled observations is a related interesting problem [93]. The model is derived from a single exemplar of the event, however, it can detect events in crowded videos. The key point is in the use of shape matching with over-segmented regions.

- *Event detection using temporal and spatial features*

  A visual event is frequently related to a moving object with constraints on its size, colour or shape. A sequence of events is represented by the tracked trajectory of the object of interest. These trajectories are further clustered to form typical trajectory templates. Therefore, the entire modelling process relies critically on the accuracy and consistency of segmentation and tracking, which are often ill-conditioned due to the presence of multiple objects, occlusion and non-linearity of the trajectories. Visual event detection and classification may be performed without explicit object-centred segmentation and tracking [94]. Events are represented and detected first at the pixel level and then at a blob level (grouped pixels) autonomously. A pixel change history is proposed to characterise pixel-wise temporal visual information in order to detect pixel-level events, and is based on the temporal history of each pixel intensity. Crucially, it is combined with an adaptive mixture background model to form a new representation for detecting and classifying pixel-level events. It also provides an important cue for characterising blob-level events which are defined on the basis of grouped pixel-level events. Blob-level events are computed by unsupervised clustering with automatic model order selection. The EM algorithm is employed to cluster events with MDL (Minimum Description Length) used for automatic model order selection. Although no explicit object-centred segmentation and tracking were performed, meaningful event clusters can be consistently formed. The detected blob-level events can be classified into meaningful classes without object-centred tracking.

  Recognizing a motion event requires choosing the most likely spatio-temporal model. Black [95] employs optical flow with parameterized spatio-temporal models for representing motion events. Within a Bayesian framework, the phase, rate, spatial position, and scale are taken into account to deal with image variations. The computational mechanism, based on the condensation algorithm, incrementally estimates a distribution over model parameters. The approach automatically detects and recognizes motion events based on image derivatives.

  Xu et al [52] detect video events by representing video motion as the responses of frames to a set of motion filters. Motions between two video frames are represented by an energy redistribution function. The redistribution function is

filtered by a set of motion filters, each of which is designed to be most responsive to a type of dominant motion. Such a filter process converts a video into a temporal sequence of filter responses in which distinct temporal patterns, corresponding to high level concepts, are presented. The motion content of a video is used to extract meaningful dynamic events by using probabilistic models. Only low-level motion features are exploited to maximise generality and increase efficiency. The motion activity models, namely the residual motion with a causal Gibbs or Gaussian mixture model, are determined by analysing the distribution of local motion-related measurements. These are derived from a weighted mean of normal flow magnitude.

In [96], events are detected from moving blobs of MPEG video in the compressed domain. Feature vectors from a video clip form a high dimensional curve, simplification of which allows one to browse the video clip at places where events have occurred. The camera motion is computed from motion vectors, and the residual vectors are regarded as moving blobs. Interesting events, such as a new object appearing, objects interacting, or an object changing shape, are detected from these moving blobs. The event detection module builds feature vectors from 2D histograms of stepwise motion vectors and finds discontinuities in the trajectories of the feature vectors. Therefore, dynamic event detection requires four main techniques: identification of camera motion, segmentation of moving blobs, tracking of moving blobs, and analysis of their respective motions for classification of their interaction. The limitations using motion vectors to detect events are: 1) the object speed cannot be fast; 2) a motion vector cannot represent the motion properly when the blobs are too small. For event identification, a vocabulary of dynamic events, based on relative motions and the size changes of tracked moving blobs, is required.

Dynamic events can be considered as being comprised of spatio-temporal atomic units, called actions. Events can be represented as a mixture of actions and the transitions among these actions. In [97], a mixture model learns an optimal combination of various components representing actions. This approach can also be interpreted as a unifying framework for combining appearance and temporal features in events. The composition of the feature content is controlled by the number of mixtures in the model.

Leonard is the first system that goes all the way from video to event classification using recovered force dynamics [98]. Event classification is efficiently performed on this preferred subset of models using prioritized cardinality and temporal circumscription. In earlier work, the maximum-likelihood approach for visual event classification was used [50]. Siskind et al proposed a technique to classify events by recovering changing support, contact, and attachment relationships between participant objects, using a kinematics simulator driven

from the output of 3D tracking. Kinematics describes object motion without considering the masses and forces that bring about the motion. To represent the idea, several movies of simple spatial-motion events were taken, including: *picking-up*, *putting-down*, *pushing* and *pulling boxes*, and *dropping erasers*. An edge detector and line finder were then applied to each of the movie images and an animated output obtained. The event recognition task is partitioned into two independent sub-tasks: a lower-level task which detects object orientation, shape, and size, and an upper-level task, which uses the 2D pose stream produced by the former, to classify an instance of a given event type.

A double threshold multidimensional segmentation algorithm is proposed to automatically decompose a complex human motion into a sequence of simple linear dynamic models, without prior knowledge of the number of dynamic models [99]. Event classification was performed using cluster analysis with the model parameters as input. The dynamic model parameters form a compact representation of the motion data, which is amenable to cluster analysis for event classification.

A time interval multimedia event (TIME) framework is presented as a robust approach for semantic event classification in multimodal video documents [48]. The presentation used in TIME extends the Allen temporal interval relationships. For automatic classification of semantic events, three different machine learning techniques are employed: the C4.5 decision tree, maximum entropy and SVM. The framework explicitly handles context and synchronization and yields a robust approach for multimodal integration. Events are presented in patterns. To model such a framework, the relationships between any two time intervals are considered. There are thirteen relationships: *precedes*, *meets*, *overlaps*, *starts*, *duration*, *furnishes*, *equals and inverse*, etc. The events in soccer videos are *goal*, *yellow card*, *red card*, *substitution*.

Another approach to event detection in video, is the use of an event-inference module. Using this approach, Haering et al [100] propose a three-level video-event detection methodology and apply it to animal-hunt detection in wildlife documentaries. The first level involves color, texture, motion features, shot boundaries and moving objects. The second level uses a neural network to determine the object class of the moving blobs. The third level detects video segments that match user-defined event models. Osadchy et al [101] use an anti-face method to detect events in both the gray scale and feature domain. The algorithm was applied to detect *activity curves* corresponding to sketched symbols in two and three dimensions. Using two basis views, it was possible to successfully detect sketches in views that substantially differ from the training set. Three advantageous features of the technique include: 1) the method is robust to rotation, scale, and speed of the event; 2) the proposed method is capable of discriminating a given word from very similar words; 3) the method is used for motion feature recognition.

- *Event detection using AV features*

  For detecting events in a meeting, the usual approach is to extract a set of standard audio-visual features from three cameras. In a meeting event scenario [60], visual features consist of head vertical centroid position and eccentricity, hand horizontal centroid position, eccentricity, and angle. The motion magnitude for head and hand blobs were also extracted. The average intensity of different images computed by background subtraction are extracted. For audio features from a microphone array, a speech activity measure was computed. Three acoustic features, namely energy, pitch and speaking rate, were then estimated on speech segments.

  Detecting semantic events from audio-visual data with spatio-temporal support is a challenging multimedia understanding problem. A duration dependent input and output Markov model to detect events based on multiple modalities was proposed by Naphade et al [102]. It provides a hierarchical mechanism to map media features to output decision sequences through intermediate state sequences. It also supports discrete non-exponential duration models for events. Combining these two features, the Viterbi algorithm is used to infer events. Kristjansson et al [103] present an extension of the forward-backward algorithm that can be used for inference and learning in event-coupled HMMs. They present results on a simplified multimedia indexing task, where the objective is to detect an event whose onset is loosely coupled in audio and video.

  Specific types of events occurring in a classroom or lecture environment, can be detected using a query-driven approach which combines visual and audio cues derived from an image and the textual content of presentation slides [9].

Visual events are detected using the displayed slide, or are captured from a video stream by region hashing. A region of a video frame can be recognized as containing a specific slide if the affine intervals of corresponding region pairs are identical. Affine coordinates of features in a region are computed first. The range in which these coordinates lie is noted in the corresponding affine interval, and the affine interval information consolidated and represented in an index structure called the interval hash tree.

Both internal AV features [104][105][106], and various types of external information sources can be utilized for event detection in team sports videos. In the case of a soccer video, the event types are *goal*, *save*, *shot-off target*, *penalty-goal*, *corner-kick*, and *free-kick*. Three fusion schemes are proposed: rule-based scheme, aggregation, and Bayesian inferences. The use of multiple sources of information based on intrinsic AV features and external knowledge helps to detect events in the soccer video. Comparisons show that Bayesian inference has the best capabilities to tackle asynchronism among the three schemes.

- *Event detection using stochastic processes*

  Events are regarded as stochastic temporal processes [107] [108], with two events being considered as similar if they could have been generated by the same stochastic process. A simple statistical distance measure between video sequences captures the similarities in their behavioral content. This measure is nonparametric and can thus handle a wide range of complex dynamic actions. A behavior-based distance measure between sequences can be used for a variety of tasks, including: video indexing, temporal segmentation, and action based video clustering. By presenting events in a nonparametric way, periodic and nonperiodic activities, isolated occurrences, and multiple repetitions can be recognized utilizing a single framework for both structured video and dynamic textures.

