UNIVERSITY
of
GLASGOW

Audio-Visual Football Video Analysis

FROM STRUCTURE DETECTION TO ATTENTION ANALYSIS

Reede Ren

September 30^{th}, 2007

For PhD Degree

IR Group, Computing Science Dept.
in Faculty of Mathematics and Information Science
University of Glasgow

Abstract

Sport video is an important video genre. Content-based sports video analysis attracts great interest from both industry and academic fields. A sports video is characterised by repetitive temporal structures, relatively plain contents, and strong spatio-temporal variations, such as quick camera switches and swift local motions. It is necessary to develop specific techniques for content-based sports video analysis to utilise these characteristics.

For an efficient and effective sports video analysis system, there are three fundamental questions: (1) what are key stories for sports videos; (2) what incurs viewer's interest; and (3) how to identify game highlights. This thesis is developed around these questions. We approached these questions from two different perspectives and in turn three research contributions are presented, namely, replay detection, *attack* temporal structure decomposition, and *attention*-based highlight identification.

Replay segments convey the most important contents in sports videos. It is an efficient approach to collect game highlights by detecting replay segments. However, replay is an artefact of editing, which improves with advances in video editing tools. The composition of replay is complex, which includes logo transitions, slow motions, viewpoint switches and normal speed video clips. Since logo transition clips are pervasive in game collections of FIFA World Cup 2002, FIFA World Cup 2006 and UEFA Championship 2006, we take logo transition detection as an effective replacement of replay detection. A two-pass system was developed, including a five-layer adaboost classifier and a logo template matching throughout an entire video. The five-layer adaboost utilises shot duration, average game pitch ratio, average motion, sequential colour histogram and shot frequency between two neighbouring logo transitions, to filter out logo transition candidates. Subsequently, a logo template is constructed and employed to find all transition logo sequences. The precision and recall of this system in replay detection is 100% in a five-game evaluation collection.

Abstract

An *attack* structure is a team competition for a score. Hence, this structure is a conceptually fundamental unit of a football video as well as other sports videos. We review the literature of content-based temporal structures, such as play-break structure, and develop a three-step system for automatic *attack* structure decomposition. Four content-based shot classes, namely, *play, focus, replay* and *break* were identified by low level visual features. A four-state hidden Markov model was trained to simulate transition processes among these shot classes. Since *attack* structures are the longest repetitive temporal unit in a sports video, a suffix tree is proposed to find the longest repetitive substring in the label sequence of shot class transitions. These occurrences of this substring are regarded as a kernel of an *attack* hidden Markov process. Therefore, the decomposition of *attack* structure becomes a boundary likelihood comparison between two Markov chains.

Highlights are what attract notice. *Attention* is a psychological measurement of "notice ". A brief survey of *attention* psychological background, *attention* estimation from vision and auditory, and multiple modality *attention* fusion is presented. We propose two attention models for sports video analysis, namely, the role-based attention model and the multiresolution autoregressive framework. The role-based attention model is based on the perception structure during watching video. This model removes reflection bias among modality salient signals and combines these signals by reflectors. The multiresolution autoregressive framework (MAR) treats salient signals as a group of smooth random processes, which follow a similar trend but are filled with noise. This framework tries to estimate a noise-less signal from these coarse noisy observations by a multiple resolution analysis. Related algorithms are developed, such as event segmentation on a MAR tree and real time event detection. The experiment shows that these attention-based approach can find goal events at a high precision. Moreover, results of MAR-based highlight detection on the final game of FIFA 2002 and 2006 are highly similar to professionally labelled highlights by BBC and FIFA.

Acknowledgements

To make a PhD is a big challenge for me, not only from research, but also from studying abroad. I am very grateful to my supervisor group and colleagues in Glasgow, who make the past four years rewarding and enjoyable. This work is also funded by faculty scholarship of FIMS and Overseas Research Students Award Scheme (ORS). Without these supports, this thesis would be impossible.

First and foremost, I thank my direct supervisor, Dr. Joemon Jose. He has been a source of excellent comments, and also provides me with sufficient intelligent freedom to pursue interesting research directions. His constant encouragements keep me forward and finally leads to this thesis. I also want to express my gratitude to Prof. Keith van Rijsbergen and other members in information retrieval group. Prof. Keith helps me to ensure the funding and his sharp insights on the demonstrability of ideas have immensely influenced my research. Many thanks to Jana Urban and Iraklis A. Klampanos for their insightful comments and discussions over years; Mark Baillie, Frank Hopfgarnter, and Martin Harvey for their kindness and help in the development of my system.

Finally, I would thank my parents and my wife, Chen Yu. It is who support me passing this special period of my life. This thesis is my best gift I could contribute to my family.

Table of Contents

List of Figures

List of Tables

1 Introduction

> "The real merit of a visual information retrieval system is its ability to allow enough extensibility and flexibility that it can be tuned to any user application."
>
> Gupta and Jain (Communications of ACM, May 1997)

This thesis is dedicated to content mining in sports videos, especially football videos. Three closely associated research topics are addressed: content-based video structure decomposition, video event detection and sports highlight identification. This work leads to an automatic indexing and retrieval framework for sports videos.

Two related techniques are developed and presented to acess video semantics, *attack* scene segmentation and psychobiological *attention*-based highlight identification. An *attack* structure refers to an attack-defence round of team movements. The *attack* structure decomposing tool simulates the variation in video production skills, e.g. close-up and field view, and thus divides a long continuous sport video into a sequence of video segments, each of which is semantically independent and carries clear and complete semantics according to the context of sports videos. Extra information resources, such as caption texts, audio and game records, are employed to index *attack* segments and result in an efficient video skimming system. The psychological approach of *attention* estima-

tion introduces psychobiological findings on the perception process of watching videos into the analysis of video contents. As the most interesting moments in a game, sports highlights always attract great attention from spectators and video viewers. Hence, local maxima of *attention* signals denote interesting parts of a game and thereby hint the appearance of highlights. This indicates that a temporal sequence analysis of *attention*-related features is able to identify domain-dependent highlights without concerning too many content details. Therefore, it is an effective and efficient approach to detect highlights by tracking viewer's feeling towards video contents. Additionally, two multiple modality *attention* fusion frameworks are developed in this thesis: the multiresolution autoregressive model (MAR) and the role-based attention model, both of which combine feature stimuli from different modalities, i.e. audio track and visual frames, multiple temporal resolutions and variant data updating rates to estimate a unified *attention* signal.

1.1 Motivation

Sports videos are enjoyed by a huge audience across the world. A recent three generation mobile service survey [3G-News, 2005] reports that sports videos, especially live football videos, are one of the most popular video genres in video-on-demand (VOD) services. This popularity stresses the financial interest around content-based sports video analysis, including content-based segmentation, context extraction, highlight identification, video indexing, content retrieval and adaptive encoding. Video content suppliers, such as BBC Sports, regard these techniques as value-added services to satisfy customers as well as a profit-making approach. Moreover, with the popularity of personal data processing devices, e.g. personal data assistants, 3G mobile phones and personal media centres, the service of personalised video-on-demand has become a prevalent interest of the media industry. Thus arises the requirement for intelligent video content processing. It is important to develop an intelligent information agent [Russell and Norvig, 2002] on sports videos, which not only cuts the cost of media storage, but also reorganises video contents to support efficient personalised browsing and video retrieval.

A convenient characteristic of sports videos is relatively simple content structure. Generally, the goal of content-based video analysis is to model video semantics by simulating syntax variation. Two aspects are involved, the selection of syntax set (Chapter 3) and the choice of semantic video model. Both aspects are closely associated with video context and video presentation formats. In general video data, content-based

video analysis is a complex task. This is because it is unrealistic to develop a generic syntax set. As the theory of knowledge presentation asserts, such a development requires an entire knowledge base of video contents. However, these semantics related problems can be partially alleviated by focusing on some certain video genres, such as sports videos in this thesis. This concentration make it possible to develop an automatic or semi-automatic content-based analysis system. The reasons are obvious as follows. Firstly, the syntax space of sports videos is usually finite. Game organisers, e.g. FIFA and UEFA, have developed a complete terminology and regulation set for the competition, each item of which refers to a video syntax or a video content. Secondly, the context of sports videos is self-evident. For example, Ekin et al. [2003] proposed a set of heuristic rules to guess video context. Burke and Shook [1996] defined a visual language to guide the composition of sports videos, although this production process was titled as a visual art and full of individual preferences. This means that a set of projection rules exist from low level audio-visual features to video semantics. Furthermore, external knowledge is available to facilitate the analysis of video contents, such as FIFA game records and web broadcasting. Some systematic knowledge networks, such as football ontology, have already been developed [Buitelaar and Ramaka, 2005] by game fact statistics and domain knowledge. These knowledge networks describe inherent links among syntax features and sports video semantics.

Content-based sports video analysis is a prototype of general video processing. This work can facilitate the development of many essential techniques, such as video concept abstraction, semantics presentation, video indexing, content-based segmentation, document normalisation and retrieval model. Moreover, a video is a complex style of knowledge presentation, which consists of multiple modalities, e.g. audio, visual and text streams, to iterate a story or an idea. Content analysis provides an opportunity to observe the various organisation of synthetic knowledge. For example, the interaction between video syntax items will enrich the understanding of the joint perception process, in which stimuli from vision, auditory and text understanding are mixed to build a unified awareness. Nevertheless, content-based video analysis is a complete circle of observation, reasoning and conclusion, one of the most important research topics in artificial intelligence. In this case, the extraction of low level features is an observation; the selection on modality syntax is a knowledge-based abstraction; the combination and balance among modalities and syntax features is an adaptive knowledge weighting; and the assumption of video content, context and semantics is a reasoning based on limited knowledge. These experiences from content-based sports video analysis, especially in the simulation of video semantics, will improve the knowledge of human intelligence.

Many content-based applications have been accepted as value-added services in the video-on-demand network, such as personalised video browsers, automatic reply inserting, statistical message highlighting and content-based adaptive encoding. A personalised video browser provides a swift skimming to review a game story, which filters out uninteresting moments and reorganises video segments according to user behaviour and video contents [Ekin et al., 2003] [Xie et al., 2004]. This technique provides a swift means of browsing game events. Automatic reply segment inserting is an active audit tool [Wang and Cheong, 2006]. This tool identifies sports highlights automatically and replays highlights from multiple viewpoints to improve viewer enjoyment. The application of online information hinting appends game statistics and background information, such as a comparison on the number of corner kicks by both teams, to explain game context. Adaptive video encoding is a new trend in commercial video encoding standards, such as H.264svc. This technique discards data packets with trivial contents to avoid meaningless rebroadcasting on a busy network and decides a proper moment to recover a broadcasting after a network jam [Gasiba et al., 2006]. A widely accepted solution is to find the most interesting moment during the missing temporal interval at the server side and provide a short online video skimming as a brief of lost video contents. Note that all these applications require: (1) the identification and segmentation of game events; and (2) the comparison of event contents .

Highlights are a special subset of game events. However, highlight identification is of equivalent importance to event detection in sports video analysis. In the Merriam-Webster dictionary, highlights are defined as the most interesting events with significant video contents, which suffice the iteration of a game [Pan et al., 2001] [Ren and Jose, 2006]. An important task of sports video directors is to develop an exciting presentation of highlights [Burke and Shook, 1996]. Therefore, the identification of highlights is an efficient alternative to event detection, which not only reduces the size of an event set, but also provides a precise game summary [Xu et al., 2001][Pan et al., 2001][Pan et al., 2002]. For instance, BBC sports develops a football goal collection to conclude a game season.

1.2 Problem Addressed

Figure 1.1 displays a pipeline of event-based content mining in sports videos, although a component of audio processing is absent. This framework helps the identification of research questions in content-based sports video analysis.

Three modality streams are extracted from a video collection, namely, visual, au-

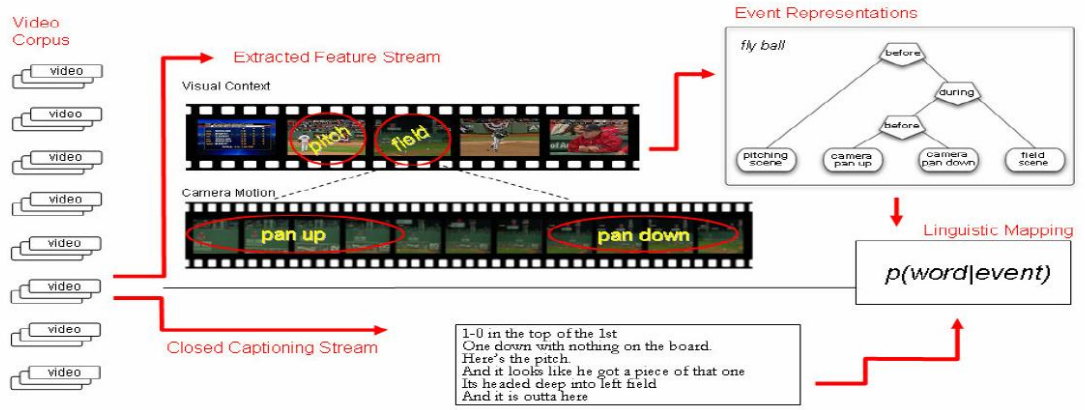


Figure 1.1: The Pipeline of Video Content Mining [Fleischman and Roy, 2007]

dio and linguistic streams. The visual stream is a sequence of visual frames; the audio stream denotes embedded audio tracks; and the linguistic stream refers to text information in a video, such as closed caption, title text, automatic speech recognition (ASR) and sometimes web broadcasting. Besides the step of feature extraction, there are three major components in a content-based analysis system: multimodality fusion, event representation (event pattern recognition) and linguistic mapping (semantic label propagation). Multimodality fusion combines domain features and produces a time sequence of content-related labels or a syntax stream. Event pattern recognition analyses the syntax stream to determine a sequential pattern of video events according to a predefined event set. Meanwhile, domain knowledge and external constraints, e.g. background music [Hua et al., 2004] and video duration, are used to ensure the justication of event patterns. The linguistic mapping projects event patterns onto video semantics or high level concepts. However, this projection usually relies on prior knowledge networks, such as a language model and word net [Fleischman and Roy, 2007]. From many aspects, a mining process of content-based video events is an autonomic understanding of game semantics from independent observations of audio, visual and linguistic streams.

The theory of artificial intelligence claims that an event-based content mining system is an information agent. Such a system enables: (1) to decompose semantic units in the target data set; (2) to decide the importance of semantic units; and (3) to adjust the performance with user requirements, if necessary. The prior two tasks require an information agent to understand video semantics and have to deal with semantic uncertainty in event detection. Note that semantics uncertainty here is more complex than seman-

tic gap in the inference from low level features to semantic concepts. Semantics is *the meanings, or the relation of meanings in a sequence* (Merriam-Webster Collegiate Dictionary, 10^{th} version). This linguistic description hints that the semantic identification of a video event is determined by video context as well as the content. Hence, event detection should take prior game stories into consideration together with the video segment it owns. Moreover, the semantic explanation of events may not be unique. For instance, goal events in football videos are furthermore categorised into penalty, free kick, dog fight and so on. This classification is contingent on the context of events, such as the position where the shoot takes place, e.g. the penalty. Nevertheless, a video event is a temporal interval of a continuous video stream, but sometimes lacks a clear boundary. Many overlapping but of different periods can be easily found to meet given semantics. For example, a goal event is made by several components, such as passing the middle line, breaking through, goal and celebration. A video clip to describe this story can be made up with goal, break-through-goal, or goal-celebration segments, all of which can be accepted to iterate such a semantic event. Therefore, some duration-related issues become essential in content-based sports videos analysis: (1) how to measure the fitness between given contents and video segments; and (2) how to find a good presentation approach and event description pattern.

Middle level content concepts or syntax features are necessary to describe video contents. These syntax features not only facilitate the inference from low-level features onto video semantics, but also define the context for semantic reasoning and content-based video importance weighting. However, the selection of syntax features remains a main problem in sports ontology, because the development of such a syntax set requires the support of systematic knowledge bases and should cover all possible contexts, e.g. the LSCOM system for news videos [Kennedy and Hauptmann, 2006]. A large amount of manual efforts are demanded.

Highlights are interesting game events. The identification of highlights is to rank event *interest* and thus select a subset of events with the strongest *interest*. The *interest* is a perceptual phenomenon relevant to personal feeling and cultural background [Osgood et al., 1957]. The decision whether an event is interesting or not, depends on the viewer and the presentation style of video contents. An event may be interesting for a specific group of viewers but not for others. For example, a home team is always cheered, since most spectators are local supports. The editing skills of replay can increase the story affection [Rui et al., 2000]. This uncertainty is known as psychological ambiguity. From the perspective of computational psychology , the *interest* is a joint reflection

against stimuli from audio, visual and text understanding. These modality stimuli are: (1) temporally correlated, and (2) conceptually independent, but take effects jointly. Nevertheless, the computational nature of these stimuli is complex. Audio signals are continuous and of real value; visual features, e.g. the zoom depth, are discrete both in time and value; the measurement of text understanding employs qualitative analysis, whose result is usually a binary vector. How to combine these signals or feature sequences to assume an unified stimulus intensity remains an open question in computing psychology and signal processing (Chapter 6). Additionally, the versatile nature of psychology determines that highlights are a sub set of events but can hardly be fully covered by any given event sets. Some game events can be highlighted with a high probability, but the probability will never reach one.

In the following sections, three major challenges within the field of sports event and highlight detection are discussed, namely event pattern mining, multiple modality fusion and semantic labelling. The idea behind our solutions and their rationales will be presented together.

1.2.1 Event Pattern Mining

A game event is an array of visual frames, a segment of audio and related text information. According to the theory of pattern recognition, an event is a time sequence pattern with mixed modality data. The detection of events is characterised with: (1) a multi-modality data set; (2) a temporal sequence mining in a long multimedia stream; (3) strong data dependence. Briefly, this problem of event pattern mining can be broken down into two sub issues,

1. Raw data representation, which can be formulated as the extraction of modality features and the choice of a subset from a pool of features. Both of them are significant for a successful recognition process due to the sheer volume, high redundancy and low signal noise ratio (SNR) in raw media data, e.g. audio. Note that the cancellation of noise is crucial in the development of a robust application. In Chapter 3, the feature collection and related extraction algorithms will be presented for sports event detection and highlight identification. Additionally, our purpose of feature selection is to find a reliable feature set rather than the reduction on feature redundancy, although the later is more popular and has been well developed in the theory of Bayesian information criterion (BIC).

2. Event pattern learning, which develops a method to extract a model of video events from proposed feature spaces, the representation of raw data. Generally,

an event is a group of video clips with similar contents, such as a goal in a football game. These events are sparsely distributed throughout the video, where plain and usually trivial moments occupy most of storage space [Babaguchi et al., 2002]. However, the similarity in video contents can hardly be detected directly in low level audio-visual features. The learning process is characterised with the following,

(a) The data sample consists of multiple modality data, which are highly correlated with time and usually of different resolutions;

(b) The pattern is a time sequence. However, sequence length varies with context. The learning process has to estimate model duration individually;

(c) Events are sparsely distributed among the whole time sequence. The frequency of pattern recurrence is low. Therefore, it is inefficient to mine an event pattern by recurrence-based statistics;

(d) Such a pattern may not be unique and usually adaptive to data.

In short, the event model learning is an extraction of time sequence patterns without a prior of model length. It challenges methods of time sequence analysis due to its low recurrence [Banerjee et al., 2004].

To ease the identification of highlights, salient features are employed to estimate the stimulus from audio and visual frames. Psychobiological feature-attention models have been developed to compute and normalise the affection of salient features on perceptions. The survey of salient features will be found in Chapter 3. We regard the selection of salient features as a problem of model learning in a given feature basket. Two feature selection algorithms are presented, multi-resolution autoregressive model (MAR) in Chapter 6 and suffix tree in Chapter 5. The MAR model is a multiresolution variation of dynamic Bayesian network (DBN). Due to the multiresolution character of media modality, the MAR model decomposes the salient feature sequence onto a series of autoregressive models (AR) at multiple scales. This model has been proven to be equivalent to a Markov process on a graph [Willsky, 2002]. Each feature's contribution to the overall information entropy is computed; and coherent features are grouped to remove random noise. The MAR model deals with the dependency between adjacent audio-visual observations from multiple resolutions by simulating the temporal relation between events. Adapting to different descriptive complexities of feature sets, the model size is automatically decided by maximising the cost function of overall entropy over feature set size. The approach of suffix tree is proposed for the recurrent nature of

video events. After symbolising features into a sequence of video editing effects, a suffix tree identifies the recurrent part in the time sequence and extends them by a Markov model. Therefore, this approach is practically useful for features with large varieties.

1.2.2 Multiple Modality Fusion

Watching a video is an enjoyment of perceptual stimulus from multiple sensors, including sight, hearing and linguistic understanding. Salient recurrent patterns in videos, such as game events, are temporal segments combining visual, audio and text information. The modelling of these experiences should work on multi-modalities. Lack of either media modality may entrap the presentation of video content. For example, though anchor shots are salient visual patterns in news videos, these similar visual segments convey entirely different stories with different audio and caption texts. Hence, event detection is a joint decision process with multiple information modalities. However, besides knowledge scarcity on multi-modality perception, the fusion mechanism across multiple modalities remains a great challenge in content-based video analysis for the following reasons.

1. Media reliability. Besides the visual stream, the availability of a media modality relies on the interest of content suppliers. A game video can go without audio, caption text or any other media modality. Moreover, there are inherent restrictions in the appearance of modalities, such as the geographic location of closed captions [Babaguchi et al., 2002], the automatic gain controlling (AGC) on audio loudness [Baillie and Jose, 2003] and the copyright of game records. One essential characteristics of commercial sports video analysis systems is to be adaptive to the change in the availability of media resources.

2. External knowledge discovery. External domain knowledge is helpful in the content identification and labelling, such as game records [Xu and Chua, 2004] and web-casting text [Xu et al., 2006]. For example, as a new phenomenon in internet, web-casting broadcasts the sequence of game events by text messages without visual data. It is an abstract of content description with time stamps, which increases the precision of event detection and helps the identification of highlights. However, the discovery of these information sources remains a question.

3. Event resolution in time and content description. The presentation period of events varies with media and their contents. The visual stream updates at thousands of bytes per second, which depicts rich details of the play field, coach area and stadium. Meanwhile, the voice of commentators is relatively slow, but brings

a brief description. A game event, e.g. a goal in the football game, may extend into several minutes of a video stream. But game commentators or spectators use only a few words to state the story. With the introduction of video editing effects, especially replay, an event can have multiple appearances in the visual stream. For instance, the goal event is usually replayed several times to convey different views, each of which is a whole appearance or description of the content. This raises the question of how to detect these various appearances so as to locate boundaries of an event. The joint segmentation of the visual stream and other modalities is necessary to match contents across media.

4. Media asynchronism. To save bit rate, modality data are encoded and decoded independently. For instance, audio in encoded MPEG-1 PAL format videos is sampled at 44kHz/16bit (MPEG-1 audio layer 2), while the visual stream is at 25 frames per second. Random delay exists among media streams. Most commercial media encoding standards, such as MPEG-1/2/4 and H.263, follow the psychobiological assumption that perception residency abides 1-2 sec audio-visual asynchrony. But such an allowance can hardly cover the temporal latency in the process of composition and broadcasting. In our test collection, the misalignment between audio and visual stream can even reach 11 seconds [Ren and Jose, 2006]. Another example of modality asynchronism arises from the resolution difference of modality events. For example, the FIFA official documents record game events in minutes, while we detect game events at the resolution of visual frames.

The above mismatch of resolution, data sampling and media alignment hint that the multimedia fusion has to be carried out on a coarser temporal resolution than visual frame or shot level. As far as we are concerned, multimedia modalities, i.e. audio and visual stream, are only synchronous on semantic events.

1.2.3 Semantics Labelling of Statistics Patterns

Understanding and symbolising video contents is the "holy grail of content-based video analysis". A multimedia pattern is meaningless unless it can be interpreted by the underlying semantics. A semantic label serves as an abstraction of video contents, which links the proposed pattern to a particular domain knowledge and reflects the characteristics of this pattern. For instance, a small red square in the closed caption denotes a red-card punishment. It is desirable to build computational models, which automatically establish semantic interpretations of syntactic patterns.

In Chapter 5, we develop a framework of automatic association between game events and text labels gathered from caption texts, commentator audio and external knowledge resources, e.g. FIFA game records. This interpreting model is made through the transcription of texts from metadata, because media information is complementary and texts are easy to understand. Some statistical models are developed to estimate the probability of associating certain text words with given audio-visual patterns (and vice versa) through co-occurrence analysis [Satoh et al., 1999] and statistical machine translation.

Besides these challenges, the limited size of available data collections is a practical problem in event model learning. Although the collection of FIFA World Cup 2002 and 2006 used in this thesis is the largest in the literature of sports video analysis to the best of our knowledge 1-0The game video collection of FIFA World Cup 2002 and 2006 contains more than 90 full games. The whole size of the collection is over 350 GB or 190 hours., this data collection is still not enough to fully simulate video event patterns.

1. Sports events are sparsely distributed in the video. The event number is usually small in the whole collection. For instance, the number of goals in 90 games is about 190. The sample size is too small to estimate a sequential video event pattern robustly. Moreover, the pattern length is a parameter in the detection, which is unknown.

2. Video composition is an art [Burke and Shook, 1996]. There are numerous video editing styles and content representation methods.

It is essential to maintain the generality of trained detectors in event pattern learning. There is a trade-off between the size of the syntax set and model generality, because these middle level concepts are close observations of raw data. A pattern with many syntax features will fit the training data well, but its adaptability to other data is questionable. Note that syntax is an evidence supporting semantics reasoning. The redundancy in syntax set may lead to a conflict, which is harmful in the reasoning. Nevertheless, the value of sports video decreases dynamically with time [Ekin et al., 2003]. The processing time is crucial in most applications, such as online summarisation and information hinting. Real time processing on event detectors is thus required.

1.3 Research Summary

The main objective of this research is: (1) to develop a set of robust and efficient techniques for football event detection, in particulate highlight identification; and (2) to

build a computational framework for automatic content-based football video analysis in order to support personalised video-on-demand services. The main objectives can be broken down into five goals: (1) a flexible and extensible model for content-based video decomposition; (2) an efficient syntax set for content presentation; (3) a computational framework for estimating the *"interest"*; (4) a reliable content model; (5) a robust algorithm for highlight identification and segmentation, which can support commercial applications with the requirement of real time processing. These works can be extended to other sports games, such as rugy, volleyball and basketball, although they are only widely tested in football videos because of data availability.

Attack is a complete competition cycle in the football game. It is a scene structure in the domain of football videos. An *attack* segmentation system has been developed by simulating the process of video production. Four types of shots are discriminated, namely play, break, replay and focus. A group of domain features are extracted automatically for shot type classification, including play field ratio, zoom depth, spectator area, and visual mean contrast. A set of video objects, such as player uniform, human face and goal post, are detected by the Foley-Sammon transform (FST) classifier (Appendix A) and Ada-boost classifier [Polikar, 2006*a*] [Polikar, 2006*b*]. The *attack* structure is labelled according to events, which take place inside the structure, such as a goal or successive attack. Based on the *attack* structure segmentation, a video index is created and a video skimming system is developed [Ren and Jose, 2005].

The literature of salient feature extraction is surveyed. A set of modality features is proposed in the context of football videos to compute affection stimuli in the psychological *attention* space. Two computation models, the role-based model and the multiresolution autoregressive model (MAR), are developed and evaluated in the application of goal event detection. The role-based fusion model analyses the reflection structure in the video production and combines affective signals from one reflector to remove reflection bias. The MAR model regards modality stimuli as sequential noisy observations of a smooth stochastic process. This model employes an autoregressive tree to simulate a random process from multiple resolutions. The peformance of MAR model is impressive in the experiment. The result is similar to some professional annotations from content suppliers, i.e. BBC Sports and FIFA. Such a multiresolution model is robust against signal noise and media asynchronism. Many applications are developed, such as attention-based video summarisation [Ren, P.Punitha, Urban and Jose, 2007]. An automatic football highlight detector has been developed by the MAR model [Ren, Jose and He, 2007] and leads to a hardware solution for real time applications [Yin and Ren,

2007].

1.4 Contribution and Limitation

This section pinpoints original contributions made by this thesis. However, a few limitations on these techniques are listed as well.

Two approaches are presented to address the problem of temporal structure mining, *attack* video structure decomposition (Chapter 5), and *attention* based highlight identification (Chapter 6). *Attack* structure is a cycle of team competition for a goal in football games. Four production techniques, i.e. field view (F), close-up (C), replay (R) and break (B), are discriminated; and a four-state hidden Markov model is trained to simulate the transition between these production techniques during an *attack*. A suffix tree is enhanced to pick out the most recurrent part, which is called as *attack* kernels, in the temporal label sequence of video production techniques, *e.g.,*"FFCFFCRCFFB ...". These *attack* kernels are extended by the four-state hidden Markov model and thus decompose a long continuous video stream into a series of *attack* structures. Structure boundaries are decided by computing the likelihood of a boundary belonging to two Markov process. This approach alleviates the difficulty of model adaptation on a long time sequence and transforms the problem of video structure segmentation into boundary identification between two neighbourhood Markov processes.

Attention-based highlight identification is an exploration of computational psychology to content-based video analysis. This approach introduces many psychological observations to measure the importance of video contents. A unified *attention* curve is computed to combine stimuli from different modalities and time resolutions. Local minima of the *attention* curve is used to divides a game video into so-called perceptually complete stories, while local maxima hint the appearance of game highlights. A significance of *attention*-based highlight identification is a high coverage rate of goal events in the top five *attention* peaks, even though no goal-related syntax, i.e. goal post and ball, is used. Moreover, this approach is able to find general highlights and the output is similar to these professionally labelled in the experiment.

Two modality fusion algorithms are designed to combine multimodality stimuli, the role-base model and the multiresolution autoregressive model. Based on the perception structure during watching and producing a sports video, the role-based model groups feature-based *attention* sequences to eliminate the latency between reaction roles. This

model alleviates the problem of modality asynchronism and reduces the number of *attention* signals to facilitate signal fusion. The multiresolution autoregressive model regards a sports video as a multi-source observation on a smooth stochastic process. This is because: (1) the content space of a sports video is limited, for example, the attack structure segmentation only employs four states or labels; (2) all contents are indexed by time; (3) the content change is slow and the state transmission only relies on the prior state. Therefore, *attention* signals from visual, audio and text streams are observations of the random process from different time resolutions. The MAR model proposes a multiresolution framework to adapt these temporal resolution difference and designs a series of AR process to facilitate the fusion of these asynchronous observations. An autoregressive tree is developed to smooth *attention* signals and estimate a unified *attention* curve. Moreover, an entropy normalisation is proposed as an alternative to general signal normalisations. This replacement not only enlarges signal-noise ratio (SNR) of *attention* signals, but also introduces an information theory explanation for *attention* estimation.

In summary, the main original contributions of this thesis are as follows.

1. An efficient and effective replay detection system;
2. A set of robust syntax feature extractors;
3. A suffix-tree for *attack* kernel identification and a hidden Markov basket for *attack* video structure decomposition;
4. An unsupervised multiresolution model for general highlight identification;
5. A fusion framework for asynchronous multiple modalities streams;
6. A self-entropy *attention* estimator.

Additionally, most techniques developed in this thesis can be employed to process other sports games besides football, such as rugby. For example, many syntax detectors, *e.g.,*game pitch and zoom depth works for all field games with unified field colour and play uniforms. The hypothesis of attention-based highlight identification is plausible in most sports activities: game highlights always excites viewers. In an experimental test, this technique can work on a volleyball video, although some salient features have to be changed, such as play field ratio. However, given the scarcity of game video collection, *attention*-based have not been intensively tested on other sports videos.

The most apparent limitation of these technique is the dependence on editing. *Attack* structure decomposition is based on the transmission of production techniques and salient features reflect viewer's feeling on given images and sounds. Lack of editing effects will make *attack* segmentation and *attention*-based approaches inefficient.

1.5 Thesis Organisation

After the introduction chapter, this thesis proceeds as follows.

Chapter 2: Links with data mining and multiple modality fusion are briefly introduced as well as the literature of sports event detection. Particular attention is given to the domain of football videos.

Chapter 3: This chapter is dedicated to feature extraction in sports videos. A set of feature extraction algorithms are presented and evaluated in a large football video collection. I discuss the difference between content-based syntax features and affective salient features and show how to generate video semantics from these syntax features.

Chapter 4: Replay is a specific video editing strategy in sports videos. To emphasise the most important or interesting moment, video directors reiterate the same game content several times by slowing down the motion or from different viewpoints. The collection of replay segments is widely accepted as a sufficient game summary. Based on the distribution of shot duration and the colour difference in video editing effects, a five-lay adaboost classifier is developed for the detection of transition logos (Section 4.3) and thereby identifies replay segments. Additional pitfalls and possible improvements are discussed.

Chapter 5: The stochastic model for *"attack"* video structure decomposition is presented. The rationale for this temporal structure and possible applications in video skimming and summarisation are shown.

Chapter 6: A sports video is emotionally inspired. Therefore, it is an efficient approach to understand video contents by analysing emotion aspects. Related psychobiological measurements and computational psychological methods are surveyed; and feature-attention models are listed. Two fusion models are developed to combine attention features into a unified signal curve, thus lead to an efficient highlight identification system.

Chapter 7: Main contributions are concluded and briefly reviewed. Possible future research directions, such as attention graph and syntax frequency, are discussed as a result of the investigations on this thesis.

2

Related Work

As one of the most popular video genres, sport videos became interest to researchers in the field of content-based video analysis with the blooming of digital video service in the late 1990s, as well as news videos and story films. Three major research questions are identified in content-based sport video analysis, namely event detection, semantic annotation and highlight identification. With these techniques, sport videos are decomposed into semantically meaningful segments; content sequences are simulated and segmented by event patterns, which lead to an automatic understanding of game stories; content-based indexes are created for efficient video retrieval and browsing. Many applications have already been proposed in the media industry to improve the quality of services (QOS) during broadcasting and video distribution. These applications include personalised sport video summarisation, content-based compression and video content recommendation. Moreover, personal data storage devices, such as PDA, 3G mobile phone, and PlayStation® portable (PSP), are widely used. These devices not only promote the sharing of sport videos, but also demand improvements on existing techniques of content-based video analysis, such as real time processing, online interaction, content comparison among multiple video streams. Therefore, content-based sport video analysis is beyond the scope of semantic labelling, and will take more practical issues into consideration, e.g. perceptual weighting of video contents.

Event detection plays a crucial role in content-based sport video analysis. This technique is a clustering of video shots under domain knowledge or a classification of video segments based on content similarity [Duan et al., 2003]. Two sub topics are involved: (1) how to identify event contents; and (2) how to locate event boundaries as well as extract the story from a video stream. Both of these topics are equally important in applications, because it is necessary to know the semantics of a game event as well as the event duration. For example, an individual needs to know both content and duration of an event to label an index item. According to the hierarchy of video structures, a game event is the scene2-0Scene is a temporal video structure, which conveys integral semantics. It usually consists of several video shots to present a complete story in the video. in sport videos. Some algorithms for scene segmentation in general video genres were employed in early works, such as fusion with respect to the coherence in motion, colour and video objects [Rasheed and Shah, 2003]. However, these approaches can hardly identify the content of clustered shots. For instance, it is difficult to title a motion-coherent video segment with clear game semantics. The detection of sport events should base on game contents rather than comparing the similarity of low level audio-visual features.

Employing a definite sports semantics collection, such as the FIFA football regulation handbook, a series of special event detectors, e.g. goal and free kick detector, have been developed by searching and combining syntax features. These syntax features consists of a goal post, game pitch boundary, keywords from commentator speech, and player discrimination [Adams et al., 2000] [Assfalg et al., 2002] [Ekin et al., 2003]. For example, Ekin et al. [2003] proposed to identify goal events by testing whether two video objects, a football and a goalpost, are overlapping in a visual frame. Although this goal event detector is semantically clear, this method is impractical because some video objects, such as the football, are too small to be found. Furthermore, Yan [2006] developed a probabilistic model, which is trained by the joint appearance possibility of several syntax features, to combine syntax features from diverse information sources. Although this model is successful in multimedia retrieval, the approach is closely associated with the appearance of syntax features and is not so effective in sports video analysis for following reasons:

- The syntax set of sports video is finite. Although the small size of syntax set makes model training easy, the limited information gain results in a weak model.
- Most syntax features are legends in a game pitch, such as goal post and middle circle. The joint probability of these syntax appearance is high.

- The semantics of sports events is closely associated with the location where these events take place.
- This model relies on the meta-data collected from visual frames. This means that the model does not take the temporal issue into consideration. It is an approach of image retrieval rather than video retrieval.

With the help of domain knowledge, syntax concepts are combined to clarify the game story and to improve the precision of event detection. However, a syntax is the snapshot of a continuous video process without information of event duration. In other words, the approach of syntax combination cannot allocate the start and end of a game story, though it facilitates the identification of game event contents. For instance, the goal detector in [Ekin et al., 2003] missed many components of a goal event, such as break-through, shot and celebration.

Additionally, a few complex temporal models have been proposed to find coherent segments. Xie et al. [2004] designed a two-state hidden Markov model to divide football videos into *play* and *break* structures. Later, we identified *attack* structure. The segmentation algorithm will be presented in Chapter 5. However, these temporal structures cannot cover all possible game contents. A reasonable solution for content-based event detection should include three steps: (1) identify *attack* structures; (2) cluster shots in an *attack* with syntax features to search boundaries of event components, e.g. break-through; (3) link event components with prior event models, for example, a goal event should include four components, namely break-through, shot, goal, and celebration.

Semantic annotation links a video segment to its semantics. It labels video objects, actions and events to assist the recognition of video contents, such as a football and a goalpost (video objects) [Ekin et al., 2003], entering a room (action) [Courtney, 1997], and an explosion (event) [Naphade, 1998]. The methodology of automatic annotation can be roughly categorised into two groups, syntax detector and complement modality annotation. The MediaMill system [Snoek, Worring, van Gemert, Geusebroek, Koelma, Nguyen, de Rooij and Seinstra, 2005] relies on prior knowledge of a video collection and develops a large set of syntax detectors, such as sky, sand, car and bicycle detectors, to assume the video content. Although these syntax detectors are usually efficient in the discrimination of specific video objects, this approach is systematically inefficient because each kind of video objects requires a specific detector. Russell and Norvig [2002] claimed that it is impossible to build up a universal detector collection for general content identification. Hence, the approach of syntax detector has to focus on applications;

and the main challenge of syntax detector is to find an application specific detector set, which is sufficient to describe all possible contents efficiently. For instance, a vivid discussion took place in the competition of TrecVid 2006 about a suitable number of syntax descriptors for news video retrieval. Due to the difference in syntax reasoning networks and content-based knowledge space, the number of proposed descriptors varies from 64 (IBM proposal) to 15,000 (CWI proposal). This huge gap indicates the difficulty in developing syntax detector collection. Complementary modality annotation seeks external information resources to carry out automatic or semi-automatic content annotation. This approach regards an annotation process as the propagation of correlated labels or maximisation of posterior probability between given labels and syntax features. The Name-It system [Satoh et al., 1999] utilised text messages in videos, including Video OCR and caption text, to name human faces in visual frames. Baillie and Jose [2003] analysed official game records to count game events, which are semantically important. Wang and Cheong [2006] discovered a new text media, web broadcasting, which publishes online comments from spectators on a website. These authors searched key words and time stamps in comments to label video segments. The main challenge of complementary modality annotation is the availability of complementary media. For example, web casting is a free internet media without a grantee of availability.

Content-based video annotation faces many challenges. Although some complement modality approaches can produce high level event-based labels, such as a goal and free kick to annotate video segments, semantic annotation mostly interprets a continuous visual stream with simple syntax information [Satoh et al., 1999] [Ekin et al., 2003] [Snoek, Worring and Smeulders, 2005], i.e. visual objects. According to the theory of artificial intelligence, these two actions are equivalent, to describe a long video segment with a set of object-based annotations and to present short term or long term memory by sensory memory. The working memory model [Baddeley and Wilson, 2002] argues that a reasoning component is necessary to complete the interpretation from sensory memory to short term memory. This raises the problem of content modelling. Additionally, semantic annotation for a visual frame or a video segment is not unique because of semantic ambiguity. For example, an image with the view of sky can be labelled as blue sky, cloudy, clear, and birds if there is a bird occasionally passing. Although some semi-automatic labelling systems [Carbonaro and Ferrini, 2007] try to formalise keywords in order to avoid annotation ambiguity, a standard annotation set does not exist till now.

Sport highlights refer to interesting game events in a sport. Note that feeling interest is

a perceptual concept. This incurs psychological ambiguity. Three approaches for highlight detection are proposed, namely special event set, replay detection and perceptual attention estimation. Given the self-evident sport semantics, it is possible to predefine a set of game events, which would make viewers feel interesting at a high probability, e.g. goals in football videos. Therefore, highlight detection can be specified into a series of specific event identifications. Assfalg et al. [2002] searched for a large event set, including goal, kick-off, free-kick, and throw-in to list all possible football highlights. Replay detection is based on the visual language in sport video production. As a unique video editing phenomenon, replay is designed to carry semantically important game contents from the perspective of professional video directors. The discrimination of replay segments is an efficient approach to find interesting video contents. An effective replay detection system will be presented in Chapter 4. Psychological *attention* estimation is a relatively new pathway. It relies on a set of feature-stimulus models or modality-stimulus models to estimate the intensity of psychological stimulus from audio, visual and text understanding. The idea of *attention* estimation will be elaborated in Chapter 6.

2.1 Summary of Literature

A broad overview is presented around the research topic of event detection in sports videos. As the first step in content-based sports video analysis, event detection mines out recurrent and semantically meaningful video patterns. This technique is an application of data mining in the multimedia data stream and closely associated with semantic perception. Therefore, a brief introduction of data mining techniques is stated. Some data mining applications, e.g. motif analysis, are compared with sports event detection. Such a comparison not only suggests possible solutions for sports event detection, but also helps to identify research challenges among sports video data, including ambiguous measurements on multimedia data, variant pattern length, and uncertainty in semantics understanding.

Subsequently, I discuss the combination scheme among multiple modalities, such as audio and vision. Although it remains a research question in psychology and artificial intelligence, multimodality fusion is a promising approach for video semantics understanding. This technique is regarded as an intuitive and efficient approach to clarify video semantics and thus improve the precision of event detection. The complementary information from multiple modalities not only provides evidence to facilitate the discrimination of video contents, but also results in a content-based video decomposition.

A conceptual fusion framework is presented, which includes three steps, audio-visual affection, audio-visual concept extraction, and text message fusion. Two practical fusion methods, early fusion and later fusion, are discussed to identify approach advantages and disadvantages.

Finally, the literature of sports event detection is reviewed. These techniques can been categorised into four classes, namely rule-based analysis, video clique discrimination, sequence learning and affection analysis. Rule-based analysis is based on game regulations and close observations of the competition process; video clique discrimination regards the event detection as a sequential pattern discrimination; the approach of sequence learning employs methods of time sequence analysis; and affection analysis introduces psychological measurements on spectator reflection. Although proposed by totally different perspectives, these techniques face common challenges: (1) how to identify event content; (2) how to segment a event clique precisely. A few sub-questions arise in the combination of multimodality information: (1) whether the algorithm is sensitive to media asynchronism, (2) how to deal with multiresolution media data. These challenges define the test bed for the evaluation of sports event detection techniques.

The following sections are organised as follows. In Section 2.2, existing works in data mining are reviewed. Particular attention is paid to define the difference between multimedia pattern mining and other mining methods in conventional applications. Section 2.3 covers the literature of multiple modality fusion. Several machine perception principles and fusion models are presented, which will guide the practice of modality fusion in content-based sports video analysis. Section 2.4 surveys the state of art in sport event detection. A short conclusion will be found in Section 2.5.

2.2 Connection to data mining

Sport events refer to semantic repetitions in a game story. The detection of sport video events closely resembles the problem of data mining, which extracts frequent recurrence in various data collections. Generally, data mining is the discovery of association rules or patterns among "transactions" in a large data collection [Agrawal et al., 1993] [Brin, Motwani, and Silverstein, 1997]. These association rules are a specific case of data correlation and implications [Brin, Motwani, Ullman and Tsur, 1997] in statistical tests, such as dependency relationship, conditional probability and likelihood ratios. For example, the shopping-basket analysis [Agrawal et al., 1993] is a traditional data mining problem, which predicts the co-occurrence part of goods bought together with

a high confidence.

Data mining in a long sequence can be regarded as a generalised technique for sports event detection, because event detection aims to find a sequential correlation or syntax implication in the combined audio-visual time sequence. For example, biological sequence analysis is a case of event detection on DNA sequences. This technique is proposed to identify interesting and relatively conserved motifs, e.g. similar short DNA or protein segments, embedded in the "noisy" background of numerous long gene sequences. Gibbs motif sampler [Lawrence et al., 1993] [Liu et al., 1995] computed the conditional distribution among protein pairs and alleviated the discovery problem as the alignment of sub-sequences at some given positions. Then a multi-nominal motif model was extracted as a sequential recurrence pattern. The MEME (motif-based hidden Markov modeling of biological sequences) [Bailey and Elkan, 1995] reduced the computational cost of Gibbs motif sampler by developing a Markov model to simulate the sequential dependency between cliques. The advantage of a MEME sampler is to relax the strict assumption of independent and identical distribution (i.i.d.) between motifs. Although the raw observation on the DNA is symbolic and uni-dimensional, the sequential pattern of motif and the background HMM are probabilistic. This is similar to the process of *attack* structure detection (Chapter 5). The temporal pattern in network traffic analysis is another mining application, which deals with time sequences and is driven by continuous measurements. This technique sampled the process of network transmission periodically to estimate a network performance model. The feature set included peak, tail, behavior and period network transportation. Several continuous state space models, e.g. autoregressive moving average model (ARMA) [Iyengar et al., 1999], were employed to catch the structure of interest (SOI) in the noisy environment. This extraction process is the same as video content modelling in many aspects, such as time driving, modelling selection, observation noise reduction and temporal model learning, though the feature space in sport videos is multiresolution and contains multiple modalities.

The purpose of event detection or video pattern mining is to find recurrent and meaningful segments in the collection of video data. The problem can be cast into the association rule detection in the joint audio-visual space of syntax concepts, video motif alignment, or video trend identification. The modelling of video contents therefore can be a Gibbs motif sampler [Lawrence et al., 1993] without explicit background knowledge or a MEME model [Bailey and Elkan, 1995] with the support of multi-modality data. However, various domain differences exist,

- The uncertainty in media feature measurement. The data collection in data mining is “clean” and unambiguous. For example, the shopping basket mining and the motif allocation process symbolic signals. The amino acrid ”G” will not be confused with ”I” in a biological motif. Multimedia data is characterised with noisy and usually ambiguous feature measurements. The same colour from different pixels in an image is likely to convey different RGB values at a high probability, because of the noise in CMOS image sensors and the complexity of vision perception. In most cases, the qualification of continuous multimedia signals is closely associated with the context, e.g. the neighbour area of a pixel. The mislabelling of syntax is usual, because there exist numerous media feature values for a syntax. This results in high computational cost for a Gibbs sampling.

- The absence of known temporal scales or event duration. The length of data pattern, such as the scale of motif and the range of candidate sequences in baskets, is known in most data mining applications. This prior knowledge significantly reduces the computational complexity of motif detectors. However, event duration in sports videos is random; there exists no clear event boundaries in most cases. The temporal range of an event varies from several minutes to a few seconds, contingent on different event contexts. Moreover, the editing style of video directors is an important factor influencing event duration. For example, the insertion of replay segments may double the presentation period of a video event. This fact partially explains the strong data dependency of many techniques for sports event detection.

- The link with semantics. In data mining, the evaluation criterion for a target pattern emphasises statistical significance, for example, how often such a pattern appears, whilst pays little attention on the meaning of a pattern. For instance, the shop habit “coffee and cream” or the browsing pattern “a high peak at noon” are stable patterns. However, a statistically significant pattern without semantics is meaningless in event detection. This is a salient difference. For instance, the frequently repeatable field-view shots tend to be ignored because of their plain contents [Xie et al., 2002] [Ren and Jose, 2005]. An event pattern should convey clear game contents, although its meaning is affected by the selection of event sets, video data collections, viewers and applications.

- Data dependence on collections. Data pattern is usually constant across collections. For example, the gene sequence ”VGIIGAG” stands for a biological signal no matter which gene collection such a sequence comes from. To ensure pattern

robustness, such a mining process always employs several collections to check the statistical significance of a pattern. As I have mentioned in the early part of this section, an event is a content-based video segment, which is closely associated with the context, viewer perception and applications. There is a strong dependence of event patterns on the data collection for the following reasons. Firstly, video production is a creation activity of visual art [Burke and Shook, 1996]. The representation style of sports games is decided by personal understanding and emotional feeling about the game story, which will vary with different video directors, content suppliers and competition organisers. Secondly, the feature distribution strongly correlates with data collection. For example, the syntax feature, play field ratio, relies on the colour distribution of play field in games, which changes with location, weather and sun light. In short, event pattern is data dependent. Sports event detection should be based on the local statistics within a game video.

Although the detection of sport events is more complex than most data mining applications, the techniques of data mining will spark ideas and suggest possible solutions in many research aspects of sports event detection, e.g. feature selection and abstraction, syntax fusion, statistical pattern extraction and content modelling.

2.3 Multimodality Fusion

Liberally, the word *"modality"* is *"a category of sensory perception"* in the Oxford English Dictionary. Sport videos combine visual frames, audio segments and text messages to present their stories. The enjoyment of sports videos contains multiple perception on multiple modalities, such as vision, auditory and language understanding.

Content-based video analysis processes and fuses these modality information to access video contents. Multiple research fields, such as computer vision, artificial audition and natural language processing, are necessary front ends, which conduct to automatic understanding of video semantics. Apparently, psychological and physiological findings on the perception process are helpful. Multiple external information resources are available for sport video content analysis, such as official game records. These information resources can also be regarded as independent modalities.

2.3.1 Vision and Audition

The perception on light and sound is a complex process. Studies on human behavior show that vision and audition can influence each other and create a joint affection. In the McGurk experiment [McGurk and MacDonald, 1976] on audio-visual hearing, the overlay of an audible syllable ("ba") onto the video tape of a speaker mouthing a different syllable ("ga") would result in the perception of "da". Another obvious proof comes from brain imaging, which allocates function divisions in the brain by measuring metabolic or electromagnetic reflective signals for a stimulus. This image indicates that the audio-visual perception process, such as lip-reading [Calvert et al., 1997], will activate an additional brain area different from the single function region of audio or vision. Therefore, the perceptual processing of audio and visual stimulus should consist of three parts, audio, visual and audio-visual joint perception.

Marr [1982] laid down the theoretical foundation for visual perception computation. He assumed that *"the complex reality in the human sight is built progressively until the fully construction is completed inside mind"*. Therefore, this continuous process of visual perception is divided into three different but not necessarily independent layers, i.e. *implementation*, *algorithm* and *computational theory*. Such a division is supported by neuron physiological evidences, e.g. neural response time [Calvert and Thesen, 2004]. Although disputes exist in both psychology and philosophy [Churchland et al., 1994], Marr's theory remains the most widely accepted computational model for visual perception.

The early perception of audio-visual stimulus is divided into two functionally distinct modes in psychology [Julesz, 1991]:

1. Pre-attentive mode, which is a spatially parallel process without specific region of interest (ROI). This mode is usually the initial stage of audio-visual perception creating an overall or environment impression in the mind. The perceptual process in the pre-attentive model is characterised with: (1) independence of the number of stimulus elements; (2) almost instantaneous execution; (3) lack of sophisticated scrutiny; (4) large sensitive field.

2. Attentive mode, in which the ROI are identified and scrutinised with extra weights of *focal attention*. The perceptual process in the attentive mode relies on implicit background knowledge and identifies content-based elements, i.e. video object. The attentive mode is usually regarded as the gate way to later perception, which accesses the semantics by reasoning.

As a conclusion, the combination of vision and audition includes three steps (Marr framework) or two stages (psychological attentive model). The fusion of audio-visual stimuli starts at the low level physical stimulus, e.g. lightness and loudness, and is completed by the mixture of syntax concepts. The major challenges in audio-visual fusion are: (1) how to match these signals and concepts from different temporal and content resolution; (2) how to identify syntax features; and (3) how to develop a fusion model to combine audio, visual and audio-visual syntax features with proper weights.

2.3.2 Natural Language Processing

Natural language understanding is high level perception of human language. As a symbolic modality, language should be faithful to the meaning or semantics. For example, the output of content-based video analysis is a collection of text tags, which clearly state the semantics of video clips. The framework of natural language understanding has been well developed [Baeza-Yates and Ribeiro-Neto, 1999], including four individual stages, namely morphological (word-stemming), syntactical (part-of-speech tagging), lexicon (word sense disambiguation), and pragmatic (discourse and dialogue analysis). These processes summarise the content and reveal the underlying intention in a text document.

A language processing module is essential in content-based video analysis. It plays the following roles: (1) the analyser for the text input, such as closed caption, video OCR and automatic speech recognition; (2) the fusion end for symbolised meta-data; (3) the reasoning network for video intention discovery. The main concern, which hampers the utilisation of natural language processing, is the scarcity of data. It remains a question how to turn a video into a text document. Moreover, there is no language model specified for video processing.

2.3.3 Psychological Fusion Scheme

Although each of these media modalities is complex to compute, it is an even more daunting task to design a fusion model to combine them for the extraction of video semantics. A draft psychological graph of audio, visual and text modality fusion is shown in Figure 2.1. The perceptual sequence of video understanding includes three stages,

1. Audio-visual affection, which is pre-attentive perception. It is carried out simultaneously with audio-visual joint affection to set up the environment context;

2. Audio-visual concept extraction, which is the attentive stage. Meaningful objects and region of interest (ROI) are identified to build the syntax space;

3. Content understanding, where syntax features are combined and explained with background knowledge. At this stage, text information is introduced as a highly symbolised modality to label audio-visual contents. Video labels are propagated between similar video segments. As background knowledge for syntax reasoning, a language model or semantic network is used to discover video intention.

This psychological framework of multiple modality fusion is conceptually plausible in many aspects. However, this framework is too complex to be implemented in most applications. Many derivations have been developed and will be presented in the following section.

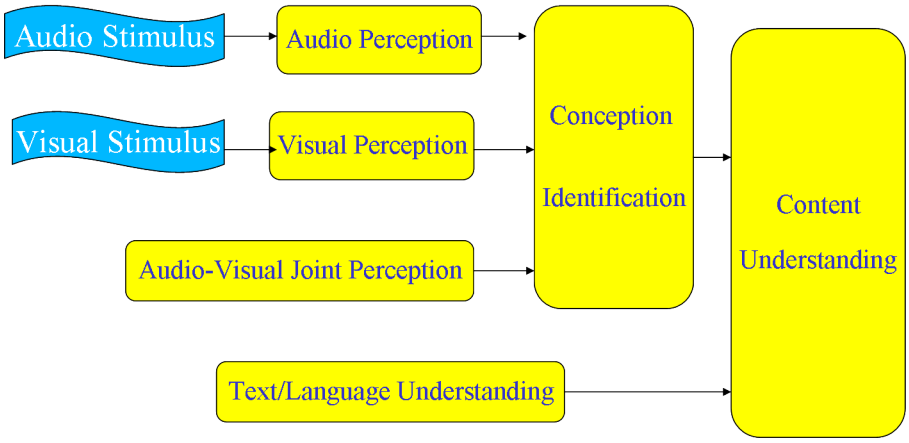


Figure 2.1: Multi-modality Affection in Perception

2.3.4 Practical Modality Fusion

Although the perception process behind multiple modality fusion has not been completely understood, many practical approaches have already been developed, e.g. the syntax-based Bayesian Network [Al-Hames and Rigoll, 2005*b*], the multi-modal graphical model [Al-Hames and Rigoll, 2005*a*], SVM (support vector machine) based concept classifier [Snoek, Worring and Smeulders, 2005], the Best-First search [Gunes and Piccardi, 2005], word net based reasoning [Fleischman and Roy, 2007], and the syntax-cinematic regulation set [Ekin et al., 2003]. All these approaches combine two or more modalities to improve the precision of syntax detection and video semantics identification.

A popular criterion for modality fusion is to maximise the posterior probability of semantic labels over modality features. A video, which contains many modality features and semantic labels, can be recorded as,

$$Modality = \{m_i, 0 \leq i \leq K^m\} \tag{2.1}$$

$$Syntax = \{s_n^i, 0 \leq n \leq K_i^{ms}\} \tag{2.2}$$

$$Label = \{c_j, 0 \leq j \leq K^c\} \tag{2.3}$$

where K^m denotes the number of available modalities, K^c refers to the number of semantic labels and K_i^{ms} is the number of syntax features in modality m_i. The feature space can be symbolised as a three-element tuple (f, s, m), where the syntax s can be detected by low level feature f in the modality m.

According to the phase of combination, fusion approaches can be categorised into two groups, namely early fusion and late fusion. Early fusion (Figure 2.2) integrates low level features before learning syntax concepts (Equation 2.4). This is a regression on multimedia features, which directly maps low level features to semantic labels.

$$c_N = \arg\max_N \prod_{0 \leq i \leq K^m, 0 \leq j \leq K^c} p(c_j | (f, m_i)) \tag{2.4}$$

where the c_N is the favoured semantic label, which maximises the posterior possibility of given semantics on modality features. In Equation 2.4, the appearance of modality features are independent.

Late fusion (Figure 2.3) introduces a middle level of syntax detection to facilitate

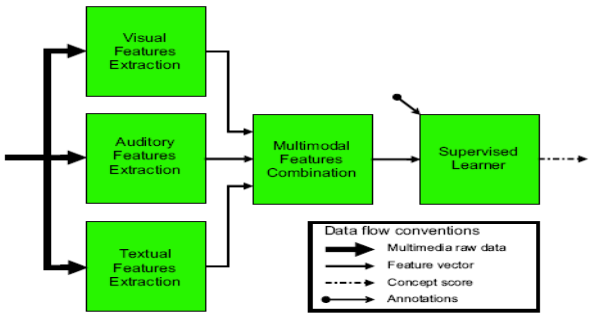


Figure 2.2: General Scheme of Early Fusion [Snoek, Worring and Smeulders, 2005]

the identification of semantic labels. This scheme includes two steps: (1) learn syntax scores individually from low level features (Equation 2.5), which may come from

multiple modalities; and (2) combine these syntax scores to assume semantic labels (Equation 2.6).

$$p(s_n^i) = \Phi(\prod(p(s_n^i|(f,m_i)p(f,m_i))) \quad (2.5)$$

$$c_N = \arg\max_N F(p(s_n^i)p(c_j|s_n^i)), \quad (2.6)$$

where $0 \leq i \leq K^m, 0 \leq j \leq K^c$. $p(f,m_i)$ is the observation of feature f in the modality m_i and the fusion function $\prod$ combines observations on modality feature distribution to estimate the probability of syntax appearance. The score function Φ weights the influence of a syntax s_n^i in semantic labelling with the syntax appearance probability $p(s_n^i)$. The conditional possibility $p(c_j|s_n^i)$ is a prior knowledge about the relation between syntax and labels, which is learnt by training. F is a discrimination scheme, which combines syntax scores and decides a semantic label. Many discrimination schemes have been proposed in the step of semantic label assumption, such as support vector machine in [Snoek, Worring and Smeulders, 2005], Bayesian Network in [Al-Hames and Rigoll, 2005*b*] and decision tree in [Ekin et al., 2003] [Fleischman and Roy, 2007].

The advantage of late fusion is the employment of a syntax score. As an extra semantic representation layer between semantics and low level features, these syntax facilitate label discrimination. Late fusion approaches are widely used in content-based video analysis. However, there are also a number of disadvantages:

- The later fusion requires a heavy learning process, not only in the syntax detection, but also in the combination of syntax scores;
- The selection of syntax is mostly heuristic, there is the potential loss of correlation in mixed feature space, although some techniques for the discovery of hidden features have been proposed [Yan, 2006];
- The uncertainty of syntax score estimation. Although a score function is proposed to catch the causal relationship between syntax and a semantic label, a syntax score is computed by the appearance probability of a syntax feature. It is uncertain whether there exists a plausible tie between syntax appearance and its influence on the semantic label. For instance, we try to detect a goal event by searching for a goal post. But it is hard to say there is a positive relation between the existence probability of a goal post and that of goal events.

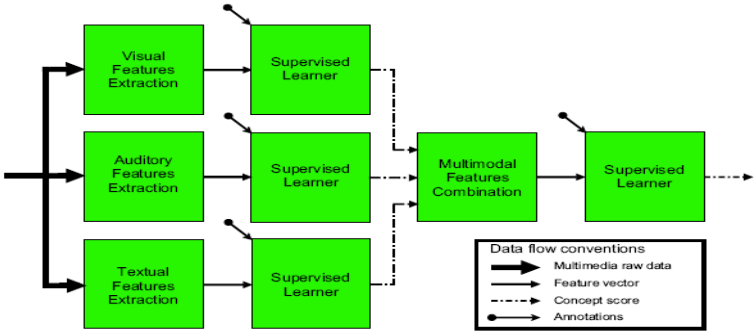


Figure 2.3: General Scheme of Later Fusion [Snoek, Worring and Smeulders, 2005]

2.4 Event Detection

The approach of sports event detection can be categorised into four groups, namely player activity analysis (Section 2.4.1), video clique classification (Section 2.4.2), sequence learning (Section 2.4.3) and affection analysis (Section 2.4.4). Since a game event can be identified by observing participant's behaviour, player activity analysis tracks the movement of players and guesses video contents by player relative positions. Video clique classification employes a sliding window to sample video data and merges visual frames as a feature vector. This approach treats event detection as a classification among feature vectors, such as two-class discrimination, event vs non-event. Note that these feature vectors are of the same length as the width of the sampling window. This means the duration of any events in video clip classification is the same. Sequence learning is a branch of time sequence analysis. Several hidden Markov models are developed to simulate a game process, which regard event detection as a problem of sequence labelling. For example, a semantic event label (the most possible hidden Markov state) is decoded from video (the observed sequence). The methodology of affection analysis differs from previous approaches. Since a game event is a meaningful video segment, the video clip of an event would incur some reflections among spectators, commentators and video directors. Affection analysis estimates the interest of video contents by psychological measurements on visual and audio stimuli. This method treats some emotional patterns as game events. The advantage and disadvantage of affection analysis are both apparent: affection analysis finds interesting game moments but cannot identify game contents. In the following sections, these techniques will be discussed in detail.

2.4.1 Player Activity Analysis

Player activity analysis relies on the surveillance of play field, a technique of computer assisted coaching. Han et al. [2002] stated that observing the play field to summarise play tactics is one of major aims in sports video analysis. Many techniques exist for play field surveillance. Yow et al. [1995] generated an image mosaic to show the action sequence and visualise game events. Seo et al. [1997] developed a player tracking scheme by recognising player uniform colour, ball and goal posts. Gong et al. [1995] divided the play field into zones with different pitch markings, such as corner flag, 18 yard box and centre circle. These authors designed a geometry model of game pitches and matched image edges to known pitch marks, such as play field boundary, in order to determine the whereabouts of the camera and the pitch. Furthermore, Assfalg et al. [2002] divided a game pitch into six zones for one half of field, namely twelve zones in total. The authors supposed that an area with the dominant colour was a part of a game pitch and proposed a field shape descriptor, which includes the line number, the line orientation, corner appearance, and the area of grass field. Then a Bayesian network was trained to discriminate the zone position.

Heuristic rules are drawn by close observation on the behavior of game players and production skills. Intille and Bobick [1999] proposed that the relative position among players, ball and geographical legends in play field, e.g. goal post, middle circle, boundary line, and corner, indicated the occurrence of certain events. For example, if an opposing player with the ball is close to the goal post, the probability of an important event occurring increases [Seo et al., 1997]. Multiple models have been developed to formalise these propositions. Assfalg et al. [2002] developed a finite state machine (FSM) to detect highlight sequences by modelling the relative position among the pitch, players, and the camera. Moreover, these authors discriminated the type of free-kick by searching penalty box and corners. Besides the position of ball, goal post and crossbar, Ekin et al. [2003] took the cinematic pattern into consideration. For example, “a break due to a goal lasts no less than 30 but no more than 120 seconds”. Liang et al. [2005] detected the change in the closed caption to guess player behavior in the baseball game, e.g. the runner and the batter. Additionally, the authors proposed a group of cinematic models according to the baseball regulation.

The detection of geographical legends, e.g. corner, and the automatic tracking of player movements are essential to this approach. But the theory of computer vision argued that it is an ill positioned question to allocate these video objects from an one-camera

visual stream (one eye system). Although this problem could be alleviated by supplying video data from multi-cameras simultaneously, it is difficult to evaluate the precision of such an allocation quantitatively. An expensive idea is proposed in [Intille and Bobick, 1999], which placed a GPS device (Global Positioning System) on players. However, this solution can hardly take place in high competitive games because these equipments may hinder players movement and cause possible injuries.

The rule set from player behavior, relative position, sports regulations, and video editing skills is heuristic. Although this rule set is easy to be extended by new rules, the robustness of the whole set is questionable. There lacks the evaluation on large test collections. For example, Gong et al. [1995] used approximately 120 key frames. Assfalg et al. [2002] employed 60 short video sequences, each of which was only 15 frames. Moreover, the searching cost of rule-based reasoning is propositional to the size of employed rule set, though some heuristic rules can occasionally decrease the computing complexity. Conventionally, the full search space for the position-based reasoning is n^{id}, where n is the number of possible actions, which d players may take at the position i. The complexity of play tactics reasoning is $O(n^d)$ at least.

2.4.2 Video Clique Discrimination

Approaches of video clique discrimination regard event detection as image sequence classification, i.e. two-class classification, event vs non-event. These methods usually consist of three components: (1) video sequence sampling, which is often carried by a sliding window [Truong et al., 2000] [Suresh et al., 2004]; (2) feature extraction, which reduces the dimension of a data space; (3) event discrimination. For example, Truong et al. [2000] used a C4.5 decision tree with a feature set, including video editing type, motion and colour, to discriminate game events. Xu and Li [2003] worked on raw image data and employed principle component analysis (PCA) to decrease data space dimension. The authors simulated the data distribution by a Gaussian mixture model (GMM) to identify commercial and interview segments within sports videos.

To improve the precision of classification, multiple modality information is employed in recent applications. Kang et al. [2004] supposed that event discrimination was an interactive reasoning across media, i.e. audio and visual streams. The authors divided the process of goal detection into two individual steps. Firstly, excited speeches were detected by audio pitch raise and energy increase. Secondly, a goal post was allocated in visual shots near those excited speeches. Sadlier and O'Connor [2005] assumed that

slow-varying audio energy was roughly synchronous with visual features. Therefore, the authors combined audio energy with visual features into a feature vector and employed a support vector machine (SVM) to detect goal events in this audio-visual joint feature space. Xu et al. [2005] developed a later fusion approach. Low-level features from multiple modalities, such as audio, caption text, and visual stream, were combined to identify a set of "middle-level content modalities" or syntax. A hierarchical hidden Markov model was developed to discriminate game events by these syntax.

Although there exist several successful applications, such as the commercial segment discrimination in [Xu and Li, 2003], the approach of video clip discrimination is generally inefficient for video event detection. This is because the sports event detection is not a typical pattern recognition problem. Firstly, there are variant game contents and numerous kinds of game events. It is therefore inapplicable to develop a specific classifier for each type of events. Secondly, as I have mentioned in Section 1.2.1, the event identification relies on contents as well as the context. It is beyond the ability of pattern recognition to model the context in a long time sequence. Another drawback of video clip discrimination is the ignorance of time for the following reasons: (1) the duration of a game event is random, which can hardly be caught by a constant observation window; (2) although information from multiple modalities is helpful, the latency between media streams is unavoidable; and (3) data from multiple modalities are of different temporal and content resolution. The approach of video clip discrimination is implemented with a single temporal resolution, and can hardly deal with these challenges. Nevertheless, the operation that merges all features into a vector, brings a high-dimension data space and results in a high cost in data classification.

2.4.3 Sequence Learning

The game story is a Markov process. This observation inspires approaches of sequence learning, which employ methods of time sequence analysis and regard recurrent patterns as video events. Xu et al. [2001] proposed two content-based structures for football videos, namely *play* and *break* (Section 5.1.1). The authors classified visual frames into two classes, *play* and *break*, by low level visual features such as dominant colour and motion field. To smooth the class label sequence, dynamic programming was used to extract *play-break* video structure. Xie et al. [2004] proposed a hierarchical hidden Markov model to simulate the variation of dominant colour and motion. They claimed that the automatically learnt temporal structure was similar to the *play-break* video structure. This means that the *play-break* video structure can be automatically learnt.

Later, *attack* structures, scenes in sports videos, are extracted by a four state hidden Markov model in [Ren and Jose, 2005] (Chapter 5). Xu et al. [2006] gathered semantic labels and event time stamps from web documents, in order to annotate game events. Since there is a delay in video broadcasting, the authors developed a four-state FSM, which simulated shot transitions in a video stream in order to locate the starts of video events.

The temporal asynchronism between different modalities is the main challenge in the fusion of multiple modality information, because the sequence learning approach relies on time sequence to draw conclusions. Some complex Markov models are proposed to remove the media asynchronism, such as the hierarchical hidden Markov model [Xu et al., 2005] and the couple Markov model. These models employ a set of closely associated Markov chains to simulate the interlaced matching between audio and visual streams. Although such an approach is plausible in the conceptual description of video contents, it is difficult to train a coupled Markov model or a hierarchical hidden Markov model on a long time sequence because of the expensive computational cost. Rather than simulating the whole video sequence, Lenardi et al. [2004] modelled shot transmission by a controlled Markov chain and took the embedded audio energy as the control token. The authors ranked highlight candidates by audio loudness and reported a high coverage of goal events in the top 5 of a candidate list.

Most sequence learning approaches employ various hidden Markov models to discover game intentions. Although the Markov model is effective in time sequence analysis, the proposal of a complex model risks losing model generality. A Markov model with a limited state number can hardly cover all game contents, while a model with many states is fragile before the production artefacts. Such a trade-off between the number of Markov states and model generality is particularly evident when a Markov model is designed for the detection of some specific events, e.g. goals in football. Although more Markov states will improve detection precision in the test collection, it makes the model inextensible to other video data. However, a Markov model with fewer states is ineffective in the presentation of video contents. It is difficult to specify an appropriate model size. Moreover, combining multimedia information for event detection is a multiresolution process (Section 1.2.2). One possible solution is to introduce more Markov states to catch modality information from different resolutions, for example, the hierarchical hidden Markov model in [Xu et al., 2005]. However, this approach is not exempt from critics because such a solution leads to a strong temporal correlation among Markov states and weakens the hypothesis of state independence.

2.4.4 Affection Analysis

Affection-based sports event detection is a recent research topic. Since viewing sports videos is enjoyable, an interesting game event will incur emotional reactions among viewers, such as exciting cheers. Therefore, it is possible to detect meaningful events by measuring affective stimulus from sports videos. Ma et al. [2002] made a hypothesis that such reflection intensity is propositional to the interest of game contents. Based on the observation of computing psychology, the authors proposed a set of individual feature-attention models (Section 6.3.1) to estimate stimuli from low level audio-visual features, such as motion-based attention. An overall stimulus strength was computed as a linear combination of these feature stimulus. Hanjalic [2005] extended the framework of feature-attention into the valance-arousal space (Section 6.2.3), in which the dimension of arousal measures the reflection intensity while the dimension of valance reflects the emotional surface, i.e. negative for rage and anger and positive for excitement and happiness. The authors supposed that event or highlight detection is to allocate a local or global maximum in the arousal dimension with a positive valence. In the estimation of arousal strength, the authors identified the media asynchronism and developed a sliding window to count the number peaks in feature-affection curves rather than combining these stimuli directly. An improved system was reported in [Hanjalic and Xu, 2005], which implemented a smooth algorithm to enhance the signal-noise ratio (SNR) of feature-affection curves. I have developed two affection fusion algorithms in Chapter 6, namely MAR and linear predictor array. These algorithms treat the combination of feature-affection curves as a fusion of time signals at multiple resolutions.

Although affection analysis has found efficient and successful applications, especially in the sports highlight detection, this approach is far from perfect from many perspectives. Firstly, affection is a psychological phenomena, which is affected by culture and personal feelings. Psychological observations maintain that high level understanding, such as text topics, plays a more important role in affection than low level audio-visual stimuli [Osgood et al., 1957]. However, it remains a challenge to include these high level issues into affection estimation. Secondly, the projection function from stimulus to affection is mostly heuristic, due to the lack of quantitative results in psychological experiments. For example, Ma et al. [2002] used the normalised feature value as affection intensity, which is without clear psychological evidence. Moreover, the combination of feature-based affective signals is a multiresolution process the same as modality information fusion. This fact challenges current simple fusion algorithms, such as sliding window and linear combination.

2.5 Conclusion

In this chapter, the literature of sports event detection is reviewed. Four approaches are addressed, including player activity analysis, video clique discrimination, sequence learning, and affection analysis. Advantages and disadvantages of these approaches are discussed on the basis of facts as follows: (1) the generality of approach; (2) the identification of video contents; (3) the allocation of event boundary; (4) the robustness against asynchronous media noise; and (5) the combination of multi-resolution modalities.

To throw light on possible approaches of sports event detection, related research domains, such as data mining and multimodality fusion, are briefly addressed. Two data mining problems, motif detection and network transportation surveillance, are compared with event detection. All of these research topics pay due attention to the extraction of sequential patterns and take component order into consideration. Multimodality fusion approaches are categorised into two groups, early fusion and late fusion, according to the phase of combination. Additionally, the approach of late fusion is widely used in the content-based video analysis.

Another perspective of this chapter is the analysis of experimental data collection. As a creation of visual art, a video is full of variance although the content may be similar. This indicates that event mining and other video analysis techniques have to avoid the strong data dependence on training collection. In this thesis, all algorithms only employ local information. For example, the grass hue model (Section 3.3) is assume by the statistics of image blocks in the same game video.

3

Feature Extraction

> One important issue for digital libraries is finding good models and similarity measures for comparing database entries. A part of this difficulty is that feature extraction and comparison methods are highly data-dependent.
>
> Minka and Picard (Pattern Recognition 1997)

This chapter is dedicated to the extraction of syntax and affective features, which is employed as a preprocessing step in the video segmentation and *attention* computation.

3.1 Introduction

Content-based video analysis aims to understand existing video records automatically. This research makes use of knowledge in the field of machine learning, pattern recognition, video processing, information retrieval, databases and artificial intelligence. The prominent problem of content-based video analysis is to find an efficient and effective description for semantically unpredictable contents. It is the first but most decisive step to extract a reliable feature set, which is general but sufficient to discriminate numerous video contents. The theory of machine learning maintains that a carefully selected feature set not only simplifies the selection of classifiers, but also promises a satisfactory

performance. Moreover, a meaningful feature set would reduce the presentation redundancy, decrease the overall complexity of semantic description, and facilitate video content management applications, such as video summarisation.

3.1.1 Syntax Feature

In linguistics, syntax refers to grammar rules and special patterns governing the order of words and symbols. Syntax features in video processing not only denote sequential patterns in video streams but also include visual or audio symbols associated with domain knowledge, such as a close-up view and the video object of a football in football video analysis. These features are sometimes described as high-level or middle-level features, which reflect semantic concepts in a video. Figure 3.1 shows a system flowchart of syntax based content analysis. Syntax features are employed for the following advantages.

- Conceal strong variation of low level features;
- Define the hypothesis of context reasoning;
- Discriminate video contents. For example, a goalpost is inevitable for a goal event in a football video. Therefore, it is effective to discriminate goal events from other video contents, such as a free-kick, by searching the video object of a goalpost.

Syntax extraction plays an essential role in content-based video analysis. For instance, the open competition for video retrieval and content analysis (*TRECVID*), considers the extraction of high level features as one of three main tasks in content-based video retrieval. However, an investigation of syntax features has to deal with the gap of semantics. Such a gap refers to the difficulties in the mapping from low level features to video syntax and domain knowledge. Anyway, some guidelines for syntax selection exist: (1) a syntax feature is semantically clear; (2) syntax features can be easily coupled to discriminate video contents with the help of domain knowledge. These propositions are based on the usage of syntax features in content analysis.

The projection from syntax to video semantics usually involves a reasoning process subject to domain knowledge and heuristic rules. For example, the discrimination of a goal and a penalty relies on the prior moment of a goal. Therefor, it is difficult to identify necessary syntax features without a systematic knowledge base. Moreover, the development of a syntax detector is an application of pattern recognition in visual and audio data, such as an adaboost classifier for human face detection [Polikar, 2006*a*].

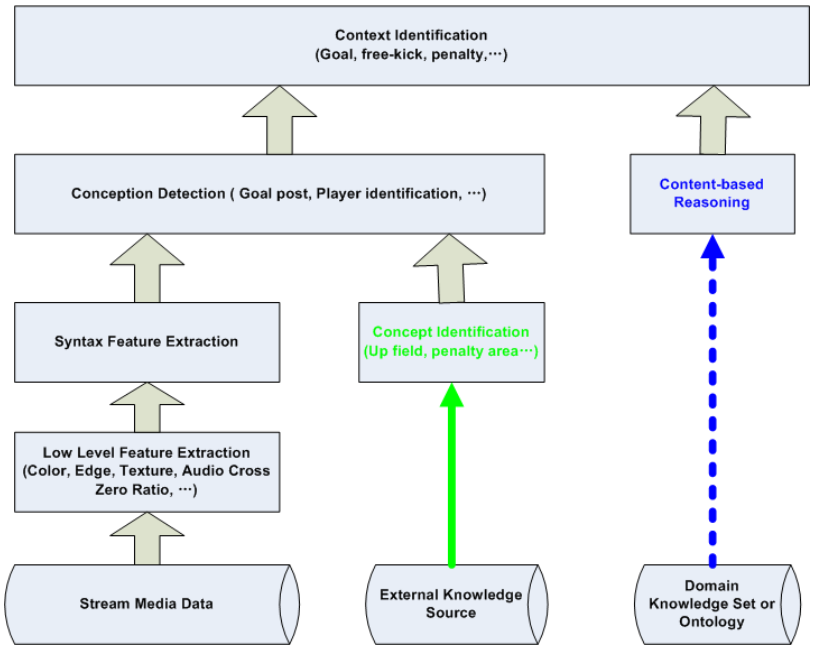


Figure 3.1: Syntax-based Sports Video Analysis

The effectiveness of such a detector is subject to the classifier and training data.

By means of self-evident sports semantics and sports regulations, many sports syntax features have been identified, including game pitch, goal posts, pitch boundary lines and players. These syntax features can be extracted by low-level visual and temporal features, such as grass hue, texture, and average contrast. The following sections present a set of syntax extraction algorithms, such as game pitch detection, shot style classification, camera zoom depth, and video object discrimination, e.g. player uniform and goal post.

3.1.2 Affective Feature

Affective features depict emotional aspects in a video, which are sometimes referred as salient features too. Since sports videos record intensive emotional scenes, such as cheers of spectators, emotion becomes a rich information resource for video content analysis, especially for the discrimination of interesting vs non-interesting contents. Many applications have reported the usage of affective features in literature, such as highlight detection in sports video [Ma et al., 2002], video genre classification [Wang and Cheong, 2006] and adaptive video summarisation [Hanjalic and Xu, 2005].

Similar to the *semantic gap* in syntax computation, the extraction of affective features

has to deal with the *affect gap* [Mittal and Cheong, 2003] [Wang and Cheong, 2006]. The affect gap denotes the uncertainty caused by the inscrutable nature of emotions, which includes two aspects: (1) there are numerous emotion types; and (2) a similar stimulus is able to incur different emotions, depending on the psychological context. Moreover, the mapping from physical signal strength, such as audio loudness, to the psychological stimulus remains a research question in computing psychology. In Chapter 6, three signal-stimulus project models will be addressed, i.e. signal normalisation, adaptive normalisation and self-entropy, for the effect estimation of physical signal intensity on emotions.

Several affect spaces have been proposed to describe human emotions, such as attention space [Ma et al., 2002], arousal-valence space [Hanjalic, 2005] [Wang and Cheong, 2006], and arousal-valence-stance space [Crary, 1999]. A brief comparison among these psychological spaces will be found in Section 6.2. As a matter of fact, two fundamental dimensions, arousal and valence, are involved in these different psychological spaces. The arousal axis stands for the strength of an emotion, while the valence axis refers to the attitude of an emotion, i.e. positive for happiness and negative for sadness. The valence axis can be ignored in affective sports video analysis. This is because of following facts: (1) a sports video director should be properly indifferent about game contents; and (2) sports videos are produced to amuse viewers rather than to anger them. This means that an one-dimension *attention* space suffices for the description of emotional variations in sports videos and is enough to apply affective sports video analysis.

The estimation of stimulus intensity is a research topic of computing psychology. This topic involves many research issues, such as stimulus effect discrimination, stimulus strength measurement and stimuli fusion. Effect discrimination identifies an stimulus effect on valence, whether this stimulus excites reaction or not. Strength measurement estimates an emotion increment on a given signal intensity. Stimuli fusion combines multiple stimuli to guess an overall emotion state of reflector. However, there are no clear conclusions reached on both research topics, stimulus strength estimation and stimulus fusion, especially in the perception process of video watching. For instance, both quick motions and fast switches of visual shots can attract attention. It is difficult to compare attention increments caused by quick motion and by shot switches, because such a psychological gain is decided by the psychological context as well as stimulus strength. Moreover, as mentioned in Section 2.3, stimuli from multiple media modality should be combined for an unified affection, because a reflector can only hold one emotion state at a given moment. For example, a human being cannot be angry and happy

at the same moment.

Table 3.1 surveys salient features and their psychological effects. The symbol set

feature	attention facts	qualitative relationship
football size	zoom depth	$+$
uniform size	zoom depth	$+$
face area	zoom depth	$+$
domain color ratio	zoom depth	$-$
edge distribution	rect of interest	$*$
goalpost	rect of interest	$*$
shot duration	temporal variance	$-$
shot cut frequency	temporal variance	$+$
motion vector	temporal variance	$*$
zoom-in sequence	temporal variance	$+$
visual excitement	motion	$+$
lighting	spatial variance	$*$
colour energy	stimuli strength	$*$
replay	temporal contrast	$*$
off-field shot	temporal contrast	$*$
base band energy	loudness	$+$
cross zero ratio	sound variation	$+$
speech band energy	sound variation	$+$
keyword	semantic	$*$
LFPC and delta	sound variation	$*$
MFCC and delta	sound variation	$*$
spectral roll-off	sound variation	$+$
spectral centroid	loudness	$+$
spectral flux	loudness	$+$
chroma and its delta	sound variation	$*$
LSTER	sound variation	$+$
octave energy	loudness	$+$
music scale	sound variation	$*$
audio type proportion	valance classification	$*$
scene affect vector	valance classification	$*$

Table 3.1: Salient Feature

of $\{+, -, *\}$ indicates the qualitative effect of salient features on arousal: "$+$" refers to a positive qualitative relation between feature and attention, where a strong signal will incur an intensive attention; "$-$" indicates a negative effect that a strong signal will pacify reflections; and "$*$" denotes that the stimulus effect relies on the psychological context. Additionally, if a feature can incur both positive and negative affections, this feature will be marked as "$*$" qualitatively unsure. Note that this list of salient features

(Table 3.1) is not adequate. Many salient feature have not been addressed in this thesis, such as linear predictor coefficients (LPC) [Sadlier and O'Connor, 2005] and Mel frequency cepstral coefficients (MFCC) [Kijak et al., 2003]. Although these features have been reported the effectiveness, it is difficult to qualify feature-based stimuli and state the link with perception. For example, MFCC is a widely employed frequency feature in audio analysis. Kijak et al. [2003] used MFCC feature to estimate the intensity of audio stimuli. However, it lacks direct psychological evidences to support such a perceptual effect from a cepstral domain.

This chapter proceeds as follows. Section 3.2 introduces the measurement of visual frame difference and the allocation of cut-style shot boundary. Section 3.3 presents the adaptive estimation algorithm of grass hue by Gaussian mixed models (GMM) in order to decide the feature of play field ratio. Section 3.5 detects player uniforms and other video objects to estimate the scale of camera zoom depth. A set of salient feature computation are stated in Section 3.6, including motion salience, colour salience, and audio salience. Summary and discussion are found in Section 3.7.

3.2 Shot Density Computing

A shot is a sequence of visual frames which keep harmony in colour or other low level visual features, e.g. motions and an edge map. Such a video structure is treated as a temporal data unit for video processing. The idea of shot originates from the process of video production [Bordwell and Thompson, 2004], where a shot is an action of video recording and the change of shots is mostly due to a switch of cameras. Hence, shot segmentation is a pre-requisite of video analysis. For instance, Lienhart [1999] regarded the extraction of shot structures as the first step in content-based video analysis.

Shot frequency (Equation 3.1) is a widely accepted salient feature [Hanjalic, 2005] [Xu et al., 2006] [Wang and Cheong, 2006]. According to the theory of visual language, a frequent switch among cameras or a high shot frequency is an efficient way to attract attention, although such an action are not necessarily welcomed by viewers [Bordwell and Thompson, 2004]. For example, an abrupt increase of shot frequency often indicates an occurrence of breath-hold exciting moments in a story film [Wang and Cheong, 2006].

$$F_{shot} = \frac{N_{shot}}{T_{period}} \sim \frac{N_{boundary}}{T_{period}} \tag{3.1}$$

where shot frequency F_{shot} is the shot number N_{shot} or the number of shot boundaries $N_{boundary}$ in a given time period T_{period}.

However, sports videos are characterised with highly similar colour and strong motion, which can hardly be processed by general shot segmentation algorithms [Xie et al., 2002] [Ekin et al., 2003]. We developed a two-pass shot segmentation algorithm with adaptive thresholds according to the distribution of shot durations. This method not only alleviates image distortions caused by strong motion, but also smoothes the time sequence of shot boundaries.

The identification of shot boundaries is a two-class hypothesis test on successive visual frames, boundary vs non-boundary, with the constraint on temporal continuity. Therefore, three issues are involved during the development of shot segmentation algorithms, namely shot transition types, correlated measurements on visual similarity and threshold decision [Lienhart, 1999]. A shot transition type is the style of connection sequence between two shots and is usually regarded as a skill of video editing. For example, an abrupt switch is called *cut*, where two shots are linked directly without any editing effects. Gradual changes can be classified into *fade-in/out* and *dissolve*. A *fade* transition consists of a gradual diminishing (*fade-out*) or increasing (*fade-in*). A *dissolve* transition is a combination of *fade-in* and *fade-out*. Correlated measurements estimate the visual similarity between two video frames. For example, colour histogram, pixel-based image distance and camera motion are widely employed visual features in shot segmentation [Ahanger and Little, 1996] [Tsekeridou and Pitas, 2001]. However, sports video is characterised with fast local motions and unified dominant colour of a game pitch. Colour histogram and motion vector are not so robust and efficient in sports videos as they are in story films and news videos. For example, the motion incurred by camera switch can hardly be discriminated from that of video objects. There are two solutions to this problem: (1) using new visual features rather than colour histogram and motion; (2) developing a robust algorithm for adaptive threshold decision. Note that the threshold decision is an empirical question, which is subject to the genre of video and the selection of visual features. Many works have been proposed in the literature, such as twin threshold [Lienhart, 2001], grey level statistics [Lienhart, 1999], information entropy [Cernekova et al., 2002], frame-based difference Bayesian model [Vasconcelos and Lippman, 2000] and frame transition hidden Markov model [Boreczky and Wilcox, 1998]. The highlight of sports video shot segmentation algorithm in this thesis is to take shot duration as an extra parameter. Besides the requisite of shot frequency computation, there are two reasons for such a consideration. Firstly,

the likelihood of a new shot boundary is highly contingent on the interlace between the prior one, especially in sports videos. There is a statistical average length of *attack* team action (http://www.soccerstats.com/), which brings an expectation on the duration of a scene in sports videos. Therefore, the precision of segmentation can be improved by taking the feature of shot duration into count [Vasconcelos and Lippman, 2000]. Secondly, a shot transition is usually less than 10 frames in sports videos. A duration-adaptive threshold will lead to a wide observation window. Such a window can abide intensive visual variations in a shot transition period and thus facilitate the identification of a shot boundary. Additionally, it is a popular post-process to merge short segments, whose length is less than 20 frames, in some practical shot segmentation systems.

3.2.1 Image Distance

Three visual features, namely region-based pixel difference, region-based motion and histogram distance, are used; four temporal scales, $1, 5, 10, 15$ frames, are employed for the computing of visual differences. The largest scale of fifteen frames (0.6 sec in MPEG-1 videos) ensures such a comparison across a period of shot transitions, which is usually shorter than 0.5 sec or 12.5 frames in MPEG-1 PAL format. The range of five frames is proposed to address the problem of double cut, in which a shot change falls into two or more frames equally [Boreczky and Wilcox, 1998]. The 10-frame range offers a smooth measurement between 5-frame and 15-frame observation.

Region-based Pixel difference

The visual data in MPEG-1 video streams are in YUV colour space but with different colour sampling patterns corresponding to video encoders, i.e. $YUV420$ and $YUV422$. All colour data are transformed into the RGB colour space.

$$R = Y + 1.1398V \tag{3.2}$$

$$G = Y - 0.3946U - 0.5805V \tag{3.3}$$

$$B = Y + 2.0321U \tag{3.4}$$

To localise strong motions, visual frames are divided into 3 ×3 regions to compute region-based pixel difference and motion (Figure 3.2).

Such a division brings benefits as follows. Firstly, a game pitch is an rectangle area with an unified grass colour. It increases detection precision of game pitches to cut a visual frame into regions. Secondly, enlarging the block size will decrease the possibil-

Figure 3.2: 3 ×3 Region Graph

ity of region mismatch and facilitate the detection of camera motion. Motion vectors on small blocks, such as 8 ×8 blocks in MPEG-1 encoding, is trivial, because of multiple fits. Such a motion field in an encoded visual stream is random distributed and can hardly be used for camera calibration. Thirdly, video producers focus on player's activities. Areas with strong motion are mostly found at the centre of visual frames, while suburb areas remain unchanged. The region-based statistics of pixel distance will enlarge the distance between local motions and a gradual transition over the whole image. For instance, a motion field comparison is shown in the histogram of image-based pixel distance distribution in Figure 3.3 and that for region-based in Figure 3.4 for the semi-final game of World Cup 2006, Brazil vs. France, respectively. The last advantage concerns the reduction of motion vector number, which is helpful the estimation of global motion and camera pan detection.

The pixel distance is the number of mismatched pixel pairs in two images, whose mean absolute difference (Equation 3.5) is greater than given threshold (Equation 3.6).

$$Diff_{i,j} = \|R_i - R_j\| + \|G_i - G_j\| + \|B_i - B_j\| \tag{3.5}$$

$$RegionDiff_C = \sum_{n,m \in C} sgn(Diff_(n,m) - T_c) \tag{3.6}$$

where n, m are the pixel pair of two visual frames in region C; the threshold T_c removes noise which is set to 15 in experiments of *RGB* colour space; and *sgn* is the sign function

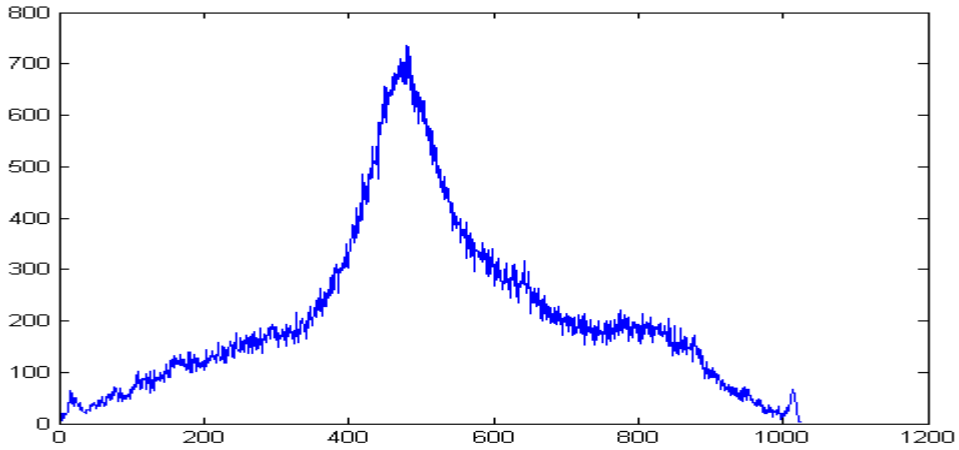


Figure 3.3: 1024-bin histogram of adjacent frame pixel distance in Brazil vs. France (World Cup 2006). The peak in bin 470^{th} hints the number of visual frames with strong local motions, while the weak peak in bin 810^{th} refers to shot transitions.