- *Event detection using other features*

  Amera et al recognise context-independent events using key-image extraction [109]. Context-independent events refer to events having a fixed meaning. Amera et al rigorously define a set of context independent events including *enter*, *appear*, *exit*, *disappear*, *move*, *stop*, *occlude*, *remove*, *depositor*, etc [110]. These are automatically detected using feature extraction following segmentation, motion estimation and object tracking. Events are automatically detected by combining trajectory information and spatial features, such as size and location. When specific conditions are met, events related to these conditions are detected.

MediaTE (Media to Everyone) is able to create videos of higher narrative or aesthetic quality with a complete mobile lifecycle [44]. It proposes an at-capture bootstrapping of event information from which all system guidance flows. Bootstrapping focuses on extracting entities from the user input. The event can have global attributes, both physical and discourse related, as well as similar attributes inherited implicitly from actors and objects that it contains. The goal of characterizing and populating an event is to enable the creation of shot suggestions specifically moulded to a user's context, and to obtain a sufficient amount of information about an event from the user at capture time in a natural manner. The bootstrap consists of three aspects: a setting attribute, relevant human actors, and relevant objects.

Robust event recognition is achieved by recovering the viewpoint transformation and time correspondence between a query action and a given action segment in the video [111]. This can be used to efficiently deal with viewpoint changes, execution style changes and occlusions.

Fern et al recognize events from video input using temporal, relational and force-dynamic representations [38]. The raw input is the video-frame sequence, segmentation and tracking components then transform this input into polygon movies in which objects are marked with polygons. A model-reconstruction component then transforms the polygon movies to a force dynamical model. Finally, an event-recognition module determines which events, from a library of event definitions, occurred in the model. The detected events are: *pick-up*, *put-down*, *stack*, *unstack*, *move*, *assemble* and *disassemble*.

4.3.1 Event detection from sports video

- *Event detection from football video*

  Babaguchi et al. [112][113] propose a combination of methods for event detection in sports video. Firstly, multimodal information is processed by tracking the dependency between media streams based on the concurrency of their related events. This process, called inter-modal collaboration, establishes links between visual and linguistic streams. Secondly, domain knowledge is exploited to extract specific visual objects. An event is detected by object occurrence analysis in the visual streams. Typical events include: *touch down*, *field goal*, *point after touch down*, *safety*, etc. In [113], four extra events are detected: *players*, *motion*, *referees gesture*, *change of score*, and *keywords from auditory*.

  In further work on inter-modal collaboration [114], the temporal correspondence between the visual and closed caption (CC) streams is exploited to improve the reliability and efficiency of video content analysis. The proposed method attempts to seek time spans in which events are likely to take place, through keyword extraction from the CC stream. These are then used to index shots in the visual stream. Detected events include: *touch down* and *field goal*. Miyauchi et al [115] also adopt inter-modal collaboration to detect semantic events in three stages: closed caption analysis, auditory analysis, and visual analysis. Key words are related to events from the CC stream and feature parameters characterising cheering and shouting from the auditory stream. Multimodal streams consist of visual, auditory and textual information.

  Three level events are detected from rugby video using layered HMMs [59]. The first of these is structural events of a shot, e.g., medium shot, medium shot low, angle close up, person in a close up, long shot, miscellaneous. The second are play events, e.g., play, non-play and replay. Lastly, are the action events, e.g., running and passing, maul, line-out, kick, penalty, scrum and try.

- *Event detection from soccer video*

  Soccer video events are detected using a three-layer event detection scheme [116]. A probabilistic framework, based on Bayesian inference, is used to reason whether interesting events are presented. A short video segment composed of consecutive frames that contain a special cue comprises an intermediate-level semantic descriptor which is at a semantic level above low-level features and shots. Six semantic units are considered: SMR, close-up, audience, caption, goalmouth and close-up & audience/caption unit. When evidences are observed, they are inserted into the network and the posterior probabilities of events are calculated using model parameters, priors and conditional probabil-

ities. *Shooting* and *red or yellow card* events in soccer are detected based on a Bayesian network. Furthermore, Tang et al [117] presents a content-adaptive transmission system for streaming reconstructed soccer goal events over networks. The reconstructed event consists of one panoramic image or a sequence of panoramic images. The system constructs a field model by detecting landmarks. Each transmission scheme defines the video content to be transmitted, how many images are to be reconstructed, and where the goal reconstruction event will take place. For each frame of the goal event sequence, the positions of the ball and players are detected, and segmentation is performed on the rectangular region around them. The positions of the extracted segments are localized on the field model and the segments are pasted accordingly.

Tactic patterns can be discovered from goal events in broadcast soccer videos based on the tactic clues extracted from players and ball trajectories. In [90], a multiobject detection and tracking algorithm is employed to obtain player and ball trajectories during a goal event. Goal events are extracted with far-view shots based on the analysis and alignment of web-casting text and broadcast video. An aggregate trajectory is constructed, based on multiple trajectories, using an analysis of the spatio-temporal interaction between players and ball. The interactive relationship, information from the playing region, and hypothesis testing for trajectory spatio-temporal distribution are exploited to analyze the tactic patterns in a hierarchical coarse-to-fine framework.

A semantic video indexing algorithm based on a FSM and low-level motion indices, extracted from the MPEG-2 compressed bit-stream, is presented in [46]. The proposed algorithm can detect sports events, such as scoring of a goal in a soccer game and other relevant events using fast pan and fast zoom-in.

A wide range of player actions and game events that are derived from a hierarchical entity-relationship model representing the prior knowledge of soccer events, is presented [118]. In this approach, information on the players and ball from multiple monitoring points is combined to derive their positions via triangulation. A list of observed events, interpreted events, and soccer actions are provided. Knowledge using in detecting the various events includes the laws of the game, understanding of how a soccer match progresses, and all the possible events which may happen during the course of a game. In [42], complex events are detected based on the detection of basic actions. Once an event is detected and marked as valid, the algorithm will invoke another set of heuristic rules to determine which specific actions are involved in the event.

One approach to sports video annotation is to integrate text and image streaming [113]. From the text and image data, actors, actions and events of each scene are extracted using linguistic cues and domain knowledge. The linguistics are segmented and used to extract parts where an event has a high

probability to have taken place. This is done by utilizing key phrases to get the elements of each story. An action is independent of the type of sport involved and is of a general nature. An event is a result of those actions and is specific to the particular type of sport being played.

- *Event detection from baseball video*

  A feature of baseball sports videos is that they usually have a well-defined structure that contains segments of pitching and batting [119]. A baseball event can be defined as the portion of a video clip between two pitches: a play is a concatenation of many events, and a baseball video is composed of a series of plays. The recognized caption is inferred to find possible semantic categories of play in the first step; visual features of the video are utilized to find out the type of play, from one of *non-hitting*, *infield*, and *outfield*, in the second step; and the resulting information is combined to find the exact semantic meaning of the play. It is then semantically classified using an algorithm that integrates caption rule-inference and visual feature analysis.

  LucentVision is a multimedia system for live and real world event detection [20]. The system can integrate between 2-8 cameras, incorporates enhanced analysis and visualization, and includes object tracking and virtual replays. LucentVision sends out broadcast grade graphics over the air, generates Virtual Reality Modelling Language environment models, and detects changes in those models throughout an event. The queries in the database include those based on scores, statistics, space, and of a historical nature. LucentVision provides live web updates to the ATP tour official website (atptour.com). The system periodically updates the site and offers a selection of LucentVision visualization options, including a map, statistics and a virtual replay. Multiple types of baseball event detection, using superimposed caption text detection and recognition in [120]. These include *out*, *run*, *walk* and *score*, and also event boundary detection. Events, and the associated game state information, are extracted using a videotext detection and recognition module. The event boundary detection is based on video view recognition. An event occurrence is detected by the caption changes. The pitching view and non-active views surrounding the event are detected to determine the beginning and end points of an event.

  Non-hitting, *in-field*, and *out-field* events from a baseball game are detected using video motion vectors [121]. These are used as features for a three-layer feed-forward neural network, which is shown to be adept at correct classification. The neural network is trained using the back-propagation algorithm. In another approach to event detection in baseball videos [122], a rule-based decision module infers what happened by checking the information changes in

the caption. With the help of official baseball rules, the rule-based decision module detects events occurring between two consecutive pitch shots. In addition, a model-based decision module further classifies events that could not be explicitly determined by checking the caption information. The four shot context features from the test sequence are classified as events by the k-nearest neighbour algorithm. The following events are detected: *hit*, *double*, *triple*, *home run*, *stolen base*, *caught steal*, *fly out*, *strikeout*, *base on balls*, *sacrifice bunt*, *sacrifice fly*, *double play*, and *triple play*.

4.3.2 Event detection from surveillance video

Event detection from surveillance and monitoring videos plays a practical role in personal security. Typical vehicle-related events in unmanned airborne vehicle surveillance [25] include: *approach checkpoint*, *stop short before arriving*, *car goes through checkpoint*, *car avoids checkpoint*, *move inside*, and *leaving*. Another example is theft at a phone-booth. The events include: *bringing to object*, *attacking a person*, *using phone*, *taking away the object*, *passing by*, etc. Temporal interval logical relationships are used to compute multiple agents, and multiple threaded events.

Anomalies in individual and interactive event sequences are an important issue in surveillance. In [71], an SMC method is employed to track an event sequence in discrete state space for anomaly detection. A Markov Random Field (MRF) is used to extend SMC for both individual and interactive events. An adaptive temporal differencing method is used to describe pixel changes, and an effective and efficient event representation approach, employing SMC and subspace methods, are combined to implement event tracking in probabilistic manifolds.

Chan et al [41] studied event recognition in a busy scene consisting of a refuelling airplane being serviced. Events recognised included: *close-to*, *contained-in*, *appear-near*, *disappear-near* and *moving*. For this application, object tracks are often fragmented, therefore, the level of track fragmentation best for event recognition was investigated. The approach was to use DBNs to model events, with observed nodes corresponding to the spatio-temporal semantic relationships between event actors and elements. Interpolation over track gaps, in both space and time, was then performed. The model represents complex events defined by interactions between multiple object tracks. The main contribution in [42] is the combination of track linking with event recognition in a joint formulation that optimizes both simultaneously. The advantage is that events can be recognized despite highly fragmented tracking due to long occlusions, in scenes with many non-involved movers, under different scene viewpoints and / or configurations.

Parking lot events are monitored in [123]. The event recognition module receives input information, such as location, tracking and classification of moving objects, and classifies an event as standard or dangerous on the basis of pre-defined object motion models. The work consisted of three parts: object classification using an adaptive high order neural tree, object tracking based by the mean shift algorithm, and recognition of normal, suspicious and dangerous events. The functionalities of a car park surveillance system usually include the online classification and detection of abandoned objects [124], and the automatic detection and indexing of video event shots showing the cause of an alarm [57]. The method for video-event shot detection and indexing is obtained by integrating the metadata of the three subsystems, as well as addressing video-object layers represented by blobs.

Events are also detected from traffic surveillance videos [125]. The detected events have three levels: low, medium, and high (traffic jam-low, lane change-medium, and traffic rule violation-high, respectively). The low-level module detects moving objects from captured images; the middle-level module analyzes the relationships between the input image and the road surface in the real world; the high-level module calculates parameters for each passing vehicle. The main feature is that no prior information of the capture conditions is required. Nishida et al [126] develop a tracking algorithm, based on a spatio-temporal MRF, in order to acquire and visualize events from traffic images with occlusion and clutter problems. The detected events are vehicle counts of traffic direction, velocities, frequent paths and so on. Jung et al [67] track moving objects using Kalman filtering and occlusion reasoning. The trajectory of the moving object is approximated by polynomial functions and is described by motion trajectory descriptors.

Events can also be detected from crowd scenes [127][56][93]. Crowd events are usually difficult situations containing highly-cluttered dynamic backgrounds. Khan et al [127] present a planar homography constraint to resolve occlusions and to

robustly determine locations on the ground plane corresponding to people's feet. The algorithm is able to accurately track people in all views maintaining correct correspondences across views. The algorithm is ideally suited to situations when occlusions between people would seriously hamper tracking, or if there are simply not enough features to distinguish between different people. The major contribution is the detection of ground plane locations of people and the resolution of occlusion using a planar homography constraint. Combining foreground likelihoods from all views into a reference view and using the homography constraint ensures that the blobs representing feet are segmented out. The feet are tracked by clustering them over time into spatially coherent worms. In [56], crowd behaviour is characterized by observing the crowd flow, with unsupervised feature extraction to encode normal crowd behaviour. The unsupervised feature extraction applies spectral clustering to find the optimal number of models required to represent normal motion patterns. Using projections of the eigenvectors in the sub-space spanned by normal crowd scenes, the proposed technique applies spectral clustering to automatically identify the number of distinct motion segments in the sequence. The features in the clustered motion segments are used to train different Multiple Observation HMMs for normal sequences, which compose a bank of models for the simulated training video.

Another application is the detection of events in commercial spaces such as retail stores [128]. The framework is evaluated in a retail environment for detecting trollies entering or leaving the back door of a store and the opening or closing of a cashier's cash drawer. Five different event classes were automatically learned, in terms of their location and temporal order, through unsupervised clustering, with the following events manually labelled: *can taken*, *entering* and *leaving*, *shop keeper*, *browsing* and *paying*. In other work, learned mixture models were utilized to recognize detected blob-level events online [94].

Events involving two-person interactions have also been the subject of research [129]. Two-person interactions are a combination of single-person actions, which are themselves composed of a human body-part gesture. Each gesture is an elementary motion event and is composed of a sequence of instantaneous poses at each frame. The method is based on a hierarchy of action concepts: static pose, dynamic gesture, single-person action, and person to person interactions. Human actions are represented by multiple triplets aligned according to spatial-temporal constraints between actions.

Detecting unusual activities by dividing the video into equal length segments and classifying the extracted features into prototypes, from which a prototype segment co-occurrence matrix is computed [130]. A correspondence relationship between prototype and video segments, which satisfies the transitive closure constraints, is sought. The main feature of this algorithm is that it utilizes extremely simple features that are automatically selected from the signal. Zhou et al [131] detect

unusual events via multiple camera mining. The unusual event detection uses two-stage training to bootstrap a probabilistic model for common events. An event not classified as common is considered unusual. Zhang et al [55] proposed a semi-supervised adaptive HMM framework, in which common event models are initially learned from a large data set, whilst unusual event models are learned by Bayesian adaptation.

The IBM Smart Surveillance System (S3) has an *open and extensible architecture* for video analysis and data-management. Its role in video analysis is to encode the camera streams and send them to the video or streaming database, and also to analyze the camera streams for events and send the resulting metadata in XML format to the metadata database. Its role in data management is in providing a human-interface layer for queries, alerts, events and real-time event statistics. The system consists of middleware for use in surveillance systems, and provides video-based behavioral analysis capabilities. S3 consists of two components: the Smart Surveillance Engine, which provides the front end video analysis capabilities, and middleware for Large Scale Surveillance, which provides data management capabilities.

4.3.3 Meeting event detection from indoors scenes

Meetings are social events where people exchange ideas. An information system for indoor-group oriented activities is provided in [132]. This involves the storage and indexing of multimedia data consisting of video, audio, PowerPoint files and other media-based information. Two steps are used for the meeting system: aspects of the event model are specified based on user requirements and domain semantics; the second involves implementation of the model. A natural approach would be to model both the static and dynamic aspects of the information simultaneously with one model.

Meeting events are detected from projected documents based on document image analysis for integrating non-temporal documents into multimedia meeting archives [133]. The approach takes advantage of the observable events related to documents that are visible during meetings. Slide changes are detected using the Synchronized Multimedia Integrating Language.

Bayesian Network, Gaussian Mixture Model, Maximum Likelihood Pixel, Radial Basis Network, and SVM classifiers are evaluated for detecting meeting events such as discussion, monologue, note-taking, white board activities and presentations [134] [61]. Segmentation and classification of meeting events are implemented using multiple classifier fusion and dynamic programming. Within the DBN, a multistream HMM is coupled with a linear dynamical system to compensate for disturbances in the data. Three audio and visual modalities are fused in the multi-stream

HMM. The DBN shows a significantly higher recognition performance compared to a single-stream HMM. Combining artificial neural networks and a HMM result in a highly discriminative system which outperformed conventional models [61].

Meeting event labelling is both laborious and time-consuming [60], since meetings are often lengthy and events are jointly defined by audio-visual patterns. A meeting is modelled as a sequence of exclusive events taken from the following set: *discussion*, *monologue*, *note-taking*, *presentation* and *white-boarding*. Given a sequence of audiovisual features extracted from a meeting, the Viterbi algorithm produces the sequence of states, in other words, events that are most likely to have generated the features.

Events in an office environment are detected in [135]. These events include *talking on a phone*, *checking voicemail*, *bringing a cup to a face*, *scratching/rubbing face*, *yawning and hand at mouth*, *putting on glasses/earphones*, and *rubbing eyes*. Temporal boosting was used to improve weak classifiers by allowing them to use the previous classifiers response in evaluating the current frame, and making use of the temporal continuity of video at the classifier and detector level. In addition, the framework is able to combine information from multiple cameras to increase overall system performance.

Chapter 5

Multi-Modal Event Computing

An event is a fundamental semantic concept in multimedia systems, which are fast becoming ubiquitous. There are strong and deep conceptual, engineering, computational and human-centred design reasons to consider events as a primary structure for organizing and accessing dynamic multimedia systems. Consequently, event-based applications are under development in different fields. In this chapter, an overview is given of these applications, in particular, event detection in surveillance and sports videos is considered. Surveillance videos mainly contain events that have happened in scenes such as car parks, airports, lobbies, traffic, checkpoints etc. Several special events, mostly related to surveillance, have also been widely studied, such as emotion events [136], exciting events [137], unusual events [55][131], and suspicious events [124]. Event detection in live broadcast videos from sports, such as baseball, soccer, football etc., has been widely studied in recent years due to the tremendous commercial potentials.

5.1 Multi-Modal Event Capturing in a Sentient Environment

The creation of an EventWeb can now be considered. Different from a traditional website, in this chapter, a comparison is made between current DocumentWeb and up-and-coming EventWeb, a concept in a new emergence area; the structure of famous and biggest event web: SEraja.com are introduced as an example, the events detected from multiple sensors with multi-mode data are stored for future retrieval and analysis. Based on the events, it is possible to detect the powerful EventBase, a prototype of an EventWeb can then be constructed. In this prototype, the sensors mounted in the office building are used for multi-mode event detection and connect

the multiple event websites together to share the EventBase. The continuous event streams will be stored in the EventWeb and present for users in real-time.