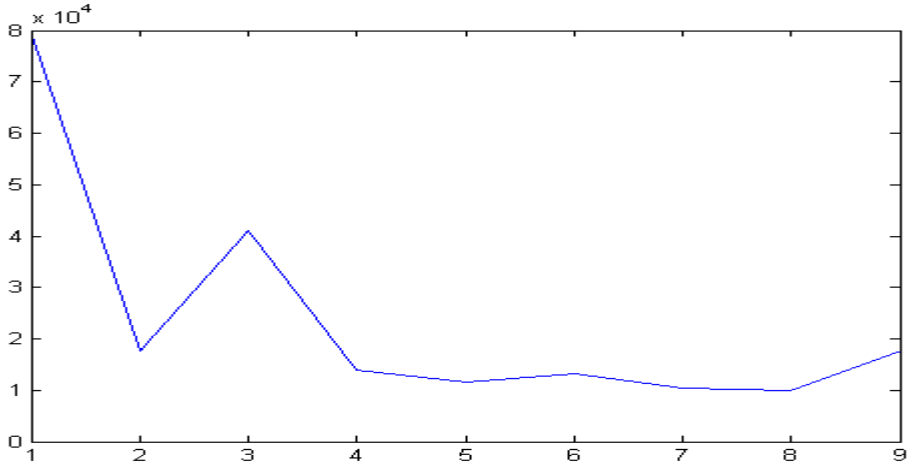


Figure 3.4: 9-bin histogram of adjacent frame's region-based pixel distance in Brazil vs. France (World Cup 2006). This histogram counts the number of changed regions and indicates that a threshold of 4 blocks is a good threshold to discriminate local motion and shot transition.

Equation 3.7.

$$sgn(x) = \frac{x}{\|x\|} \tag{3.7}$$

The region-based pixel difference of a visual frame pair is the number of regions, in which the number of changed pixels is greater than the average. Figure 3.4 shows the statistics of pixel differences in every region of the semi-final game, Brazil vs. France, World Cup 2006.

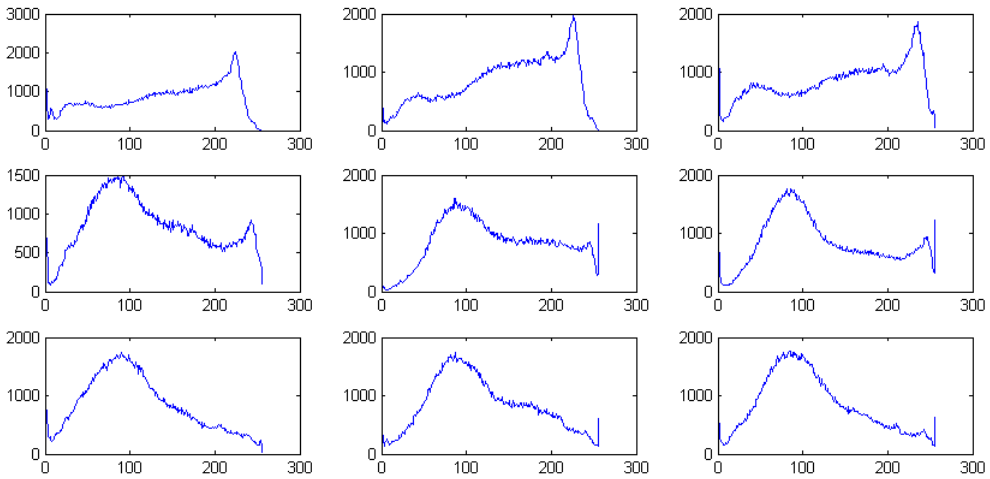


Figure 3.5: 256-bin histograms for 3 ×3 region-based adjacent frame's pixel difference distribution in Brazil vs. France, World Cup 2006

Region-based Motion

Region-based motion vectors are estimated by the 2D block translation model without deformation (Equation 3.8), because the camera zoom is rare in sports videos. An $N \times N$ block in frame k centred about the pixel $n = (n_1, n_2)$ is modelled by a shift of the same size block in frame $k+l$, where k, l are integers, a 2D block translation model without deformation is,

$$s(n_1, n_2, k) = s(n_1 + d_1, n_2 + d_2, k + l) \tag{3.8}$$

where d_1, d_2 are shifts.

The maximum matching per count (MPC) criterion is followed; the three step search algorithm [Cheung and Po, 2003] with diamond-shape pattern (Figure 3.6) is employed to find the best match of current regions in the successive frame. A visual sequence with constant global motion will be labelled as *pan* and removed from pixel difference statistics.

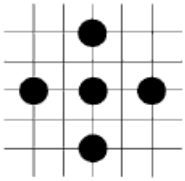


Figure 3.6: Diamond search pattern for image block match

Histogram Distance

Each colour in the *RGB* colour space is quantified into six bins. Hence the total number of histogram bins is 216. Two types of histogram distances are computed in this thesis, namely intersect distance [Swain et al., 1997] and quadratic distance [Ashley et al., 1995], because these distances can be used for later game pitch detection and orientation discrimination. Figure 3.7 displays the logistic histogram of colour distance between adjacent visual frames.

The intersect distance is widely used in colour-based image retrieval (Equation 3.9), which measures a colour similarity between two histogram h, g. Such a distance is a scalar in $(0, 1]$, where 0 denotes two histogram are totally different and 1 for the same.

$$d(h,g)_{intersect} = \frac{\sum_{r\in R}\sum_{g\in G}\sum_{b\in B} min(h(r,g,b)g(r,g,b))}{\min(\|h\|,\|g\|)} \tag{3.9}$$

Quadratic distance was proposed in the **QBIC** content-based image retrieval system [Ashley et al., 1995] as the cross-correlation measurement between histogram bins based on the perceptual similarity of colours (Equation 3.10)

$$d(h,g)_{quadratic} = (h-g)^t A(h-g) \tag{3.10}$$

where A is the similarity matrix, whose element $a_{i,j}$ is

$$a_{i,j} = 1 - \frac{d_{i,j}}{\max(d_{i,j})} \tag{3.11}$$

$d_{i,j}$ is the Euclidean distance of two colour i and j in the *RGB* colour space.

3.2.2 Threshold Decision

Shot segmentation is a prior test of a shot boundary indicator variable S over two hypothesis,

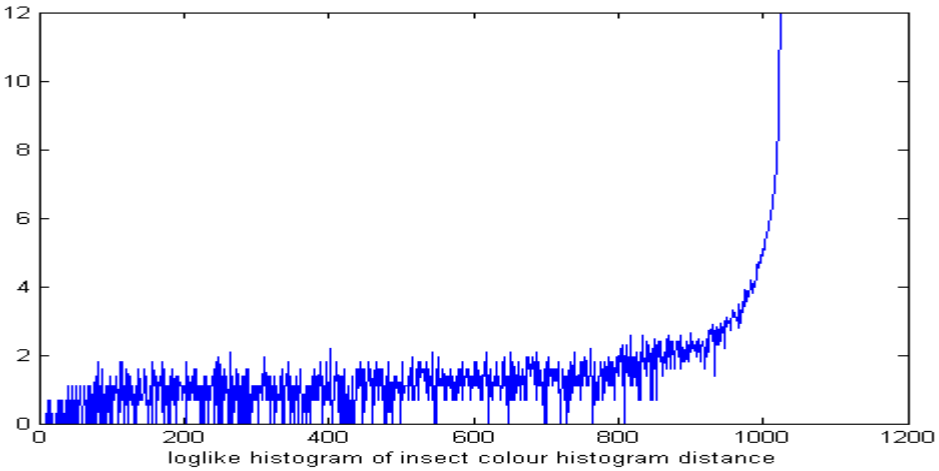


Figure 3.7: Logistic Histogram Distribution (1024 bins) of Insect Colour Histogram Distance between Adjacent Frames in Brazil vs. France(World Cup 2006)

- H_0: no shot boundary between two frames (S=0);
- H_1: a shot boundary between two frames (S=1);

In case that an optimal decision can be provided by the likelihood ratio, the hypothesis H_1 is chosen if and only if

$$L = \log \frac{P(D|S=1)}{P(D|S=0)} > 0 \tag{3.12}$$

and H_0 is favoured, otherwise.

An optimal threshold is determined by a prior distribution of distance variable D. In the case of Gaussian distribution $D_i(\mu_i, \delta), i \in \{0,1\}$ under each hypothesis, H_1 will be favoured, if the distance between two images is larger than the average of distribution expectations (Equation 3.13).

$$D > \frac{\mu_0 + \mu_1}{2} \tag{3.13}$$

One of this model's disadvantages concerns the ignorance of temporal constrains that a shot is made up by a group of continuous visual frames. The drawbacks are obvious as follows. Firstly, a single pre-defined threshold hides the statistic behaviour of visual dissimilarity during a shot transition. For example, in a gradual transition, the cumulative effect of visual changes is distributed equally into successive frames so that a shot boundary may contain several frames. The two-cut shot transition [Boreczky and Wilcox, 1998] is a *"cut"* but with a double-frame boundary. One of possible solutions is to estimate the duration of shot transitions and thus employs a wide observation

window to overcome shot boundaries. Secondly, a video is a discrete process with a given frame-rate, e.g. 25 frames per second in the MPEG-1 video stream. The inclusion of a shot duration is significant in both video segmentation and video characterisation. Vasconcelos and Lippman [2000] extended the likelihood ratio model (Equation 3.12) with a posterior likelihood of shot state transition. Therefore, let non-boundary state $S_{t,t+\delta} = 0$, where t is the start frame number of a transition and δ for the interlace between observations, then the posterior likelihood ratio between hypothesises is,

$$\begin{aligned}\frac{P(S_{t,t+\delta}=1|S_t=0,D_{t+\delta})}{P(S_{t,t+\delta}=0|S_t=0,D_{t+\delta})} &= \frac{P(D_{t+\delta}|S_t=0,S_{t,t+\delta}=1)P(S_{t,t+\delta}=1|S_t=0)}{P(D_{t+\delta|S_{t+\delta}=0})P(S_{t,t+\delta}=0|S_t=0)} \\ &= \frac{P(D_{t,t+\delta}|S_{t,t+\delta}=1)}{P(D_{t,t+\delta}|S_{t,t+\delta}=0)}\frac{P(S_{t,t+\delta}=1,S_t=0)}{P(S_{t+\delta}=0)}\end{aligned}$$

Furthermore, the posterior log-likelihood ratio L_{post} of a given observation D can be,

$$L_{post} = \log\frac{P(D_{t,t+\delta}|S_{t,t+\delta}=0)}{P(D_{t,t+\delta}|S_{t,t+\delta}=1)} + \log\frac{P(S_{t,t+\delta}=1,S_t=0)}{P(S_{t+\delta}=0)} \tag{3.14}$$

In Equation 3.14, the first part is the same as the likelihood ratio of a given observation over two hypothesises in Equation 3.12, while the second part is a prior probability of shots, whose duration is less than δ. This posterior log-likelihood shows that an optimal threshold for a shot boundary decision depends on not only an observation distribution under each hypothesis, but also a shot duration distribution.

If we regard the occurrence of a shot transition as the problem of random arrival, the distribution of shot duration will follow an Erlang distribution (Equation 3.15) or a Weibull distribution (Equation 3.18). An Erlang distribution describes the waiting time of k occurrences of independent events, which take place at an expected rate. The density function of an Erlang distribution is,

$$f(x,k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{(k-1)!}, x \geq 0 \tag{3.15}$$

and the cumulative function is,

$$F(x,k,\theta) = \frac{\gamma(k,\frac{x}{\theta})}{(k-1)!} \tag{3.16}$$

where θ is the arrival rate; k refers to the shape parameter, which is the shot length in the shot segmentation application.

Vasconcelos and Lippman [2000] gave the optimal judgement favouring H_1 hypothesis under the Erlang assumption as Equation 3.17,

$$\log \frac{P(D_t|S_t=1)}{P(D_t|S_t=0)} > \log \frac{F(t,k,\theta)}{F(t+\delta,k,\theta)-F(t,k,\theta)} \tag{3.17}$$

The Weibull distribution describes the waiting time for a new independent event, which arrives at an expected rate. The density function of a Weibull distribution is,

$$f(x,k,\theta) = \frac{k}{\theta}(\frac{x}{\theta})^{k-1} e^{-(\frac{x}{\theta})^k} \tag{3.18}$$

Therefore, the judgement favouring H_1 hypothesis becomes,

$$\log \frac{P(D_t|S_t=1)}{P(D_t|S_t=0)} > -\log(\exp(\frac{(t+\delta)^k - t^k}{\theta^k}) - 1) \tag{3.19}$$

Both Erlang and Weibull distributions are closely associated with the exponential distribution and Poisson process. These continuous distributions are not sensitive to a scale difference. Therefore, we can efficiently estimate a distribution at a coarse resolution and then employ such models again at a high resolution without obvious scale deterioration.

3.2.3 Sports Video Shot Segmentation Algorithm

In the sports video collection used in this thesis, most of shot transitions are of *cut* or *N-cut* type, though some automatic video editing effects occasionally appear, such as dissolve. A *N-cut* transition is automatically produced by nonlinear video editing tools. This transition style smoothly links two shots by a sequence of N dissolve frames ($N < 5$). Without losing generality, we assume that the duration of a shot transition rarely exceeds half a second or 12 frames in football videos.

Besides the calculation of a 4-scale frame difference array (at $2, 4, 10, 15$), the shot segmentation algorithm is proposed as follows.

Data: Four-scale frame difference array $G = (G^2, G^4, G^10, G^{15})$ with length $\|G\|$
Result: Shot boundary array BL
Compute the average frame G^{15} difference s_0;
Set two experimental thresholds, $s_h = 2.6s_0, s_l = 1.2s_0$;
$i = 0$;
while $i \leq \|G\|$ **do**
  if $G_i^4 > s_h$ **then**
    add i into BL ;
  else
    if *three continuous frame difference* $s_i, S_i - 1, s_i + 1 > S_l$ **then**
      add i into BL ;
    end
  end
end
estimate the Erlang (Equation 3.15) or Weibull model (Equation 3.18) from the set BL by EM algorithm;
clear the set BL;
$i = 0$;
while $i \leq \|G\|$ **do**
  compute the adaptive threshold s_i^a from Equation 3.17 or Equation 3.19;
  if $G_i^2 > s_i^a$ **then**
    add i into BL ;
  end
end

Algorithm 1: Two-pass Sports Video Shot Segmentation Algorithm

Algorithm 1 includes three steps. The first step is a typical two-threshold shot segmentation algorithm. Two static thresholds are set as given ratios to the average frame difference. The higher threshold is usually titled as strong threshold, while the other is a weak one. The judgement of shot boundary is as follows: if a frame difference is greater than the strong threshold, the frame will be labelled as a shot boundary; if three continuous frame difference is greater than the weak threshold, the middle frame will be a shot boundary. The second step is to estimate an Erlang or Weibull distribution from these roughly detected boundaries, and thus an adaptive threshold is computed. The third step uses the adaptive threshold to search shot boundaries.

3.2.4 Evaluation and Conclusion

Five football video clips, each of which lasts fifteen minutes long, are randomly selected: three clips from the game of Brazil vs Germany (World Cup 2002, the final game); one from France vs Korea (World Cup 2006), and one from Italy vs France (World Cup 2006, the final game). These video clips are manually labelled frame by frame in order to build a ground truth. All these clips are of 352 ×288 visual resolution in MPEG-1 PAL format.

The recall and precision measurements in [Toole et al., 1999] are employed. Recall is the proportion of shot boundaries correctly identified by the system to the total number of shot boundaries (Equation 3.20). Precision is the proportion of correct shot boundaries identified by the system to the total number of shot boundaries identified by the system (Equation 3.21).

$$Recall = \frac{Number - of - Shot_{correctly-identified}}{Number - of - Shot_{all}} \tag{3.20}$$

$$Precision = \frac{Number - of - Shot_{correctly-identified}}{Number - of - Shot_{detected}} \tag{3.21}$$

Table 3.2 and Table 3.3 compare the performance of shot segmentation algorithms with different hypothesises. The column of twin threshold refers to the first pass in the shot segmentation algorithm, which is carried out with two stable thresholds. These columns provide performance baselines. Columns of Weibull and Erlang stands for segmentation algorithms, which come with adaptive thresholds but follow different hypothesises, Weibull and Erlang distribution, respectively.

Test Video	Shot Number	Twin Threshold			Weibull			Erlang		
		Correct	Fault	Missed	Correct	Fault	Missed	Correct	Fault	Missed
0	90	74	34	16	83	9	7	79	12	11
1	93	67	59	26	79	14	14	81	16	12
2	48	46	16	2	48	6	0	47	7	1
3	77	69	16	8	73	8	4	73	7	4
4	93	82	37	11	89	17	3	89	14	3

Table 3.2: Shot Segmentation Evaluation

Table 3.2 shows that this two-pass shot segmentation algorithm reduces detection errors significantly. The usage of shot duration distribution in the threshold computa-

Test Video	Game	Twin Threshold		Weibull		Erlang	
		Precision	Recall	Precision	Recall	Precision	Recall
0	Brazil vs Germany	0.685	0.822	0.902	0.922	0.868	0.877
1	Brazil vs Germany	0.531	0.744	0.849	0.849	0.835	0.870
2	Brazil vs Germany	0.742	0.958	0.888	1.0	0.870	0.979
3	France vs Korea	0.812	0.896	0.901	0.948	0.913	0.948
4	France vs Italy	0.689	0.881	0.840	0.957	0.864	0.957
mean		0.692	0.860	0.876	0.935	0.870	0.926

Table 3.3: Shot Segmentation Precision and Recall

tion leads to an adaptive threshold, which various with the prior visual difference and decreases with time elapsed. Such a threshold avoids errors caused by strong motions. Hypothesises of Erlang and Weibull distribution do not show salient difference in experiments, although the Weibull distribution seems to fit the data collection a little better.

3.3 Play Field Ratio

The game pitch is a significant domain feature in sports videos, because such an area occupies most visual frames in a game video. Xu et al. [2001] argued that the ratio of play ground (Equation 3.22) reflected game contents and then proposed a play-break video structure for game content filtering.

$$R_{field}(t) = \frac{\|H_{field-colour}(t)\|}{\|H(t)\|} \tag{3.22}$$

where H is the colour histogram of image, $\|H(t)\|$ is the number of samples in given colour histogram bins, and $H_{field-colour}$ is the sample number with grass colour.

A grass hue model plays a key role in the estimation of play field ratio. Ekin et al. [2003] manually gathered grass pixels from multiple videos and built a prior model of grass hue in the HSV colour space. The authors found that grass colour occupied $65^o - 75^o$ interval of H dimension. However, such a model is coarse and relies on prior data collection. In most cases, the hue of a play field varies with environment conditions, such as location, weather and light. A prior colour model can hardly cope with these strong variations, while keeping a satisfactory precision on grass area detection. Xu and Chua [2004] claimed that play field colour would be the dominant colour in visual frames, because a game pitch had to be focused throughout a game. Note that dominant colour is an approved MPEG 4 visual feature and a visual content descriptor in MPEG-7. A set of extraction algorithms have already been developed in several

commercial standards. However, this assumption of dominant colour relies on data collections and the style of video production. An area with a dominant colour does not always belong to a game pitch. For instance, more than 37% visual frames have a field ratio lower than 20% in the collection of World Cup 2002. Most *focus* and *break* shots (Section 5.3) convey little game pitch area [Ren and Jose, 2005]. Obviously, the dominant colour in these visual frames are away from the grass hue distribution.

We propose an approach to extract a grass hue model from a game video automatically. The following general observations are found in football videos:

1. A play field is usually a homogenous area in colour and texture;
2. A visual frame focusing on a play field is more unified in colour and texture than others;
3. In a given game video, the distribution of play field colour will not change greatly.

A full football game video (120 minutes) contains more than 10^5 frames. Hence, there are sufficient visual data to estimate a game-specific hue model. This means that a grass hue model will be computed for each game rather than using a prior colour model for all game videos. Based on prior observations, a two-layer booster filter is developed to collect possible grass hue samples. The first layer rejects non-homogenous frames in a video, while the second layer excludes non-homogeneous regions in homogenous frames. Moreover, I select *sRGB* colour space (Equation 3.23-3.25) to remove the effect of light.

$$r = \frac{R}{R+B+G} \tag{3.23}$$

$$g = \frac{G}{R+B+G} \tag{3.24}$$

$$b = \frac{B}{R+B+G} \tag{3.25}$$

Figure 3.8 compares the distribution of game pitch colour in *RGB* and *sRGB* colour space. It is obvious that the distribution of grass hue samples is condensed in the *sRGB* colour space.

Given the block-based encoding in commercial standards, such as MPEG 1 (8 ×8 blocks), we propose block mean (Equation 3.26) and block covariance (Equation 3.27)

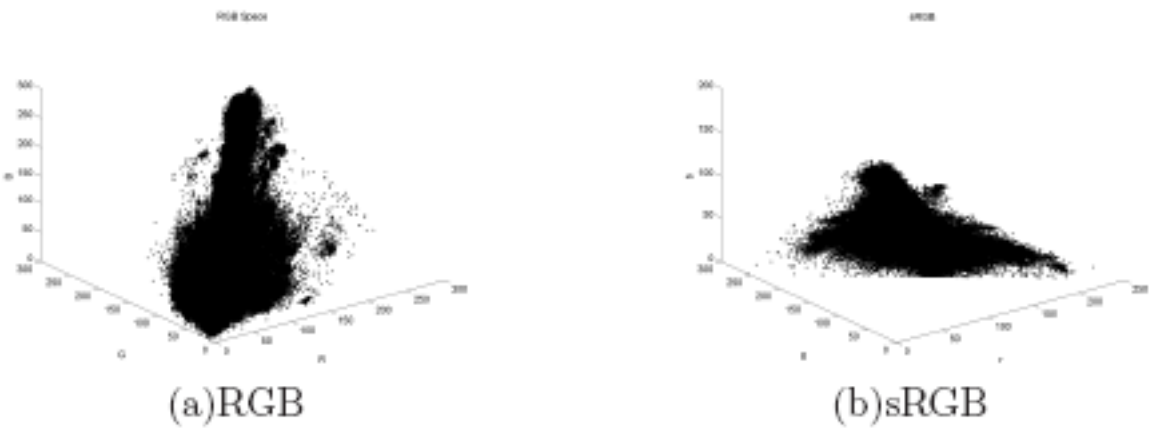


Figure 3.8: Mean block colour distribution after two-state boost in Germany vs Brazil in World Cup 2002

for $n \times n$ image block with the centre at (i, j),

$$mean(i,j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} C(i \times n + x, j \times n + y) \tag{3.26}$$

$$cov(i,j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} (C(i \times n + x, j \times n + y) - mean(i,j)) \tag{3.27}$$

where C is the pixel colour. The average covariance of a frame will be,

$$MeanCov_{frame} = \frac{1}{IJ} \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} cov(i,j) \tag{3.28}$$

where a frame contains $I \times J$ blocks. Based on the $N = 256$ bin histogram of average covariance distribution, a maximum entropy threshold is computed for homogenous vs non-homogenous image classification,

$$T_h = \arg\max_{N} \sum_{n=0}^{N} (-P_n \log(P_n)) \tag{3.29}$$

Where P_n is the portion of bin n over all histogram bins. Frames with a higher frame covariance than the threshold (Equation 3.29) will be rejected at the first layer of booster classifier. A similar threshold will be computed to remove blocks with high hue covariance in the remaining frames.

After collecting grass hue samples, a five-class Gaussian mixed model (GMM) is

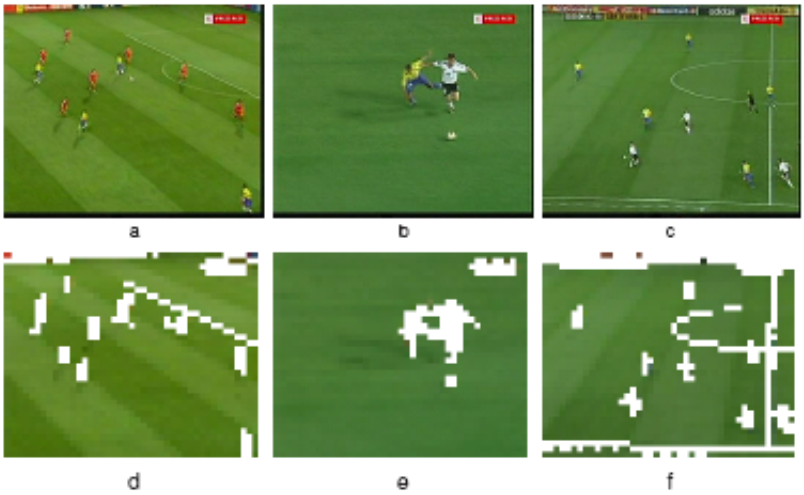


Figure 3.9: Grass area booster effect (Picture a,b,c are original images and picture d,e,f are respective result after boosting). Most non-pitch areas in these visual frames are removed after two-layer boosting, while game pitch areas are kept.

created by K-mean clustering to estimate a play field colour model. The class number is decided by the performance of shot style discrimination for video structure segmentation (Chapter 5) as shown in Table 3.4, where *play*, *focus* and *break* are shot styles for video structure decomposition.

GMM Class Number	Classification Precision			Average Precision
	Play	Focus	Break	
2	0.745	0.693	0.746	0.728
3	0.802	0.701	0.760	0.754
4	0.840	0.764	0.752	0.784
5	0.817	0.705	0.743	0.755

Table 3.4: Grass hue Gaussian mixed model with different class number

3.4 Game Pitch Orientation

The discrimination of pitch orientations includes two aspects: (1) identifying camera viewpoints; and (2) allocating the displaying area in an actual game pitch. The knowledge of camera viewpoint is helpful for video object detection, because camera viewpoints indicate possible appearances of video objects. For example, a goal post looks like a net texture from a camera behind the post or a white rectangle region in other viewpoints. Moreover, game contents are associated with the location, where an event takes place. For instance, a kick in a football game is named by its location, e.g. a

corner kick, a free kick and a penalty. Nevertheless, this orientation discrimination facilitates the estimation of projection models, e.g. a six-parameter affine model [Yow et al., 1995], which maps 2D visual images onto a 3D world and vice versa. The geometry of a game pitch has been well defined by sports regulations. This helps the identification of geometrical matching points, such as corners and pitch boundaries, to locate areas in a game pitch. The decision on the set of geometrical matching points is essential for robust estimation of projection models and the creation of a visual sports environment.

Many computer-assisted coaching systems treat the identification of game pitch orientations as an important prior step in a 2D-3D projection model estimation [Gong et al., 1995] [Yow et al., 1995]. Gong et al. [1995] divided a play field into zones with pitch legends, such as corner flags, a 18 yard box and a centre circle. The authors detected these video objects to locate the displaying area in a game pitch. The disadvantages of this approach are obvious. Firstly, many video objects are not unique. For example, there are four corner flags in a football game pitch. It is difficult to discriminate corner locations with a single corner flag. Secondly, video object detection is more difficult than area orientation discrimination, according to the theory of computer vision. This difficulty is partially due to object deformations and a low image resolution in broadcasting videos. Finally, the detection of video objects results in extra computational cost in a projection model estimation. In short, it is inefficient and ineffective to discriminate game pitch orientation by detecting pitch legends. Assfalg et al. [2002] proposed six zones for one half of a game pitch, or twelve zones in a complete game pitch. A set of geometry models were then developed to record relative positions between pitch boundary lines. The authors matched image edges to these models in order to determine the whereabouts of the camera and the pitch. Hence, an estimation of projection models is transformed into the optimisation on geometrical template fitting. The main drawback of this approach is an expensive computational cost. For example, Assfalg et al. [2002] have to try all twelve models, six boundary directions and numerous zoom depth before finding a possible optimised template. Moreover, it is inefficient and usually unnecessary for a content-based video application to build a precise virtual game pitch. This is because we just want to know what is taking place in the play field rather than how such an event happens, although it may be helpful to allocate the position where an event happens.

3.4.1 Orientation Classes

The number of possible camera viewpoints is definite in a game video. Baillie and Jose [2003] reported that more than six cameras are usually employed in a football video. These cameras are deployed around a play field and often allocated as follows: two camera behind goal posts and four parallel to pitch boundaries but on different sides. Considering pitch symmetry and variations in zoom depth, we propose six categories of pitch orientations (Figure 3.10), namely *horizontal, vertical, corner, close, low-close*, and *others* to classify visual frames. Note that these orientation classes are closely associated with game contents. In the horizontal orientation, a camera is roughly parallel to a game pitch and provides a global view of the game. The vertical orientation is for a camera located after or close to a goal post, which mostly occurs at corner kicks and goals. The corner orientation is a 45^o view and conveys group actions in a front field, especially when a corner kick takes place. Close and low-close classes are deviations of close-up shots, in which a camera focuses on a small area in a game pitch. A grass area or play field occupies most of images in the close viewpoint but only the deep bottom in the *low-close*. The *close* viewpoint is employed to display the swift motion of a small player group, e.g. break-through. The *low-close* is mostly a break in the game, such as the celebration after a goal or a throw-in on the pitch boundary. The category of *other* orientation refers to visual frames without any pitch area, for example, coach and spectator shots.

(a) Vertical (b) Horizontal (c) Corner

(d) Close (e) Low close (f) Other

Figure 3.10: Play Field Orientation

3.4.2 Region-based Grass Ratio and Classifier

The feature of region-based grass ratio is extracted for orientation discrimination. A visual frame is divided into 8×8 blocks and a grass field ratio on each block is estimated through the grass hue GMM model (Section 3.3). Hence, a 64D feature vector of region-based grass ratios is created for a visual frame. A three-layer feed-forward neural network (Figure 3.11) is developed for orientation discrimination, with 64 adalines (adaptive linear neuron nodes) at the input layer, 32 adalines at the hidden layer and 6 adalines at output layer. Each of output adalines stands for a class of orientations. The transfer function is a tangent sigmoid and the Levenberg-Marquardt back-propagation algorithm is used for training.

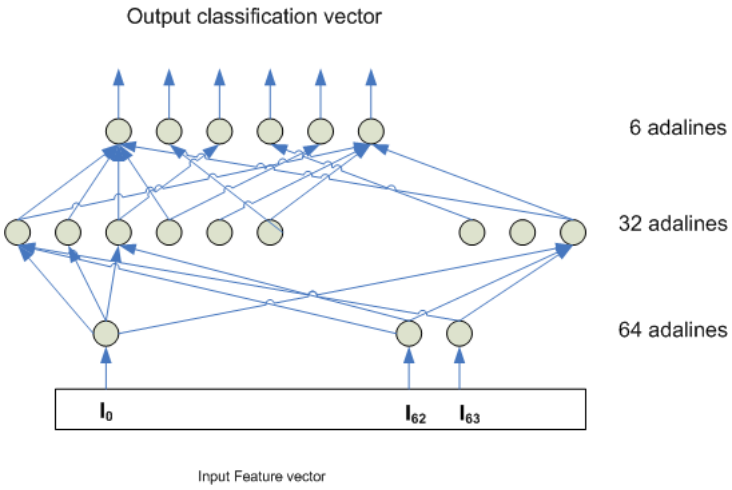


Figure 3.11: Neural network for orientation classification

3.4.3 Evaluation

The first half of the game, Japan vs Turkey and the final game, Germany vs Brazil, in World Cup 2002 are both manually labelled as the ground truth. 13462 frames in all are sampled at the rate of $1/25$, including 2782 *horizontal* frames (20.7%), 476 *vertical* frames (3.5%), 1277 *corner* frames (9.5%), 2977 *close* frames (22.1%), 1276 *low-close* frames (9.5%) and 4674 *other* frames 34.6%. We randomly selected 400 frames from the *other* category and 100 frames from each of left five classes to train a neural network (900 frames in total), and keep the remaining frames for evaluation. The training and evaluation results are displayed in Table 3.5.

	Horizontal	Vertical	Corner	Close	Low close	Other
Training Set	0.980	0.870	0.890	0.920	0.960	0.967
Evaluation Set	0.971	0.861	0.762	0.935	0.981	0.932

Table 3.5: Orientation Discrimination Precision

3.5 Zoom Depth

Zoom depth is a term from photography, which is a measurement on the quality of visual details and reflects a sight field. Ma et al. [2002] proposed that this feature is roughly propositional to *attention* intensity. This is because it is easy to attract notice by displaying more details than usual [Picard, 1997] [Bordwell and Thompson, 2004]. Therefore, zoom depth are supposed to be a salient visual feature, not only in sports videos, but also in story films.

An approach of zoom depth estimation is to compare the size of an object in visual frames with that in the real world. The size ratio of a video object over that in real world is anti-propositional to zoom depth, according to the definition of zoom depth in photography. Therefore, major challenges in the estimation of zoom depth is how to select a proper video object and how to measure the object size. Note that the shape of a video object is deformable in visual frames (Section 3.4) because of nonlinear 3D-2D projection. Although the discrimination of pitch orientations (Section 3.4) may facilitate an adjustment of object deformation by providing a rough assumption of camera position, it is difficult and usually computationally expensive to estimate an actual object size from object images. Many video objects have been proposed for zoom depth estimation, such as a human face, a football [Seo et al., 1997] and a penalty circle Ekin et al. [2003]. One of advantages these objects have is the circular shape, which is geometrically invariant during the projection from the 3D real world to a 2D image. However, these propositions have inherit limitations in applications: a football is too small to be detected and measured; the appearance of penalty circles is scarce because of their geographical locations; the detection of human faces is challenged by face rotation, besides the problem of small size and low image resolution.

We propose that a sports uniform is an obvious domain feature in sports videos. This video object has advantages for the estimation of zoom depth as follows.

1. There are guidelines for uniform dressing. For example, the FIFA regulation states that a player uniform should have bright colours and special patterns so that such a dressing can be easily discriminated by spectators. This means that

the video object of a uniform is able to be detected by the colour and texture easily.

2. Players always wear uniforms in a game pitch. This means that a uniform is a pervasive video object in sports videos.

3. The video object of an uniform is rotation invariant because of the dress code. It is impossible to wear a uniform bottom-up.

4. A player uniform is bigger than a football or a human face. In most cases, the video object of a uniform is large enough to be measured.

3.5.1 Uniform Detection

In the data collection of FIFA World Cup 2002, the size of a uniform varies significantly from 9×13 pixels to more than 180×150 pixels in 352×288 video frames. For convenience, thirteen scales (from 0 to 12) are used to measure video object sizes. A Foley-Sammon Transform classifier (Appendix A) is trained to detect player uniforms. This classifier can reach the global optimisation under the Fisher criterion, when only positive samples are available 3-0If a sample satisfies the requirement of a classifier, this sample is called as a positive sample and vice versa. In the training process, both positive and negative samples are essential to determine the ability of a classifier. However, a negative sample set is sometimes difficult to be defined. For example, a human face detector classifies images into two categorises, face and non-face. The image with *face* is easily identified by human, but images with *non-face* can hardly be counted. This problem is referred as imbalance classification, in which only one type of samples, such as positive samples, is provided.. An 11-layer pyramid is built to offer a multiple view of an image, in which every layer is 1.25 larger than its prior layer. The bottom layer is at 352×288 pixels, while the top layer is at 32×26 pixels. The uniform detector scans this image pyramid from top to down. For instance, if a polo shirt is found on a certain layer, e.g. the second layer, this frame will be labelled with a zoom size of two. If a polo shirt is not found in any layer, the zoom size will be set as zero. To improve the precision, a median filter at the scale of five is employed to smooth the sequence of uniform size.

A uniform training set is manually collected from FIFA World Cup 2002, which includes more than three hundred 9×11 pixel samples from different viewpoints. A part of the sample set is shown in Figure 3.12.

Figure 3.12: Player uniform samples

3.6 Low Level Salient Features

This section discusses low level audio-visual salient features, including motion salience, colour salience, audio short term energy, speech zero cross ratio and low short term energy ratio. Most of these features can be directly extracted from audio and visual streams. Additionally, the relationship between *attention* and these features will be addressed.

3.6.1 Motion Salience

In video analysis, the motion is a measurement of spatial changes between adjacent visual frames. For example, Gu et al. [1998] simulated motion variations to detect slow motion video segments. Moreover, the feature of motion 3-0This motion concept in perception is mostly an actual movement or relative position changes among objects. This definition is slightly different from that in video analysis. is regarded as one of most significant issues in perception. Some psychological experiments prove that swift motion is correlated with mental excitements because a quick movement in sight means danger in a wild life [Simons et al., 2003].

Motion salience is proposed to estimate the *attention* intensity caused by a motion as-

pect [Adams et al., 2000] [Duan et al., 2003]. However, there is no agreement on the computation of motion salience. Ma et al. [2002] and Hanjalic [2005] supposed that the mean of motion vectors is approximate to motion salience. But Wang and Cheong [2006] argued that a quick movement was conventionally associated with danger and excitement, and thus took the maximum amplitude of motion vectors for salience estimation. Moreover, there are two different motion descriptions in video analysis, namely motion vectors and the number of changed pixels. A motion vector contains two components (x,y) to state a 2D displacement between two similar image blocks in adjacent visual frames. The amplitude of a motion vector is defined in Equation 3.30.

$$d = \sqrt{x^2 + y^2} \tag{3.30}$$

A spatial accumulation of motion vectors is referred to as a motion field. The motion salience in [Ma et al., 2002] is

$$d_{vector} = \frac{1}{IJ}\sum_{i}^{I}\sum_{j}^{J} d_{i,j} \tag{3.31}$$

where $d_{i,j}$ is the amplitude of motion vector at the image block (i,j); I,J are the numbers of vectors in a motion field.

The calculation of motion vectors is decided by three variables, block size, tracking depth and searching pattern [Cheung and Po, 2003]. Apparently, the extraction of motion vectors is a computationally intensive task. Although motion fields in encoded visual streams are coarse, we extracted these vectors directly from compressed fields.

The number of changed pixels d_{pixel} is a measurement of image dissimilarity rather than an estimation of video object movements. To reflect a variation in perception, the computation of changed pixel number is carried out in the perceptually unified *CIE Luv* colour space. Therefore, a *Luv* distance d_{luv} between two pixels is,

$$d_{luv} = \sqrt{\frac{1}{3}(s_l(L_1 - L_0)^2 + (u_1 - u_0)^2 + (v_1 - v_0)^2)} \tag{3.32}$$

where s_l is a scaling factor to enhance the sensitivity of luminance. Pixels, whose d_{luv} exceed a given threshold, are counted by the difference measurement d_{pixel}.

The salience measurement of motion vector and changed pixels are combined to com-

pute the feature of motion salience in this thesis (Equation 3.33).

$$M = d_{pixel} \times d_{vector} \tag{3.33}$$

3.6.2 Colour Salience

Colour is one of the most important facts in vision. Valdez and Mehrabian [1994] claimed that in colour perception, brightness is propositional with valence, while the saturation is correlated with arousal. For example, a red colour incurs stronger reactions than a blue colour does. Wang and Cheong [2006] proposed a colour salience measure (Equation 3.34) to estimate the perceptual effect of a colour in the *HSV* colour space,

$$C = \sum_i \sum_j p(c_i)p(c_j) \times d(c_i, c_j) \times \sum_{k=0}^{M} E(h_k)s_k v_k \tag{3.34}$$

where c is a bin in the histogram and indexed by i, j; $p()$ is the pixel ratio of a given bin; $d(c_i, c_j)$ is a 2-norm distance between two bins (c_i, c_j); M is the number of pixels; and s_k,v_k are the saturation and lightness of a pixel at an index k.

3.6.3 Audio Base Band Energy

A loud sound always attracts attention. This fact indicates the close association between audio energy based features and auditory salience. To improve temporal discrimination, Wang and Cheong [2006] used the feature of short term energy to estimate the loudness in a short period. For a stereo audio track, this feature is defined by Equation 3.35.

$$AE = \sum_{i=0}^{I} (r_i + l_i) \tag{3.35}$$

where r and l stand for the audio energy in the left and the right track, respectively. I is a reflection period of audio stimulus [Treisman and Kanwisher, 1988]. The minimum of I is the number of samples in 0.3 sec. For example, the value of parameter I is 13230 in MPEG audio layer 2 3-0MPEG audio layer 2 is abbreviated as MP2, which is the most popular audio encoding format in MPEG videos., because the audio sample rate is at 44.1 KHz. Moreover, Hanjalic [2005] suggests a relatively long period for audio energy computation, e.g. 1 sec and 5 sec. This is because the average energy on a long period can remove signal noise and thus enhance the signal to noise ratio.

Besides the loudness, audio variation also incurs interest among audience. The fea-

ture of a low short-time energy ratio (LSTER) in Equation 3.36 is proposed to represent the variation of short term energy.

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(0.5\overline{AE_2.4} - AE(n)) + 1] \tag{3.36}$$

where N is the number of 0.3 sec audio frames; $\overline{AE_2.4}$ is the average short term energy in a 2.4 sec neighbourhood; $AE(n)$ is the short term energy in the n^{th} audio frame; $sgn()$ is a sign function.

$LSTER$ is effective in a clear audio environment, such as only commentator speech or only spectator noise. This feature is useful in the detection of abrupt loudness change, for example, a switch between stadium microphones, which causes an abrupt jump in the average audio energy.

Another interesting audio feature is the ratio of noise frames in a given audio clip [Lu et al., 2002]. The judgement of noise frames relies on the ratio calculation of the maximum local peak over normalised correlation function in an audio frame. If such a ratio is lower than a predefined threshold, e.g. 0.3, this frame is considered as a noise frame. The ratio of noise frames is reciprocal to the percent of speech and music components in a given audio clip. Hence, the feature of noise frame ratio is qualitatively anti-propositional to auditory salience.

3.6.4 Speech Zero-crossing Ratio

Zero-crossing ratio (ZCR) is the sign-change rate of a temporal signal. This feature is employed to detect unpronounced stops and speech trucks Cole et al. [1995]. Lu et al. [2002] proved the effectiveness of ZCR in audio characterisation, especially in the classification of music, speech and environment noise. Tsekeridou and Pitas [2001] developed an audio-based video index, which ranked ZCR values as a measurement of content importance. The authors supposed that quick speech indicated an excited mental state and thus hinted an important moment in a game. Since ZCR counts the number of zero passing in a given period, this feature is propositional to the speed of speech. Therefore, speech ZCR is a salient feature, which is associated with mental states.

Although ZCR is useful in speech analysis, this feature is sensitive to noise. An audio track in sports videos is a mixture of commentator speech, spectator cheers and other sounds from environments, such as music and echos in a stadium. This means

that the audio stream in a broadcasting video is not clean according to audio processing techniques. Additionally, it is a common scene in sports videos that a commentator will adjust the volume of audio inputs abruptly to ensure that comments are recognisable. Therefore, a variation of ZCR is developed, which is more discriminative and robust than exact value of ZCR (Equation 3.37).

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(ZCR(n) - 1.5\overline{ZCR_1}) + 1] \tag{3.37}$$

where n denotes an audio frame index, $ZCR(n)$ is the zero-crossing rate at the n^{th} frame, ZCR_1 refers to the average ZCR in a 2.4 sec window; $sgn()$ is a sign function.

3.7 Summary and Discussion

In this chapter, a set of feature extraction algorithms are presented and used for the later video structure detection and salience computation. These features include shot frequency, play field ratio, game pitch orientation, motion salience, colour salience, audio short term energy, low short term energy ratio, and zero-crossing rate. Related research topics are addressed, such as video shot segmentation and visual object detection. Moreover, some techniques can be employed in other applications. The FST classifier is an efficient approach for general video object identification, e.g. corner flags and human faces [Guo et al., 2003], though the performance requires extra careful evaluations. Weibull-based and Erlang-based temporal adaptive weighting helps the estimation of video content importance by predicting event duration and event occurrence frequency. For instance, an interview TV program contains shots of interviewee and interviewer. These two types of shots are switched regularly; and shot durations reflect the importance of video contents. Thus a temporal distribution hypothesis can be used to detect important talk topics.

These features addressed in this chapter are not sufficient to cover the literature of salient fusion. The context of perception and audio-visual stimuli is so complex that many advanced techniques have been developed to deal with information ambiguity. Such an ambiguity refers to the uncertainty in emotion estimation. For example, an auditory reflection varies with sound genres. Audio classification and segmentation will provide useful and sometimes decisive messages for both content understanding and salience computation [Lu et al., 2002]. A classification of speech, music, silence and environment noise [Pfeiffer et al., 1996] helps the choice of *attention* models and de-

fines the context for salience fusion, such as enjoying music and understanding speech. In the situation of music enjoyment, an analysis of audio pitches and music tempos are of prominent importance in salience computation [Hua et al., 2004], while key word detection and content reasoning play a prime role in the mode of speech understanding [Xu et al., 2003] [Kang et al., 2004]. However, this thesis does not concern these research topics, partially because the semantics of sports videos is relatively simple and data collection is full of noise in audio tracks. For example, an audio track in football videos is a mixture of comments, spectator cheers, environment noise and stadium echoes. It is usually inefficient to discriminate these noisy audio segments.

In the following chapters, I present my work in the identification of content-based video structures, including replay segments (Section 4), attack structures (Section 5) and attention decomposition (Section 6). One of major differences between video retrieval and text retrieval is the retrieval output. In text retrieval, we type in a set of key words and get a whole document as a result. However, such a scene in video retrieval can hardly be repeated. There are two obvious reasons: (1) it is inefficient to supply a whole video because such a video document may be hundreds of mega bytes; and (2) a small video clip conveys most retrieval requirements. This indicates that an knowledge element in video retrieval is only a small part of a long video rather than an entire video. In text retrieval, we compute many semantic distribution features, e.g. term frequency, in a document, because we assume that the fundamental knowledge unit is a whole document rather than a section or a chapter. Hence, a syntax or semantic concept statistics on an entire video might be meaningless for retrieval. This is because such a computation are across multiple knowledge elements. Comparing with text retrieval, such a deed is ridiculous to compute a single term frequency of many non-related documents. Therefore, it is prerequisite to find knowledge units in a video for the computation of syntax frequency and other semantic distribution features. This means that the identification of content-based structures are of importance in video retrieval.

4

Replay Detection

This chapter presents a system for replay detection in sports videos. Replay is a unique video editing style in sports videos, which is only employed to convey essential video contents. It is an efficient and effective approach for sports video summarisation to accumulate replay segments in a sports video [Gu et al., 1998] [Xu et al., 2001] [Pan et al., 2002]. Therefore, the detection of replay segments becomes an interesting research topic in sports video analysis.

4.1 Introduction

Replay is a special video editing style, which popularly exists in sports videos whilst is rare in other video genres, such as news videos and story films. To some extent, this editing style is unique for sports videos. There are several reasons to explain this specification. Firstly, sports videos are characterised with strong and fast group competitions. It is difficult to completely present video contents at a normal display speed because of quick local motions. Secondly, the employment of replay provides a second chance to reiterate a past story at the cost of current normal contents. However, such a cost can hardly be afforded in many video genres, such as news videos. Nevertheless, a replay segment inserts extra video segments into a continuous video sequence and thereby breaks the temporal continuity of video contents. This may hinder viewer

understanding. Therefore, video directors have to carefully choose replay segments and avoid a frequent employment.