Event is symbolic abstractions for semantic segmentation of happenings in a specific spatio-temporal volume of the real world. Events have the outstanding attributes to naturally abstract and summarize the descriptions and further for retrieval, management, inference and exploration. The current WWW is really a DocumentWeb in which each node is a document and connected to others using manually created referential links. Due to the emergence of digital media devices and technology, the development phase of EventWeb is current in force. EventWeb [2, 3, 5] organizes all data in terms of events and experiences, allows users to widely access from various users' perspectives, it is possible to incorporate with dynamic, temporal and live data. Of course, it allows users to access any event nodes from anywhere.

5.1.1 Event

Life occurs in three-dimensional space, this space is time based and continuous. To refer to space, different types of abstractions are used. Some of the abstractions are based on natural characteristics like the concept of universe, planets, stars, continents or oceans. Others are artifact of human nature. Some of these are natural and some man-made but they all get strongly linked to space because they remain relatively fixed in space. These objects have fixed position in space but initially, space is defined in terms of these objects. It is common to specify location in terms of position on a road, or in a city, or in a country. With popularity of digital maps, the terms to describe space in absolute terms using latitude and longitude have become more popular, but the abstractions based on natural and man-made objects and structures remain more common, these abstractions are more naturally part of a language.

Time has the similarity to space, it is continuously ticking. There are three very general abstractions on time, known as: past, present, and future. Structures are imposed on this continuous and seemingly infinite timeline. Calendars were developed in different parts of the world to provide us the structured way to specify and partition time. Obviously the aspects of this calendar are based on natural factors, like the seasonal or daily cycles, while others like hours, minutes and seconds are just a way to partition time.

5.1.2 DocumentWeb and EventWeb

The Internet has emerged as a mass communications and information medium. The Web today, however, is based on documents and can be called DecumentWeb. Each

node of this DocumentWeb is a page that is prepared to be a node on the page so that when a visitor wants to access this page, it can be sent to the user and displayed on his computer.

Since inception of the Web, in addition to enormous advances in the Web related technology, rapid advances have taken place in several other areas such as available bandwidth, wireless, sensing devices, storage, processing, audio and video processing. The Web as it was conceived in early 1990s is going to evolve into the Web that will provide us experiences using multiple sensory modalities in addition to providing us information just in the form text pages.

Table 5.1: Comparison of DocumentWeb and EventWeb.

DocumentWeb	**EventWeb**
Authoring systems	Event entry systems
Crawlers	Crawlers and NG-RSS
Document directory	EventBase
(Keyword-based) Search engines	Event exploration environment

For each event, all the data and information from sensors, documents, and other sources is united and available to the user independent of the media. The user then experiences the preferred parts of a particular event in the preferred medium. In this vision, following true Web philosophy, all events are treated equally. An EventWeb will be of great interest to current web users for many applications. Sporting events, meetings, lectures, concerts, and numerous other events are currently captured using only a few photos and sparse text pages, and they will use rich media and correspondingly provide rich experience to users. Sensors are now being connected to form networks for various Internet applications. In short, the beginnings of the EventWeb are being witnessed, just as about a decade ago the DocumentWeb emerged.

EventWeb can not become a reality like DocumentWeb until some important tools are made available. Currently these tools are not well developed. Media production environments and media search are progressing fast and are receiving attention both from research as well as business community. Media production environment is essential for people to easily produce the content that will be part of the EventWeb. Effective production tools will encourage users to put their events on the web to share with others. Experience with DocumentWeb has demonstrated that search tools become essential to locate content of interest.

5.1.3 SEraja Structure

SEraja.com has gone live. Users can enter any events from any place so soon events from all over will start appearing on the EventWeb, it looks very good and very promising. For the first time, there is a place for people to go beyond the 'calendar of events' and make events real experiences. And this is evolving into a web of events in which people will be able to immerse themselves and not only experience but also gain experiential insights.

Seraja.com is the biggest event website on the world so far. It collects events from user manual input in cyberspace. The events are uploaded in the following entities: title, description, start, time, end time, location, user ratings and relevant documents, etc. Users can upload their events from PC and Mobile environments.

Figure 5.1. shows the basic structure of SEraja event system. After users upload their events to the event database, the seraja.com will manage the events, the event miner will explore new events in the database, event aggregator starts to combine the similar events together and reduce the redundancy, event crawler will be working without stopping to report the event indexing servers for their findings.

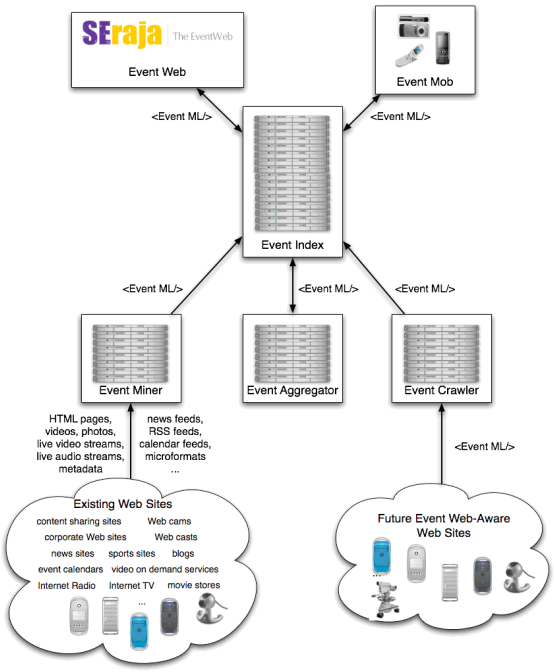

Figure 5.1: SEraja.com structure.

5.1.4 Event-based Analysis of Multi-Modal Data

In an experiential environment, multiple sensors will be owned for event collection, these sensors have a wide range of functionalities from camcorders to microphones for visual and audio information capturing; from smoking sensors to thermometer for environment monitoring; from motion sensors to RFID, from card readers to people counters, from fingerprint identification system to face recognition system, etc. Data is received from multiple sources in an experiential environment. What has been experienced is that the signals are not only for visual information, but also for audio signals and other information. Facing to so much multi-modal information, an answer has to be given on how to assimilate them, how to filter them for utilization, because using the information from multiple sources will make the problems much easier to be solved.

Use The Best Combination of Sources for Event Detection

After the proper sources for an experiential environment are selected, they should be combined as these sources play a vital role in event capturing. The criterion of best combination is the synthetic signal should continuously and completely include one or multiple events. From the combined signals, one can easily detect events without loosing. The simple combination is based on linear blending, people like to mix signals from different sources use the weights they assigned; most of time, average weights are employed. This obviously ignores the importance of the signals from different sources and is unreasonable because there is a constraint among the different information channels. The correct design is to provide a time based dynamic system. It is possible to select the most important signals with temporal variations. The best combination undoubtedly should be designed based on probability analysis.

Because the information combination is for event detection purpose, hence, the following question should be answered: where and when does the event happen? who and what is the event about? What is the content description of the event? Thus, in information combination, combining these components belonging to one event together, and find out the real description of the event. The redundancy of an even should be removed, the empty entities should also be filled up.

Know What Information Needs to be Extracted

In a sentient environment, information including events will continuously reach us from multiple sensors simultaneously, it is not possible manually to select those expecting channels, the dynamic and meaningful signals should be selected. Since the purpose is to capture events, only the significant information will be stored. The information should be properly integrated together to describe an event, it should include the content of one event at least. Thus, the information extracted should include one or many entire events.

In information extraction, the information extracted should correctly describes the basic components of an event. Namely, the extracted information should have great contributions for event capturing.

Use of Multiple Sources Makes Event Detection Much Easier

In a sentient environment, the data from single sensor or only one type of sensors mostly are not sufficient for us to judge whether an event exists or not since data of one sensor only describes the event from one aspect. The ideal way is to take multiple sensors into account and capture the data to describe events from multiple aspects, such as cameras capture the visual information, microphones capture the audio

information, motion sensors and APID capture moving information of human bodies. If multiple sensors are used, it will provide multiple sources of event description; their complement will be helpful to fill-up the event entities and their intersections are very useful for us to judge the events happened in the real world and enhance the confidence of event detection. It will be possible to understand the events in the real world from various aspects.

5.2 Approaches

5.2.1 Capture Events Using Requirement-Specific Sensors

In order to capture events, multiple types of detectors are required for event detection. As shown in Figure 5.2, at least three types of events are needed, such as: video capturing, text capturing and other sensors to detect events from a sentient environment. In the environment, all events have the correlations, the correlations can help us to capture multi-modal events.

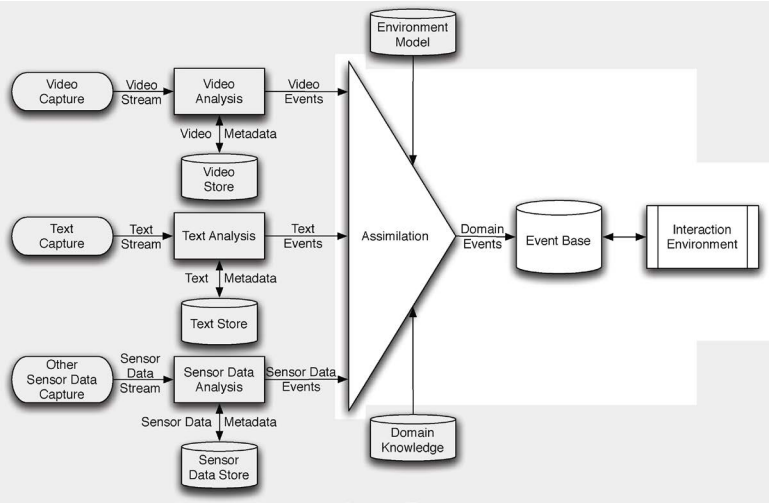

Figure 5.2: Event capturing.