Sports videos are characterised by relatively simple and predictable contents. This indicates that the loss of a small part of the video sequence will not trouble the enjoyment on a complete game story. Hence, replay is regarded as an efficient approach to iterate game stories. This is because: (1) replay segments can provide more visual details than general video segments; (2) as a prerequisite, only sports highlights will be replayed, a frequent occurrence of replay segments indicates the attractiveness of game contents; (3) replay extends the temporal duration of highlights and keeps viewers excited. Therefore, replay detection is a reliable approach for sports highlight discrimination. Ekin et al. [2003] identified game highlights by the occurrence of a replay segment. Pan et al. [2002] treated an accumulation of replay segments as an acceptable video summary. However, replay is a skill of video editing. With the advancement of video editing tools, more and more new styles of replay segments are introduced, such as original segment reinsertion, slow-motion replay, high speed recorded segment insertion, and multiple viewpoints switch.

The remaining chapter is organised as follows. Section 4.2 surveys the literature of replay detection and explains the proposal to take logo transition detection as a replacement of replay detection. An adaboost classifier for logo transition detection is presented in Section 4.3. Experimental results and conclusion are found in Section 4.4 and Section 4.5, respectively.

4.2 Related Work

Replay detection is a specific research question in sports video analysis. Gu et al. [1998] categorised replay segments into four groups according to visual appearance: replays from the same camera (exact replay); replays from another camera; slow motion replays from the same camera; and slow motion replays from another camera. This classification was based on video sources and the usage of slow motion skill, although many new visual aspects have been introduced in the composition of replay segments. Several approaches were proposed to detect these replay classes individually. Exact replay is a copy of prior shots. Therefore, the detection of exact replays is a direct comparison on visual similarity between shots. Gu et al. [1998] claimed that if a shot could be matched at least twice, this shot would be an appearance of exact replay. Moreover, to speed up this comparison, a two-stage matching algorithm was developed in the MPEG com-

pressed field to search a local minimum of frame sequence difference sum. However, it is difficult to predefine a threshold to discriminate similar shots. In sports videos, a game pitch usually occupies a large ratio of visual frames. This indicates that many visual shots are highly similar in both colour and texture. Given image noise and encoding bias, the decision of visual similarity threshold is mostly empirical and usually fragile. Nevertheless, this exact replay becomes rare in sports video broadcasting with the advance of video production techniques.

There is another approach for replay detection by identifying slow-motion segments. Gu et al. [1998] claimed that a slow-motion was actual multiple insertion of a same visual frame, which decreased the visual frame updating rate, e.g. from 25 frames per second to 5 frames per second. Therefore, the authors counted the ratio of **P** and **B** inter-coded frames in the MPEG compressed stream, because this repetitive insertion results in a lot of similar image blocks and leads to the appearance of inter-coded **P** and **B** frames. However, high speed cameras become more and more popular in sports video production [Baillie and Jose, 2003]. Such a repetitive insertion of previous visual frames are replaced by a normal speed replay of video segments which have been recorded by high speed cameras. Hence, the statistics on inter-coded frame ratio is not so effective as before. Kobla et al. [2000] proposed a zero crossing ratio to evaluate the amplitude of fluctuations in the time sequence of frame difference. The authors employed linear classifiers to discriminate slow-motion clips. Later, Pan et al. [2001] analysed the temporal structure of replay segments and developed a sandwich pattern for replay detection, including a logo-transition, slow-motion segments, normal speed replay, and another logo-transition (Figure 4.1). Therefore, a four-state Markov model

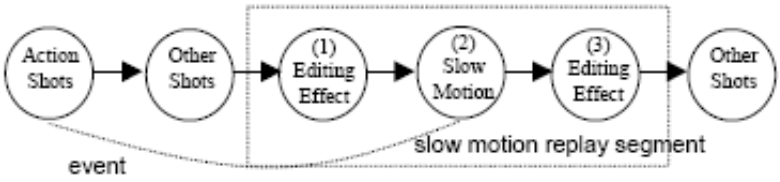


Figure 4.1: Structure of slow-motion replay [Pan et al., 2001]

(Figure 4.2) is designed to simulate that this temporal pattern.

With the advance of composition skills, the production of replay segments becomes more and more complex. Firstly, more shots are involved in a replay segment than before. In the collection of FIFA World Cup 2006, a replay segment usually consists of several shots from different viewpoints and from different replaying speeds. For ex-

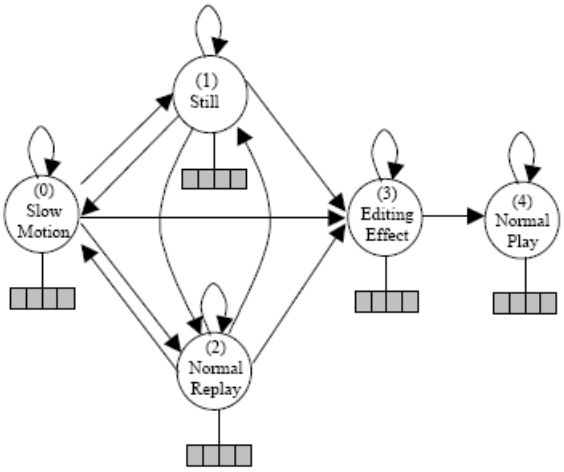


Figure 4.2: Hidden Markov Model for Replay Detection [Pan et al., 2001]

ample, in the final game, Italy vs. France, a replay for a goal event is made up by two or three viewpoints, i.e. behind goalpost, close-up, and by-side, and some slow-motion or normal speed segments occasionally. This challenges above techniques for replay detection.

Pan et al. [2002] proposed an alternative approach of transition logo detection (Figure 4.3) as a post-processing step, which improves the precision of slow-motion detection. The authors searched similar images before and after slow-motion segments to build up a template of transition logo. However, it is unnecessary for video compositors to attach a logo transition just before and after a slow motion segment, as shown in Figure 4.2. Additionally, a replay segment may not contain any slow motion clips in many cases.

Figure 4.3: Logo Transition in World Cup 2002

4.3 Logo Transition Detection

A logo transition is an artificial visual effect sequence (Figure 4.3). Such a sequence marks the start and the end of a replay segment and is widely employed to facilitate viewers to identify replay segments in video collections of FIFA World Cup 2002/2006 and UEFA Champion 2006. Therefore, it is effective to detect replay segments by allocating logo transitions.

The transition duration is usually less than 1 sec or 25 frames 4-0The duration of a logo transition is mostly constant in a game video, although a small frame number changes exists between game videos and collections. For example, a logo transition is 20-22 frames in World Cup 2002, about 21 frames in World Cup 2006 and 18-20 frames in Champion 2006. and constant in a game video. A logo is characterised with apparent colour and occupies most area of a visual frame (Figure 4.3). Hence, five features are proposed for logo detection as follows.

- Shot duration. Logo transitions are shots, whose length is less than 1 sec. Xu et al. [2006] claimed that logo transitions were of the same frame number in a game video because these segments were automatically inserted by an editing software.

- Sequential colour histogram. The colour of a logo is rare, such as silver yellow. Moreover, a logo is big enough to affect the distribution of a colour histogram. Therefore, a sequence of colour histograms is a vector $\overrightarrow{H} = \{h_i, i = 0...K\}$, where K is the sequence length, and h_i denotes the colour histogram of image i. The sequential colour histogram distance is defined in Equation 4.1.

$$dist(\overrightarrow{H_1}, \overrightarrow{h_2}) = \sum_{i=0}^{K} Inter(h_{i,1}, h_{i,2}) \tag{4.1}$$

  where $Inter()$ refers to the intersect distance between histograms (Section 3.2.1).

- Shot frequency between two transitions. A replay segment is sandwiched by two logo transitions. Since a replay shows many different viewpoints and thus contains many shots in a relatively short period, shot frequency in a replay segment is significantly higher than average.

- Average motion. A replay usually contains slow-motion and close-up segments, in which the magnitude of motion vectors are small. Therefore, average motion is computed as the mean of magnitude of motion vectors in a shot.

- Game pitch ratio (Section 3.4). A logo occupies most area of visual frames in a logo transition, which results in a very low game pitch ratio.

However, the segmentation of shots is an empirical and complex task (Section 3.2). It is meaningless to measure the duration of a logo transition or a shot at the precision of a visual frame. This is because: (1) video decoder bias, there is one or two frame random difference in decoding between different video decoders; (2) a logo transition is highly similar to a fly-in/out shot transition and thereby ignored as a dissolve (Section 3.2). Therefore, a classifier cascade or an Adaboost classifier is developed to detect logo transitions, in which each classifier only uses one feature. The system framework is shown in Figure 4.3.

We employ a simple two-class classifier with a constant threshold to discriminate

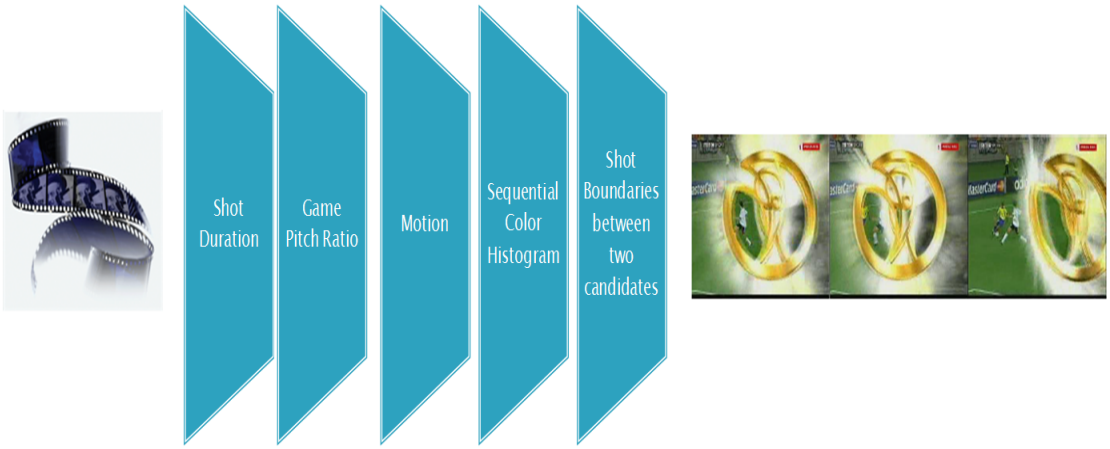


Figure 4.4: Logo Transition Detection Framework

features, including shot duration, pitch ratio and average motion. Most thresholds are calculated as a ratio over the average statistics in a video. Logo transition shot duration is less than 1 second; pitch ratio in a logo frame is below 0.4 of average game pitch ratio; the average motion in a logo transition shot is larger than 2.0 times of average motion. Before sequential colour histogram comparison, all video segments less than 10 frames are removed. We take the maximum of sequential histogram distances to measure colour similarity between two different length video segments. A Gaussian model is created to simulate the distribution of sequential histogram distance. All segments out of 0.7 credit will be rejected. Shot frequency between two logo transitions is employed as an evaluation for replay detection.

4.4 Experiment

Four games from World Cup 2002 are selected for logo-transition evaluation, namely Brazil vs Germany(final), Japan vs Turkey, Germany vs Korea(semi final) and England vs Sweden. Note that the logo-transition detector in this thesis requires little training.

Table 4.1 shows the result of logo transition detection. Note that the false alarm is very low. This indicates that we can reach a 100% precision by appending a post-process step, which extracts a logo transition template from classifier results and then allocates all logo transition segments.

	Ground Truth	Detected	False	Miss
Brazil-Germany	66	66	0	0
Japan-Turkey	68	63	0	5
Germany-Korea	82	81	0	1
England-Sweden	90	89	1	2

Table 4.1: Logo Transition Detection

Table 4.2 compares the precision and recall of replay segment detection among algorithms in [Pan et al., 2002], [Ren and Jose, 2005] and this thesis. The algorithm in [Pan et al., 2002] is regarded as the base line for replay detection, which developed a five-state hidden Markov model to simulate slow motion segments. Another base line is our old system in [Ren and Jose, 2005], which extracted a sequential template of logo transition by slow motion detection in order to improve recall of replay detection. Both algorithms need intensive training. However, the adaboost classifier proposed in this thesis significantly out-performances these prior systems.

	Slow motion HMM		Mixed algorithm		Logo-transition Detection	
	Precision	Recall	Precision	Recall	Precision	Recall
Brazil-Germany	0.37	0.86	0.42	1.0	1.0	1.0
Japan-Turkey	0.59	0.92	0.67	1.0	0.97	0.91
Germany-Korea	0.54	0.96	0.58	1.0	0.98	0.98
England-Sweden	0.44	0.89	0.45	1.0	0.95	0.98

Table 4.2: Replay Detection Performance

4.5 Conclusion

In this chapter, we present an automatic effective approach for replay detection. Avoiding complex slow motion simulation, this system is based on logo transition detection. An efficient five-layer Adaboost classifier is created and filters visual segments with shot duration, average game pitch ratio, average motion, sequential colour histogram and shot frequency. A high precision and recall of logo transition detection are achieved in the experiment as well as replay detection. Therefore, this system is employed as a successful replay detection module in the following chapters. Nevertheless, an important lesson is gained in this chapter, how to deal with the time issue in the content analysis. It is a difficult task to detect slow motion in a visual frame sequence. This is because visual frames are temporal samples which are taken at a given interval, and hardly reflect the speed variation of motions. Relatively, the classifier cascade developed in this thesis is simple and efficient. The main advantage is that this classifier set pay attention to the temporal accumulative effect of replay segments. This indicates that a different temporal resolution may lead to a simple solution to many problems in content-based analysis, such as video content identification and content presentation.

However, replay is a editing skill. Numerous editing styles exist and vary with director preference. It is hard to describe replay structure precisely. For example, a replay segment can ignore any components in the sequential model (Figure 4.1), such as slow-motion, and fast viewpoint switch shots. Moreover, in UEFA Champion 2007, a new replay style was developed, which replaced logo transitions with two slow motion segments. This style is beyond the capability of the system addressed in this chapter. Although it is possible to develop some new techniques for the detection of general replay segments, such as attention analysis (Chapter 6), the most efficient approach is to involve a content-based tag of replay in the standard of video production, e.g. MPEG-7.

5

Attack Temporal Structure

> The football video is a loose structured time sequence with relatively simple contents.
>
> *Jürgen* Assfalg (Semantic Annotation of Sports Videos 2002)

A sports video is a personalised record rather than an evident iteration of game stories. This is because intensive individual understandings from camera recorders, video directors and compositors are mixed in such a video: camera recorders watch an on-site game and decide a convenient camera angle and a proper presentation method, such as pan and zoom-in, to record game facts; video directors gather camera videos from multiple recorders and select an appropriate video sequence by switching shots; compositors sit in an auditory room and re-organise these records by inserting editing effects, e.g. slow motion, replay and closed captions, to complete a broadcasting video and related highlight documents. The process of making a sports video is a aggregation work of individual creations, which make a game video more interesting. Therefore, the production of sports videos has to follow some general guidelines to guarantee video integrity so that various producer preferences will not hinder the presentation of game contents. This fact explains many interesting observations: (1) content-based video structures are available in finalised game videos rather than raw data; (2) content-based video structures can be detected by statistics on an appearance sequence of camera shots

[Xie et al., 2004] [Tjondronegoro et al., 2004*a*]; (3) game video structures are highly similar as long as these videos are composed by the same group of producers; (4) the composition of these structures vary or depend on data, because of variations in game contents and producer preference; (5) a game event is a special video structure, which interrupts a general video presentation sequence and inserts its own style. Baillie and Jose [2003] proposed a conceptual image of football video organisation by close observation. The authors developed a transition graph (Figure 5.1) among video contents and encapsulated a temporal flow between major broadcasting components, such as game, studio and advert sequences. Several extra sub-structures, e.g. reports, interviews and general events, are labelled in the graph together.

Figure 5.1: Temporal flow of broadcasting sports videos [Baillie and Jose, 2003]

Although such a content graph (Figure 5.1) is mostly a conceptual abstract, a plenty of temporal patterns are easily observed. For example, a football video starts with an opening title, such as a flash of the game logo with strong music and some adverts, before moving into a studio. In a studio section, an interviewer or an anchor makes a brief introduction, typically a list of players and coaches. There could be extra reports, player interviews or a game directly. A game sequence records a half of the match, which lasts 45 minutes, but an additional temporal duration may occur, which varies from fifteen to thirty minutes including other post-processing effects. Sequentially, an advert or a studio segment appears because of an inter-section break. Such a video sequence, consisting of a title section, an auditing segment and a game sequence, repeats until the appearance of closing titles which ends a broadcast. In short, this conceptual graph

displays an overall structure of a broadcasting football video and proves the availability of content-based video structures. This indicates structure-based approaches are effective in both the discrimination of video contents and the detection of content events. However, this conceptual model is difficult for employment, because this model is too general to convey any detailed game contents. Some advanced video structures are necessary in content-based video analysis.

The remaining sections are organised as follows. Section 5.1 surveys the literature of structure analysis, including video content simulation, temporal video model, event pattern and related model extraction techniques, such as controlled hidden Markov model and hierarchical hidden Markov model. The challenges in sports video content modelling are addressed in Section 5.2. The proposition of *attack* structure and structure components are presented in Section 5.3. Section 5.4 is dedicated for the segmentation algorithm and the performance can be found in Section 5.5. Several applications are introduced in Section 5.6, which demonstrate how to employ *attack* structures in semantic video retrieval and content-based video retrieval. Conclusion and a short discussion are found in Section 5.7.

5.1 Related Work

Structure analysis mines similar content-based sequential patterns or video structures among video streams. This technique is useful to (1) clarify of video content indications; (2) divide a long continuous video into a series of relatively independent video segments; and (3) facilitate the discovery of content events. Additionally, these sequential patterns can be deterministic or stochastic, because of the complexity in content presentation and video production scheme.

A transition between two game events experiences little temporal delay. This character shows that temporal patterns of game events are densely distributed and can cover a sports video entirely. Two additional advantages are found as follows.

- A long data stream can be sufficiently represented as an alternative sequence of temporal structures;
- Each of constituent structures can be modelled with a unified parametric class.

According to the temporal scope of structure models, video structures can be categorised into two classes: a generic model which covers an entire game video, and an

event pattern which describes special moments in a game. A generic model only exists in a densely distributed structure data. This model is usually a unified stochastic network, such as a hierarchical hidden Markov model [Xu et al., 2005], a coupled Markov model, a two-state *play-break* Markov chain Xie et al. [2002], and a hierarchical hidden Markov model with Bayesian basket [Xie et al., 2004]. Additionally, the four-state *attack* hidden Markov model [Ren and Jose, 2005], which is presented in this chapter, is a generic model for football videos. To some extent, the conceptual model in [Baillie and Jose, 2003] is a special case of generic models, although Figure 5.1 presents a common time sequence in video production and is too general to discriminate video contents. The main challenge of generic modelling is the trade-off between model generality and efficiency. The generality requires that a content model can adapt to game videos as many as possible. This requirement favours a model with a small state number. However, efficiency demands a large number of model states to describe complex video contents. Therefore, a general model is usually inefficient to discriminate content-based video events, because the requirement on model generality is met at the cost of model fitness on data. The conceptual model in Figure 5.1 is an extreme example. Such a model keeps its fitness for any broadcasting sports videos, but loses the ability of game event identification.

Event patterns limit their scope in certain moments of a game, such as goals and corner kicks, rather than a complete video. Such an approach avoids the complexity in the simulation of a long stochastic process and increases discrimination rate for semantic events. This method leads to a temporal pattern for event detection whilst ensuring an acceptable model generality. Numerous event patterns or detectors have been developed, e.g. audio-visual keywords matching for goal detection Kang2004, a controlled Markov chain for highlight detection [Lenardi et al., 2004] and an audio-visual event sequence for goal detection [Huang et al., 1998]. However, there are numerous game events. This indicates that a large number of event patterns have to be developed for the description of game contents.

Both generic models and event patterns are successfully applied in content-based video analysis. Generic models introduce prior domain knowledge and filter out unimportant video contents. These models facilitate the detection of sudden video events. For example, the play-break model [Xu et al., 2001] [Xie et al., 2002] [Xie et al., 2004] efficiently removes non-game video segments and could save up to 60% of the storage space [Xie et al., 2002]. An event pattern is a temporal-spatial detector for semantic events. Such a pattern allows the simulation of a temporal sequence of knowledge transition among

multiple information modalities. For instance, Kang et al. [2004] employed audio, visual and text information during the search for goal events. This approach observes a complex content-based video event carefully and allows heuristic regulations to improve detection performance. A set of event patterns are able to create a content space for video indexing and leads to a methodology of event-based video analysis, which treats a game video as a sparse set of game events. In other words, a content-based reasoning network becomes a necessary postprocessing step in event-based video analysis, because we have to link these events to describe game contexts.

5.1.1 Generic Video Model

In a football video, game contents are densely distributed. This indicates a game video can be fully decomposed into video segments with certain semantics or content-based video structures. It is possible to develop a content-based video model in the domain of sports videos. Although a model containing multiple modalities is helpful in content analysis, most generic video models limit their scope on a single media modality, either visual stream or audio. This is due to the complexity in content presentation and media asynchronism between modalities.

One of the most important video structure sets is the visual pattern of *play* and *break* (P-B structure) [Xu et al., 2001], where *play* refers to normal game clips, and *break* represents a stop in a game, such as field-away shots on coach and spectators. Xu et al. [2001] employed a simple low-level visual feature, grass pixel ratio (Section 3.3), to discriminate visual frames into play and break classes by a Gaussian mixed model (GMM). This is because video producers have to focus on the game pitch to record game status and vice versa. Hence, the ratio of a game pitch in a visual frame can be used to discriminate visual contents. The P-B structure was originally proposed for content filtering [Xu et al., 2001]. Later, Tjondronegoro et al. [2004*a*] found such a visual classification was helpful in the sports highlight detection and content indexing. The authors counted the duration and frequency of *break* clips as a prior possibility for sports events. A set of heuristic rules were developed to detect highlights from P-B segmentation. This work prove the effectiveness of a P-B structure in content discrimination.

To improve the precision of P-B structure segmentation, Xie et al. [2002] advanced the GMM classifier with a post-processing step of dynamic programming (Figure 5.2). This post-processing step has merits as follows. Firstly, the labelling operation of P-B

structure allows multiresolution analysis on visual segments as well as visual frames. For instance, we can calculate the average grass ratio over a shot and smooth the label sequence at the temporal resolution of visual shots. Secondly, dynamic programming is a high order Markov chain. The depth of dynamic programming reflects the shortest duration of a video structure. Such an operation takes the context of a video structure into consideration and removes errors in visual frame classification if the error sequence is short.

Another advantage of P-B structure lies on the fact that this video structure can be learnt automatically from video data. Xie et al. [2004] improved a hidden Markov model to multiple levels (hierarchical hidden Markov model) for unsupervised learning of P-B structures. The authors claimed that high level states in the hierarchical Markov model referred to an abstraction of contents or game events. Thus, state transitions in a high model level are another Markov chain, which describes the relationship between game events. Xie et al. [2004] stated that output segments from the hierarchical hidden Markov model were similar to P-B structures, if a set of visual features, including grass ratio, average motion displacement and edge map, were used for a model fit. However, such a hierarchical model is a simplification of a Markov graph. The computational cost for model training and fitting is high. A simplified algorithm will be found in Section 5.1.3 for the construction and training of a hierarchical hidden Markov model.

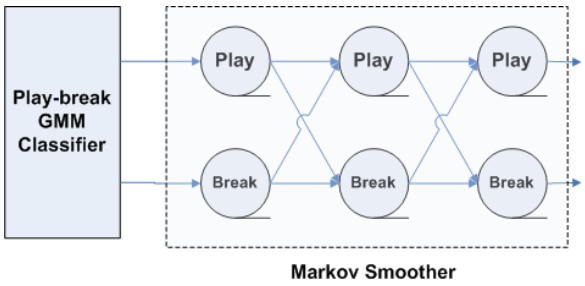


Figure 5.2: Play-break Structure Decomposition

Ekin et al. [2003] proposed another visual structure set based on the zoom depth of cameras: global, middle-close, close-up and field-away. A global segment displays an overall view of a play field; a field-away is a video clip outside game pitch; a close-up focuses on one player, especially on player's face [Tjondronegoro et al., 2004*b*]; a middle-close is a class between close-up and global, which shows actions among a small group of players. A two-pass classification is proposed. The feature of grass ratio over visual frames was employed to discriminate field-away, global view and close view,

and then close views are classified into middle-close and close-up by human face detection. The authors supposed that these structures indicated different game states and the appearance sequence of video structures recorded a tempo of game stories. However, Ekin et al. [2003] did not develop a stochastic model based on these video structures to simulate an entire game video.

Another interesting video structure segmentation is reported in [Kang et al., 2004], which developed an object-based video *grammar*. The authors labelled video clips by video objects. For example, a goal post was regarded as the stop symbol, because most highlights take place around goal posts. [Ekin et al., 2003] and [Intille and Bobick, 1999] follow a partially similar approach. These authors detected middle circles or play field corners to mark video segments. In the audio phase, Rui et al. [2000] segmented baseball videos by the sound of hit or touchdowns for highlight detection. However, audio tracks are usually noisy in sports videos.

5.1.2 Event Patterns

Unlike general models, event patterns focus on the detection of a small content-based event set, such as a goal and a corner kick in a football game. These models focus on a relatively short media sequence, and thus alleviate the uncertainty of video contents during the simulation of a long time sequence. Moreover, external domain knowledge can be introduced into the discovery of event patterns. Intille and Bobick [1999] and Gong et al. [1995] tracked the movement of a football and analysed interactions between players in order to detect a set of game events, such as a goal, a free kick, a shot and a corner kick. The authors introduced many heuristic rules to facilitate the detection. For example, a goal event is discriminated from a shot by testing whether a football passes a bottom line.

Most event patterns combine information from multiple modalities, e.g. audio, caption texts, official game records and some external information resources, such as webcasting text. According to the fusion stage, these event patterns can be further classified into feature combination (early fusion) [Naphade, 1998] [Naphade and Huang, 2001] [Sadlier and O'Connor, 2005], decision fusion (later fusion) [Lenardi et al., 2004] [Xu et al., 2006] [Wang and Cheong, 2006], and two-pass decision [Chang et al., 1996] [Duan et al., 2003] [Xu et al., 2003] [Kang et al., 2004]. Sadlier and O'Connor [2005] supposed that the feature of audio energy was roughly synchronous with visual features. The authors merged audio and visual features into a feature vector and used a support

vector machine (SVM) to identify goal events in an audio-visual joint feature space. [Xu et al., 2001] coupled low level features from audio, caption texts, and visual frames to discover so-called middle-level concepts. A hierarchical hidden Markov model was built to cluster these middle level concepts for game event detection.

The approach of decision fusion deals with audio and visual streams individually. Lenardi et al. [2004] proposed a controlled Markov model to simulate shot transmissions around goal events. The authors took the feature of embedded audio energy as a controlling token. The detail of controlled Markov chain can be found in Section 5.1.3. Game highlights were identified by ranking candidates with audio loudness. The authors evaluated the system performance with a coverage of goal events among the top five highlight candidates. The advantage of decision fusion is that these approaches can match asynchronous audio-visual modalities events and thereby improve detection precision.

The two-pass decision refines decision fusion with a bi-directional search step. Since random delay exists between modality events due to network transition and video encoding, this bi-directional search alleviates the constraint on modality event order. For example, a controlled Markov chain in [Lenardi et al., 2004] indicates the visual event of zoom-in should occur before the increase of audio energy. Kang et al. [2004] developed a two-pass search for goal events. The authors seek for video segments with excited speech by counting pitch numbers and measuring audio energy; and then tried to detect the video object of a goal post around these audio clips. This work assumed that a video clip would be a goal event, if such a clip contains both excited speech in an audio track and a goal post in a visual frame. Additionally, such a search sequence can be inverse.

An event pattern focuses on sparsely distributed events rather than understanding a complete game video. This focus brings many benefits in content-based video analysis at the cost of context-related knowledge. Most patterns originate from a close observation on a short data sequence but ignore video context. For instance, a loud background noise could be an excited cheer or a buzz. The event pattern in [Kang et al., 2004] and [Lenardi et al., 2004] can hardly discriminate these different video contents. Such a problem of content ambiguity may be solved by increasing pattern length and employing a large feature set. However, both of solutions weaken the generality of event patterns and incur strong data dependency. Additionally, it remains a challenge to discover an event pattern directly from raw data in video analysis as well as in pattern recognition.

5.1.3 Mining Visual Patterns

A sports video is a record of a Markov process; audio and visual streams are sequential observations of a content-based Markov chain from different sensors which are of multiresolution and multi-sampling rate.

The output of structure discrimination is a continuous time sequence of structure labels, such as a *PPB...PPP* sequence in the play-break detection, or a discrete event label with a time stamp, such as a goal and the moment when the goal takes place. Ignoring the difference in contents, each label stands for a game state. A transition between labels only relies on prior game states and the game content at that moment. Hence, the sequence of label updating is a Markov process. A hidden Markov model with a sufficiently large state space can asymptotically converge to the distribution of these label sequences, regardless of the actual memory capacity. Moreover, this conclusion can be extended to any content process of sports videos. If we regard game events as states, a content process is an experience of state transitions, in which a game story changes from the state it was in the moment before. A transition between game events only depends on the prior team actions, though such a sequence itself is mostly random.

Although a game content sequence can be described as a Markov process, this state transition sequence can hardly be simulated by any Markov model with given states. This is because (1) game semantics are mostly unpredictable and (2) content presentation approaches are various and complex. Moreover, a trade-off exists between model ability and model generality. Involving more model states can improve model performance in the simulation of a given content sequence. However, too many states will make such a model inextensible. Additionally, video production is full of artefacts [Xu et al., 2005]. A Markov model with too many states is fragile in most cases. Nevertheless, modality events from different media are closely associated with each other and lead to a strong time correlation among hidden Markov states. Therefore, techniques like Markov Chain Monte Carlo (MCMC) will be ineffective in the estimation of model parameters and the adaption of a model to a given data collection are computationally expensive.

Additionally, audio and visual sensors have different resolution and sampling rate. This indicates that multiple resolution and media asynchronism are essential aspects of content-based audio-visual fusion. Audio is a generally brief media style; a visual stream carries rich but sometimes trivial details. A loud shout “goal” from a commenta-

tor costs several minutes of visual data to reiterate the same story. All these mismatches from resolution, data sampling and media alignments, hint that multi-modality fusion has to be carried out at a coarse temporal resolution. To some extent, audio and visual streams are only synchronous on semantic events. Markov states in the audio and visual stream are asynchronous in most cases except at the semantic event level.

Two successful Markov frameworks for football content modelling have been presented in the literature: controlled Markov chain (CMC) and hierarchical hidden Markov model (HHMM). The controlled Markov chain is a temporal evolution of neighbouring states under strong time constraints, such as an occurrence of a zoom-in shot transition in the visual stream and an increase in audio loudness [Lenardi et al., 2004]. There are four components in a CMC, { state, initial transition probability, controlled transition probability, controlling input }. A CMC is equivalent to a Petri-network with coloured token rings. A hierarchical hidden Markov model is a technique of knowledge extraction rather than an algorithm for structure decomposition. The topology of a HHMM reflects knowledge abstracted from data. Learning approaches of a non-loop HHMM can be roughly categorised into three classes,

1. Supervised learning when manually segmented training sequences are available. The topology of a HHMM has already been manually selected. Each sub-tree in this HHMM is learnt individually on separated segments. Cross-level transitions are decided by a transition statistics among training segments. In [Xu and Chua, 2004], an example was presented for the training of a three-layer HHMM model.

2. Unsupervised learning. All parameters including cross-level transition are learnt jointly. This is the most difficult case.

3. A mixture of supervised and unsupervised learning. The state or the topological structure at high levels are known in a HHMM, whilst parameters for cross-level transition are to be estimated. Some applications can be found in automatic speech recognition with word-level annotation [Chien, 1999], and hand writing recognition with text parsing support [Fine et al., 1998].

The decomposition of an *attack* structure belongs to the mixed supervised and unsupervised case. Given the similarity in the video creation process, we define a low-level structure from domain knowledge and leave the high level structure learnt automatically. The following sections will bring a brief introduction of a CMC and a HHMM.

Controlled Markov Chain

In sports videos, a semantic event always incurs multiple modality accidents in a short time interval. For example, a goal event will lead to visual shots of a goal, an audio clip of excited cheers and a replay segment. To describe these concurrent issues, a controlled Markov chain model is developed by extending a Markov model with additional features of controlling inputs or tokens. These controlling signals affect transition possibilities between Markov states [Leonardi et al., 2002]. Let $s(t)$ be a state variable of CMC at time $t \in T = \{0, 1, \ldots, \}$. For instance, $s(0)$ is the initial state at time $t = 0$. The evolution of $s(t)$ from time t to $t+1$ is decided by a possibility P_t, which is a function of an input signal $u(t)$ in the discrete space U. This controlled transition function is formally defined in Equation 5.1.

$$P(s(t+1) = s' | s(t) = s, u(t) = u) = p(s, s', u) \tag{5.1}$$

where $\forall s, s' \in S, u \in U, t \in T$, $p : S \times S \times U \rightarrow [0, 1]$. If the signal space U is of cardinality one, a controlled Markov chain will reduce to a standard homogeneous Markov chain.

Lenardi et al. [2004] assumed that a goal event would incur a zoom-in transition between two nearby shots. The authors proposed a CMC (Figure 5.3) to simulate a sequential shot transition, where the state space S is made up of camera types of visual shots, such as camera pan and zoom. An appearance of a shot boundary is regarded as a controlling signal U, where $u = 1$ refers to the arrival of a boundary and $u = 0$ for none. The frame number indexes all model states as a time stamp t. Therefore, each state is a two-component tuple, $s(t) = (x(t), q(t)) \in S = X \times Q$, where $q(t) \in Q := \{0, 1\}$ is an operative mode or a coloured token in the CMC, and $x(t)$ takes value in a discrete camera action set $X = \{lack - of - motion, fast - pan, fast - zoom, fast - pan - zoom, Other\}$. The operative mode $q(t)$ of a CMC remains the same until a shot transition takes place (Equation 5.2).

$$p((x, q), (x', q'), u) = 0; if (u = 0, q \neq q') or (u = 1, q = q') \tag{5.2}$$

The evolution of a CMC in a given operative mode is the same as a standard Markov model (Equation 5.3),

$$P_q = p((x, q), (x', q), 0) \tag{5.3}$$

Hence, a controlled Markov chain can be trained as a set of individual Markov chains sequentially. All transition possibilities switches with the change of operative modes. For example, the CMC model for goal detection is characterised by two sets of transition probabilities P_0 and P_1 with respect to the evolution mode $q = 0$ and $q = 1$ (Figure 5.3).

The approach of CMC modelling can be extended to other general content-based

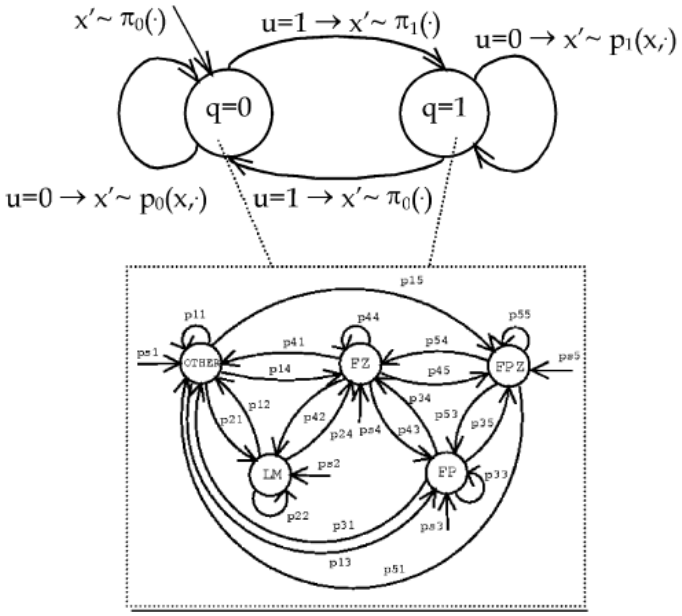


Figure 5.3: Controlled Markov model for goal detection [Lenardi et al., 2004]

events, only if these events would incur a change in video production, e.g. a shot switch and zoom depth variations. Therefore, it is possible to build a basket of CMC models and to discriminant game contents by maximising the likelihood of a CMC. Such a process can be described as follows.

1. Extract low level features around changes in the video production.
2. Decide a state space S by feature values and a controlling input U for a possible change set, e.g. {shot change, no shot change} in [Lenardi et al., 2004].
3. Divide a video into a set of clips which only contains two neighbour controlling states.

4. Compute the likelihood of these clips to a CMC basket and thereby decide a proper CMC model by maximising a likelihood.

For example, two classes of CMC models are employed for goal event detection, namely goal model and non-goal models. If the likelihood of a video sequence to the goal model is greater than that to non-goal models, this video sequence is determined to be a candidate of a goal event. Later, other modality information is introduced to annotate these visual sequences. For example, the feature of audio energy is used to rank these goal candidates in [Lenardi et al., 2004]. Although such a ranking seems naive, Lenardi et al. [2004] reported a high coverage of goal events in the top five candidates.

Hierarchical Hidden Markov Model

A hierarchical hidden Markov model is a hierarchical controlling structure over standard Markov chains. Every state in a high level manages a number of symbols or states in low level Markov chains (Figure 5.4A). A transition from a low level state to a high level state is only invoked when this low level Markov chain enters an emission state.

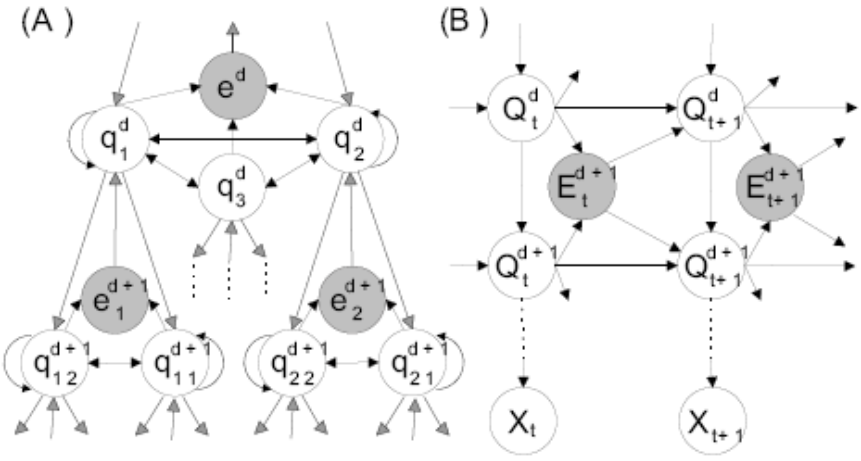


Figure 5.4: Graphical representation of a hierarchical hidden Markov model at levels d and $d+1$: (A) tree structure with bar labels (B) dynamic Bayesian Network. X_t is the observation at the bottom, shaded nodes are emission states which jumps across levels at time t [Xie et al., 2004]

For convenience, we use a bar annotation (Equation 5.4) to record the configuration of a HHMM from the top (level 1) to d^{th} level with a Q-ary d-digital integer, where Q

denotes the maximum of the state size in sub-HMMs.

$$q^d = \overline{q_1 q_2 \cdots q_d} = \sum_{i=1}^{d} q_i Q^{d-i} \tag{5.4}$$

where $1 \leq q_i \leq Q; i = 1,;d$. For example, in a two-level HHMM with two top-level states and three sub-states at the bottom of each top-level state, the second state in the second model will have $q = 4$.

The parameter space Θ of a HHMM (Equation 5.5) includes three parts,

1. Within-level transition probability matrix A_q^d;
2. Prior probability vector π_q^d, which denotes the probability of child states or the observation distribution at the bottom;
3. Exiting probability vector e_q^d, which measures the chance of jumping into the high level.

Without losing generality, we suppose there is only one state at the root.

$$\Theta = \bigcup_{d=2}^{D} \bigcup_{i=0}^{Q^{d-1}-1} \{A_i^d, \pi_i^d, e_i^d\} \tag{5.5}$$

The complexity of parameter estimation is closely associated with the topological structure of a HHMM. Roughly, the number of parameter is $O(dN^2)$, where d is a topology constant less than three, and N is the overall number of Markov states. The graphical structure in Figure 5.4B can be factored into a generalised chain without loops. Hence, an EM algorithm can be used for model inference. Therefore, the overall computational complexity is $O(D(dN)^{2D})$, where D is the level number. However, the training of a HHMM combines the estimation of parameters and the adaption of model onto observations or a data collection. Xu et al. [2005] claimed that a three-layer configuration of a HHMM structure with limited hidden state number is a good balance between training complexity and model ability of data presentation. Although such a configuration depends heavily on the prior knowledge and can hardly be extended, the authors show that this three-layer topology can meet the requirement of sports content extraction.

A general learning approach is proposed in [Xie et al., 2004]. The authors developed

an iterated dynamic framework for generic HHMM training (Figure 5.5). Each training iteration includes four procedures,

1. Parameter updating by the EM algorithm, which will not change the structure configuration and the number of parameters.

2. State splitting. This is a random repartition of direct children of a given state to generate a new state at the same level. This operation introduces a new Markov state into the topology by dividing an existing state.

3. State merging. This operation merges two states at the same level into one state and collapses their children. It is the inverse move of *state splitting*.

4. State swapping. Exchange the parents of two states but keep children unchanged. Since a sport video is made up by recurrent team competitions, such as an *attack* structure, the temporal structure of different video segments may be very similar5-0Sometimes such a similarity on temporal structures is called as fractional similarity. This character can speed up the estimation of a complex Bayesian network.. A high ratio of sub-models have the same size in the state space, but with different high level or multi-level structure. This *swap* operation can speed up model adaption.

The training approach in [Xie et al., 2004] is a simplification of Bayesian network training, although the split-merge step replaces the birth-death step in [Andrieu et al., 2001]. For the simulation of video contents, this minor improvement avoids the problem of knowledge innovation and thereby saves the computation cost on the repartition of sub-structures. Additionally, the content structure of a sports video is usually invariant because of recurrent semantics. For example, the target of a football game is to goal. Hence, the topology of a sports video content model is with a constant layer number. This indicates that the freedom for topological reconstruction is unnecessary in a HHMM in the case of football content extraction. Moreover, the method of Markov chain Monte Carlo method can speed up the searching in a model parameter space and decrease training cost (Appendix B).

5.2 Challenges in Video Modelling

Video modelling is a complex task in content-based analysis. The reason is obvious. Firstly, this simulation process is proposed to extract desired sequential patterns or se-

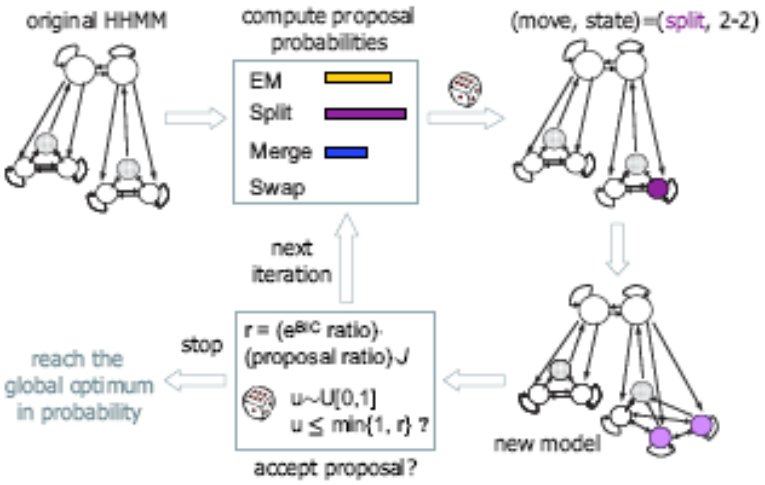


Figure 5.5: Dynamic framework for hierarchical hidden Markov model learning [Xie et al., 2004]

mantic events in a long temporal sequence of video contents. Hence, this process is closely associated with automatic semantics understanding of a large video collection. Most existing approaches rely on low-level features or middle-level syntax. This means that the process of video modelling is to complete a mapping from low-level features to high level semantic end. The semantic gap is unavoidable. It is why this thesis is limited in the scope of sports videos which have a clear domain context and facilitate the discrimination of video contents. Secondly, video modelling should be self-organised and flexible. The production of a video is an art of knowledge presentation. Although temporal patterns actually exist as a guideline for visual arts [Burke and Shook, 1996], a sports video is full of individual creations, such as the preference of video directors. A successful video model has to abide these numerous variations and identify video contents precisely. An effective video pattern and its mining process should be extendible or adaptive to given data collections. Thirdly, such a video model has to support semantic annotation. However, the classification of visual information remains a research target in the project of MPEG-7 and SmartWeb (http://www.smartweb-project.de/). Nevertheless, although video modelling is proposed to ease the manual labour of video labelling [Baillie and Jose, 2003], the ability of automatic identification of video contents still requires a huge amount of labelled data to train a useable system. However, such data collection is rare because the labelling process needs a huge amount of manual work. Additionally, most labelled video data are marked informally and can not be employed directly in the training of a content-based video model. Finally, there lacks a formal method to evaluate the performance of content-based video models. Although the precision and recall of event detection have been widely accepted [Lenardi et al., 2004][Xie

et al., 2004][Kang et al., 2004][Wang and Cheong, 2006][Hanjalic and Xu, 2005], this evaluation is indirect and insufficient to test a content-based stochastic model.

Controlled Markov chains and hierarchical hidden Markov models are states of the art in the domain of content-base pattern mining. These methods simulate a sequence of video contents from different resolution and multiple modalities. However, there are drawbacks. A controlled Markov chain detects a content change in the production sequence and uses a shot boundary as a driving token for model evolution. Such a prior requirement is a limitation on the observation and shortens the model length of a CMC. This indicates a CMC model pays an expensive cost for robustness and model generality.

1. A CMC cannot decide the duration of an event. Note a CMC is not a balanced model and is evoked by a control token. It cannot be run inversely in order to allocate event boundaries.

2. A CMC is of a single temporal resolution. It only works at visual shot level, though other modalities can be introduced in the post-processing or as a controlling token.

3. The employment of a shot boundary limits the simulation period of a CMC. This may lead to an information scarcity in the implementation of a CMC pair, such as goal vs non-goal.

4. A CMC is an observation of a content event rather than an knowledge abstraction, such as HHMM.

5. The information coupling from multiple modalities is weak in a CMC. For example, audio information has to be introduced in the post-processing step [Lenardi et al., 2004].

6. A CMC model is difficult to extend. This is because the observation sequence is predefined and new states can hardly be added into a CMC model.

A hierarchical hidden Markov model is a self-organised approach of knowledge abstraction. This model discovers video patterns by optimising a topological structure of a Markov hierarchy. The split-merge step in model learning adds new Markov states in order to find an efficient presentation of complex knowledge. However, this dynamic structure indicates that the model adaption and parameter estimation in HHMM is a computational intensive task. Nevertheless, the criteria for a HHMM adaption on given

data collection, such as Bayesian information criterion (BIC) and the shortest description length (SDL), are inefficient in model learning, because these criteria can hardly be involved in the model training process.

In the proposition of a video content model, there is a trade-off between model dynamics and learning costs. For example, a HHMM is a general framework for video pattern mining and can be easily extended to cope with content variations. However, the learning of a HHMM is a difficult job with a computational complexity over $O(N^d)$, where d is the number of Markov states. This is because (1) such a model usually hold a large number of Markov states (Equation 5.5); and (2) the training process will modify model topology by adding or deleting Markov states, which results in high complexity of reiterative training. Compared with a HHMM, a CMC limits its scope in a short period of given events so that a CMC-based model can be easily studied. However, the semantics of a CMC is predefined. A CMC can hardly be extended for extra contents.

The scale of a Markov state space is closely associated with the data presentation ability of a Markov model. It is essential for a video content model to select a proper state size, which should be small but sufficient for the discrimination of proposed game semantics. Note that sports videos are characterised by a definite content set (Section 2.4); the style of visual shots can be numbered in sports videos; and the production technique is similar [Burke and Shook, 1996].

5.3 Attack Structure

In this section, I present my work in *attack* structure decomposition, including structure proposition, visual shot classification, structure kernel detection and a Markov basket for video segmentation.

5.3.1 Structure proposition

A sports video is a loose simple-structured temporal sequence. Assfalg et al. [2002] described repetitive video segments as a video pattern or a video structure. Early propositions of sports video structures are based on visual similarity among image frames, such as play-break [Xie et al., 2002], or the appearance of semantic video objects, such as a field corner and a middle circle [Ekin et al., 2003]. However, it is rare to combine visual and content similarity into a proposition of a video structure simultaneously. To facilitate sports video indexing, we try to define a content-based video structure, which

conveys an entire and independent game story. Such a video structure is equivalent to a **scene** temporal structure. In content-based video analysis, a **scene** is defined as the semantic element of a video, which conveys clear and complete contents. Generally, a **scene** is made up by several sequential shots.

According to the terminology of sports videos, a football game is made up of a sequence of group competitions called *attack*, which results in a goal event [Burke and Shook, 1996]. Such a team action can be further categorised into *attack* and *defence*, according to the positions of players in a competition. An *attack* structure is a fundamental unit in professional game analysis, such as game comment and sports coaching. The repetition of *attack* structures results in a content-based video tempo, which repeats and can be easily identified in a broadcasting video. Video producers have developed a group of camera skills to describe game contents in different phrases of an *attack* structure. Therefore, some temporal patterns of shot transitions exist as a production strategy in order to present game stories formally and smoothly. For instance, Ekin et al. [2003] reported their observation in a small game video collection, which consists of three games from World Cup 2002.

1. When an attack begins, a global view is used until the ball passes the centre circle.
2. When a ball comes into a front field, a middle view appears to display a detailed scene of group competitions.
3. When a ball enters or is close to a penalty area, a close-up view occurs to show interactions between a small group of players and camera recorders are ready to catch possible highlights.
4. If a highlight or an essential semantic event takes place, a replay segment will be employed to reiterate the story.

Figure 5.6 describes a transition of shots which take place in an *attack* structure according to the observation statement in [Ekin et al., 2003]. Three conclusions are drawn from above observations and discussions as follows.

1. In a sports video, production techniques, e.g. field-viewing, and close-up, dictate embedded game semantics and thus follow content structures, i.e. *attack*.
2. *Attack* structures are densely distributed in a game video. Hence, a game video can be fully simulated or decomposed by a temporal model, e.g. a hidden Markov model, based on *attack* structures.

3. *Attack* is a video unit, which conveys an entire and independent video story. Such a structure is equivalent to a **scene**.

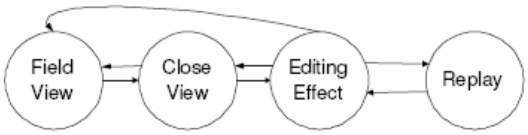


Figure 5.6: Video Production Sequence in Attack

It is useful to extract *attack* structures in content-based sports video indexing. Two main targets are urged in this segmentation process as follows.

- Each *attack* structure contains one and only one semantically meaningful game event, according to the definition. This is because an *attack* is an entire round of making goal, which is ended by a goal or the transmission of ball controlling.
- An event-based video index is created for video retrieval and skimming based on *attack* segments. Meaningful patterns are estimated by the statistics of shot sequences in an *attack* structure and are employed to discriminate game events.

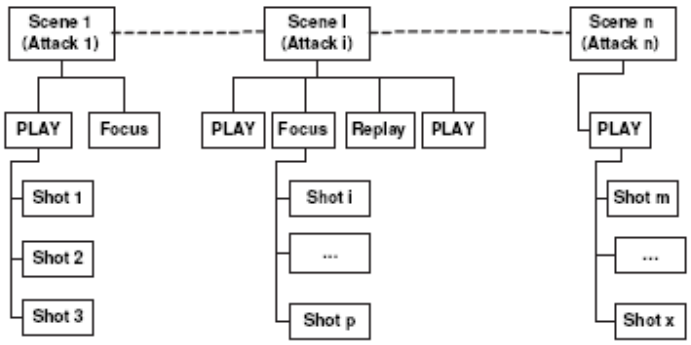


Figure 5.7: Video Structure Hierarchy

5.3.2 Four State Attack Model

As we have mentioned in Section 5.3, an *attack* video structure is drawn from the strategy of video productions. Four types of camera techniques are identified in an *attack* structure: *play*, *focus*, *replay* and *break*.

Definition (Play Camera Type) The camera type of *play* displays a global status in the game pitch, which is usually long and medium shots or a field view.

Definition (Focus Camera Type) The camera type of *focus* traces a player or a small group of players to describe swift actions.

Definition (Break Camera Type) The camera type of break denotes non-game video clips, such as coach shots, field-away, interview and adverts.

Definition (Replay Camera Type) The camera type of replay refers to slow-motion or normal replay clips.

The definition of *play* here is similar to the "Play"structure in the play-break framework [Xie et al., 2004]. *Focus* is a short stop which is called as player close-up in the terminology of video production. Compared with the two-class play-break scheme, the above four-class camera type definition brings advantages as follows.

1. These shot classes are closely associated with game contents;
2. These shot classes reflect game states directly;
3. These shot classes are exclusive in both time sequence and video semantics.
4. These shot classes are directly relevant to actual production techniques, such as focus and replay;

These merits mean that: (1) these shot classes can be identified by visual and temporal features directly; (2) a video index based on this shot classification avoids neighbourhood search caused by semantic ambiguity.

With the discrimination of the above shot classes, an *attack* structure can be simulated by a Markov model in Figure 5.8: four Markov states are involved, namely (0) **Break**, (1) **Play**, (2) **Replay** and (3) **Focus**; and this Markov model starts at a **Break** state and ends with a **Break** state.

5.4 Attack Segmentation

In the above four camera types, a *replay* is a replenish of prior visual frames, while *play, focus* and *break* are characterised with camera view points and zoom depth.

A two-pass classifier (Figure 5.9) is developed to process this difference. The output of classification is a label sequence with respect to camera types. The first pass is a Gaussian mixed models (GMM) on a visual feature set, including play field ratio (Section

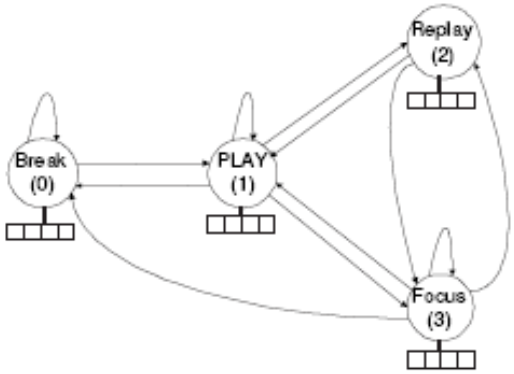


Figure 5.8: Attack hidden Markov model

3.3) and zoom depth (Section 3.5), to discriminate *play*, *focus* and *break*. The output is smoothed by a dynamic programming step. The second pass detects *replay*, which has been discussed in Chapter 4. The whole process of video structure classification is shown in Figure 5.9.

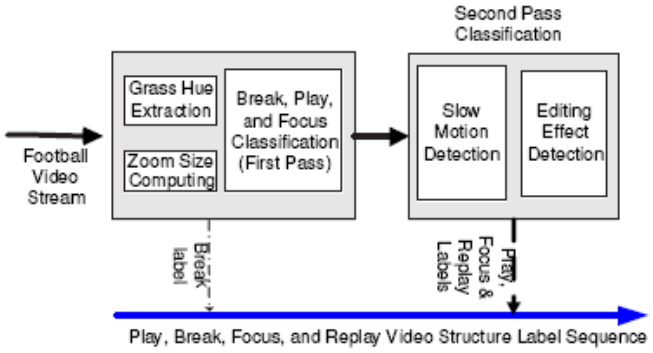


Figure 5.9: Attack Segmentation Flowchart

5.4.1 Structure Decomposition

A label sequence is extracted by camera type discrimination, such as " ...BPFPFPRP... " , where **B**, **P**, **F**, and **R** denote a *break*, a *play*, a *focus*, and a *replay*, respectively. This label string records a process of video production and contains necessary information for *attack* segmentation. However, only grouping these labels is not enough for a successful decomposition. As a semantic unit, an actual *attack* sequence varies with the context and the style of video production. Since a composition of camera shots is an art, the composition of a semantic component in a video is mostly dynamic in term of sudden ideas and tempos. For example, an *attack* structure may not experience all states in

the four-state Markov model (Figure 5.8), while other states are missed, such as *replay* and *focus*. It is a challenge in the training of a Markov model that all Markov states inside a model can be an exit state. Apparently, such a complex situation is beyond the capability of the four-state Markov model in Figure 5.8. A possible solution is to insert new states into the four-state Markov model. However, this action not only increases the complexity of model training, but also is questionable in content-based video analysis. Several reasons are easily identified. Firstly, it is difficult to estimate necessary Markov states from the observation. The possible appearance of an *attack* structure is numerous in sports videos as a mathematical description of visual arts. Secondly, there is a trade-off between the scale of a Markov state space and model generality, as we have discussed in Section 5.1. With more states, a Markov model can depict a stochastic process more precisely but is harder to extend.

A Markov basket is developed to solve this problem in Markov simulation and time sequence decomposition. This solution includes three steps, prior Markov model training, structure kernel detection and structure boundary identification by Markov basket competition. The prior Markov model for an *attack* structure is based on the four-state Markov model (Figure 5.9), because this model is a generalised description of an *attack* structure. The structure kernel detection utilises the repetitive nature of *attack* structures. A suffix tree is employed to count repetitive parts in a label string in order to allocate structure kernels. Therefore, the segmentation of *attack* structures becomes an identification of model boundaries between two Markov processes inside a Markov basket, in which each Markov chain is an appearance of the prior *attack* Markov model on a given structure kernel.

5.4.2 *Attack* Markov model training

This training process consists of two steps: (1) prior model training; and (2) model adaption on given data collections. A set of label sequences, which contains a single *attack* structure, are manually collected to train the prior Markov model. This training step results in a prototype of the *attack* structure model. The model adaption is proposed to improve model fitness in a given sports video.

In the production of sports videos, a *replay* segment interrupts the sequence of game content presentation as does a *break*. Hence, both *replay* and *break* are regarded as stop marks in the content presentation of a sports video. Symbols of R and B divide an entire label sequence into a set of sub-strings which start and end with a label of R

or B, although these string may contain more than one *attack* sequences. Additionally, an *attack* structure is equivalent to a *scene* of a sports video and should be the largest semantic video structure (referring to Figure 5.7). Therefore, an *attack* structure is the longest common repetitive substring among these string.

These observations indicate a self-learning approach for the training of an *attack* hidden Markov model. Given the variation of production strategies and occasional artefacts, it will be helpful to adapt an *attack* model entirely to each game. However, data samples from a sports video is not enough to support effective model training. Therefore, a sub-optimised solution is developed, which improves this general video content model via these roughly segmented substrings.

5.4.3 Structure kernel detection

The *attack* structure is the longest temporal repetition in a sports video. This means that an *attack* is the longest substring in a given label sequence. Therefore, structure kernel detection turns into a problem of the longest common repetitive substring allocation in a known long string. Note that this assumption is robust against production errors and artefacts, although these production variations may decrease the length of the longest common repetitive substring.

Let alphabet $\Sigma = \{F, B, P, R\}$ and T be a string over Σ. The problem of longest common repetitive substring extraction, therefore, is defined as follows.

Definition (Normal Repeat and Super Maximal Repeat) A string p is called a normal repeat of T, if $p = T[i..i+\|p\|-1]$ and $p = T[i'..i'+\|p\|-1]$ for $i \neq i'$. A super maximal repeat is a maximal repeat that never occurs as a substring of any other repeat.

Definition (KN Common Repeat) Given a set of strings $U = \{T_0, T_2, ..., T_N\}$, the (k, N) longest common repeat problem is to find the longest normal repeat which is common to k strings in U for $1 \leq k \leq N$.

A generalised suffix tree [Hui, 1992] is employed to speed up this search task. Such a tree stores all suffixes of a set of strings as a suffix tree (ST) does for a string. Figure 5.10 is an example of the generalized suffix tree (GST) for a string $T_1 = BBPFP$ and $T_2 = BPFPPFP$. Each leaf node has an ID denoting the original string where this suffix comes from.

Besides the construction of a suffix tree, the algorithm for *attack* structure kernel

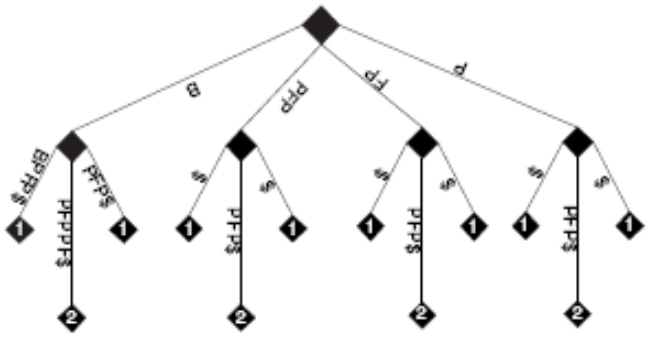


Figure 5.10: Generalised Suffix Tree for $T_1 = BBPFP$ and $T_2 = BPFPPFP$

detection is presented in Algorithm 2. Algorithm 2 is an application of general suffix

Data: shot class label string S and substrings s_i separated by **R** and **B**
Result: structure kernel array SK
Build a suffix tree ST_n for a substring s_n ;
Build a general suffix tree $GST(ST_0, ST_1, ..., ST_N)$;
$i = 100$;
while $i > 1$ **do**
    Find the maximal repeat t_n for each branch in $GST(ST_0, ST_1, ..., ST_N)$;
    $i = \max_{n=0}^{N} \|t_n\|$;
    Remove super maximal repeat branch from $GST(ST_0, ST_1, ..., ST_n)$ and build a GST of super maximal repeats GST_{smr};
end
Find the longest branch S_{kernel} in GST_{smr};
Search S_{kernel} in S and record every appearance in SK ;

Algorithm 2: General suffix tree for structure kernel detection

tree on a given string set. By merging suffix tree of each string, a duplicated suffix tree is created to count all repetitive parts. Then these repetitive parts with enough appearance counts are extracted to build a suffix tree. Apparently, the longest branch of this suffix tree satisfies the requirement.

5.4.4 Structure boundary searching

After the structure kernel detection, we allocate an array of *attack* structure kernels in a string of production techniques. Each kernel stands for a possible *attack* structure and thereby a part of an experience of the *attack* Markov Model in Section 5.4.2. Note that such an *attack* model is bidirectional because there is only one exit state. These kernels can be extended both directions, backward and forward, by this Markov model. Hence,

a structure boundary is identified when two experiences of this *attack* Markov model meets (Figure 5.11). A structure boundary is decided by the possibility ratio between two Markov chains, both of which aim to take labels on the edge.

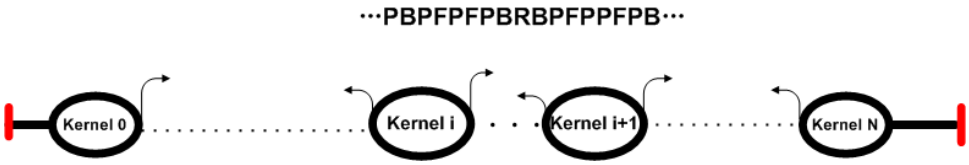


Figure 5.11: Extend structure kernels bidirectionally until they match

5.5 Experiments

The experiment data set includes two complete game videos in the World Cup 2002 collection from BBC, the final game, Germany vs. Brazil and the semi-final game, Japan vs. Turkey. It is about 320 minutes long or more than 400000 visual frames at the resolution of 352×288, which occupies over 4.3 GB, containing interviews, celebrations and commercial adverts. Both games were divided into halves, Final I,II Japan-Turkey I and Japan-Turkey II. The first half of Japan-Turkey (Japan-Turkey I) and final game were labelled manually to set up ground truth. 13462 frames were sampled at the rate of $1/25$, including 4535 *play* frames (33.7%), 4253 *focus* (31.6%) and 4674 *break* 34.6%. There were 33(19/14)5-019 replays in the first half and 14 in the second half *replay* segments in the final game and 34(18/16) in the Japan-Turkey game. The training set included 2000 frames (about 15% in all, 400 from *play*, 1000 from *focus*, and 600 from *break*), which were randomly selected from the marked samples for the classification of four-class video structures. The remaining samples were kept for test.

We measured classification accuracy as the number of correctly classified samples over the total number of samples (Equation 5.6.

$$Precision = \frac{\|f_{correct}\|}{\|f\|} \tag{5.6}$$

where $f_{correct}$ is a frame which is classified correctly.

The discrimination of *play*, *break* and *focus* segments were conducted in two steps (Figure 5.9). The first pass was trained by the visual frame training set and the classification accuracy of Gaussian mixed model is shown in Table 5.1. The smoothing

Markov model was trained by a half game while other three clips were kept for evaluation. This test process repeated for each video clips as the training set. Training and testing accuracies are shown in Table 5.2. Average generalisation performance (avg-gen) is the mean of precision when this clip is used as a training set.

Test Set	Play	Break	Focus
Final I	0.894	0.840	0.823
Final II	0.786	0.664	0.708
Japan-Turkey I	0.862	0.877	0.853
Japan-Turkey II	0.880	0.860.8	0.836
avg-gen	0.856	0.811	0.805

Table 5.1: Play, Focus, Break GMM Classification Precision

Test Set	Training Set											
	Final I			Final II			Japan-Turkey I			Japan-Turkey II		
	Play	Break	Focus	Play	Break	Focus	Play	Break	Focus	Play	Break	Focus
Final I	*0.963*	*0.944*	*0.905*	0.913	0.852	0.830	0.948	0.920	0.877	0.933	0.917	0.891
Final II	0.824	0.730	0.721	*0.863*	*0.817*	*0.824*	0.835	0.713	0.773	0.824	0.711	0.765
Jap-Tur I	0.887	0.930	0.910	0.872	0.880	0.863	*0.890*	*0.952*	*0.917*	0.887	0.925	0.912
Jap-Tur II	0.905	0.892	0.870	0.897	0.870	0.845	0.930	0.905	0.887	*0.971*	*0.901*	*0.903*
avg-gen	0.894	0.874	0.852	0.886	0.855	0.840	0.898	0.873	0.864	0.904	0.867	0.868

Table 5.2: Precision of Play, Focus, Break Segment Classification After HMM smoothing

Final II is noted for the lowest precision. This is because a long celebration video clip seriously garbles the classifier. In the celebration clip, a large group of people wearing player uniforms moved around in the game pitch; various production techniques were employed to highlight this triumph. These visual frames are compliant with *play* and *focus* in the proposed feature space, though we labelled these frames as *break*, according to video contents. The skim-how average precision and recall of *play*, *focus* and *break* classification is displayed in Table 5.3. Additionally, the accuracy of segment discrimination is 89.6% (91.4% in *play*, 89.9% in *break*, and 87.6% in *focus*) besides Final II.

We employed the measurements of precision and recall measurements for video segmentation in [TRECVID, 2003] to evaluate the performance of *attack* segmentation. The precision is the time ratio of correctly identified segments over entire videos; the

Test Set	Average Precision				Average Recall			
	Play	Break	Focus	Over all	Play	Break	Focus	Over all
Final I	0.931	0.896	0.866	0.898	0.940.8	0.926	0.883	0.917
Final II	0.827	0.718	0.753	0.766	0.894	0.879	0.864	0.879
Japan-Turkey I	0.882	0.912	0.895	0.896	0.902	0.907	0.872	0.893
Japan-Turkey II	0.930	0.889	0.867	0.895	0.955	0.896	0.875	0.909
Mean	0.893	0.854	0.845	0.864	0.923	0.902	0.873	0.899

Table 5.3: Average Precision and Recall of Production Skill Classification

	Attack Number	Precision	Recall
Final I	32	0.732	0.890
Final II	40	0.541	0.794
Jap-Tur I	31	0.762	0.846
Jap-Tur II	44	0.710	0.803

Table 5.4: Attack Segmentation Performance

recall is the time ratio of correctly identified segments over the total time of reference segment. Table 5.4 displays the performance of *attack* structure identification.

5.6 *Attack*-based Applications

An *attack* structure is an equivalence of **scene** in sports videos. This indicates that this structure is useful for content-based video analysis and thus have numerous possible applications. In this thesis, I present two applications for football video retrieval and management, namely syntax frequency and browser index. Two demo system for interactive football video skimming and summarisation are displayed in Figure 5.13.

5.6.1 Syntax Frequency

One of main challenges in video retrieval is how to present a video efficiently and effectively. Given the huge data size, it is difficult to deal with video data directly in a video query. An abstraction of raw video data, therefore, is necessary. Moreover, most queries care for video semantics (Chapter 1), e.g. looking for a video clip about the film ``star war´´[TRECVID, 2003]. This indicates that the abstraction of video contents for retrieval should be closely associated with video semantics. However, it is impossible to identify all possible semantics of a video clip. This is because such a semantic understanding is relevant to not only video contexts, but also culture background [Osgood et al., 1957]. Hence, we have to limit the query scope into (1) direct semantics that a semantic concept can be searched if and only if such a concept occurs in a video or (2) a

definite semantic domain, e.g. football videos. One widely accepted solutions is a two-pass strategy [Snoek, Worring and Smeulders, 2005]. This approach detects a group of syntax from low level features and then creates a graph of syntax reasoning with prior knowledge, such as a directional graph from Word Net and LSCOM [Kennedy and Hauptmann, 2006]. Therefore, a search process on video data involves a comparison between video syntax and a match between syntax reasoning graphs. In this case, a video is presented by a two-element tuple, a syntax set and a reasoning graph. However, such a retrieval approach is inefficient, because of high computational cost caused by reasoning graph creation and matching. Additionally, the creation of a reasoning graph requires knowledge on related domains which is hard to be defined. For example, a LSCOM system is a knowledge base and an ontology of a given noun entity set [Kennedy and Hauptmann, 2006]. It costs a few years to build such a system.

To alleviate this problem caused by syntax reasoning, it is worth developing a new presentation method for video semantics. Note that word net and LSCOM are unpopular solutions in the application of text retrieval, although both techniques are proposed to search text documents. It may shine possible solutions to video retrieval by comparing video and text retrieval. Generally, a text retrieval system indexes a large collection of text documents with key words frequency (term frequency) and inverse document frequency [Crestani et al., 1995], accepts a set of keywords as a semantic description of a query and then outputs a list of related text documents. We observe: (1) an entire document is the element of information abstraction and the target of retrieval; (2) the statistics of key word represents the semantics of a document rather than a whole document collection; (3) a document is discriminated by key word distribution across a collection. A visual syntax can be regarded as a key word in a video. However, an entire video can hardly be treated as a semantic element for retrieval. This is because (1) an entire video is usually too huge for a query; (2) a long video is made up of a series of independent semantic stories or **scenes** in the terminology of video analysis; (3) people only look for a small video clip rather than a whole video. In short, an entire video is not either an processing unit or a retrieval target in video retrieval. Therefore, we have to identify a fundamental video unit, which conveys independent and complete semantics and is able to cover most video queries.

In sports video, an *attack* structure is a round of competition. Thus, the semantics of this structure is complete, according to game contents. This indicates that such a structure can fulfil most requirements in sports video retrieval. Hence, an *attack* structure can be regarded as the element of information abstraction or a document in text

retrieval. We compute syntax frequency as the ratio of syntax occurrences in an *attack* structure (Equation 5.7). Therefore, an *attack* structure can be represented by a vector of syntax frequencies.

$$sf_i = \frac{n_i}{\sum_k n_k} \tag{5.7}$$

where n_i denotes the number of visual frames holding a given syntax, and the denominator is the number of occurrences of syntax.

Similar to inverse document frequency, an inverse *attack* frequency is proposed as the ratio of the *attack* number in a game collection, such as FIFA World Cup 2002, over the number of *attack*s which contain a given syntax (Equation 5.8).

$$iaf_i = \log \frac{\|D\|}{\|d : sf_i \in d\|} \tag{5.8}$$

where $\|D\|$ is the number of *attack* structures in a video collection, and d refers to an *attack* structure in D.

An experimental system has been built on the game collection of World Cup 2002, which consists of 46 games and more than 3260 *attack* structures. Given the availability of syntax detectors, we try a small syntax set, including goal post, corner, middle circle, game pitch boundary, human face number, player uniform number, and audience. Among these features, the numbers of human face and player uniform are vectors of four elements, each of which refers to a syntax occurrence number (0-3). Therefore, the syntax frequency vector is of thirteen dimensions. A complete *attack* segment is used as a query (query-by-example). The retrieval output is a video clip list of *attack* structures which are similar in video contents. Although such a retrieval system is trivial and needs further developments in many aspects, such as the size of data collection and syntax set, we find the performance of a pro-type system is satisfactory when processing goal event queries.

5.6.2 Browser Index

An indexing scheme called browser index is developed for football videos based on *attack* structures. Such an index organises *play* and *focus* around *replay* segments, and thus generates a hierarchical content tree (Figure 5.12). This is because *replay* covers game highlights [Pan et al., 2001] [Wang et al., 2004]. A congregation of *replays* is regarded as an acceptable video summary [Pan et al., 2002] [Wang et al., 2005]. Therefore, a *replay* segment is able to represent the story of an *attack* briefly. In another word,

an *attack* can be ignored in video summarisation and indexing, if such a structure has no highlights. Video segments of *play* and *focus* are assigned to a *replay* segment in the same *attack* and thus make up a middle layer of a browser index. These segments contain detailed game information around a *replay*. Then these non-replay segments are further decomposed into shots, which are the bottom layer of an browser index. Figure 5.12 displays such an index structure.

A non-linear video browser and an interactive video summarisation system are devel-

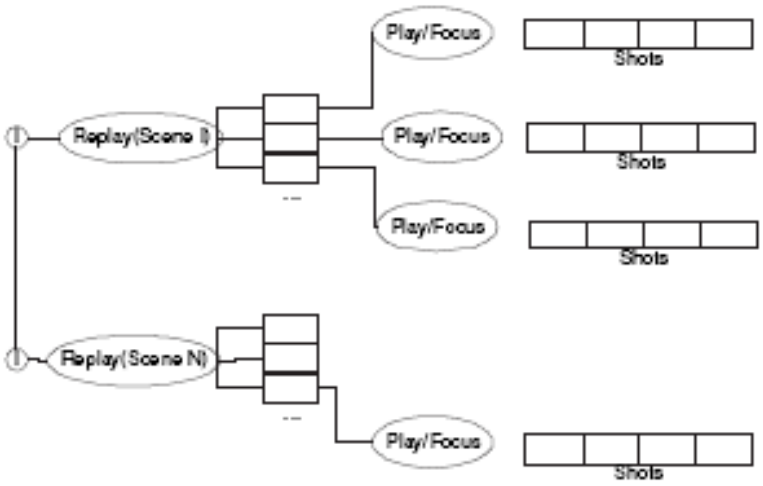
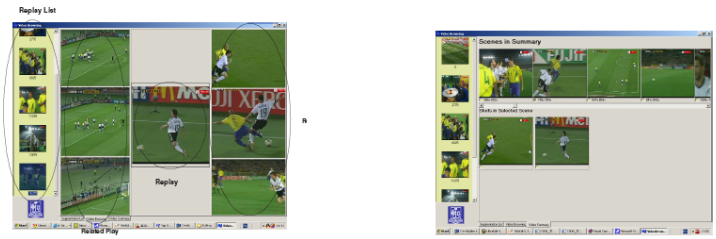


Figure 5.12: Video Browser Index

oped with this browser index. These demo systems can improve video summaries with intensive user interactions. Two main interfaces are provided, related video browser (Figure 5.13a) and summary browser(Figure 5.13b).

Related video browser retrieves *replay* segments and integrates *play* and *focus* seg-



(a) (b)

Figure 5.13: Football Video Skimming (a) Relation Browser (b) Summary Browser

ments. This interface provides two interactive panels, namely a *replay* segment list and a related segments panel. The related segments panel displays the top n(n=3) closest *play* and *focus* to a selected *replay*. A user chooses a *replay* segment from the *replay* list and decides whether to add this highlight and related content segments into a personalised summary. Additionally, a double-click on icons in both the *replay* list and the related segment panel will lead to a replay of a video clip in a stand alone media player window.

Summary browser displays all video segments in a personalised video summary. Such an interface enables nonlinear video editing. A user can insert and remove shots conveniently. The upper right region (Figure 5.13b) browses video segments, which are chosen in **related video browser**. As a default, all *replay* segments are included. When a user selects a video segment in the list, shots relevant to this segment will be shown in the bottom right region for editing.

5.7 Conclusion and Discussion

In this chapter, we identified a semantic video structure called *attack*. Such a structure records a team competition for a goal event and is equivalent to a **scene**. A three-step segmentation algorithm is developed, including prior model training, structure kernel detection and boundary allocation. A four-state hidden Markov model is trained to simulate an *attack* process as a prior knowledge and then improved by local data. A general suffix tree is employed to count and allocate appearances of this Markov processes, each of which stands for an *attack* structure. A Markov basket is created, which consists of all these *attack* Markov processes. Therefore, the *attack* boundary identification turns into a likelihood comparison of a video segment to two neighbour Markov chains in the basket. Experiments show that this high-level *attack* structure can be extracted with a high accuracy.

The segmentation of *attack* structures provides a semantic decomposition for a long game video. Such a structure defines a fundamental temporal unit for syntax statistics and content abstraction. The list of *attack* structures is helpful in content-based video indexing and many other applications. Several experimental systems are introduced, such as a syntax frequency based video retrieval system and the browser index for sports video skimming and summarisation. These applications hint the potency of an *attack* structure in content-based video retrieval and video data management.

However, many drawbacks exist in the proposition and segmentation of *attack* structures. *Attack* structure is based on the visual transition in a game video. The segmentation algorithm only utilises visual information whilst ignoring other media modalities. This indicates that such an approach is vulnerable for artefacts. For instance, we found that a record for local football competitions only use the camera type of *play* without *replay* and *break*. Although such a producer preference can be detected by prior model adaption on local data (Section 5.4.2), this action may seriously decrease the performance, because the Markov state *focus* is absent. A possible solution is to introduce other modalities to improve a rough segmentation on local data (Section 5.4.2) so that the *attack* Markov model can be self-trained by local data. Moreover, the extraction of *attack* structures is a temporal decomposition. Such a segmentation lacks the ability of content weighting. This means we can not identify which structure is of most importance, although some heuristic rules may help, such as the availability of *replay* (Section 5.6.2). A reliable pathway is to take other modality information into consideration. For example, Lenardi et al. [2004] ranked visual segments of highlight candidate with audio energy in a post-processing step. Although it may be efficient to deal with modalities individually at different processing steps, we suppose a unified framework is valuable, which combines visual, audio and other modalities simultaneously into both video segmentation and content weighting. Therefore, a research question rises, what is the temporal structure, which is effective on all modalities, e.g. audio and visual streams. This question leads to my work on sports video *attention* analysis, which manages audio, visual and other modalities from different content and temporal resolutions at the same time.

6

Attention Analysis

> Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought ... It implies withdrawal from some things in order to deal effectively with others.
>
> William James, (Principles of Psychology 1890)

In this chapter, two attention-based approaches for sports video event detection and highlight identification are presented, namely, role-based attention model and a multiresolution autoregressive model. Many aspects of affective analysis are discussed, including emotional space selection, modality synchronisation, feature-based attention estimation, attention fusion, attention-based event segmentation and game highlight identification. Three emotional spaces are compared, including arousal-valence-dominant space, arousal-valence space and attention space. The attention space is employed in the subsequent video content analysis. Modality features, such as speech and visual syntax, are asynchronous because of different updating rate, variant temporal resolution, reflection bias, transmission delay and production artefacts. To synchronise modality features becomes a important issue in modality fusion. Two approaches are developed in this thesis. The role-based attention model clusters modality stimuli ac-

cording reflectors in order to remove reflection bias and matches 15 sec long attention segments by maximising signal correlation. The MAR framework smoothes attention segments at coarse resolutions to find a synchronous temporal resolution. A self entropy operator is proposed to estimate attention intensity incurred by a stimulus, which is compared with a normalisation operator, an adaptive normalisation operator and a peak-enhanced normalisation operator. Local attention maxima are detected as game events and ranked to identify game highlights. A deep-first search algorithm is used to allocate event boundaries.

6.1 Introduction

Generally, game videos convey various emotional aspects, such as happiness and sadness. As a visual art, these videos reflect individual understanding and feeling in response to game contents. For example, spectator cheers and excited comments are recorded in an audio track for a goal event. These theatrical affections create a vivid environment for video watchers around TV sets. Therefore, most of the emotional aspects, which take place in a stadium and auditory room, are kept in broadcasting videos. Therefore, it becomes an effective approach for video content filtering and importance weighting to analyse emotional stories in a game video. For instance, a cheer in an audio track shows spectator's excitement and suggests game highlights. It is reported that strong emotional variations are closely associated with content interests [Ma et al., 2002]. Moreover, the magnitude of an *attention* variation is propositional with the interest of video segments [Lesser and Murray, 1998]. This indicates that highlights, the most interesting part of a game, can be identified by detecting local peaks in an *attention* time sequence or an *attention* curve.

This chapter is organised as follows. Section 6.2 presents a brief introduction of the psychological background of attention analysis, such as the definition of attention, temporal characters of attention, the reflection model, and emotion space. Two important works of attention analysis in sports video content understanding are addressed in Section 6.3, including the user attention model [Ma et al., 2002] and affective video content representation [Hanjalic, 2005]. Many aspects about attention computation are carefully discussed, such as the selection of salient features, the computation of feature-based attention intensity, the combination of feature-based attention and the detection algorithm for sports highlights. Section 6.4 describes a perceptual environment during sports video watching, identifies possible salient features in sports videos, proposes a perceptual structure to describe the perceptual process of video watching, and compares

four attention extraction operators, which assume attention intensity from salient features. Two contributed attention models are addressed, the role-based attention model in Section 6.5 and the multiresolution autoregressive model in Section 6.6. Experiment results and discussions are found in Section 6.7 and Section 6.8, respectively.

6.2 Psychological Background

Affective analysis is a topic in computing psychology rather than computer science. Especially, the research on *attention* can even be traced back to the dawn of modern psychology. William James, the founder of modern psychology, offered a widely accepted definition of *attention* in his original book "Principles of Psychology". *Attention* is the ability of selective cognition, i.e. concentrating on one thing while ignoring others. Compared with many other cognitive processes, e.g. decision making, memory, and emotion, *attention* is tied more closely to perception and sometimes regarded as a gateway to cognition. Another definition of *attention* comes from neural science that *attention* is an enhancing fire of the neural correlation and that the difference in neuron firing is attributed to the mental state as well as stimulus. Some interesting facts from psychology and cognitive neural science are presented as follows.

- *Attention* can be split into covet and overt, because it is usual to notice a set of different events at the same time. Both covet and overt attention imply a state of mood and a switch between these mode states is very fast.

- *Attention* is a cultural phenomenon. What people pay attention to is related to their evolutionary and cultural history. An example is the experiment of odds identification and recall. In this experiment, a box is filled by various everyday odds and ends from different walks of life and different cultures, such as incense sticks and chips; people are asked to look into the box and write down what they have seen after an interval of several minutes. It is reported that the majority of the objects which people remembered are the unusual objects. This can be explained that normal objects tend to attract little attention but odds attract great attention.

- *Attention*s from different people are independent from each other. This is because of the difference in culture, personal experience and expectations.

6.2.1 Attention Temporal Model

In psychology, attention is regarded as a discreet temporal process. People notice something at this moment and other things later. Many continuous or discreet temporal models have been proposed to simulate such a perception process. Lesser and Murray [1998] stated two interest-attention differential formula to quantify a relationship among interest, attention, and human activity in an unknown environment as follows.

$$\begin{aligned}
\frac{dx_{i,j}}{dt} &= (bf(x_{i,j}+wx_{i,j}^2)+\frac{b(1-f)}{4}((x_{i-1,j}+wx_{i-1,j}^2) \\
&\quad +(x_{i+1,j}+wx_{i+1,j}^2)+(x_{i,j-1}+wx_{i,j-1}^2) \\
&\quad +(x_{i,j+1}+wx_{i,j+1}^2)))(1-\frac{\sum_{i',j'}x_{i',j'}e^{-\rho d(i,j,i',j')}}{N\sum_{i',j'}e^{-\rho d(i,j,i',j')}})-mx_{i,j} \\
\frac{dy_{i,j}}{dt} &= sf(x_{i,j}y_{i,j}+wy_{i,j}^2)+s\frac{1-f}{4}((x_{i-1,j}y_{i-1,j}+wy_{i-1,j}^2)) \\
&\quad +(x_{i+1,j}y_{i+1,j}+wy_{i+1,j}^2)+(x_{i,j-1}y_{i,j-1}+wy_{i,j-1}^2) \\
&\quad +(x_{i,j+1}y_{i,j+1}+wy_{i,j+1}^2)
\end{aligned}$$

where N refers to the attention intensity, x denotes the interest intensity of an observer, y is the activity of a reflector, f stands for the basal rate of interest excitation, ρ refers to the decay factor in resource overlapping with distance. $d(i,j,i',j')$ indicates the distance between two "interest"states, $x_{i,j}$ and $x_{i',j'}$. b,s,m and w estimate the rate at which an attention becomes an interest, an interest leads to an activity, a decay raises, and positive feedback incurs, respectively.

Therefore, an audience's interest and behaviour can be predicted in the context of sports videos. This indicates the possibility of finding a direct solution of attention intensity. *Attention* is proportional to the strength of a stimulus, and broadly proportional to the interest of video contents, where the stimulus rises. Another conclusion is from information theory that *attention* is propositional to the amount of information drawn during a given temporal interval. This is because a stimulus strength is propositional to the speed of information pan-out.

6.2.2 Stimulus-Response Model

Attention is a natural reaction to the outside world. The investigation of signal sources or stimulus in the terminology of psychology, remains an active research topic. Computing psychology introduces many quantitative methods to measure the strength of a stimulus and has found several applications in active cognition systems, such as ac-

tive vision. Response depicts an immediate jump of *attention* intensity against a given stimulus but ignores the time cost for such an increment. The stimulus-response model studies a quantitative relationship between stimulus and response, which establishes a mathematical function for an expected response Y against a given stimulus x. One of the most common stimulus-response model is Equation 6.1.

$$E(Y) = \alpha + \beta x \tag{6.1}$$

where α is the threshold triggering a response and β refers to a reaction parameter. However, a log-like formula (Equation 6.2) is also plausible in many applications, such as audible stimulus [Milanese et al., 1995] and colour affection [Engel et al., 1997].

$$E(Y) = \alpha + \beta \log(x) \tag{6.2}$$

6.2.3 Emotion Space

Emotion space is a quantitative descriptive approach for human emotions. By discriminating facial expressions, Ekman [1987] identified a basic set of emotions, which consists of *happiness, surprise, anger, sadness, fear*, and *disgust*. Osgood et al. [1957] proposed a Valance-Arousal-Dominance (VAD) dimensional system to depict emotion variations. Valance measures the intrinsic attractiveness (positive valence) or aversive state (negative valence) of an event, object, or situation [Detenber et al., 1997]. For example, anger and fear are usually treated as "negative valence", while joy has "positive valence". Arousal is referred as a psychological state of being awake, involving the reticular activation in the brain or an autonomic nervous system. A strong arousal stimulus leads to a condition of sensory alertness, such as an increase of heart rate and blood pressure. Hence, arousal ranges in a continuous scale from energised, excited, alert, calm, drowsy, to peaceful. Dominance stands for the tolerance of a reflector, which ranges from "no control"to "full control". Dominance is particularly useful in distinguishing emotional states which have similar arousal and valence, such as grief and rage. Osgood et al. [1957] claimed each emotion could be represented by a point in such a 3D space.

As a matter of fact, valence and arousal count for most independent emotional variances in an enjoyment process of pictures, television, radio, computers and sounds [Creenwald et al., 1989] [Dietz and Lang, 1999]. As such, a valence-arousal (VA) emotion space is proposed as a simplification of this VAD space by ignoring the dominance dimension. This emotion space is widely employed to visualise a relationship between

different emotions. For example, Hanjalic [2005] tracked emotion variations in story films by demonstrating a *VA* trajectory. Later, Wang and Cheong [2006] projected low level audio and visual features into the VA space and claimed that films, such as love films and action films, can be discriminated by feature distributions in the VA space.

Attention space is a further simplification, which only keeps the arousal dimension in the VA space. In sports videos, the measurement of valance is usually meaningless. This is because (1) watching a sports video is an enjoyable experience, which rarely incurs negative valance, e.g. sadness and rage; (2) sports videos are similar to video documentary, in which producers, such as video directors, should keep neutral on contents to ensure the precision of actual story recording. Ma et al. [2002] developed a group of feature-*attention* models to estimate *attention* variation in a sports video. The authors supposed that such an approach was efficient and effective in sports highlight detection. Moreover, our works [Ren and Jose, 2006] [Ren, Jose and He, 2007] have shown the effectiveness of *attention* space in sports video analysis.

6.3 Attention-based Sports Video Analysis

A salient feature captures or estimates the intensity of attention from raw video data and thus projects media features into the psychological attention space. Itti and Koch [2001] surveyed computational models for an active vision system and proposed a bottom-up framework to compute image-based visual attention. The authors drew a topographical salient map by combining multiple image features [Itti and Koch, 1998, 1999], so that the rectangle of interest (ROI), which attracts most attention in an image, was identified as the area with maximum salient intensity. Later, Itti and Koch [2001] discussed several different models to simulate the spatial distribution of static attention in an image, such as a Gaussian distribution model. Bollmann et al. [1997] claimed that dynamic features, such as motion, were important in *attention* estimation. The authors regarded these dynamic features as an amplification of static salience or a controller of gaze shifts. Ma et al. [2002] reviewed these early works in the field of active vision and proposed an *attention*-based approach for sports video summarisation. The impact of media features on perception was isolated and a series of feature-based attention models were developed, such as motion attention model, static attention model, and audio salient model, to estimate feature affection. For a video, the authors estimated *attention* on every visual frame and thus computed a time sequence of *attention* observations named an *attention* curve.

Since the *attention* state is unique at any given moment, a unified "viewer attention" has to be estimated from multiple modality *attention* estimations. Ma et al. [2002] and Hanjalic [2005] suggested a linear combination scheme to fuse these feature-based attention curves. However, isolated *attention* estimations from modality features incur too much noise. With the increase of the feature number, the so-called actual *attention* signals are occasionally overwhelmed by noise in [Ma et al., 2002]. Therefore, Hanjalic [2005] selected a relatively small feature set, including average block motion vector, shot cut density and audio energy, and smoothed these feature-based *attention* curves with a 1-minute long Kaiser window. The authors counted the peak number of these feature-based attention curves in a given temporal interval to predict an *attention* peak in the unified *attention* curve. Later, an adaptive filter was proposed in [Hanjalic and Xu, 2005] to increase the signal noise ratio (SNR) in feature-based *attention* estimation. Furthermore, we developed two fusion schemes, role-based attention estimation (Section 6.5) and multiresolution autoregressive framework (Section 6.6).

Many applications have been developed based on *attention* estimation. Ma et al. [2002] collected peaks of the so-called "viewer attention" curves to create a video skim and a video summary. This is because *attention* intensity is propositional to content importance in sports videos. Hanjalic and Xu [2005] and Ren, Jose and He [2007] treated the allocation of *attention* or arousal peaks as an effective general approach of sports highlights detection. Moreover, such an assumption is sometimes useful in generic videos. Ren, P.Punitha, Urban and Jose [2007] developed an *attention*-based video summarisation system for the TRECVid rushes collection.

However, the computing of *attention* intensity is confronted with many psychological uncertainties as follows.

- The quantitative relationship between low level features and psychological *attention* is uncertain. Although computing psychology has presented several effective approaches, such as a salient map, to estimate the reflective nature of a *stimulus* 6-0Note that stimulus is a combination instead of single feature reaction. qualitatively, there is no clear quantitative measurement for modality affection. Although many stimulus-response models have been proposed in Crary [1999] (Section 6.2.2), the selection of stimulus-response models relies on prior knowledge or a psychological context. This indicates that it is difficult to decide which model is optimal in a given psychological scene, such as a sports video watching.

- Hanjalic and Xu [2005] assumed an *"average"* viewer, who had a "standard" re-

sponse against given game contents. Such a hypothesis violates the independence of individual attention [Crary, 1999; Treisman and Kanwisher, 1988]. The actual reaction from a specific viewer is psychologically uncertain.