When capturing events from videos, the visual sensors include video cameras, handy camcorders, and wireless video camera, etc. These sensors can report us the visual information and object motion in the environment. Typically, pictures can tell us the event at one moment, while video cameras can show us the continuously and entire story. Normally, some photos are used to find out one event, as well as some video shots to find one video event. The visual events occupy more than 80% information of the Human Perceptual System.

Audio events should also be captured. The typical audio sensors are used to detect the information of sound or speech. The speech can be converted to text / scripts, people may use the tools of lexical processing to detect what events happened in the real world. Mining from the text stuff is a reliable way to detect events, however, because of detecting errors in lexical event detection, the detected events may not correctly used in event detection, the noises need filtering.

Event detection from audio signals directly is also possible, audio recognition knowledge should be used to determine the types of events. It is possible to detect events using the frequency and spectrum analysis, finally convert audio signals to events.

Events also can be captured from text. Text event detector is absolutely powerful. Currently most of the media data are saved as text, most of electronic publications are in the text form, such as web pages, news and the electronic publication books. Most of lexical events have a very good structure, such as calenderer, news, paper references, meeting or conference fleets. The stuff includes the basic structure of events, such as time, venue, people, content, etc. These information can be directly stored into the event database once the event components are verified to belong to one event.

In short, although the detectors are various for event detection, some events can be obtained independently from the media. The detected events will be saved into the EventBase for users to widely browse and comments.

5.2.2 Store All Events From a Specific Environment in an Event Server using E

When significant events are captured, they are stored into multiple servers in the uniform format for indexing and retrieval. Of course, an EventBase saves event items, each event will be allocated a unique identification for recognition, the basic components (temporal and spatial information) of an event also can be found in the database. The referential relationships and links of an event are stored for further organizing. Further more, the EventBase provides the complementary information of the links for users to share. For pictures, the low features such as histogram, texture, moments and edges information are available beside the metadata EXIF data if available. For video and audio clips, the basic video and audio parameters of those footages are also available; for text events, the text information and the analyzing results are also available too. The users can easily access the events and relevant information using a browser.

Different from achieved data, for the living events from a sentient environment, the events are coming in real-time. All the events will come continuously and will be ingested. The EventBase provides the sufficient space to store them even if it has to be stored into two separate EventBase. On the other hand, the events from one environment have the possibility to be saved into different event databases, the events from different environments may be stored into different event bases. The eventbase should have an indexing mechanism to deal with these events from different databases. The different databases will be connected by the indexing servers via internet, the collaboration of such EventBase is extremely expected. Correspondingly, there should be some EventWeb should be properly connected to the different EventBases for event presentation. The EventWeb has the presentation tool to connect the event contents scattered in the various servers together and show a powerful way to show the event content.

5.2.3 Develop Aggregation / Crawling Approach to Create a Situation Model for the Environment

The events in EventBase are various since they are detected from different sentient environments. In order to effectively organize the events in the ideal way, a crawler is used to traversal events and organize them in the indexing servers. The similar events should be concatenated together to get the precise description by using aggregation. The relationships between the components of two events are various including overlapping, partial overlapping, concatenation, separate (no relationship). For the multiple relationships, they are placed in the event pool and deal with them

during a specific period of time. The relationships amongst components of events are assimilated to consist of one concrete event.

The crawlers have the responsibility to report the indexing servers the new uploading events and inform the system the latest upgrading in the indexing database, and timely provide this information for the system to present the events for users. For the visual information, the crawlers may have the responsibility to report the servers the low features of images and videos as well as features of audios.

5.2.4 Develop Event Interaction Environment

Once events are captured and stored in the event database, an interaction environment should be developed for users to access the events from EventWeb and show users the available events in the interaction environment. Users can use this interaction environment to input and output events; search, query and mine events; and get the system rapid response after an operation. Once an event is selected, the end-user can get the event content and relevant links from the interface conveniently.

- *Search* Once EventWeb for a real-time sentiment environment has been built up, millions of events are poured into the database system continuously. Search becomes a very powerful tool to get those relevant events from the database. In order to improve the search efficiency and precision, the search results should be stored on the server side for user to access. With a single EventBase, this is very easy to be implemented, however, with EventWeb constructing on multiple event websites, the crawlers should be utilized to collect event information uninterruptedly and stored the search results on the indexing servers, and the crawlers should dynamically update the relevant records on the event indexing servers for the special key words timely. All the possible searching results of the possible keywords should be ready for use. Those results for the keywords with high search frequency are extremely examined for users to retrieval, each item of the search results is ranked in the descent order, the highest score item is listed at the top of the list. When a user searches for an event from anywhere on the world, no mater which keyword (s)he has given, the system can render the results to a user timely and correctly.

- *Query* Normally people query the events with one or more fields of database in a given order, the query is based on one database. The query includes two ways: basic query and advanced query, like many have experienced in a Library. However, EventWeb is working on multiple event database, the feedback of a query should be very complex. When a user submits a query to the system, the system has to give him the response from multiple servers

and websites, all the query results should be submitted to the user interface, and the results should be well organized. Therefore, the system should have the indexes of all EventBase, and query results should be organized in the proper way for users to access. The important way is that the query should be properly saved for the next user. If a user sends the same query, it is better to provide them the existing event items which have been queried by the previous users. Definitely, the query results should be dynamically updated in case that new events are added into the system.

- *Mining* Event mining is to explore new events from those existing ones. Those new events have the same contents of the existing events while they present events from different aspects. In order to mine events, clustering and indexing are used to find the events in the EventWeb. EventWeb consists of multiple event databases, event mining is to seek the relevant events from these event databases. The event miner should explore new events from every database first, then the mined results can be properly merged together with different criterion. The mined results should have very close logic relationships for the events in the event web, no matter the relationships are based on time, space, people or content.

5.3 Event Detection Systems

5.3.1 Event Environments

The visual monitoring environment is shown in Figure 5.3. There are three cameras fixed on the same side of wall with various altitudes, another camera is located at the opposite wall. The coffeepot and the accessories are all put on the kitchen table under the three cameras. From these LinkSys wireless web cameras, users are able to clearly observe the events taking place in the coffee room.

5.3.2 Types of Events to be Detection

- *A. Scenario 1: Coffee Room* Events happening in an environment are connected to the objects in the environment. In the coffee room, the possible objects and the possible events are listed below.

- *B. Scenario 2: Group Meeting* Meeting is a very popular event, all the research groups may have their meeting every week. In the meeting room, it normally includes the items such as tables, chairs, laptops and projectors, white board

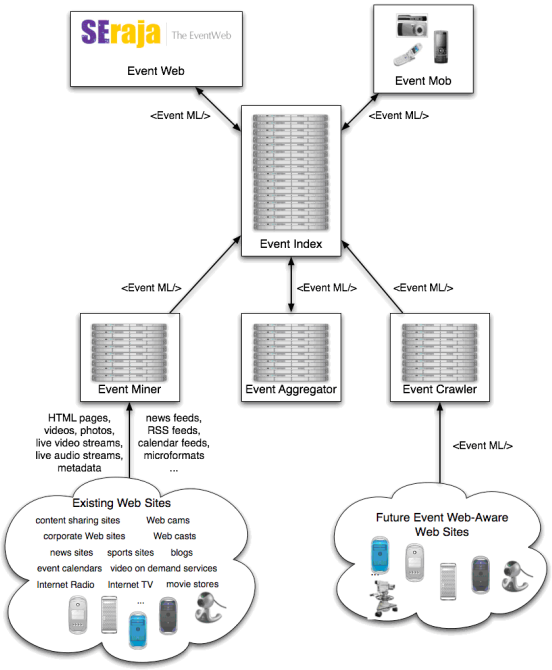

Figure 5.3: visually monitoring environment in coffee room.

and color pens and human bodies. The events will be happened near these objects include in the table 3.

- *C. Scenario 3: Corridor* In the corridor out of the office, there are two cameras, one motion sensors, and some sofa for resting. The high frequency events may include in table 4.

5.3.3 Sensors are Sources of Event Information

From event detection point of view, the information and experiences are required to have rich event components since the aim is to capture events from the sensory data. As previously mentioned, multiple sensors can be employed for event detection, and all the sensors are the sources of event information, they provide the complementation for us to detect the right events. Sensors are sources of event information means that no sensor can be ignored in event detection, they all play a vital role in event detection. Some sensors may only provide one aspect of an event, some sensors may provide another aspect of the event. Assimilation of the sensory data will assist us to fill-up the event entities in the event database.

In a sentient environment, the data is captured from sensors continuously. Sensors are with various attributes, such as digital cameras have their parameters of adjustment. So far, the types of sensors include cameras, microphones, thermometers, smoke sensors, people counters, ALID in an office environment. These sensors are the event sources. Some sensors can directly provide the event information, some sensors provide the event information indirectly, need us to refine information and detect events from the data. These sensory data have wide range of usages in events.

5.3.4 Environment Model

In a sentient environment, there are a lot of objects, it is assumed that most of events are closely related to the objects (certainly some events are out of the environment, such as whether and earthquake). Currently, these objects occupied certain space in a sentient environment, and the objects have their spatial and temporal relationships. The environment model consists of the objects and the relevant events. The objects connect to each other and impact each other in the community. Using this environment, the relationship amongst different events can be observed, the events detection is reasonably based on the environment model.