- Modality asynchronism and multiresolution. Commercial video encoding and decoding techniques, such as $MPEG-1/2/4$ and $H.263$, deal with modality data independently to save transportation bandwidth. This results in random temporal delays between audio and visual stream in an encoded video. Furthermore, the methodology of football video production makes this situation more complex. This is because audio track and visual stream come from totally different observers or reaction objects in psychological terminology. For instance, video directors watch a game, edit camera videos and compose video according to personal understanding. Meanwhile, automatic microphones record sounds from spectators and commentators. There is a reaction bias or the discrimination of attention periods. Hence, peaks from audio-based and visual-based *attention* curves rarely appear at the same time. Nevertheless, random network delay may enlarge such an observation bias, because audio and visual stream are transferred individually.

In the following sections, two *attention* estimation frameworks are discussed in detail, namely the user attention model (Section 6.3.1) and the affective video content representation (Section 6.3.2). Topics cover salient feature selection, the estimation model of *attention* intensity, and the *attention* fusion scheme.

6.3.1 User Attention Model

A sports video is a composite of an image sequence, audio track, and textural information. These media modalities exert different effects on perception and content understanding. Therefore, three perceptual domains need carefully study in order to describe an *attention* variation process whilst watching a video, namely visual, audio and linguistic attention (Chapter 2). Ma et al. [2002] suggested an average user and proposed a general user attention model (Figure 6.1). The authors assumed that *attention* impact from different modality features were independent or acted individually. Besides linguistic *attention*, a set of feature-attention computing models were developed to estimate feature affection. This is because linguistic attention is a problem of nature language understanding rather than a simple stimulus-response process.

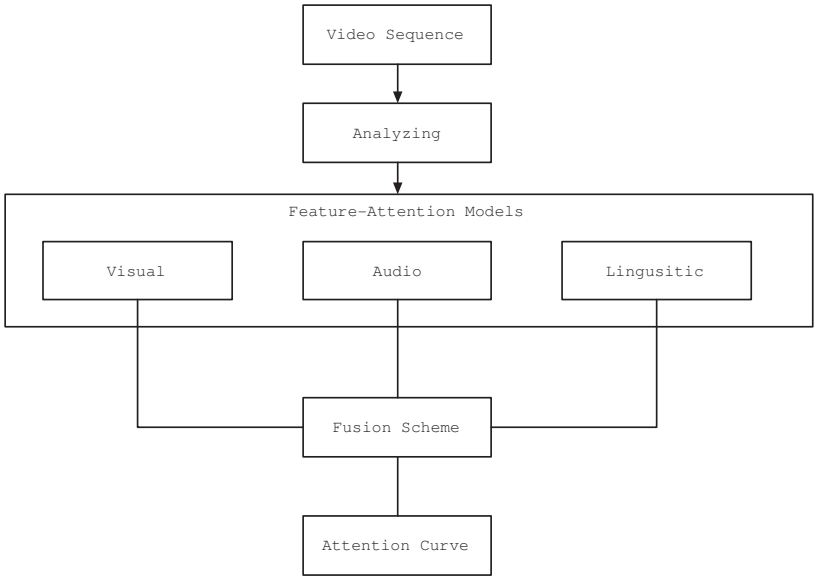


Figure 6.1: User Attention Model [Ma et al., 2002]

Visual Attention Computing

Ma et al. [2002] categorised visual attention into four classes, motion, static, face and camera. Motion attention refers to the temporal contrast in a frame sequence; static attention denotes an *attention* impact from a single visual frame; a face is one of the most salient human characters and attracts notice; camera attention estimates attention effect caused by camera transition. Note that face and camera attention are proposed based on behavior analysis during video watching.

Motion Attention Motion attention is computed in the compressed field. Ma et al. [2002] developed three "inductors"6-0Ma et al. [2002] used this terminology for feature-affection projection function. to estimate motion affection on different attention aspects, namely an intensity inductor (Equation 6.3), a spatial coherence inductor (Equation 6.5) and a temporal coherence inductor (Equation 6.7).

$$I(i,j) = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{MaxMag} \tag{6.3}$$

where $(dx_{i,j}, dy_{i,j})$ denotes components of a motion vector, and $MaxMag$ is the maximum magnitude.

The spatial coherence inductor (Equation 6.5) is based on the spatial phase consistency

of motion vectors in a given spatial macro block $w \times w$.

$$p_s(t) = \frac{SH^w_{i,j}(t)}{\sum_{k=1}^{n} SH^w_{i,j}(k)} \tag{6.4}$$

$$Cs(i,j) = -\sum_{i=1}^{n} p_s(i)\log(p_s(i)) \tag{6.5}$$

Where $SH^w_{i,j}(t)$ is the spatial phase histogram whose probability distribution is $p_s(t)$ and n is the number of histogram bins.

Similar to spatial coherence inductor, the temporal coherence inductor is extracted from a sliding window of L frames,

$$p_t(t) = \frac{TH^L_{i,j}(t)}{\sum_{k=1}^{n} TH^L_{i,j}(k)} \tag{6.6}$$

$$Ct(i,j) = -\sum_{i=1}^{n} p_s(i)\log(p_t(i)) \tag{6.7}$$

Where $TH^L_{i,j}(t)$ is the temporal phase histogram whose probability distribution is $p_t(t)$, and n is the histogram bin number.

Three salient maps are computed, an intensity salient map, a spatial salient map and a temporal salient map, when a motion vector field passes these inductors, respectively. A motion salient map is a combination of these salient maps as Equation 6.8.

$$B = I \times Ct \times (1 - I \times Cs) \tag{6.8}$$

Therefore, the average of a motion salience map is defined as motion attention of a given visual frame.

$$M_{motion} = \frac{1}{N_{MB}} \sum_{\Lambda} \sum_{q \in \Omega} B_q \tag{6.9}$$

where B_q is the salient value of a macro block; Λ refers to an area set with non-zero motion salience; Ω denotes the set of macro blocks in a salient area and N_{MB} is the number of macro blocks in motion vector field.

Static Attention Static attention estimates stimuli from colour contrasts, intensity contrasts and orientation contrasts [Itti and Koch, 2001], to assume the affection of im-

age background. Hence, static attention involves a group of topographic feature maps from multiple resolutions. The computation of static attention is displayed in Figure 6.2.

For a RGB colour image, let R,G and B denote red, green and blue colour channels.

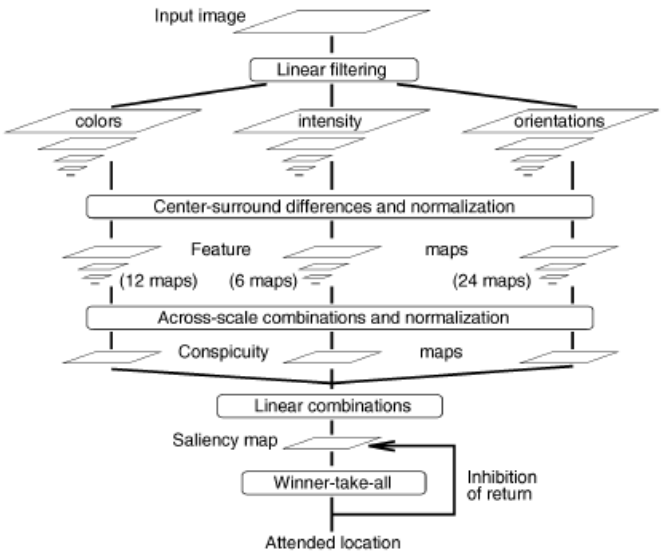


Figure 6.2: Static Salient Computing Architecture [Ma et al., 2002]

Therefore, the intensity $I = \frac{R+G+B}{3}$; r, g, and b are colour channels value normalised by I in order to decouple hue from intensity. However, normalisation is only necessary at image areas where I is larger than $1/10$ of its maximum over the entire image and remaining areas hold zero value for r, g, and b. This is because hue variations are not perceivable at very low luminance. Hence, low luminance can be ignored in salience computation.

For convenience, we transfer a *RGB* image into a $r'g'b'y'$ colour space as follows.

$$\begin{aligned} r' &= r - \frac{g+b}{2} \\ g' &= g - \frac{r+b}{2} \\ b' &= b - \frac{r+g}{2} \\ y' &= \frac{r+g}{2} - \frac{\|r-g\|}{2} - b \end{aligned}$$

if $y' < 0$, y' will be set to zero. Four Gaussian pyramids $r'(\delta)$, $g'(\delta)$, $b'(\delta)$, and $y'(\delta)$ are computed on the corresponding colour channels. The intensity contrast is computed

in Equation 6.10,

$$I(c,s) = \|I(c) \ominus I(s)\| \tag{6.10}$$

where c and s refers to different scales on an Gaussian pyramid I. $\ominus$ is a linear "centre-around"operator, which is proposed to simulate the perceptual process on a visual receptive field. Such an operator is implemented as the difference between a fine and a coarse visual resolution. For example, for a circle $\ominus$ operator, the region centre is a pixel; an observation scale is $c \in \{2,3,4\}$; the surround area is corresponding pixels inside the scale $s = c + \delta$ with the radius $\delta \in \{3,4\}$. Therefore, a $\ominus$ operation is obtained by three steps: (1) interpolating a coarse surround area to a finer scale; (2) point-by-point subtracting in the fine surround area; and (3) summing up all residues. Note that this operation may be carried out on multiple scale rates and a normalisation step is sometimes used. A common usage of $\ominus$ operator is to estimate a rectangle of interest (ROI) in an image. Leventhal [1991] explained this intensity contrast stimulus by neuron simulation. The authors claims that neurons, which are sensitive either to dark centres with bright surroundings or to bright centres with dark surroundings, are excited simultaneously.

The perception of colour is a little more complex. Neurons in the centre of receptive fields are excited by one color and inhibited by another colour, while the converse is true in the surroundings. Such a spatial and chromatic opponent exists for color pairs, such as red/green (r'/g'), blue/yellow (b'/y') and visa verse in human primary visual cortex [Engel et al., 1997]. Therefore, the salient map of color contrast is defined as follows.

$$\begin{aligned} RG(c,s) &= \|(r'(c) - g'(c)) \ominus (g'(s) - r'(s))\| \\ BY(c,s) &= \|(b'(c) - y'(c)) \ominus (y'(s) - b'(s))\| \end{aligned}$$

Texture orientation can be estimated from an intensity graph I by an oriented Gabor pyramids $O(\delta,\theta)$, where δ represents a pyramid scale and $\theta \in \{0°, 45°, 90°, 135°\}$ denotes the preferred direction [Leventhal, 1991] [Greenspan et al., 1994]. Hence, a local orientation contrast $O(c,s,\theta)$ is the $\ominus$ difference between the centre and surround scales (Equation 6.11).

$$O(c,s,\theta) = \|O(c,\theta) \ominus O(s,\theta)\| \tag{6.11}$$

A sum salient map B is created to combine above salient maps. Thus a static attention model is defined in Equation 6.12, which is a weighted normalisation of the sum salient

map.

$$M_{static} = \frac{1}{A_{frame}} \sum_{k=1}^{N} \sum_{(i,j)\in R_k} B_{(i,j)} \cdot w_{pos}^{(i,j)} \tag{6.12}$$

where $B_{i,j}$ refers to the salient sum at a pixel (i, j) in a ROI R_k, N denotes the number of ROIs, A_{frame} is the area of a visual frame, and $w_{pos}^{(i,j)}$ stands for a normalised Gaussian template with a centre at the pixel (i, j).

Face attention Face attention counts the attention increment because of human actor appearance, such as a broadcaster in a news video. This attention estimation is designed for video semantics rather than perceptional prominence. In face detection systems [Polikar, 2006*a*], the location and the size of a human face can be decided in a visual frame. Face attention is a product of face size and a prior location weighting template (Equation 6.13).

$$M_{face} = \sum_{k=1}^{N} \frac{A_k}{A_{frame}} \cdot \frac{w'_{pos}}{8} \tag{6.13}$$

where A_k stands for k^{th} face size, A_{frame} for the area of frame, and w_{pos}^{i} is a prior face location weight, which is defined in Figure 6.3.

1/3	1/3	1/3	
1	2	1	3/12
4	8	4	4/12
1	2	1	5/12

Figure 6.3: Face Location Weighting in Video Frame [Ma et al., 2002]

Camera Attention Camera attention deals with attention variation caused by production style change, such as zoom depth and global motion pattern. To reduce the computational complexity, a six-parameter affine model is employed to estimate global motion 6-0Since zoom depth is an important issue in the camera attention estimation, the eight-parameter project model may be better than such an affine model. This is because a change of zoom depth causes image deformations on z axis.. Moreover, several heuristic guidelines [Ma et al., 2002] are listed as follows, which emphasise or neglect some video objects or a video segment in certain cases in order to improve camera attention estimation.

1. Zoom and camera rolling always attract attention.
2. The faster the speed, the more important the content is.
3. Zooming-in attracts attention to details, while zooming-out presents an overview.
4. Horizontal panning incurs neglect of video objects. The faster the pan speed is, the less important the content is. Additionally, vertical panning is rare because such an editing causes unstable feeling.
5. Other camera motions, besides zoom, rolling and pan, have no obvious intention.
6. If the speed of camera motion changes too frequently, such a video segment is labelled as random or unstable motions, which are meaningless in attention estimation because of instability.

Audio Attention Computation

Ma et al. [2002] stated that an audio track attracted attention by two means, audio content and physical nature of audio energy. For example, speech and music are semantically meaningful for perception, while their physical characteristics, such as loudness and temporal variation, grab human attention. Hence, the estimation of audio attention consists of three parts, speech attention, music attention, and audio salient attention. Salience of speech (Equation 6.14) and music (Equation 6.15) are measured by the temporal ratio of a speech or a music component in an audio segment. Therefore, the first step for speech and music attention estimation is to discriminate speech and music clips from background noise. This is a typical audio processing problem. A large group of features, including Mel-frequency cepstral coefficients (MFCC), short time energy, zero crossing rates, sub-band power distribution, brightness, bandwidth, spectrum flux, linear spectrum pair divergence distance, band periodicity and the pitched ratio 6-0A pitched ratio is the number of pitched frames over the number of frames in a sub-clip., are extracted from audio and a support vector machine (SVM) is employed to classify audio segments into four classes, speech, music, silence and others [Ma et al., 2002]. To improve discrimination precision, a label smoother is used to remove abrupt label change and count the duration of a sub-segment.

$$M_{speech} = \frac{N_{speech}}{N_{total}} \tag{6.14}$$

$$M_{music} = \frac{N_{music}}{N_{total}} \tag{6.15}$$

where M_{speech} and M_{music} are speech attention and music attention, respectively; N_{speech} refers to the number of speech sub-segments, while N_{music} stands for the number of music sub-segments; N_{total} denotes the sub-segments number in the entire audio clip.

A loud sound and sudden loudness variation always catch human attention. Hence, audio salient attention is described by Equation 6.16,

$$M_{as} = \overline{E_a} \cdot \overline{E_p} \tag{6.16}$$

$$\overline{E_a} = \frac{E_{avr}}{\max E_{avr}} \tag{6.17}$$

$$\overline{E_p} = \frac{E_{peak}}{\max E_{peak}} \tag{6.18}$$

where E_{avr} and E_{peak} refer to the average energy and energy peak of a given audio segment, respectively. A sliding window is used to compute audio salient attention along segments.

Linear Multimodality Attention Fusion

Ma et al. [2002] adopted a linear combination function to fuse above feature-attention curves. All of these feature-based attention intensities are normalised to the interval $[0,1]$. Therefore, an overall attention intensity is estimated as,

$$A = w_v\overline{M_v} + w_a\overline{M_a} + w_l\overline{M_l} \tag{6.19}$$

where w_v, w_a, w_i are weights for visual, audio and text attention6-0Ma et al. [2002] called text attention as linguistic attention., while $\overline{M_v}, \overline{M_a}, \overline{M_l}$ refer to normalised media

attention, respectively. These media attention are computed as follows.

$$M_v = (\sum_{i=1}^{P} w_i \cdot \overline{M_i}) \times (\overline{M_{cm}})^{S_{cm}} \quad (6.20)$$

$$M_a = (\sum_{j=1}^{q} w_j \cdot \overline{M_j}) \times (\overline{M_{as}})^{S_{as}} \quad (6.21)$$

$$M_l = \sum_{k=1}^{r} w_k \cdot \overline{M_k} \quad (6.22)$$

where w_i,w_j and w_k are combination weights; $\overline{M_i}$,$\overline{M_j}$ and $\overline{M_k}$ refer to a normalised attention component for static salience, audio salience and linguistic salience, respectively. In the visual attention computation, $\overline{M_{cm}}$ denotes the normalised camera attention, which acts as a model magnifier, while S_{cm} can be regarded as a switch of this magnifier. If S_{cm} is greater than 1, the magnifier of $\overline{M_{cm}}$ is open. Similar to the role of camera attention, $\overline{M_{as}}$ is the normalised audio salient attention and acts as a magnifier for audio attention. In [Ma et al., 2002], both of $\overline{M_{as}}$ and $\overline{M_{cm}}$ were normalised into [0,2].

6.3.2 Affective Video Content Representation

Hanjalic [2005] proposed an affective video content model to describe variations in viewer emotions. The authors defined affective contents, such as emotional intensity and feeling type, which were expected to occur when individuals watch a video. The authors employed the arousal-valence emotion space to estimate the attraction of a video.

Arousal Computing

Three excitement-related low-level features are selected to estimate arousal stimuli in multiple modalities. Hanjalic [2005] supposed that these features should be qualitatively propositional to arousal intensity6-0Here propositional means that an value increase of feature leads to an increase in the user excitement level..

- Frame-based motion activity (Section 3.6.1), which is obtained on the basis of overall motion activity measured between consecutive video frames [Simons et al., 1999][Dietz and Lang, 1999];
- Shot or video production rhythm (Section 3.2), which can be extracted by investigating shot length variation in a video [Adams et al., 2000];

- Audio track energy in an interval of video frames (Section 3.6.3), which is computed as overall energy in a sound track [Picard, 1997][Murray and Arnott, 1993].

Motion vectors $v_i(k)$ are computed by block-based motion estimation between adjacent video frames k and $k+1$. The motion salience $m(k)$ is calculated as the average magnitude of all motion vectors $\overrightarrow{v_i(k)}$ (Equation 6.23).

$$m(k) = \frac{1}{N|\overrightarrow{v_{max}}|}(\sum_{i=1}^{N}|\overrightarrow{v_i(k)}|) \tag{6.23}$$

where $|\overrightarrow{v_{max}}|$ is the absolute maximum of motion vectors, which normalises a motion salience into (0,1].

The influence of a shot transition rate on viewer's arousal is estimated by Equation 6.24,

$$c(k) = e^{((1-(n(k)-p(k)))/\delta} \tag{6.24}$$

where $p(k)$ and $n(k)$ refer to the shot boundary ratio in a given temporal interval before and after a visual frame k, respectively; the parameter δ is a predefined constant.

Audio energy is a power spectrum in consecutive segments of an audio track. As mentioned in Section 6.3.1, peaks of audio energy play a significant role in an arousal time curve. However, the computation of audio energy is significantly affected by recording volume levels. Note that a video producer, such as a commentator, is able to adjust recording volume level according to personal preference. Therefore, the salience of audio energy $G(k)$ has to be a tradeoff between absolute audio energy $\widetilde{e}(k)$ and period normalised audio energy mean (Equation 6.27).

$$e_n(k) = \frac{\widetilde{e}(k)}{\max_k(\widetilde{e}(k))} \tag{6.25}$$

$$\overline{e}_n = \frac{1}{W}\sum_k e_n(k) \tag{6.26}$$

$$G(k) = e_n(k)(1-\overline{e}_n) \tag{6.27}$$

where $e_n(k)$ denotes the normalised audio energy; $\overline{e}_n$ is the average normalised audio energy in a time interval W.

To smooth the above arousal curves and remove fluctuations caused by shot transitions, a Kaiser window $K(l,\beta)$ with length l and shape parameter β is employed (Figure 6.4). Hanjalic [2005] claimed that smoothed arousal curves, such as motion salience

$\widetilde{m}(k) = m(k)K(l,\beta)$, were able to mimic variations of viewer excitement better than original arousal curves.

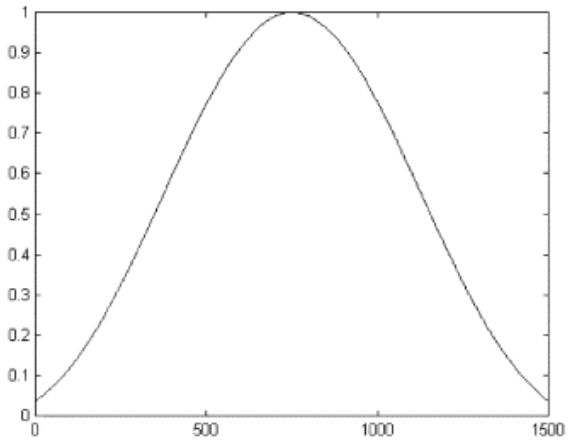


Figure 6.4: Kaiser Window Smoother

Valence Estimation

Valence is useful in the discrimination of video genres, such as love films vs. disaster films [Wang and Cheong, 2006]. Although this emotional dimension is ignored in sports video analysis, a brief description is provided about the computation of valence in this section.

Dietz and Lang [1999] stated that the intensities of arousal and valence are subject to some constraints in the actual emotion scale. The authors called this character of emotion behavior "emotion region"(Figure 6.5). Hence, it is possible to estimate a valence range through an arousal intensity, although many constraints exist. Hanjalic [2005] divided the computation of valence into two steps, the estimation of valence range and the calculation of valence variation from a given range. The estimation of valence range is given by Equation 6.28.

$$r(k) = a(k)sign\{H(D_j(k), j = 1..M\} \tag{6.28}$$

where $r(k)$ refers to the valence intensity at a visual frame k; $a(k)$ denotes the respective arousal intensity before Kaiser window smoothing; $D_j(k)$ is valence signal nature revealed by a modality feature j; and H stands for a fusion function of feature affections on valence nature. In [Hanjalic, 2005], this fusion function is a voting machine. Therefore, the valence scale $r(k)$ is determined by an arousal intensity, while the nature

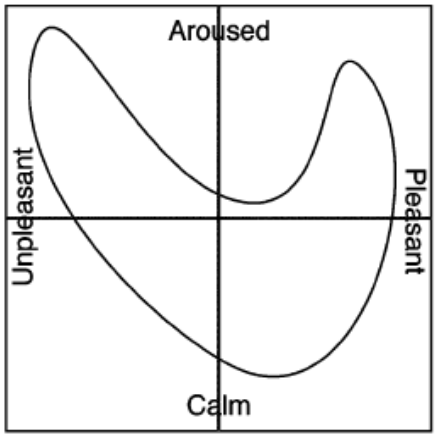


Figure 6.5: 2D Emotion Region([Dietz and Lang, 1999])

of valence is given as an accumulated feature affection.

The valence variation from a given scale $g(k)$ is estimated by Equation 6.29.

$$g(k) = \frac{n}{100} \max_k a(k) \frac{H(D_j(k), j = 1,..,M)}{\max_k \|H(D_j(k)), j = 1,..,M)\|} \tag{6.29}$$

where n denotes a amplitude limitation parameter. In computing psychology, n indicates the amplitude of perceptual ratio between valence and arousal, and is regarded as an measurement on personal behavior. Hanjalic [2005] assigned 10 to this parameter.

Hence, the valence intensity is computed by Equation 6.31, which sums up valence scale and the variation. This is because only the maximum of a valence is addressed in the emotion analysis.

$$v(k) = r(k) + g(k) \tag{6.30}$$

$$\widetilde{v}(k) = K(l, \beta) * v(k) \tag{6.31}$$

$$V(k) = \frac{\max_k \|v(k)\|}{\max_k \|\widetilde{v}(k)\|} \widetilde{v}(k) \tag{6.32}$$

Where $K(l, \beta)$ denotes a Kaiser window smoother with length l and shape parameter β; $\widetilde{v}(k)$ refers to the valence intensity after being smoothed; $V(k)$ is the normalised valence intensity.

However, psychological valence is semantically ambiguous, given the complexity of human perception. For example, black humour is a sad story but makes people laugh.

Hanjalic [2005] only reported one low-level audio feature, pitch frequency, for valence computation (Equation 6.33).

$$p(k) = P_L(k) - N \tag{6.33}$$

Where $P_L(k)$ is the pitch number in a video segment with length L; N is the so-called *neutral feeling* frequency, which maps low pitch frequency to corresponding negative valence. In [Hanjalic, 2005], N was set to 150. Murray and Arnott [1993] claimed that pitch frequency was useful in distinguishing positive and negative emotions, such as happiness (high-pitch frequency) and sadness (low-pitch frequency).

As well as other salient features, valence estimation from pitch frequency $p(k)$ is smoothed by a Kaiser Window for a pitch-frequency valence component $D_1(k)$ as follows.

$$\widetilde{p}(k) = p(k) * K(l,\beta) \tag{6.34}$$

$$D_1(k) = \frac{max_k \|p(k)\|}{\max_k \|\widetilde{p}(k)\|} \widetilde{p}(k) \tag{6.35}$$

Hanjalic [2005] took this component $D_1(k)$ as the valence intensity of a video.

Arousal Curve Fusion

In the above sections, three arousal component curves are extracted from motion field, shot transition and audio energy. However, peaks in these curves are asynchronous. This means that these peaks rarely appears at the same time, although they may refer to the same game events. The reasons are as follows.

1. time resolution difference for arousal component computation. All arousal components are average during a given time interval;

2. time delay between modalities. For example, a goal event may incur a sequence of media events like that: spectators cheers (a peak in audio energy arousal), then a camera zoom-in or a camera switch occurs (a peak in motion arousal), and finally a replay segment appears, which switches several shots swiftly (a peak in shot transition arousal);

3. arousal reflection bias. Arousal components extracted from motion and shot transition reflect game content understanding from video producers. It takes time for these reflectors to understand game contents and thereby select proper methods to iterate a game story. However, the arousal component from audio energy comes

from an automatically recorded sound track. This indicates there are temporal delays between arousal peak from different modality features.

Moreover, a peak in an arousal curve is a weak proof for sports highlights, because such a peak may be incurred by other trivial issues. For instance, a balloon passing the sky in a stadium may cause a fast camera switch and result in a peak in shot transition arousal. Therefore, it can improve the robustness and effectiveness of game event detection to fuse these arousal components. Hanjalic [2005] did not combine all arousal components into a unified arousal curve. The authors designed a sliding temporal window which was able to cover two neighbour shots, and used the sum of peak numbers in this window to guess the possibility of a highlight. This combination scheme is simple but effective. Another approach which the authors suggested, was to calculate a weighted average of arousal intensity over a long time interval so that neighbouring local maxima of arousal components could be merged.

Additionally, the combination of arousal components affects the estimation of valence, because these psychological measurements are not entirely independent. As mentioned in Section 6.3.2, there is a constraint between valence scale and arousal intensity to ensure that an emotion curve can meet the requirement of a parabolic shape in the emotional region. Hence, the fusion of arousal components confronts many internal psychological problems.

6.3.3 Discussion

The estimation of video attractiveness is a complex task. There are three issues requiring careful consideration, namely, salient feature selection, component attention modelling and attention fusion. Given the pervasive nature of human emotion, numerous stimuli from physical signals, linguistic understanding and emotional interactive affection, can incur emotional responses. It is necessary to decide the scope of emotion before selecting salient features. Component attention models simulate an response process incurred by a salient feature in order to estimate a possible emotional increment. The main challenge in the component attention model proposition is that these models should be supported by clear psychological evidence. Attention fusion combines component attentions to facilitate the identification of the most interesting events in a video.

In this section, two affective models have been introduced, namely user attention model [Ma et al., 2002] and affective video content representation [Hanjalic, 2005] [Hanjalic and Xu, 2005]. Working in the attention space, user attention model is from a well

developed background of active vision. A set of feature-based attention component models is developed from individual modality features. These attention components are linearly combined into a unified attention or so-called average viewer attention [Ma et al., 2002]. The model of affective video content representation is constructed in the arousal-valence space. Since a video is a temporal continuous stream, this model pays attention to temporal signal variations rather than static contrast. This leads to a small but effective salient feature set. Such a representation model identifies the temporal delay between component emotional curves, and employs a sliding time window for the counting of arousal peaks. Note that the number of arousal peaks is assumed to be propositional to the probability of highlight occurrence in a given time interval.

However, both models are with apparent drawbacks. The user attention model ignores modality asynchronism and the physical nature of salient signals. For example, static and face attention is computed for every visual frame, while motion attention is calculated between frames. There are differences in signal scales. Hence, the complexity of attention component fusion increases dynamically with the number of salient features. The unified estimation of attention intensity is sensitive to modality noises, although all attention components are normalised. Moreover, too many parameters are introduced, which results in expensive computation costs. Additionally, the hypothesis of average user attention conflicts with the psychological assertion of independent attention [Osgood et al., 1957]. The extra valence dimension plays a contradictory role in the model of affective video content representation. The computation of valence helps the discrimination of psychological events. But arousal and valence are not completely independent psychological characteristics. This psychological ambiguity results in extra complexity. Hence, valence estimation is fragile.

6.4 Attention When Watching Sports Videos

Although Ma et al. [2002] and Hanjalic and Xu [2005] used a general attention framework for sports video analysis, it is worth surveying the particular attention process when watching sports videos. Such an analysis clarifies this psychological story in order to identify useful salient features and avoid possible misunderstandings.

A sports video is a compact record of a game. Moreover, a successful sports video increases the ratio of interesting moments by inserting various video editing effects, such as replay and zoom-in. The target of a sports video is to attract attention and keep viewers before a TV set as long as possible. Given the simple video content, such as a

goal in a football video, a sports video is mostly an attention or arousal guided video. Therefore, valence is usually meaningless, because both positive and negative valence are incurred simultaneously by the nature of competition in a game. Hence, the attention emotional space is enough to describe the emotion variation during watching of a sports video.

6.4.1 Football Video Perception Structure

A broadcasting football video involves several perception roles during the recording. Three major reaction roles can easily be identified from the visual stream and audio track, namely spectators, commentators and video directors (Figure 6.6).

Figure 6.6: Perception Roles in Football Video

These individual understandings affects the feeling of viewers from numerous aspects, such as a cheer in the audio track (spectators), a word of goal (commentators) and a fast camera zoom-in (video directors). Video directors edit camera videos, decide shot style and insert video editing shots following their own understandings of game contents. Therefore, video directors dominate the visual stream. Many visual patterns are developed and utilised in game content analysis [Ren and Jose, 2005] [Babaguchi et al., 2002] [Tjondronegoro et al., 2004*b*]. Some salient features, such as dominant color ratio [Xie et al., 2002][Xu et al., 2001] and zoom size [Ren and Jose, 2005] are calculated to guess the content importance of a video segment, although the authors did not link these features with psychological effects. For example, a closer view can bring more details inside a game pitch and thus assigns more prominence as well as attracts a little more attention. Replay is another example. Pan et al. [2001] extracted replay segments to build video summary, because these segments were designed to iterate important moments. Note that a replay segment invert the time sequence and results in a strong

temporal contrast, which definitely incurs strong attention. Nevertheless, shot duration is a common measurement of action pace, which indicates the intensity of arousal and is used to discriminate action films and love films [Adams et al., 2000]. Besides these low level features, content-based structures in a sports video reflect the attention period of directors. Xie et al. [2002] and Ren and Jose [2005] identified video structure by dedicated hidden Markov models on the feature shot style, which is a significant measurement of director behavior as well as shot switch frequency. Additionally, Hanjalic [2005] claimed that a close view, replay segments and shot length reflected arousal intensity, the emotional equivalence of attention. It can be easily understood that video directors employ a close view and keep a long shot duration indicates an increase of director attention.

The responses from spectators and commentators can be identified in an audio track as well as the visual stream occasionally. As a group, spectators cheer for exciting moments and keep quiet in other time. They attract attention from video viewers by loud plaudits. Moreover, visual shots on spectators are relatively rare, but impress upon the atmosphere in the stadium. There are two cases for spectators to appear in a visual stream. After a highlight, such as a goal, video directors pay one or two field view on the spectator area to display an exciting scene. Otherwise, a close view on one or a small group of spectators indicates that video directors supposes the behavior of spectators is more interesting than a game. The role of commentators is complex. As a part of the job, commentators explain a game with a neutral attitude but occasionally with personal preference. Keywords in the comments can be detected to annotate game events [Xu and Chua, 2004]. However, commentators are also a group of professional spectators and excited just the same as are other spectators.

6.4.2 Attention Signal

As a psychological measurement, attention describes human behavior before stimuli and measures the period to reach and keep an stage of focus. Treisman and Kanwisher [1988] claimed that the reaction period exceeded $0.384sec$ against a strong and simple stimulus, such as a flash inside a dark room. Moreover, the authors reported that attention transmission from one stage to another stage would require similar duration. This means that a temporal interval between two attention peaks will be 0.7 sec at least. Additionally, Hanjalic [2005] utilised an 1-minute long low-pass window filter to smooth arousal signals. The successful experiments indicate that the spectrum of an attention signal is narrow or in a low bandwidth. Therefore, an attention signal is oversampled

at the updating rate of visual frames. It is possible to reduce noise and save computing cost by enlarging the temporal observation window on attention signals.

Note that attention variation is a perceptual reflection of game contents. Different attention signals in a video should be conceptually similar. This means these signals are of similar trends and temporal structures. For example, an interesting event incurs attention peaks in different components synchronously. Ma et al. [2002] and Hanjalic [2005] stated that attention peaks or crests indicated the appearance of an interesting game content. However, there is a temporal delay among these different component peaks. In a football video, spectators cheer for a successful goal and there is a spectator attention crest. Video directors notice the goal event and switch cameras to display this story. An attention increment is found in the shot frequency. Some video editing effects, such as replay, are inserted, which result in a director attention crest of replay segments. Exciting spectators are sometimes displayed, which leads to an increment of the off-field spectator attention. False alarm in sports event detection can be avoided by removing this temporal delay and aligning attention component peaks. Possible solutions are listed as follows.

1. Align peaks in the finest resolution. This is a direct approach but sometimes fragile because of modality noise.

2. Accumulate attention curves by integral. A inflexion of this sum curve allocates the best match moment. However, this approach requires that these signal are continuous and that the delay is short.

3. Carry out this crest matching at a coarse temporal resolution. Hanjalic [2005], Hanjalic and Xu [2005] and Xu and Chua [2004] showed that the football highlight detection would be effective at a coarse temporal resolution, although the authors did not decide the best temporal resolution for sports event analysis.

6.4.3 Role-related Salient Features

Treisman and Kanwisher [1988] and Lew [1996] stated that variation, contrast and stimulus energy were major issues attracting attention. Since a sports video is produced in a closed environment, *e.g.,*a game pitch and a stadium, temporal variations and contrasts play a significant role in order to stimulate viewers. This is why video production techniques, such as zoom-in, replay, and swift shot switch are widely employed in the depiction of game events [Zettl, 1990]. Table 6.1 lists a summary of salient feature from the perspective of video directors. Note that replay and field-away shots interrupt

a continuous perception process of watching videos and might hinder the understanding of video contents. Therefore, these video production methods are limited for essential game aspects. As such, the length of replay segments and field-away shots monotonically increases rather than deceases attention intensity in contrast to normal shots [Adams et al., 2000]. Additionally, we treat the occurrence of a replay segment as a boolean attention switch for highlight identification. If a replay segment appears, this segment will be labelled as a highlight. Another aspect is the rectangle of interest (ROI), which measures the size of an interesting area in a visual frame. This aspect involves two salient issues: camera zoom depth and semantic video object size. The deeper the zoom, the larger the ROI appears. Most of the salient feature extraction algorithms can be found in Chapter 3.

Reactions from spectators and commentators are mixed in an audio track. Absolutely loud and greatly varying sounds always attract attention, if ignoring linguistic perception. Table 6.2 lists a group of spectator-related and commentator-related salient features used in this thesis.

feature	attention facts	qualitative relationship
football size	zoom depth	$\wedge$
uniform size	zoom depth	$\wedge$
face area	zoom depth	$\wedge$
domain color ratio	zoom depth	$\vee$
line mark distribution	rect of interest	
goalpost	ROI	
penalty box	ROI	
shot duration	temporal variance	$\vee$
shot cut frequency	temporal variance	$\wedge$
motion vector	temporal variance	
zoom-in sequence	temporal variance	$\wedge$
replay	temporal contrast	
off-field shot	temporal contrast	

Table 6.1: Director-related Attention Features, $\wedge$ stands for a propositional qualitative relationship between feature and attention, while $\vee$ refers to an anti-propositional between feature and attention.

6.4.4 Feature Attention Operator

Salient features come from different modalities, such as audio, video and text. Therefore, these features are of different physical and psychological characteristics, e.g. sig-

feature	attention aspects	qualitative relationship
short time audio energy	loudness	$\wedge$
cross zero ratio	sound variation	$\wedge$
speech band energy	sound variation	$\wedge$
First order variation of MFCC	sound variation	
keyword	semantic understanding	
speech ratio		

Table 6.2: Spectator-related and Commentator-related Attention Features, $\wedge$ stands for the propositional qualitative relationship between feature and attention, while $\vee$ refers to anti-propositional between feature and attention.

nal updating rate and affective effects. To conceal these variations, it is necessary to find an approach to estimate a unified attention intensity contributed by these salient features individually. For example, Ma et al. [2002] directly normalised the range of all feature-based attention components into $[0,1]$. However, Hanjalic [2005] claimed that the stimulus-response was a noisy process. The authors proposed an operator (M_0 in Equation 6.36) to smooth normalised arousal signals by introducing a sliding Kaiser window.

$$M_0 : M_0(s) = K(l,\beta)N_p(s) \tag{6.36}$$

where $N_p()$ is a normalisation operator, which rescales the range of a signal into $[0,p]$; s refers to a salient signal and $K(l,\beta)$ denotes a Kaiser window smoother with length l and shape parameter β. In [Ma et al., 2002] and [Hanjalic, 2005], p is set to 1. Later, Hanjalic and Xu [2005] supposed that it can improve signal-noise ratio(SNR) to sharpen strong signal peaks in a salient signal. Therefore, the authors employed a sequential normalisation operator, M_1, in [Itti and Koch, 1999] which promote a small number of strong activity peaks by suppressing numerous similar activity peaks. This operator M_1 consists of three steps,

1. normalise the scale of a signal to a fixed range $[0..L]$;
2. find the location of the global maximum L_{max} and compute the average $\bar{L}$ of all other local maxima;
3. globally multiply the sequence by $(L_{max}-\bar{L})^2$.

In contrast to these signal enhancement processes in [Hanjalic, 2005],[Hanjalic and Xu, 2005] and [Wang and Cheong, 2006], we propose an approach based on information theory to estimate the intensity of attention and thereby alleviate algorithm dependency on data collection. As far as cognition psychology is concerned, attention is the ability

of information consuming. In a neutral situation in which people keep neutral or feel interested or uninterested in all active information sources, the pan-out speed of messages will decide the distribution of attention. This leads to two helpful evidents for attention estimation. Firstly, it is intuitive to introduce self-information (Equation 6.37) to measure the message sent out from salient audio-visual features. In information theory, self-information is defined as the amount of information that knowledge about (the outcome of) a certain event, adds to the overall knowledge.

$$Entropy = -\log_2(P_i) \tag{6.37}$$

where P_i is the appearance probability of a feature at the given value i. Moreover, the self-information can be robustly calculated by a histogram estimation of feature distribution. Some features have a known prior distribution. For example, shot updating is a stochastic process of event arrival at an expectation rate. The shot frequency follows the Weibull distribution [Vasconcelos and Lippman, 2000]. We use an EM algorithm to find the best fit and then compute the self entropy. Secondly, the unified attention state can be regard as follows.

$$I_{attention} = \overrightarrow{A}\,\overrightarrow{E} \tag{6.38}$$

where $\overrightarrow{A}$ denotes a vector of attention ratio over modalities; E refers to the modality contribution which indicates an information increment contributed by a given modality. This equation (Equation 6.38) suggests that we can develop a Kalman filter like technique to estimate the unified attention intensity with the temporal smooth constraint on perception and feature updating. Therefore, a multiresolution autoregressive framework is developed in Section 6.6. However, $\overrightarrow{A}$ is unknown in most cases of perception processes.

Therefore, a self-entropy operator M_2 (Equation 6.39 is proposed in [Ren, Jose and He, 2007].

$$M_2 : M_2(s) = N_1(-\log_2 P_i(s)) \tag{6.39}$$

where N_1 denotes a $[0,1]$ normalisation operator, $P_i()$ is the probability of a signal s at the i^{th} range scale. For computational convenience, the self-entropy M_3 (Equation 6.40) is sometimes employed directly.

$$M_3 : M_3(s) = -\log_2 P_i(s) \tag{6.40}$$

It is difficult to evaluate the performance of attention operators. Since a goal is the most important event in a football game, the attention intensity ratio of goal events

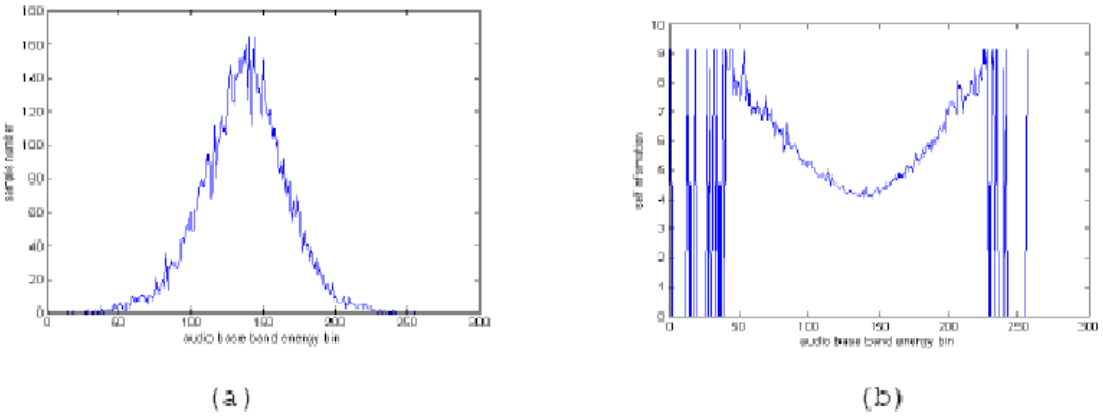


Figure 6.7: Audio based band energy distribution in the final game of World Cup 2002 (a)audio energy 256-bin histogram (b)audio energy self-entropy 256-bin histogram

over the average can reflect the effectiveness of these operators and therefore become a measurement of signal noise ratio (Equation 6.41). A higher signal noise ratio indicates that the better performance is an operator in the signal enhancement.

$$SNR_{goal} = 20\log\frac{\overline{I_{goal}}}{\overline{I_{all}}}(dB) \tag{6.41}$$

where I_{goal} and I_{all} refer to the average attention intensity over goal events and a complete game halve at 1 minute resolution, respectively.

We employed five games for operator effectiveness evaluation, including Brazil vs Germany (the final game in World Cup 2002), Brazil vs Turkey (a semi final game in World Cup 2002), Germany vs Korea (a semi final game in World Cup 2002), France vs Italy (the final game in World Cup 2006) and Arsenal vs Barcelona (UEFA Champion 2006). The salient feature collection includes average audio energy, zoom depth, and play field ratio. Table 6.3 displays the average attention intensity on goal events and a complete game halve under different attention operators. The *SNR* is the signal noise ratio computed by Equation 6.41. Note that the operator M_2 achieved a satisfying performance in all five games, although this evaluation is not enough to undercover the actual behavior of these operators in a large sports video collection.

In this thesis, we employ the operator M_2 or M_3 to estimate feature-based attention components unless other operators are specialised in computation. Both operator are based on an information theory explanation to attention phenomenon that attention is the ability of consuming information.

	M_0			M_1			M_2		
	Goal	All	SNR	Goal	All	SNR	Goal	All	SNR
Brazil vs Germany II	0.842	0.685	1.792	0.780	0.631	1.841	8.827	4.122	6.614
Brazil vs Turkey I	0.960	0.620	3.798	0.995	0.625	4.039	9.277	4.132	7.023
Germany vs Korea II	1.000	0.708	2.999	1.000	0.647	3.782	8.679	5.211	4.431
Italy vs France I	0.972	0.655	3.429	0.989	0.639	3.794	8.970	5.409	4.393
Arsenal vs Barcelona I	1.000	0.587	4.628	1.000	0.612	4.265	9.148	4.783	5.633
Arsenal vs Barcelona II	0.895	0.600	3.473	0.952	0.622	3.697	8.374	4.833	4.774
average	0.945	0.643	3.353	0.952	0.629	3.570	8.880	4.748	5.478

Table 6.3: Attention Peak Average Ratio for Normalisation Operator Evaluation. The sequence number I and II refer to the first half and the second half of a game, respectively. Three salient features are employed, average audio energy, grass ratio and zoom depth.

6.5 Role-based Attention Model

The following sections present two attention models which we proposed originally in this thesis, namely a role-based model [Ren and Jose, 2006] and a multiresolution autoregressive framework [Ren, Jose and He, 2007]. These models simulate the attention process in a sports video from different viewpoints. Role-based attention model simulates the perception structure in the sports video production (Section 6.4.1). This model argues for a two-layer attention fusion framework: (1) combine attention components into attention of directors, spectators, and commentators in order to remove the reflection bias among attention signals; (2) fuse director, spectator, and commentator attentions to detect game highlights. The multiresolution autoregressive framework regards salient feature curves as multiple observations of a noisy temporal process. Therefore, feature-based attention components are a set of signals, which vary with similar trends, such as the appearance of signal peaks, but contain intensive noise. Attention fusion, which combines multiple noisy attention components into a unified attention curve, corresponds to a signal estimation from multiple noisy observations at multiresolution. The multiresolution autoregressive framework develops a three-step algorithm to complete this estimation and information fusion, including a bottom-up multiresolution moving average (MA), a top-down multiresolution autoregressive (AR) and a signal-noise ratio estimation which decides a proper temporal resolution for sports video analysis.

The role-based attention model aims to combine attention components from the same observer and hence remove reflection bias. As mentioned in Section 6.4.1, reaction bias is one of main temporal asynchronous issues among attention signals. Moreover, such a categorisation classified attention components into three classes. Limiting the

number of attention components helps the fusion and estimation of a unified attention curve over a long video sequence. As such, salient features are grouped according to reflectors; three role-based attention curves are proposed for directors, spectators and commentators, respectively.

A system framework for role-based attention analysis is shown in Figure 6.8, which is made up by three main steps: role-based attention estimation, role attention curve purification and unified attention fusion. The estimation defines a set of modality features to extract role-based attention curves and thereby decides system reliability in highlight detection and event segmentation. The purification removes signal noises and excludes replay and field-away segments to facilitate attention fusion. This is because these video segments convey only director attention rather than a mixture. In the attention fusion, a unified attention curve is estimated for a game video, which combines attention intensity from different reaction roles, removes media asynchronism and tracks attention variation precisely.