The multi-modal events can be captured with multiple sensors in this environment. From the environment, the model is built which correctly reflects the temporal and spatial relationship between different objects, the interactions among the objects and the multiple sensors amounted in this environment will be bounded together, and the multi-modal events are detected from this environment modal and the sensor signals.

5.4 Detecting Events

The footstone of the event web is to detect and explore events from the sensor data. After this step has been successfully taken, events can be stored into the event database. Only based on this preparation, event search and retrieval becomes true.

5.4.1 Detecting Methods

Event detecting depends on sensors and media types. There is no uniform and standard method for event detection. However, no matter which method a user takes, the target of event capturing is to fill-up the entities of an event record in an EventBase. Therefore, in event detection, the method to seek the temporal clues, spatial clues and content description of various events is the uniform way to detect events from various media. Motion sensors directly provide the event information, however some sensors such as cameras and microphones cannot provide the event data directly. The event entities should be inferred from the media content and use human knowledge to explore the events from these sensor data.

5.4.2 Automatic or Semi-Automatic

Automatic event detection is the ultimate goal, however, it does not work all the time. Events are captured from multiple sensors and events detected from the data of multiple views. The semi-automatic is a reasonable way for users to detect events from the complex environments, because most of time, the detected events have errors, human has to correct the detect wrong events. Therefore, the ideal way for event detection is based on combination of automatic and semi-automatic way. Most of time, the system works automatically, if errors are found or improper events, manual correction is performed. The administrators have the right to remove the improper records from EventBase. Some events can be added if they should appear there clearly.

5.5 Reports

Once events are uploaded to EventBase, the servers should report the users the latest updates. The updates include the events and the relevant information. Different media should have different events for user to browse.

5.5.1 Videos, Audio, and Sensory Data

The EventWeb could timely report users the events existing in the database, this also shows links between events and also the content of each media. The end users have the privilege to get the relevant events and their related media.

In the report, events link to media. This not only includes the media content, but also the metadata and the relevant information which have links to the events. Video events are closely related to the features of videos, such as duration, shot, frame numbers, trajectories, streaming bits per seconds, etc; audio events are closely related to volume, duration, clips, channels, bit rate etc. Pictures are closely related to EXIF and other metadata. Sensory data emphasizes the time synchronizing, frame size, bandwidth and transmitting rate of communication channels. These event links explain the content and attributes of an event, but also notice the relationship between this media and others.

5.5.2 Privacy Issues

In EventWeb, not all the events are allowed to be published publicly, not all the information could provide for the users to browse. Because most of the sensory data are at real-time for the real world, it has the potential to be utilized for identifying a person's traces once he is tracked. Undoubtedly, it will interrupt one's privacy. In the EventWeb, different levels of privileges should be employed for different persons who read the events in different ways. Only the top level users, administrators and supervisors have the right to access the events, others' privilege will be confined in a certain limitation. The better management will be properly utilized the events for users to construct a perfect system.

Bibliography

[1] S. Haynes and R. Jain, "Low level motion events, trajectory discontinuities," in *Proc. of The First Conference on Artificial Intelligence Applications*, San Diego, USA, Dec. 1984, pp. 251–256.

[2] P. Appan and H. Sundaram, "Networked multimedia event exploration," in *Proc. of ACM Multimedia 2004*, New York City, USA, Oct. 2004, pp. 40–47.

[3] Y. Ma, M. Bazakos, B. Miller, and P. Buddharaju, "Activity awareness: from predefined events to new pattern discovery," in *Proc. of ICVS'06*, 2006, p. 11.

[4] "National Institute of Standards and Technology," http://www.nist.gov/index.html.

[5] R. Nevatia, J. Hobbs, and B. Bolls, "An ontology for video event representation," in *Proc. of CVPRW'04*, vol. 9, no. 27, Washington, USA, June 2004, p. 119.

[6] A. Vassiliou, A. Salway, and D. Pitt, "Formalizing stories sequences of events and state changes," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 587–590.

[7] N. Peyrard and P. Bouthemy, "Detection of meaningful events in videos based on a supervised classification approach," in *Proc. of IEEE ICIP'03*, Sep. 2003, pp. 621–625.

[8] A. Quinton, "Objects and events," *Mind*, vol. 88, no. 350, pp. 197–214, Apr. 1979.

[9] T. Syeda-Mahmood and S. Srinivasan, "Detecting topical events in digital video," in *Proc. of ACM Multimedia'00*, Marina del Rey, Los Angeles, USA, Oct. 2000, pp. 85–94.

[10] P. K. Atrey, M. S. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Springer/ACM Multimedia Systems Journal*, vol. 12, no. 3, pp. 239–253, Dec. 2006.

[11] N. H. Gehani, H. V. Jagadish, and O. Shmueli, "Composite event specification in active databases: Model & implementation," in *Proc. of VLDB'92*, Vancouver, Canada, Aug. 1992, pp. 327–338.

[12] S. Hongeng, "Unsupervised learning of multi-object event classes," in *Proc. of the 15th British Machine Vision Conference (BMVC'04)*, London,UK, 2004.

[13] U. Westermann and R. Jain, "Toward a common event model for multimedia applications," *International Journal on Semantic Web & Information Systems*, vol. 14, no. 1, pp. 19–29, Jan. 2006.

[14] M. F. Worboys and K. Hornsby, "From objects to events: Gem, the geospatial event model," in *Proc. of GIScience'04*, Adelphi,USA, Oct. 2004.

[15] M. Hopkins, "Strategies for determining causes of events," in *Proc. of AAAI'02*, Palo Alto, California, Mar. 2002, pp. 546–552.

[16] S. Dai and A. P. Dhawan, "Adaptive learning for event modeling and characterization," *Pattern Recognition*, vol. 40, no. 5, pp. 1544–1555, Oct. 2007.

[17] M. Teisseire, P. Poncelet, and R. Cicchetti, "Towards event-driven modelling for database design," in *Proc. of VLDB'94*, Santiago de Chile, Chile, Sep. 1994, pp. 285–296.

[18] A. Demers, J. Gehrke, M. Hong, M. Riedewald, and W. White, "A general algebra and implementation for monitoring event streams," Cornell University, Tech. Rep. TR2005-1997, 2005.

[19] J. Baulier, S. Blott, H. F. Korth, and A. Silberschatz, "A database system for real-time event aggregation in telecommunication," in *Proc. of VLDB'98*, New York, USA, Aug. 1998, pp. 680–684.

[20] G. S. Pingali, Y. Jean, A. Opalach, and I. Carlbom, "Lucentvision: Converting real world events into multimedia experiences," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 1433–1436.

[21] D. Pack, R. Singh, S. Brennan, and R. Jain, "An event model and its implementation for multimedia information representation and retrieval," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 1611–1614.

[22] A. Thawani, S. Gopalan, and S. V, "Event driven semantics based ad selection," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 1875–1878.

[23] A. R. J. Franois, R. Nevatia, J. R. Hobbs, and R. C. Bolles, "VERL: An ontology framework for representing and annotating video events," *IEEE Multimedia*, no. 76, pp. 269–288, Nov. 2003.

[24] A. Hakeem, Y. Sheikh, and M. Shah, "Casee: A hierarchical event representation for the analysis of videos," in *Proc. of AAAI'04*, San Jose, USA, Jul. 2004, pp. 263–268.

[25] S. Hongeng and R. Nevatia, "Large-scale event detection using Semi-Hidden Markov Models," in *Proc. of IEEE ICCV'03*, Nice, France, Oct. 2003, pp. 1455–1462.

[26] E. Malaia, "Event structure representation in ontological semantics," in *Proc. of MLMTA (International Conference on Machine Learning Models, Technologies & Applications)*, Las Vegas, USA, Jun. 2006, pp. 36–42.

[27] J. Pinzon, R. Singh, W. Taube, and J. Galan, "Designing interactions in event-based unified management of personal multimedia information," in *Proc. of IEEE ICME'06*, Canada, Jul. 2006, pp. 337–340.

[28] J. Tesic, S. Newsam, and B. Manjunath, "Scalable spatial event representation," in *Proc. of IEEE ICME'02*, Lausanne, Switzerland, Aug. 2002, pp. 229–232.

[29] R. Hamid, A. Y. Johnson, S. Batta, A. F. Bobick, C. L. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams," in *Proc. of IEEE CVPR'05*, San Diego, USA, Jun. 2005, pp. 1031–1038.

[30] T. Okadome, "Event representation for sensor data grounding," *International Journal of Computer Science and Network Security*, vol. 6, no. 10, pp. 129–162, Oct. 2006.

[31] H. Veeraraghavan, N. Papanikolopoulos, and P. Schrater, "Learning dynamic event descriptions in image sequences," in *Proc. of IEEE CVPR'07*, Minnesota, USA, Jun. 2007, pp. 1–6.

[32] S.-W. Joo and R. Chellappa, "Attribute grammar-based event recognition and anomaly detection," in *Proc. of CVPRW'06*, New York, USA, Jun. 2006, pp. 107–115.

[33] K. Alahari and C. Jawahar, "Discriminative actions for recognising events," in *Proc. of ICVGIP'06 (LNCS 4338)*, India, Jun. 2006, p. 552563.