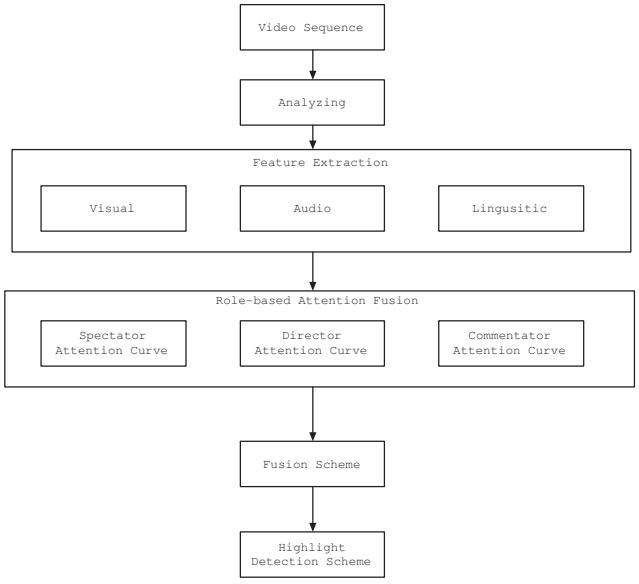


Figure 6.8: Role-based Attention Analysis Framework

The attention of video director is extracted from a visual stream. This attention curve includes two components, static attention and temporal variation $\overline{T_n}$ (Equation 6.44). The former component consists of zoom depth $\overline{Z_n}$ (Equation 6.42) and ROI C_n (Equation 6.43) in Table 6.1. Therefore, three salient curves are detected, namely, zoom depth,

ROI ratio, and temporal contrast.

$$Z_n = \frac{N_1(UniformSize_n) + N_1(FaceArea_n)}{N_1(DomainColorRatio_n) + 1} \quad (6.42)$$

$$C_n = N(N_1(GoalPostArea_n) + N_1(PenaltyArea_n)) \quad (6.43)$$

$$T_n = \frac{N_1(ShotFrequency_n) * N_1(Motion_n)}{N_1(ShotDuration_n) + 1} \quad (6.44)$$

where N_1 denotes a $[0,1]$ normalisation operator. The size of an uniform is estimated in Section 3.5; the calculation of domain color ratio is found in Section 3.3; a goal post and a penalty area are detected by a FST classifier (Appendix A); shot frequency and shot duration are extracted by the shot segmentation algorithm in Section 3.2. These time sequences of attention components are normalised and added up into a director attention curve VD_n as Equation 6.45. Figure 6.9 displays a director attention curve at 0.3 sec resolution in the game Germany vs Brazil in FIFA World Cup 2002.

$$VD_n = \frac{1}{2}(N_1(N_1(Z_n) + N_1(C_n)) + N_1(T_n)) \quad (6.45)$$

As mentioned in Section 6.5, spectator attention is propositional to the loudness of

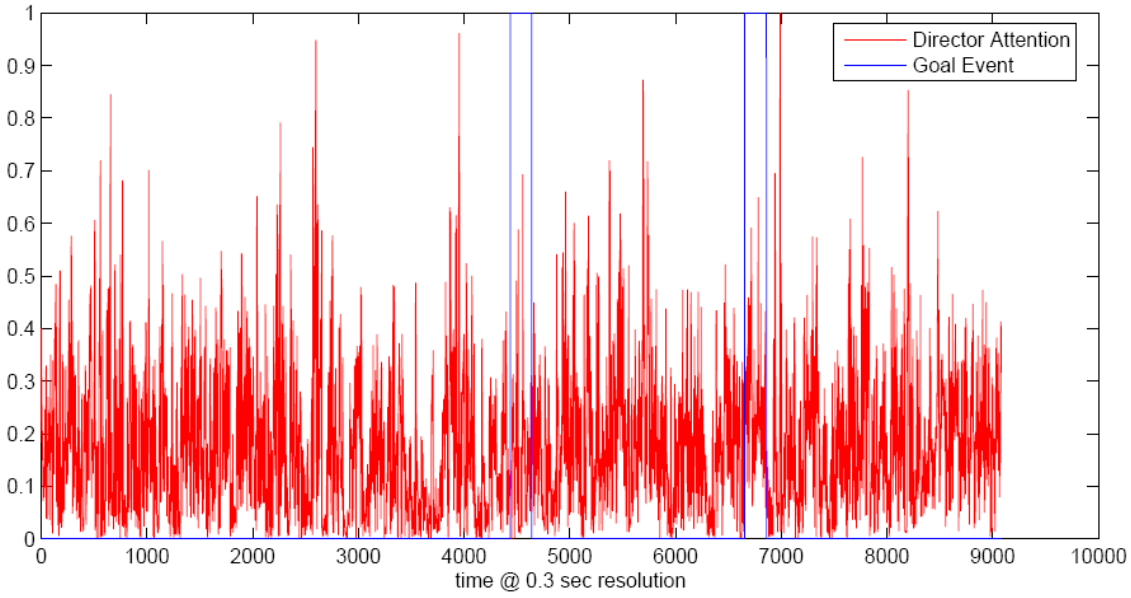


Figure 6.9: Director Attention Curve @ 0.3 sec in Brazil vs Germany, World Cup 2002. Blue lines denote time intervals of goal events in the FIFA record.

background noise. Hence, short time energy in an audio track and the absolute difference of audio energy in a short time interval are used to estimate VA. Four scales

are taken, including 1 sec, 5 sec, 30 sec and 1 minute. Hence, spectator attention VA_n is a five element vector, $(E_0, D_1, D_5, D_{30}, D_{60})$, where n is the time stamp of spectator attention; E_0 denotes the audio energy in 0.3 sec; D_1, D_5, D_{30}, D_{60} refer to the absolute difference from 1 sec, 5 sec, 30 sec, 60 sec mean audio energy, respectively. For computational convenience, the intensity of spectator attention is transformed into a scalar in $[0, 1)$ (Equation 6.46).

$$VA = \frac{1}{5}(N(E_0) + N(D_1) + N(D_5) + N(D_{30}) + N(D_{60})) \tag{6.46}$$

It is difficult to estimate commentator attention precisely. This is due to the complexity of commentator behavior in the sports video production. As an iterator, the excitement of commentators can be detected by speech speed and voice loudness. Xu and Chua [2004] claimed that the low band of *LPCC* (from 0 to 3) and zero cross ratio (ZCR) were effective to estimate commentator attention. Therefore, a commentator attention model VC is proposed in Equation 6.47. However, an audio track is a noisy mixture rather than a clean speech in an auditing room, which combines background noise, comments and music clips. It is difficult to separate these sounds. Moreover, the feature of ZCR is sensitive to noise. Such a commentator attention model is fragile because of strong noise in an audio track of sports videos. Additionally, commentators are professional spectators. Their behavior is likely to be similar to other spectators in the stadium. It sounds reasonable to estimate both spectator and commentator attention in a unified model rather than two individual models. Therefore, an audio attention model is proposed by Equation 6.46. Figure 6.10 displays the audio attention curve at 0.3 sec time resolution in Brazil vs Germany, World Cup 2002.

$$VC = N_1(N_1(\sum_{i=0}^{3} LPCC_i) + N_1(CrossZeroRatio)) \tag{6.47}$$

Although these role-based attention models try to remove asynchronous noise, Figure 6.9 and Figure 6.10 show that attention signals themselves are noisy, especially at the temporal resolution of 0.3 sec. To improve the robustness and efficiency of sports highlight segmentation, several processes are necessary, such as role-based attention curve fusion, multiresolution attention signal analysis, and signal quality analysis.

Role-based Attention Fusion By Signal Correlation

Role-based attention fusion is designed to combine role-based attention curves in order to estimate a unified attention intensity. Note that we focus on how to extract a complete

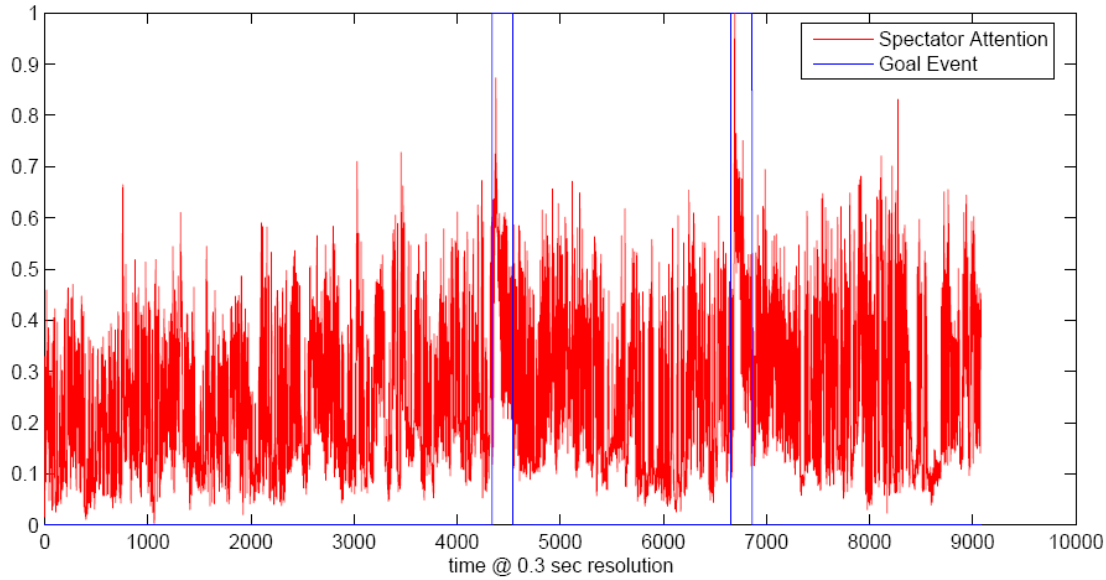


Figure 6.10: Spectator Attention Curve @ 0.3 sec in Brazil vs Germany, World Cup 2002. Blue lines denote time intervals of goal events in the FIFA record.

game event rather than a moment. Entire stories are what we look for, each of which should include a start point, an end point and several content iterations, such as replays from different viewpoints. Therefore, the approach in [Hanjalic and Xu, 2005] is not suitable, which counts the number of arousal peaks in a sliding window to estimate the possibility of highlights in a game. Moreover, a unified attention curve is necessary to facilitate the segmentation of game events.

Attention fusion is a combination of two long noisy temporal sequences, which are of the same signal scale and updating rate, although random delay exists between signals. We cut these sequences into short clips, in which the temporal delay could be treated as a constant roughly. Since these attention curves observe the same content process and are supposed to have a similar variation trend 6-0Durbin-Watson test value in the final game of World Cup 2002 is smaller than 1.4, which hints a positive correlation between audio and visual attention curves., we match and mix relevant signal segments by maximising signal correlation (Equation 6.52). Blank intervals, due to the removal of replay segments, provide natural stop marks in attention curves. To improve signal noise ratio, a Kalman filter is employed to smooth the combined attention curve and a re-sampling to 1 sec is used as well as shortening sequence length.

Let $C(role,t)$ denote role-based reaction matrix and $x(t)$ be an intensity of content in-

terest, a role-based attention $z(role,t)$ can be estimated as follows.

$$z(role,t) = C(role,t)x(t) + v(role,t) \tag{6.48}$$

where $v(role,t)$ is the observation noise with zero mean for a reflect role at a moment t. To detail, a director, spectator, and commentator attention in i^{th} attention segment can be stated as,

$$z_{director}(i,t) = C(dir,t)x(t) + v_{dir}(t) \tag{6.49}$$
$$z_{spectator}(i,t) = C(aud,t+\Delta t_a)x(t) + v_{aud}(t) \tag{6.50}$$
$$z_{commentator}(i,t) = C(com,t+\Delta t_c)x(t) + v_{com}(t) \tag{6.51}$$

where Δt refers to a reaction bias. An observation windows at the width of ± 30 sec is employed to compute signal correlation between 30 sec long attention segments. This is because FIFA recorded game events at the temporal resolution of 1 minute in official documents.

$$\Delta t = \arg \max_{\Delta t \in [-30,30]} \sum_t z_{director}(t) z_{audience}(t+\Delta t) \tag{6.52}$$

Therefore, the matching between visual and audio attention segments is completed by the maximisation of signal correlation. A unified attention curve is estimated by linearly combining matched audio and visual attention segments.

$$Z(i,t) = \sum_n \eta_n z_n(i,t) \tag{6.53}$$

where η_n is the weighting coefficients with $\sum \eta_n = 1$.

This is an intuitive solution to combine attention curves by maximising signal correlation. As such, attention peaks and valleys between different attention curves are roughly synchronised. Signal-noise ratio of attention signal over observation noise is thereby increased according to the proposition of attention analysis (Section 6.2.2) and role-based attention model. However, this approach is conceptually acceptable but faces many problems in processing. This is because a noise model can hardly be defined in the psychological process of watching a video. Moreover, correlation computation can enhance signal noise as well as attention signals. The estimation of a unified attention curve is not so robust in the case of strong noise. Moreover, it is necessary to estimate an optimised temporal resolution for attention clip segmentation during the matching process, which balances the tradeoff between the allowance of observation noise and the precision of event boundary detection. On one hand, a long attention segment pro-

vides plenty of observations which help the smoothing of signals and thereby removing strong noise. An attention segment should be long enough to grantee the robustness of a fusion algorithm. On the other hand, attention segments employed for modality matching, should be short in order to keep a roughly constant reaction delay and meet the temporal precision of event boundary detection. Hence, a careful balance is necessary in the decision of audio segment length. Nevertheless, the computation of signal correlation is a processing at a single time resolution. This means that this algorithm can hardly employ the multiresolution nature of attention signals which may lead to an efficient representation method of game contents. To solve these problems, we proposed a multiresolution autoregressive framework to complete the attention fusion and thereby developed a new attention model for sports videos.

6.6 Multiresolution Autoregressive Model

A role-based attention model is a direct description of sports video enjoyment at the phase of video production. This model is proposed to remove reaction bias and subsequently to develop an efficient attention simulation approach. However, such a model is too specialised to be extended to other video genres. A general attention model is required to make the attention-based content analysis suitable for general video genres. These work leads to the framework of a multiresolution autoregressive attention model.

Generally, attention curves are psychological reactions to a smooth temporal process, which illustrates a well organised reasonable game story. This indicates that attention signals are temporally smooth and can be described by some Markov chains. Moreover, attention intensity peaks are propositional to the possibility of game highlight occurrences, according to the attention proposition in Section 6.2.2. Note that game events are a relatively long temporal interval and related attention peaks can be observed from different resolution. This characteristic is important because these attention peaks which refer to game events, can be discriminated from strong random noise by multiresolution analysis.

The multi-resolution autoregressive model (MAR) is a multi-scale recursive linear dynamic model [Chou et al., 1994]. This model simulates a random process by a series of autoregressive (AR) models at multiple scales and is able to combine heterogeneous data in multiple spectral bands and at different spatial or temporal resolutions following a given criteria. Willsky [2002] proved that a MAR was equivalent to a Markov on graph, which is of the same computational complexity as the video content modelling.

However, attention ratio distribution $\overrightarrow{A}$ or attention gain on different modalities, such as visual attention, is unknown. It is unrealistic to combine modality based attention signals directly. Note that audio and visual stream are two independent observations on the same message production process (Section 6.4.2). We can employ one media attention curve as a measurement to the other and thereby combine all modality based attention curves by a MAR model. In other words, audio and visual attention curves are independent and rough observations, and a better estimation can be drawn by a MAR tree. This assumption can be formalised as follows.

Denote the set of resolutions by $\mathbb{R} = \{1, ..., R\}$, with $r = R$ being the finest resolution. Extracted visual salient features are of different resolutions. For example, shot frequency is meaningless if the width of an observation window is less than the minimum of shot duration. Hence, we set the finest combination resolution as 1.4 times of the longest shot duration. In experiments, the observation windows is about 50 sec, whose width is very close to 1-minute window [Hanjalic, 2005]. The node N at scale r is $N_n^{(r)} = \{1 : 2^r\}$ in the case of a bi-tree. Let $x(s)$ be the observation vector of visual attention on a node s and $y(s)$ denote the audio part, the discrete-time attention process can be described by a linear stochastic difference equation as follows.

$$y(s) = \frac{1}{N} H x(s) + v(s) \tag{6.54}$$

We assume the contribution of visual salient features are of the same importance6-0It is a widely accepted practice in affective analysis. For example, affective effect from different salient modalities are combined with the same weight in [Hanjalic, 2005][Hanjalic and Xu, 2005] and [Wang and Cheong, 2006].. H is a vector of $\{1, ..., 1\}$ and N is the normalisation parameter. $v(s)$ is Gaussian noise on the tree. We use a binary tree in the MAR model, the projection from finer resolution to coarse resolution will be

$$x(s) = [0.5, 0.5]^T x(s|s-) + w(s) \tag{6.55}$$

where $x(s|s-)$ is the sub-tree under node (s), $w(s)$ is the Gaussian noise. The Rauch-Tung-Striebel (RTS) smoother can produce the best estimation of this temporal process [Willsky, 2002].

Therefore, a two-step algorithm is developed to estimate a MAR model from data. The

first step of fine-to-coarse swapping is a generalisation of a Kalman filter but works for a multi-scale tree. A three-step recursion is involved, namely prediction, measure updating, and the fine-to-coarse merge when moving up to a coarse resolution. The second step of coarse-to-fine smoothing distributes AR model parameters, which are obtained in the former fine-to-coarse step, such as the covariance matrix at coarse resolution. These information is used to improve prior estimations at a finer resolution. This estimation algorithm will be presented by following sections in detail.

6.6.1 Fine-to-coarse Sweep

In the fine-to-coarse sweep, $\hat{x}(s|s)$ denotes the optimal estimate of $x(s)$ at each node s, which is computed by data in the sub-tree rooted at node s, together with $P(s|s)$, the error covariance in the estimation.

Initialisation

Initialise at the finest resolution. For each finest scale leaf node s, the estimation of $\hat{x}(s|s-)$ and the covariance $P(s|s-)$ from the sub-tree are

$$\hat{x}(s|s-) = 0 \tag{6.56}$$

$$P(s|s-) = P_x(s) \tag{6.57}$$

Measure Updating

Measurement updating is identical to the analogous equations in a Kalman filter.

$$\hat{x}(s|s) = \hat{x}(s|s-) + K(s)v(s) \tag{6.58}$$

where v(s) is the measurement innovations,

$$v(s) = y(s) - H\hat{x}(s|s-) \tag{6.59}$$

which is zero-mean with covariance,

$$V(s) = HP(s|s-)H^T \tag{6.60}$$

and where the gain $K(s)$ and the updated error covariance $P(s|s)$ are given by,

$$K(s) = P(s|s-)H^T V^{-1}(s) \tag{6.61}$$
$$P(s|s) = [I - K(s)H]P(s|s-) \tag{6.62}$$

Sub-tree fusion

The second step is the fusion of estimates from immediate children at node s. Specifically, let $\hat{x}(s|sa_i)$ be the optimal estimate at one of children sa_i of node s and v_{sa_i}, the sub-tree rooted at sa_i, and $P(s|sa_i)$ for the corresponding error covariance, the fusion step is,

$$\hat{x}(s|s-) = P(s|s-)\sum_{i=1}^{K_s} P^{-1}(s|sa_i)\hat{x}(s|sa_i) \tag{6.63}$$
$$P^{-1}(s|s-) = P_x^{-1}(s) + \sum_{i=1}^{K_s}[P^{-1}(s|sa_i) - P_x^{-1}(s)] \tag{6.64}$$

Fine-to-Coarse Prediction

To estimate $\hat{x}(s|sa_i)$ and the error covariance for each child of s, a one-step prediction step is proposed similar with Kalman filter.

$$\hat{x}(s|sa_i) = F(sa_i)\hat{x}(sa_i|sa_i) \tag{6.65}$$
$$P(s|sa_i) = F(sa_i)P(sa_i|sa_i)F^T(sa_i) + U(sa_i) \tag{6.66}$$

where

$$F(s) = P_x(s\bar{r})A^T(s)P_x^{-1}(s) \tag{6.67}$$
$$U(s) = P_x(s\bar{r}) - F(s)A(s)P_x(s\bar{r}) \tag{6.68}$$

6.6.2 Coarse-to-Fine Sweep

When the fine-to-coarse sweep reaches the root, the covariance and estimation at all nodes are available. Note that the fine-to-coarse step experiences all possible time delay. If the temporal resolution is coarse enough, attention from different modalities, i.e. audio and visual, are synchronous. This is because both modalities observe the same content movement. In particular, the coarse-to-fine step fuses a node s with the optimal

smoothed estimates and covariance at its parent $s\bar{r}$.

$$\hat{x}_s(s) = x(\hat{s}|s) + J(s)[\hat{x}_s(s\bar{r}) - \hat{x}(s\bar{r}|s)] \quad (6.69)$$

$$\hat{P}_e(s) = P(s|s) + J(s)[P_e(s\bar{r}) - P(s\bar{r}|s)] \quad (6.70)$$

where

$$J(s) = P(s|s)F^T(s)P^{-1}(s\bar{r}|s) \quad (6.71)$$

6.6.3 Unified Attention Estimation

In the prior two steps, the information from modality features are combined. The knowledge from the audio stream is distributed into visual attention sequences. The unified attention is estimated as a mean of all available visual attention on a given resolution (Equation 6.72). The algorithm for highlight allocation is a tree search process and is carried out at multi-scales (Algorithm 3).

$$A_i(s) = \frac{1}{N} H_i x_i(s) \quad (6.72)$$

where $x_i(s)$ is the visual attention vector at the resolution i. H is a vector of $\{1,...,1\}$ and N is the normalisation parameter.

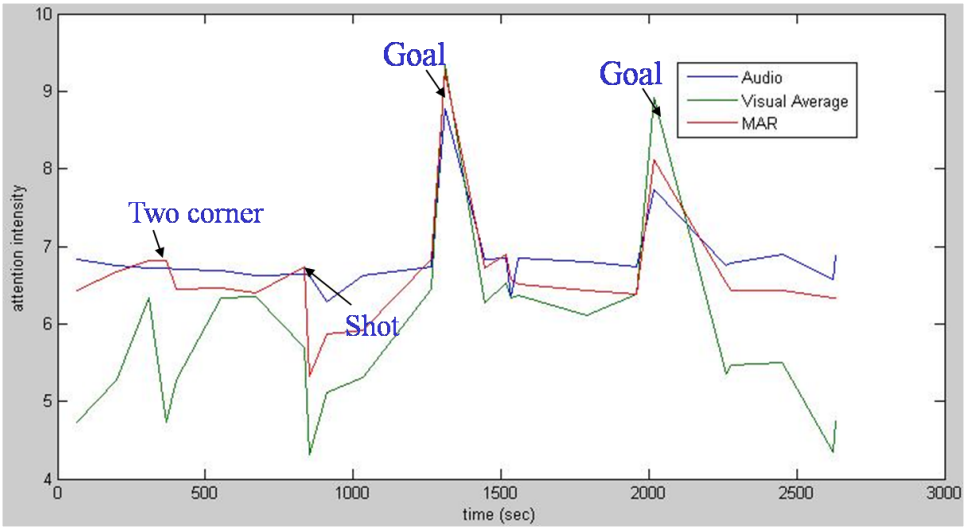


Figure 6.11: Unified attention curve @ 5 minute resolution in the second half of Germany vs Brazil, the final game of World Cup 2002

6.6.4 Highlight Segmentation in MAR

As we have mentioned in Chapter 2, game highlights are distributed sparsely across a game and attention patterns for game events are significantly different from those for normal game contents. A node in a coarse resolution stands for a video segment, in which attention varies smoothly and regularly. Moreover, attention peaks which refer to game highlights can be observed from all time resolutions. Therefore, a variation of the deep-first search algorithm is developed for event segmentation (Algorithm 3). The

Data: Event number N, a MAR tree(n,S), $\|n_s\|$ node number at layer s;
Result: Event Segment List S
$s = 0$;
while $\|n_s\| \leq N$ **do**
  s++ ;
end
sort n_s by attention intensity at each node;
for *int* $i = 0$; $i < N$; $i++$ **do**
  get *nc*, the children of node n_i ;
  divide *nc* into two groups, *ncl* and *ncr* denoting children before and after n_i, respectively ;
  while *ncl is not a leaf* **do**
    select *ncll* the children of *ncl* with lowest attention intensity; *ncl* = *ncll*;
  end
  while *ncr is not a leaf* **do**
    select *ncrr* the children of *ncr* with lowest attention intensity; *ncr* = *ncrr*;
  end
  Add event(ncl, ncr) into S;
end

Algorithm 3: Event Segmentation in a MAR tree

event list *S* is sorted by attention maximums in each event segments and the top five are regarded as game highlights.

6.7 Experiment

The experiment data set is selected from game collections of FIFA World Cup 2002, World Cup 2006, and Champions League 2006. Six games are involved, three from World Cup 2002, Brazil vs Germany (final), Brazil vs Turkey(semi final), and Germany vs Korea(semi final); one from World Cup 2006, Italy vs France(final); and two from Champions League 2006, Arsenal vs Barcelona, and AC Milan vs Barcelona. These videos were recorded from BBC and ITV in MPEG-1 PAL format with visual resolution at 352×288 and audio at 224kbit/s. To set up ground truth, we collected game

records from the FIFA official website to define the ground truth list of content-based game events and a highlight collection from BBC Sports website and FIFA highlight videos. Note that the temporal resolution of official game records is of minute, and there is a starting point misalignment between broadcasting and real games. We set a 30*sec* allowance for matching official records and experiment results. Each of the games were divided into halves, e.g. Brazil-German I for the first half of the final game in World Cup 2002 and II for the second half, to remove interview clips in the middle break. Other broadcasting aspects were kept, including player entering, celebrating, and the coach information board at the beginning of games.

We labelled all content-based events in the second half of the final game in World Cup 2002, according to the FIFA game record. Table 6.7 lists the attention intensity at events. We observed that attention intensity at events is significantly greater than that on normal contents. Moreover, as the most interesting part in a football game, goal events convey the maximum attention intensity over all. Table 6.5 compares the difference of attention intensity under multiple resolution. In the Bra-Ger II, the maximum of average event attention appeared at the temporal resolution of 76*sec*, but the maximum of signal noise ratio appeared at the temporal resolution of 5*min*(304*sec*). This result shows that the observation window of 5*min* width is the best choice for event detection and we should employ the 1*min* wide window for event segmentation. It is interesting that the result meets some facts in the statistics of game and video production: an effective shot appears about every 5 minutes and the replay duration is about 1 minute. Table 6.5 compares average attention intensity over different resolutions. Some interesting conclusions can be drawn: (1) the maximum of average event attention appears at the temporal resolution of 76 sec; (2) the maximum of signal noise ratio is reached at the resolution of about 5 min (304 sec). These findings indicate that the observation window with 5 min width is the best choice for event detection and the 1*min* for event segmentation.

As the hypothesis of affective approaches, game events catch more *attention* than general content does. The average *attention* ratio between highlights and the whole game video can evaluate the performance of a modality fusion algorithm efficiently (Equation 6.73). A good fusion algorithm will enhance the ratio. However, it is difficult to agree on the selection of highlights except goal events. We only use the attention intensity on goals.

$$R_{attention} = \frac{E(A_{events})}{E(A)} \sim \frac{E(A_{goal})}{E(A)} \tag{6.73}$$

Time	Event	Attention @ 1 sec resolution		
		Audio	Visual Average	Estimated
1'04	shot	6.832	4.733	6.431
3'18	free kick	6.748	5.272	6.670
5'08	corner	6.713	6.340	6.819
6'10	corner	6.713	4.733	6.819
6'40	shot	6.708	5.272	6.451
9'15	shot	6.687	6.340	6.469
11'10	shot	6.616	6.345	6.399
13'56	free kick	6.639	5.700	6.742
14'10	offside	6.619	4.320	5.330
15'10	shot	6.287	5.121	5.877
17'15	shot	6.631	5.312	5.910
21'05	free kick	6.740	6.450	6.833
21'50	**goal**	**8.774**	**9.327**	**9.255**
24'04	shot	6.835	6.273	6.714
25'15	shot	6.844	6.520	6.89
25'35	corner	6.355	6.332	6.573
25'59	corner	6.843	6.374	6.510
29'50	corner	6.798	6.113	6.432
32'40	corner	6.729	6.377	6.388
33'38	**goal**	**7.728**	**8.923**	**8.112**
37'40	shot	6.754	5.352	6.527
37'58	shot	6.786	5.473	6.435
40'50	shot	6.897	5.497	6.435
43'43	free kick	6.581	4.345	6.327
43'52	shot	6.875	4.740	6.340
mean over events		6.829	5.964	6.628
mean over all		4.374	4.731	4.020

Table 6.4: Event list of $2^n d$ half in Brazil vs German, World Cup 2002

where E is the expectation function, and A_{events}, A_{goal}, A stand for the attention intensity on events, goals and the whole game, respectively.

The linear combination algorithm in [Ma et al., 2002] was taken as the baseline and we employed the feature set {average block motion, shot cut density, base band audio energy} in [Hanjalic, 2005] and [Hanjalic and Xu, 2005]. In Table 6.6, six methods are compared: Linear I Ma et al. [2002] is the baseline, which directly adds up normalised feature values; Linear II linearly combines normalised feature values with the optimised weights from the fine-to-coarse step in MAR fusion; MAR I compares attention intensity on the leaves of MAR tree (0.3 sec), while MAR II on 1-minute resolution; Linear III and MAR III are similar to Linear I and MAR II respectively, but work on the seven-

Resolution	1.2	38	76	152	304	600
event mean	6.628	6.628	**6.807**	6.743	6.671	6.563
average	4.020	3.974	4.122	3.532	3.432	**3.342**
delta	2.608	2.654	2.685	3.211	**3.239**	3.221

Table 6.5: Attention intensity under different resolution in $2^n d$ half in Brazil vs Germany, World Cup 2002

feature set. The performance of MAR framework is better than linear combination at all resolutions and MAR III gets the highest attention ratio.

As used in [Lenardi et al., 2004] and [Hanjalic and Xu, 2005], we counted the number

	Linear I	Linear II	MAR I	MAR II	Linear III	MAR III
Ger-Bra II	1.522	1.874	1.802	1.997	1.213	2.141
Bra-Tur II	1.671	1.944	1.972	2.187	1.371	2.245
Ger-Kor II	1.142	1.326	1.411	1.563	1.074	1.665
Mil-Bar II	1.377	1.700	1.741	2.043	1.176	2.226
Ars-Bar I	1.274	1.427	1.419	1.778	1.143	1.912
Ars-Bar II	1.192	1.325	1.422	1.760	1.051	1.732
Ita-Fra I	1.302	1.377	1.420	1.723	1.014	1.658

Table 6.6: Attention Ratio (Goals vs. General Contents) in Games for Fusion Algorithm Evaluation

of goal events in the top five *attention* peaks in Table 6.7 and recorded related average of *attention* intensity (Table 6.8).

It is interesting to find that many replay segments of goal events are detected without using any temporal sequence information (4). As a special video editing technique, replay segments invert temporal sequence and are used to depict important game events. Hence, replay segments perceptually attract *attention*. But such a phenomenon did not be reported in other affective approaches [Ma et al., 2002][Hanjalic, 2005]. We owe this to the measurement of self-information entropy, because manual composition techniques incur the rare appearance of feature values, i.e. silence clip in an audio track and a swift increment of shot frequency, which will gain a high self information entropy value.

As a case study, we compare the highlight list from BBC Sports and FIFA official website for the game Italy vs. France with *attention* in Table 6.9. Note that the coverage of these manually selected highlights in the top five of *attention* peak list is 100%.

	Goal Number	Detected Goal Events	Rank
Ger-Bra I	0	-	-
Ger-Bra II	2	2	1,2,3,4,5*
Bra-Tur I	0	-	-
Bra-Tur II	1	1	1,2*
Ger-Kor I	0	-	-
Ger-Kor II	1	1	1
Mil-Bar I	0	-	-
Mil-Bar II	1	1	2
Ars-Bar I	1	1	1
Ars-Bar II	2	2	2,3
Ita-Fra I	2	2	1,2,4*
Ita-Fra II	0	-	-

Table 6.7: Performance of Goal Detection (*goal events are replayed for several times)

	Appearance Number	Mean Attention @ 1 min	
		Goal Events	All
Ger-Bra II	5	8.827	4.122
Bra-Tur II	2	9.277	4.132
Ger-Kor II	1	8.679	5.211
Mil-Bar II	1	9.506	4.270
Ars-Bar I	1	9.148	4.783
Ars-Bar II	2	8.374	4.833
Ita-Fra I	2	8.970	5.409

Table 6.8: Goal and general contents attention

FIFA	BBC Sports	Rank
Players enter the field	-	3(I)
Penalty	Zidane Penalty	1(I)
Goal	Goal	2,4(I)
-	Zidane expulsion	3(II)
Italian Triumph	-	1(II)

Table 6.9: Game Highlights and Attention Rank in France vs Italy (I,II for the game part)

6.8 Conclusion and Discussion

The *attention*-based approach is an exploration from computing science towards psychology. Utilising the psychological explanations of the perceptual process during watching a video, the *attention*-based method shows the efficiency in the application of semantics discovery and content importance weighting, such as highlight identification [Hanjalic, 2005], key video object selection, and video genre discrimination [Wang and Cheong, 2006]. We propose two attention models for sports video analysis, namely, a role-based attention model and a MAR fusion algorithm. These models combine feature-based *attention* curves and estimate an optimised overall assumption on video attractiveness from multiple modalities.

The role-based attention model clusters attention modalities into three groups, namely director attention, spectator attention and commentator attention. Therefore, the combination of modality attention features is transformed into a two-step fusion process: (1) fuse modality features into role-based attention curves; and (2) estimate a unified attention curve from role-based attention curves. Video events are detected by ranking attention intensity of local peaks in the unified attention curve.

The MAR model is originally proposed for image reconstruction [Willsky, 2002]. It is the first time to introduce this framework into the time sequence analysis for video event detection. The advantages of the MAR framework are as follows.

- The employment of information at coarse resolutions. The coarse resolution data are usually meaningless in content-based video analysis because most content-based approach are of single temporal resolution. For example, few syntax can

be defined for a 3-minute long strong background noise. However, sports events can be observed from multiple resolutions because of the relatively long temporal duration, special attention patterns and semantic importance. Coarse resolution information can facilitate the discrimination of content-based events. The MAR framework can extract and combine coarse resolution information efficiently.

- The multiresolution framework. As such, different modalities on multiple resolutions are automatically matched and media asynchronism are alleviated. Note that combination weights are adjusted by joint modality distribution in the coarse-to-fine step;
- The extensibility for a large feature space. The MAR framework is better than the linear combination at all resolutions and is robust in handling noisy modality data.

Another original contribution is the introduction of self information entropy to *attention* computation. Avoiding the uncertainty in the mapping between feature-based stimuli and *attention* intensity, e.g. linear and log-like stimulus-reflection function, the self-information entropy is a measurement of information pan-out speed in a neural situation and, therefore, is propositional to *attention* intensity. This measurement is mostly a statistics factor instead of a heuristic content descriptor. Hence, such a measurement can weaken algorithm dependence on specific video collections. Moreover, although video production patterns vary with directors as a visual art with personal preferences, an overall information gain is similar no matter where a sports video comes from. Additionally, the performance of the self-information measurement is interesting. This measurement is plausibly against the general assumption of affection analysis that a strong stimulus incurs a strong reaction [Ma et al., 2002][Wang and Cheong, 2006], which offers a high credit to a rare rather than a strong stimulus. Actually, attention is caused by stimulus contrast rather than stimulus strength [Osgood et al., 1957]. For example, in the final game of World Cup 2002, Germany vs Brazil, video directors occasionally switched off the audio stream in replays. This action brought a very low audio energy but credited a high self-information entropy. Thinking the deviation hypothesis in perception 6-0The deviation hypothesis argues that only the change in stimulus will attract *attention* or invoke consciousness. For example, people will be unmindful if keeping in a noisy situation but suddenly alarmed when it turns quiet., self-information is better and more robust than the direct measurement of signal strength.

The regression of attention over content-based events is a statistics pathway for event-based video analysis. This approach discovers interesting clips and subsequently help

the annotation of video contents. Such a procedure is similar to text retrieval, where document information is discovered firstly from the statistics of terms and then from the content and semantics.

Both of the attention models are far from complete. The analysis on emotional aspects leads to an efficient presentation of video contents, because normal video segments can be discarded in the context reasoning. A further step is to introduce video semantics into attention model, just like we have done in the attack structure segmentation (Chapter 5). However, more labelled data are required and a systematic video annotation approach is necessary.

7

Conclusion and Future Works

In this final chapter, conclusions are drawn together on various aspects of sports event detection and highlight identification. Three individual but closely associated techniques have been presented in this thesis, namely, *replay* detection, *attack* segmentation and *attention* weighting. These techniques try to find a fundamental temporal structure and compare the content-based importance of these structures in order to iterate game stories efficiently and effectively. Multiple perspectives are discussed in the identification and decomposition of content-based events, especially game highlights in a sports video.

We have addressed the main problems in content-based sports video analysis, i.e. the uncertainty of event pattern, multimodality fusion, and semantic annotation. Video events which are defined by sports semantics, such as a goal, can hardly be described by a group of clear and unified low level features or syntax. Both the duration of an event pattern and the content components of events change with instances and video context. Hence, motif-based mining approaches are ineffective, because a simple sequential template cannot meet the numerous variations in the model length and template compositions.

Sports videos are characterised by multiple modalities. Besides audio and visual streams,

there exist many external information resources, such as game records and web casting. These various information modalities of sports video documents differ in the format and medium, and require different methods of feature identification, extraction, representation and syntax specification. Moreover, although the syntax set of sports videos is finite, the characterisation of sports video semantics poses an element of difficulty. This is because the identification of syntax, which is meaningful for the representation of not-text media, remains a research problem. Multimedia feature set, which a fusion algorithm is processing, is not only polymorphous, but also usually inefficient for content presentation. Therefore, multimodality fusion is not a simple matter of credit voting on a concept, but a selection on possible syntax sets. The combination of modality features has to be conducted on a meta-data stream, which not only consists of text and non-text data, but also records asynchronous messages from different time resolutions.

Replay detection is based on the observation that sports highlights are usually replayed to please viewers. It is an efficient approach to collect game highlights by detecting replay segments. Another advantage of replay detection is that a replay segment is a video structure with clear boundaries, i.e. logo transitions. This indicates that a replay segment is a complete content-based video unit and that context reasoning is unnecessary when replay segments are employed to present video highlights.

Attack segmentation tries to find a conceptual structure of sports videos, which conveys a semantically complete video story as a replay segment does. Since a sports game is made up by a group of repetitive team actions, such as the effort to make a goal in football games, these structures are named as *attack*. This segmentation system utilised the local statistics of shot transitions and thereby extracts recurrent temporal sequences. Four video production techniques and video structures, i.e. *play*, *break*, *focus* and *replay*, are identified to describe content changes. Additionally, *break* and *replay* are regarded as stop marks according to the contents these structure denote. Subsequently, a sports video decomposition system is developed to divide a sports video into semantically independent and complete *attack* segments, each of which is equivalent to a **scene** video structure in sports videos. Moreover, structure boundaries of *attack* segments outline the temporal scope of game events. An *attack* structure can be employed for syntax statistics, an efficient description approach for video semantics.

Attention estimation is the main contribution of this thesis, which provides an efficient pathway to identify and allocate interesting events in a game video. Rather than an-

notating semantic objects, this technique computes the stimulus strength directly from modality streams and treats stimulus peaks as possible game highlights, according to a well established psychological hypothesis. Two original attention models are proposed for the estimation and combination of modality attention signals.

Semantic annotation is the final step of content-based video analysis, which labels video segments with proper text tags to define video contents. However, this task is not so difficult in sports videos as it is in general video data. This is because the semantics of sports video is usually limited and self-evident. Hence, this thesis does not concern this research issue.

The main contributions of this thesis are concluded: (1) a set of feature extraction algorithms, including syntax and salient features, are developed and evaluated by a large football video collection in Chapter 3; (2) an *attack* segmentation system employs syntax features to classify video shots and thereby identify video contents; (3) *attention* models estimate and combine stimuli from salient features in order to identify game highlights; (4) Several applications, such as syntax frequency for content-based video indexing, browser index for video skimming, attention graph for video content presentation, are proposed and demonstrated in Chapter 5 and Chapter 6, respectively. Moreover, a general framework of sports event detection and highlight identification is developed by combining *attack* segmentation and *attention* computation. The *attack* segmentation divides a game video into a series of video clips, each of which contains only one semantically important and complete event. The module of *attention* computation estimates the perceptual importance of these video clips as well as plots an attention curve of the game. Game highlights are local maximums of the attention curve. As such, this system not only picks out game events, but also provides a credit on event importance. Nevertheless, an attention curve can be used to speculate the integrity of content presentation. This leads to a multiresolution presentation of event-based video contents. This character is of great value in many applications, e.g. video summarisation, and has not been reported before.

The rest of this chapter is organised as follows. A brief *attack* structure decomposition and *attention* estimation and fusion, are presented in Section 7.1 and Section 7.2, respectively. The discussion and future work are found in Section 7.3.

7.1 Attack Segmentation

Attack is a semantic video structure for sports videos, especially football videos. This structure is equivalent to a **scene** in general video data, and is useful in video summarisation and indexing. The segmentation algorithm utilises video production conventions and a four-state hidden Markov model is trained to simulate a transition sequence of video production styles. The advantages of this algorithm are listed as follows.

1. The introduction of structure kernel discrimination, which locates each structure roughly. This discrimination transforms the problem of content-based video segmentation into the classification of a sequential label collection or a string set;

2. The Markov-based structure kernel extension. After the structure kernel discrimination, the four-state Markov model is employed to extend these kernels on both directions, forward and backward, to make complete video structures. When two Markov chains meet, a boundary is identified and thereby a sports video is divided. This process eases the temporal structure segmentation as the fitness or likelihood comparison between two Markov processes. Note that there are no predefined thresholds;

3. The employment of suffix tree. Most video segmentation algorithms need intensive training to afford media noise and estimate a reasonable threshold. However, video production is an art, which is full of individual creations. This means that the prior knowledge introduced by a large training set is a double-edged sword, which may decrease the performance of structure detection by incurring observation noise. As a symbolised statistical method, the suffix tree works on local data and needs few training.

However, much space is left for the improvement of *attack* structure segmentation. Although small video segments are usually meaningless and can be ignored in content analysis, the over-cut is one of major problems in video structure segmentation, which unnecessarily finds too many small segments. This is because of unpredictable variations in the video production. Small structure kernels are allocated but can hardly to be extended. An efficient approach to merge these small segments is to introduce video syntax, such as the goal and audio whistle, if these syntax can identify content similarity among these small segments. Furthermore, *attack* is a content-based video structure. There may be other embedded video structures, which may facilitate content analysis, such as the sub-shot structure. This algorithm of general suffix tree requires improvement to count the recurrence of video patterns in order to provide an efficient approach to identify these possible structures.

7.2 Attention Computation

Attention estimation is an application of computing psychology to content-based video analysis. Till now, most approaches in content-based video analysis rely on the identification of video syntax. These methods divide this analysis process into two steps, syntax extraction from low level audio-visual features, and semantic reasoning based on video syntax. There are two fundamental problems in the methodology, the ambiguity of syntax and the complexity of reasoning. The selection of syntax set and the creation of reasoning network demand an entire systematic knowledge base. However, such a support is unavailable in most interest domains, though some efforts have been down, such as LSCOM for news videos [Kennedy and Hauptmann, 2006]. The technique of attention estimation brings psychology conclusions and facilitates the understanding of multimedia streams, in particular, those relevant with emotion and feeling. For example, sports videos are characterised with strong emotional aspects. Game events, especially highlights, always attract attention and incur emotion variation. This means that the approach of attention estimation provides an efficient filter to discriminate general contents and game events. Moreover, it is a direct pathway to identify game highlights by computing video stimuli, which requires little supports from video syntax.

Two psychological emotion spaces are compared, attention and arousal-variance space; then the *attention* space is selected for the psychological analysis of sports videos. Salient features in video attention computation are surveyed in Table 6.1. Three feature-attention computation models are addressed together, namely normalisation, self-entropy and adaptive normalisation. However, the approach of attention estimation has the problem of multimodality stimulus fusion. Although there are many different stimuli from different modality features, such as light, colour and motion, people can only hold one emotion state. In other words, all modality stimuli should be combined into a unified signal. But the incomplete psychological image of multiple perception fusion has difficulty in guiding this combination process. Generally, modality stimuli are asynchronous signals from multiple resolutions. There are several issues in the fusion: (1) prior emotion state; (2) video syntax; (3) reflection role. Two fusion algorithms are proposed, the role-based model and the multiresolution autoregressive framework. The role-based model is based on the observation of reflection roles during the video production and enjoyment. This model eliminates reaction bias by grouping attention signals from one reflector and takes emotional states into consideration. The multiresolution autoregressive (MAR) framework simulates the fusion process with a series of autoregressive models on different time resolutions. This simulation includes two passes,

bottom-up and top-down, which not only smoothes signals at all time resolutions, but also distributes knowledge gained at coarse resolutions to high resolution. This means that a proper time resolution is able to be found for a given application and noise caused by resolution difference and signal asynchronism is removed. These results from the MAR framework are useful in content-based video analysis. For example, we can find an appropriate time resolution to discriminate game events rather than that at visual frames or shots, which leads to an effective abstract string of event symbols and helps content-based video modelling.

The advantage of attention computation is as follows.

1. Identify sports highlights without complete syntax understanding;
2. Find proper time resolutions to discriminate game contents;
3. Assign a psychological weight on video contents

However, disadvantages of this psychological approach are clear. Most psychological conclusions are qualitative. For example, Walters et al. [1982] conducted an psychological experiment, which tried to find the relationship between the colour and incurred arousal. These authors argued that "different colors are arousing or relaxing, and color choice indicates arousal preference". This claim shows the relationship between modality features and psychological measurements cannot be defined quantitatively. Such a psychological ambiguity makes the computational projection from modality features to psychological measurements questionable, although many computation models have been developed and evaluated. Moreover, the event found by psychological approaches is not equivalent to a content-based video event, although certain links exist between psychological measurements and video contents. Video syntax is necessary to semantically clarify psychological events.