[34] W. P. Krzysztof Cios and R. Swiniarski, *Data mining methods for knowledge discovery*. Kluwer Academic Publishers, 1998.

[35] H. M. D. Hand and P. Smyth, *Principles of Data Mining*. MIT Press,Cambridge, USA, 2001.

[36] G. P.-S. W. Frawley and C. Matheus, "Knowledge discovery in databases: An overview," *AI Magazine*, pp. 213–228, 1992.

[37] E. Schwalb, K. Kask, and R. Dechter, "Temporal reasoning with constraints on fluents and events," in *Proc. of AAAI'94*, Seattle, USA, Aug. 1994, pp. 1067–1072.

[38] R. G. Alan Fern, Jeffrey Mark Siskind, "Learning temporal, relational, force-dynamic event definitions from video," in *Proc. of AAAI'02*, Palo Alto, California, Mar. 2002, pp. 159–166.

[39] A. Fern, R. Givan, and J. M. Siskind, "Specific-to-general learning for temporal events," in *Proc. of AAAI'02*, Palo Alto, USA, Mar. 2002, pp. 152–158.

[40] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen, "Detecting rare events in video using semantic primitives with HMM," in *Proc. of IEEE ICPR'04*, Cambridge, UK, Aug. 2004, pp. 150–154.

[41] M. T. Chan, A. Hoogs, Z. Sun, J. Schmiederer, R. Bhotika, and G. Doretto, "Event recognition with fragmented object tracks," in *Proc. of IEEE ICPR'06*, HongKong, China, Aug. 2006, pp. 412–416.

[42] M. T. Chan, A. Hoogs, R. Bhotika, A. G. A. Perera, J. Schmiederer, and G. Doretto, "Joint recognition of complex events and track matching," in *Proc. of IEEE CVPR'06*, New York, USA, Jun. 2006, pp. 1615–1622.

[43] H. Gu and Q. Ji, "Facial event classification with Task Oriented Dynamic Bayesian Network," in *Proc. of IEEE CVPR'04*, Reno, USA, Jul. 2004, pp. 870–875.

[44] B. Adams and S. Venkatesh, "Situated event bootstrapping and capture guidance for automated home movie authoring," in *Proc. of ACM Multimedia'05*, Singapore, Nov. 2005, pp. 754–763.

[45] H. Kawashima and T. Matsuyama, "Integrated event recognition from multiple sources," in *Proc. of IEEE ICPR'02*, Quebec, Canada, Aug. 2002, pp. 785–789.

[46] A. Bonzanini, R. Leonardi, and P. Migliorati, "Event recognition in sport programs using low-level motion indices," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 2127–2130.

[47] Y. Satoh, H. Tanahashi, C. Wang, S. Kaneko, Y. Niwa, and K. Yamamoto, "Robust event detection by Radial Reach Filter (RRF)," in *Proc. of IEEE ICPR'02*, Quebec, Canada, Aug. 2002, pp. 623–626.

[48] C. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 638–647, 2006.

[49] Z. Ghahramani, *Adaptive Processing of Sequences and Data Structures, Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, 1998, ch. Learning Dynamic Bayesian Networks, pp. 168–197.

[50] J. M. Siskind and Q. Morris, "A maximum-likelihood approach to visual event classification," in *Proc. of ECCV'96(LNCS 1065)*, London, UK, Jun. 1996, pp. 347–360.

[51] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Procedings of IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[52] G. Xu, Y.-F. Ma, H. Zhang, and S. Yang, "Motion based event recognition using HMM," in *Proc. of IEEE ICPR'02*, Quebec, Canada, Aug. 2002, pp. 831–834.

[53] M. Naphade and T. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proc. of IEEE ICIP'02*, 2002.

[54] S. Boykin and A. Merlino, "Machine learning of event segmentation for news on demand," *Communication of the ACM*, vol. 43, no. 2, pp. 35–41, Feb. 2000.

[55] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *Proc. of IEEE CVPR'05*, San Diego, USA, Jun. 2005, pp. 611–618.

[56] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modeling crowd scenes for event detection," in *Proc. of ICPR'06*, Hong Kong, China, Aug. 2006, pp. 175–178.

[57] P. Remagnino and G. Jones, "Classifying surveillance events from attributes and behaviour," in *Proc. of British Machine Vision Conf*, Manchester, UK, 2001, pp. 685–694.

[58] M. Al-Hames and G. Rigoll, "A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 45–48.

[59] M. Barnard and J.-M. Odobez, "Sports event recognition using layered hmms," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 1150–1153.

[60] D. Zhang, D. Gatica-Perez, and S. Bengio, "Semi-supervised meeting event recognition with adapted HMMs," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 1102–1105.

[61] S. Reiter, B. Schuller, and G. Rigoll, "Segmentation and recognition of meeting events using a two-layered hmm and a combined mlp-hmm approach," in *Proc. of IEEE ICME'06*, Canada, Jul. 2006, pp. 953–956.

[62] A. Hakeem and M. Shah, "Multiple agent event detection and representation in videos," in *Proc. of AAAI'05*, Pittsburgh, USA, Jul. 2005, pp. 89–94.

[63] G. Piriou, P. Bouthemy, and J.-F. Yao, "Learned probabilistic image motion models for event detection in videos," in *Proc. of IEEE ICPR'04*, Tokyo, Japan, Aug. 2004, pp. 207–210.

[64] D. Lee and M. Yannakakis, "Principles and methods of Testing Finite State Machines - A survey," *Proceedings of The IEEE*, vol. 84, no. 8, pp. 1090–1122, Aug. 1996.

[65] M. Bertini, A. D. Bimbo, R. Cucchiara, and A. Prati, "Object-based and event-based semantic video adaptation," in *Proc. of IEEE ICPR'04*, Cambridge, UK, Aug. 2004, pp. 987–990.

[66] G. Welch and G. Bishop, "An introduction to the Kalman Filter," in *Proc. of ACM SIGGRPH'01*, Los Angeles, USA, Aug. 2001.

[67] Y.-K. Jung, K.-W. Lee, and Y.-S. Ho, "Content-based event retrieval using semantic scene interpretation for automated traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 2, no. 3, pp. 151–163, Sep. 2001.

[68] C. J. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, pp. 121–167, 1998.

[69] G. Qiu, X. Feng, and J. Fang, "Compressing histogram representations for automatic color photo categorization," *Pattern Recognition*, vol. 37, pp. 2177–2193, 2004.

[70] D. Sadlier and N. E. OConnor, "Event detection based on generic characteristics of field-sports," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 759–762.

[71] P. Cui, L. Sun, Z.-Q. Liu, and S. Yang, "A sequential monte carlo approach to anomaly detection in tracking visual events," in *Proc. of IEEE CVPR'07*, Minnesota, USA, Jun. 2007.

[72] J. H. Piater, S. Richetto, and J. L. Crowley, "Event-based activity analysis in live video using a generic object tracker," in *Proc. of Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, 2002, pp. 1–8.

[73] C. Xu, J. Wang, Y. Li, K. Wan, and L.-Y. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proc. of ACM Multimedia'06*, Santa Barbara, CA, USA, Oct. 2006, pp. 221–230.

[74] J. C. Engle and A. Odutola, "Control field event detection in a digital video recorder," *U.S.Patent 5699124*, Oct. 2006.

[75] A. G. Matthew D. Cooper, Jonathan Foote and L. Wilcox, "Temporal event clustering for digital photo collections," in *Proc. of ACM Multimedia'03*, Berkely, USA, Nov. 2003, pp. 364–373.

[76] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM Transactions on Computing, Communication and Applications*, vol. 1, no. 3, pp. 269–288, Aug. 2005.

[77] J.-H. Lim, Q. Tian, and P. Mulhem, "Home photo content modeling for personalized event-based retrieval," *IEEE Multimedia*, vol. 10, no. 4, pp. 28–37, Oct. 2003.

[78] A. C. Loui and A. E. Savakis, "Automatic image event segmentation and quality screening for albuming applications," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 1125–1128.

[79] ——, "Automated event clustering and quality screening of consumer pictures for digital albuming," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 390–402, 2003.

[80] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, CO, USA, June 2005, pp. 178–187.

[81] M. Naaman, S. Harada, and Q. Wang, "Context data in Geo-referenced digital photo collections," in *Proc. of ACM Multimedia*, New York, NY, USA, June 2004, pp. 196–203.

[82] L.-J. Li and F.-F. Li, "What, where and who? classifying events by scene and object recognition," in *Proc. of IEEE ICCV'07*, Rio de Janeiro, Brazi, Oct. 2007.

[83] C.-H. Li, C.-Y. Chiu, C.-R. Huang, C.-S. Chen, and L.-F. Chien, "Image content clustering and summarization for photo collection," in *Proc. of IEEE ICME'06*, Canada, Jul. 2006.

[84] T. Mei, B. Wang, X.-S. Hua, H.-Q. Zhou, and S. Li, "Probabilistic multimodality fusion for event based home photo clustering," in *Proc. of IEEE ICME'06*, Canada, Jul. 2006, pp. 1757–1760.

[85] S. Haynes and R. Jain, "Event detection and correspondence," in *Proc. of Optical Engineering*, San Diego, USA, Dec. 1984, pp. 251–256.

[86] N. Johnson and D. C. Hogg, "Learning the distribution of object trajectories for event recognition," in *Proc. of the 6th British conference on Machine vision*, Surrey,UK, Aug. 1995, pp. 583–592.

[87] C. Rao, M. Shah, and T. Syeda-Mahmmod, "Invariance in motion analysis of videos," in *Proc. of ACM Multimedia'03*, Bekerley, USA, Nov. 2003, pp. 518–527.

[88] C. Rao and M. Shah, "View-invariant representation and learning of human action," in *Proc. of IEEE Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, Jul. 2001, pp. 55–63.

[89] T. Syeda-Mahmood and A. Vasilescu, "Recognizing action events from multiple view points," in *Proc. of IEEE Workshop on Detection and Recognition of Events in Video 2001*, Las Palmas,USA, May 2001, pp. 64–72.

[90] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Trajectory based event tactics analysis in broadcast sports video," in *Proc. of ACM Multimedia'07*, Augsburg, Germany, Oct. 2007, pp. 58 – 67.

[91] Z. Zhang, K. Huang, T. Tan, and L. Wang, "Trajectory series analysis based event rule induction for visual surveillance." in *Proc. of IEEE CVPR'07*, Minnesota, USA, Jun. 2007.

[92] Y. Ke, "Efficient visual event detection using volumetric features," in *Proc. of IEEE ICCV'05*, Beijing, China, Oct. 2005, pp. 166–173.

[93] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. of IEEE ICCV'07*, Rio de Janeiro, Brazi, Oct. 2007.

[94] T. Xiang, S. Gong, and D. Parkinson, "Autonomous visual events detection and classification without explicit object-centred segmentation and tracking," in *Proc. of British Machine Vision Conference*, Cardiff, UK, Sep. 2002, pp. 685–694.

[95] M. J. Black, "Explaining optical flow events with parameterized spatio-temporal models," in *Proc. of IEEE CVPR'99*, Ft. Collins, USA, Jun. 1999, pp. 326–332.

[96] K. Yoon, D. DeMenthon, and D. S. Doermann, "Event detection from MPEG video in the compressed domain," in *Proc. of IEEE ICPR'00*, Singapore, Nov. 2000, pp. 1819–1822.

[97] K. Alahari and C. Jawahar, "Dynamic events as mixtures of spatial and temporal features," in *Proc. of ICVGIP'06 (LNCS 4338)*, India, Jun. 2006, p. 540551.

[98] J. M. Siskind, "Visual event classification via force dynamics," in *Proc. of AAAI'02*, San Diego, USA, Jun. 2002, pp. 149–155.

[99] C. Lu and N. J. Ferrier, "Repetitive motion analysis: Segmentation and event classification," *IEEE Transactions on PAMI*, vol. 26, no. 2, pp. 258–263, Jan. 2004.

[100] N. C. Haering, R. J. Qian, and M. I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 10, pp. 857–868, Sep. 2000.

[101] M. Osadchy and D. Keren, "A rejection-based method for event detection in video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 14, pp. 534–541, April 2004.

[102] M. R. Naphade, A. Garg, and T. S. Huang, "Duration dependent input output markov models for audio-visual event detection," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Jun. 1997, pp. 369–372.

[103] T. S. H. Trausti T. Kristjansson, Brendan J. Frey, "Event-coupled Hidden Markov Models," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 385–388.

[104] H. Xu and T.-S. Chua, "The fusion of audio-visual features and external knowledge for event detection in team sports video," in *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, USA, OCT. 2004.

[105] ——, "Fusion of AV features and external information sources for event detection in team sports video," *ACM TOMCCAP*, vol. 2, no. 1, pp. 44–67, 2006.

[106] H. Xu, T.-H. Fong, and T.-S. Chua, "Fusion of multiple asynchronous information sources for event detection in soccer video," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 1242–1245.

[107] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. of IEEE CVPR'01*, Hawaii,USA, Dec. 2001, pp. 123–130.

[108] ——, "Statistical analysis of dynamic actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1530–1535, Sep. 2006.

[109] A. Amer, E. Dubois, and A. Mitiche, "Context-independent real-time event recognition: Application to key-image extraction," in *Proc. of IEEE ICPR'02*, Quebec, Canada, Aug. 2002, pp. 945–948.

[110] A. Amera, E. Duboisb, and A. Mitichec, "Rule-based real-time detection of context-independent events in video shots," *Real-Time Imaging*, vol. 11, no. 3, pp. 244–256, Jun. 2005.

[111] T. Syeda-Mahmood, "Retrieving actions embedded in video," in *Proc. of ACM Multimedia'02*, Juan Lins Pins, France, Nov. 2002, pp. 513 – 522.

[112] N. Babaguchi, S. Sasamori, T. Kitahashi, and R. Jain, "Detecting events from continuous media by intermodal collaboration and knowledge use," in *Proc. of IEEE ICMCS'99*, Firenze, Italy, Jun. 1999, pp. 782–786.

[113] N. Nitta, N. Babaguchi, and T. Kitahashi, "Extracting actors, actions and events from sports video - a fundamental approach to story tracking," in *Proc. of IEEE ICPR'00*, Barcelona, Spain, Sep. 2000, pp. 4718–4721.

[114] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 12, no. 3, pp. 68–75, Dec. 2002.

[115] S. Miyauchi, A. Hirano, N. Babaguchi, and T. Kitahashi, "Collaborative multimedia analysis for detecting semantical events from broadcasted sports video," in *Proc. of ICPR'02*, Tokyo, Japan, Aug. 2002, pp. 1009–1012.

[116] X.-F. Tong, H.-Q. Lu, and Q.-S. Liu, "A three-layer event detection framework and its application in soccer video," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 1551–1554.

[117] Q. Tang, I. Koprinska, and J. S. Jin, "Content-adaptive transmission of reconstructed soccer goal events over low bandwidth networks," in *Proc. of ACM Multimedia'05*, Singapore, Oct. 2005, pp. 271–274.

[118] V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: Towards a complete solution," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 1551–1554.

[119] W.-N. Lie and S.-H. Shia, "Combining caption and visual features for semantic event classification of baseball video," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 1254–1257.

[120] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. of ACM Multimedia'02*, Juan Les Pins, France, Nov. 2002, pp. 315–318.

[121] W.-N. Lie, T.-C. Lin, and S.-H. Hsia, "Motion-based event detection and semantic classification for baseball sport videos," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 1567–1570.

[122] W.-T. Chu and J.-L. Wu, "Integration of rule-based and model-based decision methods for baseball event detection," in *Proc. of IEEE ICME'05*, Amsterdam, The Netherlands, Jul. 2005, pp. 137–140.

[123] G. L. Foresti, C. Micheloni, and L. Snidaro, "Event classification for automatic visual-based surveillance of parking lots," in *Proc. of IEEE ICPR'04*, Cambridge, UK, Aug. 2004, pp. 314–317.

[124] G. L. Foresti, L. Marcenaro, and C. S.Regazzoni, "Automatic detection and indexing of video event shots for surveillance applications," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 459–471, 2002.

[125] A. Yoneyama, C. H. Yeh, and C. C. J. Kuo, "Robust traffic event extraction via content understanding for highway surveillance system," in *Proc. of IEEE ICME'04*, Taiwan, Nov. 2004, pp. 1679–1682.

[126] T. Nishida, S. Kamijo, and K. Ikeuchi, "Automated system of acquiring and visualizing track event statistics from track images," in *Proc. of IEEE ICME'01*, Tokyo, Japan, Aug. 2001, pp. 169–172.

[127] M. S. Saad M. Khan, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proc. of ECCV'06*, Graz, Austria, May 2006, pp. 133–146.

[128] A. Mustafa and I. Sethi, "Detecting retail events using moving edges," in *Proc. of AVSS 2005*, 2005, pp. 626– 631.

[129] S. Park and J. K. Aggarwal, "Event semantics in two-person interactions," in *Proc. of IEEE ICPR'04*, Taiwan, Jun. 2004, pp. 227–230.

[130] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. of IEEE CVPR'04*, Washington, D.C., USA, Jun. 2004, pp. 819–826.

[131] H. Zhou and D. Kimber, "Unusual event detection via multi-camera video mining," in *Proc. of IEEE ICVR'04*, Cambridge, UK, Aug. 2004, pp. 1161–1166.

[132] S. N. Sinha and M. Pollefeys, "Synchronization and calibration of a camera network for 3D event reconstruction from live video," in *Proc. of IEEE CVPR'05*, San Diego, USA, Jun. 2005, p. 1196.

[133] A. Behera, D. Lalanne, and R. Ingold, "Looking at projected documents: event detection & document identification," in *Proc. of IEEE ICME'04*, Taiwan, Jun. 2004, pp. 2127–2130.

[134] S. Reiter and G. Rigoll, "Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming," in *Proc. of IEEE ICPR'04*, Cambridge, UK, Aug. 2004, pp. 434–437.

[135] P. Smith, N. da Vitoria Lobo, and M. Shah, "Temporalboost for event recognition," in *Proc. of IEEE ICCV'05*, San Diego, CA, USA, Jun. 2002, pp. 733–740.

[136] H.-B. Kang, "Analysis of scene context related with emotional events," in *Proc. of ACM Multimedia'02*, Juan Les Pins, France, Nov. 2002, pp. 311–314.

[137] Q. Ye, Q. Huang, W. Gao, and S. Jiang, "Exciting event detection in broadcast soccer video with mid-level description and incremental learning," in *Proc. of ACM Multimedia'05*, Singapore, Nov. 2005, pp. 455–458.