7.3 Discussion and Future Work

This thesis partially solves the research question of what constitutes semantic video structure with a complete content story in sports videos. This question leads to a different methodology of video retrieval, which focuses on the temporal accumulation or repetitiveness in the presentation of video semantics. Compared with image retrieval, the introduction of a time dimension provides extra information and alleviates the problem of information scarcity. However, this information resource can hardly be employed directly and seems to be redundant because of the similarity between visual

frames. It therefore may be meaningless to search image features frame by frame, such as image syntax, although a double-check may increase the precision of syntax detection. Therefore, a approach in video retrieval is widely accepted, which employes various image retrieval methods to complete a video retrieval task. This approach leads to a temporal decomposition and a complex image syntax reasoning: a long video is cut into shots; key frames are selected from shots; syntax or harmony regions are detected from key frames and labelled by some classifiers; these syntax are organised to guess key frame semantics, shot semantics and thus the semantics of a video clip. However, a fundamental problem cannot be answered is whether a video is a simple accumulation of visual frames or not. From many aspects, the answer is definitely no. Hence, this approach which employes image retrieval methods in video retrieval is unlikely to be successful.

Note that replication is not only a redundancy but also an approach to present ideas. The repeatable syntax sequence is a direct description of video semantics. For example, the syntax of goal post appears frequently in a video story of a shot. This syntax frequency can discriminate a football shot event from other events, such as a free-kick and a corner kick. Therefore, the vector of syntax frequency is able to present video contents and is used for retrieval game videos by a video sample (Section 5.6). In this case, the problem is how to decide the duration of a video story or the temporal interval of a complete semantic unit, which plays the role of retrieval targets, such as a document in text retrieval. A new video segmentation system is required, which is different from traditional shot-scene hierarchy. A shot can not convey complete semantics, while a scene involves so many shots that a scene can hardly be segmented and regarded as a retrieval unit. This is why we propose an *attack* structure.

From replay detection, attack segmentation, and attention estimation, we look for a proper temporal unit for video content analysis. Two units are tried, temporal repetitive structure *attack* and psychological *attention* peaks. It seems effective to decompose a video according to the attention intensity. The time interval with a relative high attention refers to an attention-based video event, because these events are when people pay attention. The frequency of multiple syntax are computed in attention events and thus create a retrieval subspace, which is similar to the computation of term frequency in text retrieval. Section 5.6 presents a prototype system for goal event search, which takes a goal video clip and returns a list of goal video clips, although these videos are from different view points, dominant colour, and audio loudness. It indicates our approach based on syntax frequency can work in a small test bed.

7.3.1 Attention Graph and Multiresolution Semantics Presentation

An attention graph is to a topology of attention segments in a sports video. Each attention segment is described by a vector of syntax frequency, which presents the semantics of a video segment. Therefore, an attention graph is a semantic description of an entire sports video. Compared with the traditional hierarchy of shot and scene, this graph is efficient for video retrieval and annotation. I will evaluate this idea and develop an algorithm for attention graph matching and participation.

7.3.2 Video Annotation

Video annotation is the final step of content-based video analysis. Different from syntax-based reasoning, we try to employ syntax frequency to annotate sports videos. This approach introduces extra parameters such as syntax appearance probability and appends current ontology systems with a fitness measurement to given video data. This indicates a possible retrieval model and a video-based retrieval system, which accepts video clips as input and outputs a ranked list of video clips.

The future work is to extend these temporal structure based approaches and make attention-based video content understanding framework available for general video content analysis. Some research targets are listed as follows.

- Improve the estimation algorithm of attention. The attention is a complex psychological phenomena. Many approaches have been developed to estimate an attention intensity or an attention area in a video. How to combine these estimations and develop a robust fusion scheme remain a research question for general video genres. This is because of the complexity in the combination of qualitative and quantitative feature vectors.

- A retrieval unit is defined by user rather than a retrieval system, although it is possible to guess what users look for. Therefore, it is necessary for a video retrieval system to take user context into consideration.

- Syntax is decided by video domain. However, the current syntax set we employ is small for an efficient retrieval, because of the scarcity of video collection. It seems difficult to invent all syntax manually. This results in a question how to handle missing syntax.

- Syntax reasoning is a useful approach for query extension in current image-based video retrieval. Such reasoning is a domain knowledge based logic system rather

than an accumulation of possible key words. How to combine these knowledge into a video retrieval model is call for further consideration.

- Real time. Multimedia data attracts great interest from industry. Real time algorithms are eagerly required in consuming electrics and media distribution industry. Many new hardware platforms have been developed, such as media centres and personalised video services. These devices provide strong computational abilities. It is possible to make current video processing algorithms run in real time by rescaling these algorithms onto these hardware frameworks. This leads to a new research topic, how to decrease the computational complexity of video processing algorithms.

Bibliography

3G-News [2005], '3g football best mobile service'.
URL: *http://www.3g.co.uk/PR/Jan2005/9006.htm*

Adams, B., Dorai, C. and Venkatesh, S. [2000], Novel approach to determining tempo and dramatic story sections in motion pictures, *in* 'ICIP 2000', Vol. II, pp. 283–286.

Agrawal, R., Imielinski, T. and Swam, A. N. [1993], Mining association rules between sets of items in large databases, *in* P. Buneman and S. Jajodia, eds, 'ACM SIGMOD International Conference on Management of Data', ACM, ACM Press, Washington, D.C., pp. 207–216.

Ahanger, G. and Little, T. [1996], 'A survey of technologies for parsing and indexing digital video', *Journal of Visual Communication and Image Representation* **7**(1), 28–43.

Al-Hames, M. and Rigoll, G. [2005*a*], A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos, *in* 'ICIP', pp. III:421–424.

Al-Hames, M. and Rigoll, G. [2005*b*], A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data, *in* 'ICME', pp. 45–48.

Andrieu, C., de Freitas, N. and Doucet, A. [2001], 'Robust full bayesian learning for radial basis networks', *Neural Computation* **13**, 2359–2407.

Ashley, J., Barber, R., Flickner, M., Hafner, J. L., Lee, D., Niblack, W. and Petkovic, D. [1995], Automatic and semiautomatic methods for image annotation and retrieval in query by image content (qbic)., *in* 'Storage and Retrieval for Image and Video Databases (SPIE)', pp. 24–35.

Assfalg, J., Bertini, M., Colombo, C. and Bimbo, A. D. [2002], 'Semantic annotation of sports videos', *IEEE MultiMedia* **9**(2), 52–60.

Babaguchi, N., Kawai, Y. and Kitashi, T. [2002], 'Event based indexing of broadcasted sports video by intermodal collaoration', *IEEE Trans. Multimedia* **4**, 68–75.

Baddeley, A. and Wilson, B. [2002], 'Prose recall and amnesia: implications for the structure of working memory', *Neuropsychologia* **40**, 1737–1743.

Baeza-Yates, R. and Ribeiro-Neto, B. [1999], *Modern Information Retrieval*, Addison Wesley ACM Press.

Bailey, T. L. and Elkan, C. [1995], 'Unsupervised learning of multiple motifs in biopolymers using expectation maximization', *Machine Learning* **21**(1-2), 51–80.

Baillie, M. and Jose, J. M. [2003], Audio-based event detection for sports video., *in* 'CIVR', pp. 300–309.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. [2004], *Hierarchical Modeling and Analysis for Spatial Data*, Vol. 101 of *Monographs on Statistics and applied probability*, Chapman and Hall.

Bollmann, M., Hoischen, R. and Mertsching, B. [1997], *Integration of Static and Dynamic Scene Features Guiding Visual Attention*, Springer.

Bordwell, D. and Thompson, K. [2004], *Film Art: An Introduction*, 7th edn, New York: McGraw-Hill.

Boreczky, J. and Wilcox, L. [1998], A hidden markov model framework for video segmentation using audio an image features, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Brin, S., Motwani, R., and Silverstein, C. [1997], Beyond market baskets: Generalizing association rules to correlations, *in* J. Peckham and E. Tucson, eds, 'ACM SIGMOD International Conference on Management of Data', ACM, ACM Press, Arizona, pp. 265–276.

Brin, S., Motwani, R., Ullman, J. and Tsur, S. [1997], Dynamic itemset counting and implication rules for market basket data, *in* J. Peckham and E. Tucson, eds, 'ACM SIGMOD International Conference on Management of Data', ACM, ACM Press, Arizona, pp. 255–264.

Buitelaar, P. and Ramaka, S. [2005], Unsupervised ontology-based semantic tagging for knowledge markup, *in* 'Workshop on Learning in Web Search at the International Conference on Machine Learning', Bonn,Germany.

Burke, B. and Shook, F. [1996], *Television production and reporting*, Longman Publisher.

Calvert, G. A. and Thesen, T. [2004], 'Multisensory integration: methodological approaches and emerging principles in the human brain', *Journal of Physiology* **98**(1-3), 191–205.

Calvert, G., Bullmore, E. T., Brammer, M. J., Campbell, R., C.Williams, S., McGuire, P. K., Woodruff, P. W., Iversen, S. D. and David, A. S. [1997], 'Activation of auditory cortex during silent lipreading', *Science* **276**(5312), 593–596.

Carbonaro, A. and Ferrini, R. [2007], Ontology-based video annotation in multimedia entertainment, *in* 'Consumer Communications and Networking Conference, CCNC 2007', IEEE Publication, pp. 1087–1091.

Cernekova, Z., Nikou, C. and Pitas, I. [2002], Shot detection in video sequences using entropy-based metrics, *in* 'IEEE Int. Conf. Image Processing'.

Chang, Y., W.Zeng, Kamel, I. and Alonso, R. [1996], Integrated image and speech analysis for content-based video indexing, *in* 'IEEE International Conference Multimedia Computing and Systems'.

Cheung, C. and Po, L. [2003], 'A novel cross-diamond search algorithm for fast block motion estimation', *IEEE Trans. Circuits System on Video Technology* **4**(12), 1168–1177.

Chien, J.-T. [1999], 'Online hierarchical transformation of hidden markov models for speech recognition', *IEEE Transactions on Speech and Audio Processing* **7**(6), 656–667.

Chou, K. C., Willsky, A. S. and Benveniste, A. [1994], 'Multiscale recursive estimation, data fusion and regularization', *IEEE Trans. on automatic control* **39**(3), 464–478.

Churchland, P., Ramachandran, V. and Sejnowski, T. J. [1994], A critique of pure vision, *in* C. Koch and J. L. Davis, eds, 'Large-scale neuronal theories of the brain', Bradford Books MIT Press.

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. [1995], *Survey of the State of the Art in Human Language Technology*, Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA. USA.

Courtney, J. D. [1997], 'Automatic video indexing via object motion analysis', *Pattern Recognition* **30**(4), 607–625.

Crary, J. [1999], *Suspensions of Perception: Attention, Spectacle and Modern Culture*, Cambridge, MA: MIT Press.

Creenwald, M., Cook, E. and Lang, P. [1989], 'Affective judgement and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli', *J.Psychophysiology* **3**, 51–64.

Crestani, F., Ruthven, I., Sanderson, M. and van Rijsbergen, C. J. [1995], The troubles with using a logical model of ir on a large collection of documents, *in* 'Proceedings of TREC-4, Fourth Text Retrieval Conference', Washington, US.
URL: *citeseer.ist.psu.edu/article/crestani95troubles.html*

Detenber, B., Simons, R. and Bennett, G. [1997], 'Roll'em! the effects of picture motion on emotion responses', *J.Broadcasting and Electron, Media* **21**, 112–126.

Dietz, R. and Lang, A. [1999], Affective agenents: Effects of agent affect on arousal, attention, liking and learning, *in* 'Proc. Cognitive Technology'.

Duan, L., Xu, M., Chua, T. and Xu, C. [2003], A mid-level representation framework for semantic sports video analysis, *in* 'ACM Conference on Multimedia 2003'.

Ekin, A., Tekalp, A. and Mehrotra, R. [2003], 'Automatic soccer video analysis and summarization', *IEEE Trans. on Image Processing* **12**(8), 796–807.

Ekman, P. [1987], 'Universals and cultural differences in the judgements of facial expressions of emotions', *J.Personality Social Psych.* **54**(4), 712–717.

Engel, S., Zhang, X. and Wandell, B. [1997], 'Colour tuning in human visual cortex measured with functional magnetic resonance imaging', *Nature* **388**, 68–71.

Fine, S., Singer, Y. and Tishby, N. [1998], 'The hierarchical hidden markov model analysis and applications', *Machine Learning* **32**(1), 41–62.

Fleischman, M. and Roy, D. [2007], Situated models of meaning for sports video retrieval, *in* 'Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers', Association for Computational Linguistics, Rochester, New York, pp. 37–40.

Gasiba, T., Stockhammer, T., Samad, W. A., Xu, W., Jenkac, H. and Schierl, T. [2006], A weighted layered broadcasting scheme for scalable video transmission with multiple site reception, *in* 'MobiMedia'.

Gong, Y., Sin, L., Chuan, C., Zhang, H. and Sakauchi, M. [1995], Automatic parsing of tv soccer programs, *in* 'Proc. Internat. Conf. on Multimedia Computing and Systems(ICMCS 95)', Washington, DC.

Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S. and Anderson, C. [1994], Overcomplete steerable pyramid filters and rotation invariance, *in* 'Proc. IEEE Computer Vision and Pattern Recognition'.

Gu, L., Bone, D. and Reynolds, G. [1998], 'Replay detection in sports video sequences'. **URL:** *citeseer.ist.psu.edu/618849.html*

Gunes, H. and Piccardi, M. [2005], Affect recognition from face and body: early fusion vs. late fusion, *in* 'IEEE International Conference on Systems, Man and Cybernetics', pp. IV:3437– 3443.

Guo, Y.-F., Li, S.-J., Yang, J.-Y., Shu, T.-T. and Wu, L.-D. [2003], 'A generalized foley-sammon transform based on generalized fisher discriminant criterion and its application to face recognition.', *Pattern Recognition Letters* **24**(1-3), 147–158.

Han, M., Hua, W., Xu, W. and Gong, Y. [2002], An integrated baseball digest system using maximum entropy method, *in* 'ACM MULTIMEDIA 2002', ACM Press, New York, NY, USA, pp. 347–350.

Hanjalic, A. [2005], 'Adaptive extraction of highlights from a sport video based on excitement modeling', *IEEE Trans. on Multimedia* **7**(6), 1114–1122.

Hanjalic, A. and Xu, L. [2005], 'Affective video content repression and model', *IEEE Trans on Multimedia* **7**(1), 143–155.

Hua, X.-S., Lu, L. and Zhang, H.-J. [2004], 'Optimization-based automated home video editing system', *IEEE Transactions on Circuits and Systems for Video Technology* **14**(5), 572–583.

Huang, J., Liu, Z. and Wang, Y. [1998], Integration of audio and visual information for content-based video segmentation, *in* 'Proceedings of IEEE Conference on Image Processing'.

Hui, L. [1992], Color set size problem with applications to string matching, *in* 'Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching', pp. 230–243.

Intille, S. S. and Bobick, A. F. [1999], A framework for recognizing multi-agent action from visual evidence., *in* 'AAAI/IAAI', pp. 518–525.

Itti, L. and Koch, C. [1998], 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20**, 1254–1259.

Itti, L. and Koch, C. [1999], A comparison of feature combination strategies for saliency-based visual attention systems, *in* 'SPIE Human Vision and Electronic Imaging IV'.

Itti, L. and Koch, C. [2001], 'Computational modeling of visual attention', *Nature Reviews Neuroscience* **2**, 194–203.

Iyengar, A., Squillante, M. S. and Zhang, L. [1999], Analysis and characterization of large-scale web server access patterns and performance, *in* 'World Wide Web', pp. 85–100.

Julesz, B. [1991], 'Early vision and focal attention', *Rev. Mod. Phys.* **63**(3), 735–772. American Physical Society.

Kang, Y., Lim, J., Kankanhalli, M., Xu, C.-S. and Tian, Q. [2004], 'Goal detection in soccer video using audio/visual keywords', *ICIP2004* **3**, 1629 – 1632.

Kennedy, L. and Hauptmann, A. [2006], Lscom lexicon definitions and annotations version 1.0, dto challenge workshop on large scale concept ontology for multimedia, ADVENT Technical Report 217-2006-3, Columbia University.

Kijak, E., Gravier, G., Gros, P., Oisel, L. and Bimbot, F. [2003], Hmm based structuring of tennis videos using visual and audio cues, *in* 'IEEE Int. Conf. Multimedia Expo', pp. 309–312.

Kobla, V., DeMenthon, D. and Doermann, D. [2000], Identifying sports videos using replay, text and camera motion features, *in* 'SPIE'.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. [1993], 'Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment', *Science* **8**(262), 208–214.

Lenardi, R., Migliorati, P. and Prandini, M. [2004], 'Semantic indexing of soccer audio-visual sequence: A multimodal approach based on controlled markov chains', *IEEE Trans. on Circuits and System for Video Technology* **14**, 634–643.

Leonardi, R., Migliorati, P. and Prandini, M. [2002], Modeling of visual features by markov chains for sport content characterisation, *in* 'EUSIPCO', Toulouse, France, pp. 349–352.

Lesser, M. and Murray, D. [1998], Mind as a dynamical system: Implications for autism, *in* 'Durham conference Psychobiology of autism: current research and practice'.

Leventhal, A. [1991], *The Neural Basis of Visual Function: Vision and Visual Dysfunction*, Boca Raton, Fla.:CRC Press.

Lew, M. S. [1996], *Principles of Visual Information Retrieval*, Springer.

Liang, C.-H., Chu, W.-T., Kuo, J.-H., Wu, J.-L. and Cheng, W.-H. [2005], 'Baseball event detection using game-specific feature sets and rules', *IEEE International Symposium on Circuits and Systems* **4**(2), 3829–383.

Lienhart, R. [1999], Comparisons of automatic shot boundary detection algorithms, *in* 'Proc of SPIE Storage and Retrieval for Image and Video Database', Vol. 3656, pp. 290–301.

Lienhart, R. [2001], Reliable dissolve detection, *in* 'Proc of SPIE Storage and Retrieval for Image and Video Database', Vol. 4315, pp. 219–230.

Liu, J. S., Neuwald, A. F. and Lawrence, C. E. [1995], 'Bayesian models for multiple local sequence alignment and gibbs sampling strategies', *Journal of the American Statistical Association* **90**(432), 1156–1170.

Lu, L., Zhang, H. and Jiang, H. [2002], 'Content analysis for audio classification and segmentation', *IEEE Trans. on Speech and Audio Processing* **10**(7), 504–516.

Ma, Y., Lu, L., Zhang, H. and Li, M. [2002], A user attention model for video summarisation, *in* 'ACM Multimedia 02'.

Marr, D. [1982], *Vision. A Computational Investigation into the Human Representation of Visual Information*, Freeman, New York.

McGurk, H. and MacDonald, J. [1976], 'Hearing lips and seeing voices', *Nature* **264**(5588), 23–30.

Milanese, R., Gil, S. and Pun, T. [1995], 'Attentive mechanisms for dynamic and static scene analysis', *Optical Eng.* **34**, 2428–2434.

Mittal, A. and Cheong, L.-F. [2003], 'Framework for synthesising semantic level indexes', *Multimedia Tools Application* **20**(2), 135–158.

Murray, I. and Arnott, J. [1993], 'Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion', *J. Acoust. Soc. Amer.* **93**, 1097–1108.

Naphade, M. [1998], Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems, *in* 'ICIP 1998'.

Naphade, M. and Huang, T. [2001], 'A probabilistic framework for semantic video indexing, filtering, and retrieval', *IEEE Transactions on Multimedia* **3**(1), 141–151.

Osgood, C., G.J.Suci and P.H.Tannenbaum [1957], *The measurement of meaning*, University of Illinois Press.

Pan, H., Beek, P. V. and Sezan, M. [2001], Detection of slow-motion replay segments in sports video for highlights generation, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Pan, H., Li, B. and Sezan, M. [2002], Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Pfeiffer, S., Fischer, S. and Wheaton, J. [1996], Automatic audio content analysis, *in* 'ACM Multimedia', ACM, pp. 21–30.

Picard, R. [1997], *Affective Computing*, Cambridge, MA: MIT Press.

Polikar, R. [2006*a*], 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine* **6**(3), 21–45.

Polikar, R. [2006*b*], 'A tutorial article on ensemble systems including pseudocode, block diagrams and implementation issues for adaboost and other ensemble learning algorithms', *IEEE Trans. Circuits and Systems Magazine* **6**(3), 21–45.

Rasheed, Z. and Shah, M. [2003], A graph theoretic approach for scene detection in produced videos, *in* 'Multimedia Information Retrieval Workshop, ACM SIGIR 2003', Toronto, Canada.

Ren, R. and Jose, J. [2005], Football video segmentation based on video production strategy, *in* 'ECIR 2005', pp. 433–446.

Ren, R. and Jose, J. [2006], Attention guided football video recommendation system on mobile device, *in* 'MobiMedia 2006', Alghero, Sardinia, Italy, pp. 202–212.

Ren, R., Jose, J. and He, Y. [2007], Affective sports highlight detection, *in* 'the 15th European Signal Processing Conference', Poznan, Poland, pp. 728–732.

Ren, R., P.Punitha, Urban, J. and Jose, J. [2007], Attention-based video summarisation in rushes collection, *in* 'ACM Multimedia', ACM, Ausburg, Germany, pp. 89–95.

Rui, Y., Gupta, A. and Acero, A. [2000], Automatically extracting highlights for TV baseball programs, *in* 'ACM Multimedia', pp. 105–115.

Russell, S. J. and Norvig, P. [2002], *Artificial Intelligence: A Modern Approach(2nd)*, Prentice Hall.

Sadlier, D. and O'Connor, N. [2005], 'Event detection in field sports video using audio-visual features and a support vector machine', *IEEE Trans on Circuits and System for Video Technology* **15**, 1225–1233.

Satoh, S., Nakamura, Y. and Kanade, T. [1999], 'Name-it: Naming and detecting faces in news videos.', *IEEE MultiMedia* **6**(1), 22–35.

Seo, Y., Choi, S., Kim, H. and Hong, K.-S. [1997], Where are the ball and players? soccer game analysis with color based tracking and image mosaick, *in* 'ICIAP', pp. 196–203.

Simons, R., Detenber, B. H., Roedema, T. and Reiss, J. E. [2003], 'Attention to television: Alpha power and its relationship to image motion and emotional content', *Media Psychol.* **5**, 283301.

Simons, R., Detenber, B., Roedema, T. and Reiss, J. [1999], 'Emotion-processing in three systems: The medium and the message', *Psychophysiology* **36**, 619–627.

Snoek, C. G., Worring, M. and Smeulders, A. W. [2005], Early versus late fusion in semantic video analysis, *in* 'ACM Multimedia'.

Snoek, C. G., Worring, M., van Gemert, J., Geusebroek, J.-M., Koelma, D., Nguyen, G. P., de Rooij, O. and Seinstra, F. [2005], Mediamill: Exploring news video archives based on learned semantics, *in* 'Proceedings of ACM Multimedia'.

Suresh, V., Mohan, C. K., Swamy, R. K. and Yegnanarayana, B. [2004], Content-based video classification using support vector machines, *in* N. Pal, ed., 'ICONIP', LNCS, pp. 726–731.

Swain, M. J., Frankel, C. and Athitsos, V. W. [1997], An image search engine for the world wide web, *in* 'IEEE Computer Vision and Pattern Recognition Conference'.

Tjondronegoro, D., Chen, Y.-P. P. and Pham, B. [2004*a*], The power of play-break for automatic detection and browsing of self-consumable sport video highlights, *in* 'MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval', ACM Press, New York, NY, USA, pp. 267–274.

Tjondronegoro, D., Chen, Y. P. and Pham, B. [2004*b*], Classification of self-consumable highlights for soccer video summaries, *in* 'ICME 2004'.

Toole, C., Smeaton, A., Murphy, N. and Marlow, S. [1999], Evaluation of automatic shot boundary detection on a large video test set, *in* 'Challenge of Image Retrieval'.

TRECVID [2003], 'Analysis and presentation of soccer highlights from digital video'.

Treisman, A. M. and Kanwisher, N. G. [1988], 'Perceiving visually presented objects: recognition, awareness, and modularity', *Current Opinion in Neurobiology* **8**, 218–226.

Truong, B., Venkatesh, S. and Dorai, C. [2000], Automatic genre identification for content-based video categorization, *in* 'ICPR'.

Tsekeridou, S. and Pitas, I. [2001], 'Content-based video pasing and indexing based on audio-visual interaction', *IEEE Trans. on Circuits and System for Technology* **11**(4), 522–535.

Valdez, P. and Mehrabian, A. [1994], 'Effects of color on emotions', *J. Experimental Psych.: General* **123**(4), 394409.

Vasconcelos, N. and Lippman, A. [2000], Bayesian video shot segmentation, *in* 'NIPS', pp. 1009–1015.

Walters, J., Apter, M. J. and Svebak, S. [1982], 'Color preference, arousal, and the theory of psychological reversals', *Motivation and emotion* **6**(3), 1573–6644.

Wang, H. L. and Cheong, L.-F. [2006], 'Affective understanding in film', *IEEE Trans. Circuits and System for Video Technology* **16**(6), 689–704.

Wang, J., Xu, C., Chng, E., Duan, L., Wan, K. and Tian, Q. [2005], Automatic generation of personalized music sports video, *in* 'ACM MULTIMEDIA 2005', ACM Press, New York, NY, USA, pp. 735–744.

Wang, J., Xu, C., Chng, E., Wah, K. and Tian, Q. [2004], Automatic replay generation for soccer video broadcasting, *in* 'ACM MULTIMEDIA 2004', ACM Press, New York, NY, USA, pp. 32–39.

Willsky, A. [2002], Multiresolution markov models for signal and image processing, *in* 'Proceedings of the IEEE', pp. 1396–1458.

Xie, L., Chang, S., Divakaran, A. and Sun, H. [2002], Structure analysis of soccer video with hidden markov models, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing'.

Xie, L., Xu, P., Chang, S.-F., Divakaran, A. and Sun, H. [2004], 'Structure analysis of soccer video with domain knowledge and hidden markov models.', *Pattern Recognition Letters* **25**(7), 767–775.

Xu, C., Wang, J., Wan, K., Li, Y. and Duan, L. [2006], Live sports event detection based on broadcast video and web-casting text, *in* 'ACM Multimedia 2006'.

Xu, G., Ma, Y., Zhang, H. and Yang, S. [2005], 'An hmm-based framework for video semantic analysis', *IEEE Trans on Circuits and System for Video Technology* **15**, 1422–1433.

Xu, H. and Chua, T. [2004], The fusion of audio-visual features and external knowledge for event detection in team sports video, *in* 'MIR 2004'.

Xu, L. and Li, Y. [2003], Video classification using spatial-temporal features and pca, *in* 'ICME 2003'.

Xu, M., Maddage, N., Xu, C., Kankanhalli, M. and Tian, Q. [2003], Creating audio keywords for event detection in soccer video, *in* 'ICME 2003'.

Xu, P., Xie, L., Chang, S.-F., Divakaran, A., Vetro, A. and Sun, H. [2001], Algorithms and system for segmentation and structure analysis in soccer video., *in* 'ICME'.

Yan, R. [2006], Probabilistic Models For Combining Diverse Knowledge Sources in Multimedia Retrieval, PhD thesis, Carnegie Mellon University.

Yin, H. and Ren, R. [2007], Online sports video summarisation method, Patent Review CN8-2007-0229, IBM Research China.

Yow, D., Yow, B., Yeung, M. and Liu, B. [1995], Analysis and presentation of soccer highlight from digital video, *in* 'Proc. of 2nd Asian Conference on Computer Vision', pp. 499–503.

Zettl, H. [1990], *Sight, Sound, Motion: Applied Media Aesthetics*, Wadsworth, Belmont CA.

A

Generalised Foley-Sammon Transform Classifier

The Foley-Sammon transformation (FST) is widely regarded as one of the best methods in terms of linear feature discrimination. Based on the fisher discriminant criterion, the FST tries to map all data samples onto a point in the high dimension feature space and thereby maximises fisher criterion (Equation A.7). The computation of the Foley-Sammon optimal discriminants is follows.

1. Construct the orthogonal complementary space of the subspace spanned by the discriminant vectors calculated before;

2. Choose the vector that maximises the Fisher criterion function as the present discriminant vector from the orthogonal complementary space. This discriminant vector spans a one dimentional subspace where the sample set has the minimum within-class scatter and the maximum between class scatter.

However, it can hardly conclude that this algorithm can optimise scatter matrices in the subspace spanned by all discriminant vectors after the FST, although each discriminant vector is supposed to find the best span in 1D subspace. This algorithm can not grantee the subspace extracted can minimise with-in scatters whilst maximises between-class scatters.

Therefore, a generalized optimal set of discriminant vectors (GFST) is proposed (Equation A.9). The main improvement is the criterion on discriminant vector selection. Different from seeking for an optimisation in the complementary space, which try to minimise with-in scatters and maximise between-class scatters at the same time, the GFST is a projected set of the training sample set in the subspace spanned by the vector and the other discriminant vectors previously calculated, which lead to the maximum ratio between the between-class distance and the within-class distance. Note that this projected set on a GFST subspace is not proposed for the best separability in the global yet.

The FST classifier employed for the detection of uniform and other video objects (Chapter 3) is based on the algorithm in [Guo et al., 2003], which is a GFST classifier but improved for a better sample separability. This algorithm is formalised as follows. Let $w_1, w_2, ..., w_m$ be m known pattern classes, $X = \{x_i\}, i = 1, 2, ..., N$ be the set of $n-$dimensional samples. Each sample x_i in X belongs to a class w_j. Suppose the mean vector, the covariance matrix and a prior probability of a class w_i are m_i,c_i,$P(w_i)$, respectively. Therefore, a between-class scatter matrix S_b, a within-class scatter matrix S_w, and a population scatter matrix S_t are determined in the following formulae.

$$S_b = \sum_{i=1}^{m} P(w_i)(m_i - m_0)(m_i - m_0)^T \tag{A.1}$$

$$S_w = \sum_{i=1}^{m} P(w_i) E\{(x - m_i)(x - m_i)^T / w_i\} \tag{A.2}$$

$$S_t = S_b + S_w = E\{(x - m_0)(x - m_0)^T\} \tag{A.3}$$

$$\tag{A.4}$$

where $m_0 = E\{x\}$ is the mean vector of the population distribution of samples.

Hence, the Fisher criterion is reiterated by Equation A.5.

$$J_f(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \tag{A.5}$$

where φ is an arbitrary $n-$dimensional vector. Let φ_1 be the unit vector which maximise $J_f(\varphi)$, and thereby become the first vector in the FST optimal discriminant vector set. Moreover, the i^{th} vector of Foley-Sammon optimal discriminant vectors are estimated

by optimizing Equation A.6.

$$\max_{\varphi_j^T \varphi_i=0, \|\varphi_i\|=1} J_f(\varphi_i), j = 1,2,...,i-1 \tag{A.6}$$

Therefore, let $S = \{varphi_i\}$,$\Phi = (\varphi_1, \varphi_2, ..., \varphi_r)$ a FST can be formulated as Equation A.7.

$$y = \Phi^T x \tag{A.7}$$

The ratio between the within-class and between-class distance is,

$$J(\Phi) = \frac{tr(\Phi^T S_b \Phi)}{tr(\Phi^T S_w \Phi)} = \frac{\sum_{i=1}^{r} \varphi_i^T S_b \varphi_i}{\sum_{i=1}^{r} \varphi_i^T S_w \varphi_i} \tag{A.8}$$

It is clear that the transformed set has the best separable ability in global sense when $J(\Phi)$ reach maximum.

$$J(\widetilde{\Phi}) = \max_{\Phi} J(\Phi) \tag{A.9}$$

where $\widetilde{\Phi} = (\widetilde{\varphi_1}, \widetilde{\varphi_2}, ..., \widetilde{\varphi_r})$ and $\widetilde{\varphi_1}, \widetilde{\varphi_2}, ..., \widetilde{\varphi_r}$ are unit orthogonal column vectors in n dimensional space. Therefore, $J(\widetilde{\Phi})$ is called as the generalized Fisher discriminant criterion and the respective transformation is called generalized FST (GFST).

An iterative algorithm is employed to find optimal discriminant vectors which constitutes a GFST. For convenience, we assume the population scatter matrix S_t is non-singular. Obviously, S_t is positive-definite and S_t^{-1} exists. The detail process is stated in Algorithm 4.

When S_t is singular, we employ the support set of S_t. Suppose $S_t^{-1}(0) = span\{\alpha_1, ..., \alpha_k\}$, $\overline{S_t^{-1}(0)} = span\{\beta_1, ..., \beta_{n-k}\}$, where $\alpha_1, ..., \alpha_k, \beta_1, ..., \beta_{n-k}$ are both orthogonal unit vectors. Hence, the between-class distance of the projected set on α is zero. Optimal discriminant vectors should be selected from $\overline{S_t^{-1}(0)}$. $\forall \beta \in \overline{S_t^{-1}(0)}$, $\beta = a_1\beta_1 + a_2\beta_2 + ... + a_{n-k}\beta_{n-k} = P\beta$. In the function of $\widetilde{\Phi}$, let $\varphi_l = P\hat{\varphi}_l, l = 1..r$, we get

$$\widetilde{J(\Phi)} = \frac{\sum_{l=1}^{r} \hat{\varphi}_l{}^T (P^T S_b P) \hat{\varphi}_l}{\sum_{l=1}^{r} \hat{\varphi}_l{}^T (P^T S_w P) \hat{\varphi}_l} \equiv \widetilde{J(\hat{\Phi})} \tag{A.10}$$

where $(\hat{\Phi}) = ((\hat{\varphi}_1), ..., (\hat{\varphi}_1))$. which is non-singular. Therefore, analogous to non-singular population scatter matrix, $(\hat{\Phi})$ can be calculated by Algorithm 4. The optimal

Data: between-class scatter matrix S_b, Non-singular population scatter matrix S_t,
Result: discriminant vector set λ, and optimal discriminant vectors
$\widetilde{\varphi_1}(\lambda), \widetilde{\varphi_2}(\lambda), ..., \widetilde{\varphi}_r(\lambda)$
$S_t^{-1}(0) = \Phi$;
$a = 0, b = 1, \lambda = (a+b)/2$;
$stop = false$;
while !$stop$ **do**
  $S = S_b - \lambda S_t$;
  $\lambda_1, \lambda_2, ..., \lambda_r$, the first r eigenvalues of S;
  $\varepsilon_1 = \sum_{i=1}^{r} \lambda_i$;
  if $\varepsilon_1 = 0$ **then**
    $\widetilde{\varphi_1}(\lambda), \widetilde{\varphi_2}(\lambda), ..., \widetilde{\varphi}_r(\lambda)$ the orthonormal eigen vectors corresponding to $\lambda_1, \lambda_2, ..., \lambda_r$;
    break;
  else
    if $\varepsilon_1 < 0$ **then**
      $b = \lambda$;
    else
      $a = \lambda$;
    end
  end
  $stop = (1 + r\mu/\Delta|\lambda - \lambda_0|) \leq (1 + r\mu/\Delta|a - b|)$;
end

Algorithm 4: Discriminant vector computation for GFST (I)

discriminant vectors are $\widetilde{\varphi}_l = \|\widetilde{P(\hat{\varphi}_l)}\|$

B

Overview of Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a group of sampling algorithms for high dimensional optimisation. A Markov chain is constructed, which has the desired distribution as the equilibrium distribution, to sample a data set from probability distributions. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of this sample improves as a function of the number of steps. The approach of MCMC has been proven successful in the learning of Bayesian statistical models.

A MCMC is a stochastic algorithm and consists of a large number of independent iterations, each of which tests a set of parameters. Generally, such an iteration includes two steps,

1. Proposal step, which gives a new model according to given distributions or just randomly. This step relies on former tested models and the statistics on given data set.

2. Decision step, which computes a fitness or an acceptance probability of the proposed model by maximising posterior or other criterions. As such, this proposed model are decided to be accepted or rejected with a credit probability.

If the benign constraints meet proposal distributions, a MCMC is supposed to converge to a global optimum. However, the speed of convergence depends on the *goodness* or *fitness* of the proposals.

Another application of MCMC is the model selection with the technique of reverse-jump. A reverse-jump MCMC constructs reversible moves between parameter spaces of different dimensions. Andrieu et al. [2001] applied this reverse-jump MCMC to learn a radial basis function (RBF) neural network. The authors proposed birth-death and split-merge steps to reconstruct RBF kernels and experienced all possible solutions.

The main limitations of MCMC are: (1) the number of steps, which leads to a converge, is unknown. Although the residue after modelling fitting can be used to stop this repetitive test-evaluation procedure, it is hard to avoid this MCMC to stop at a local maximum rather than a global maximum; and (2) the assumption that all model propositions and data samples are independent, which allows an individual observation and evaluations. A MCMC is inefficient when there is a strong correlation between data distributions.

C

Attention Curves in Games

This appendix chapter displays a part of estimated attention curves in game collections of World Cup 2002, World Cup 2006 and UEFA 2006. Each figure group includes a self-entropy curve of audio attention (top) and the unified attention curve estimated (bottom).

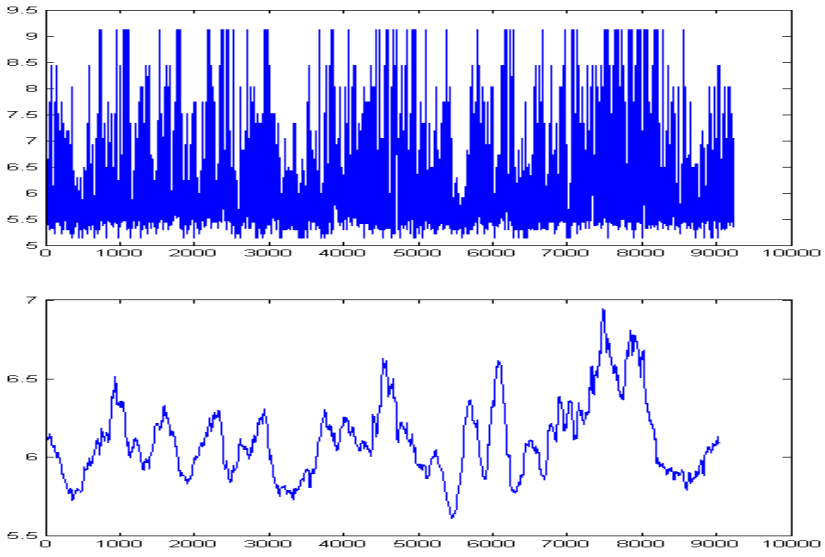


Figure C.1: 1^{st} Half of AC Milan vs Barcelona in UEFA 2006

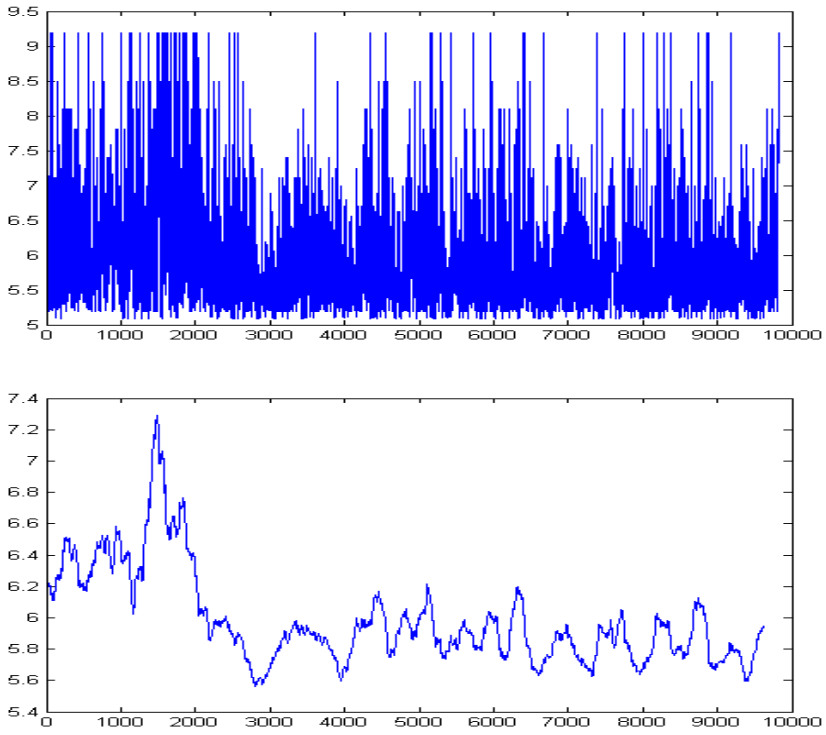


Figure C.2: 2^{nd} Half of AC Milan vs Barcelona in UEFA 2006

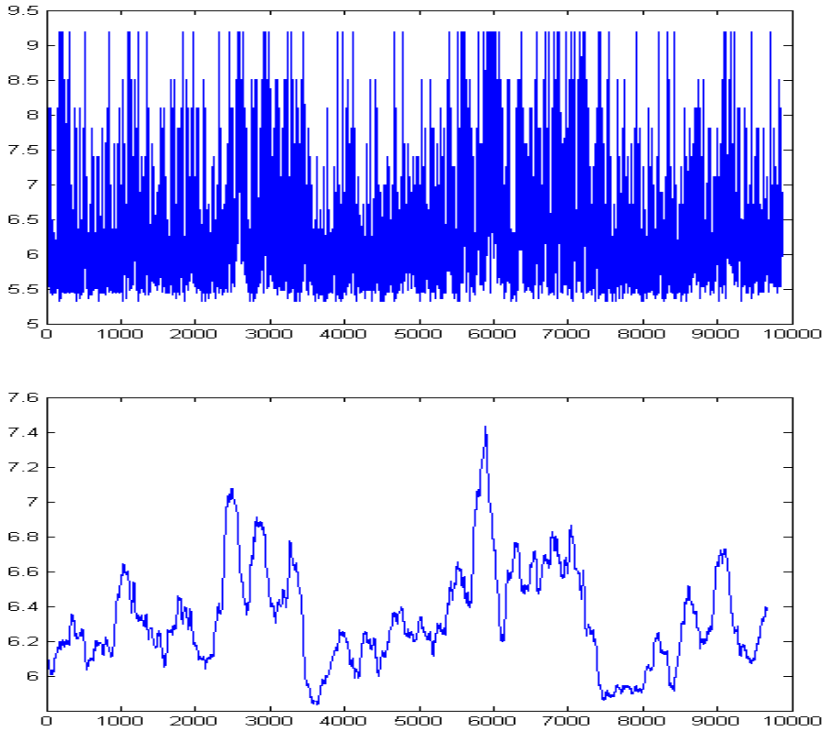


Figure C.3: 1^{st} Half of Arsenal vs Barcelona in UEFA 2006

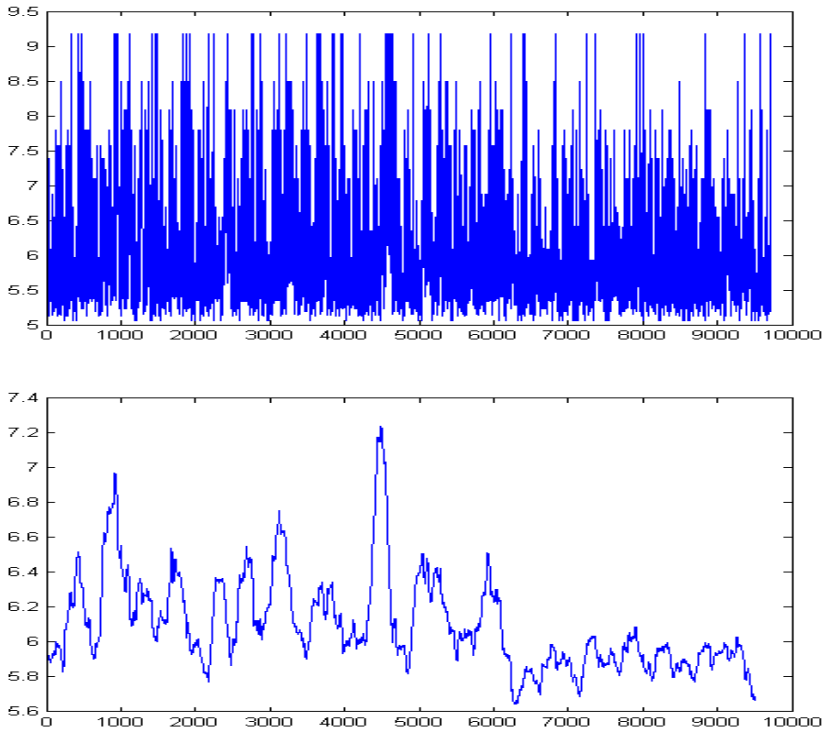


Figure C.4: 2^{nd} Half of Arsenal vs Barcelona in UEFA 2006

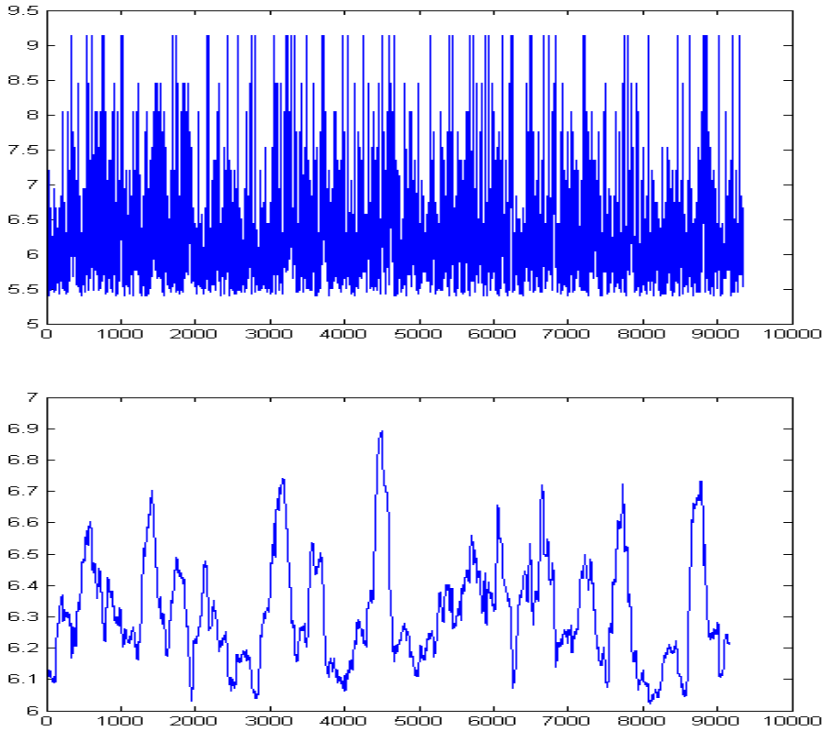


Figure C.5: 1^{st} Half of Italy vs France in FIFA World Cup 2006

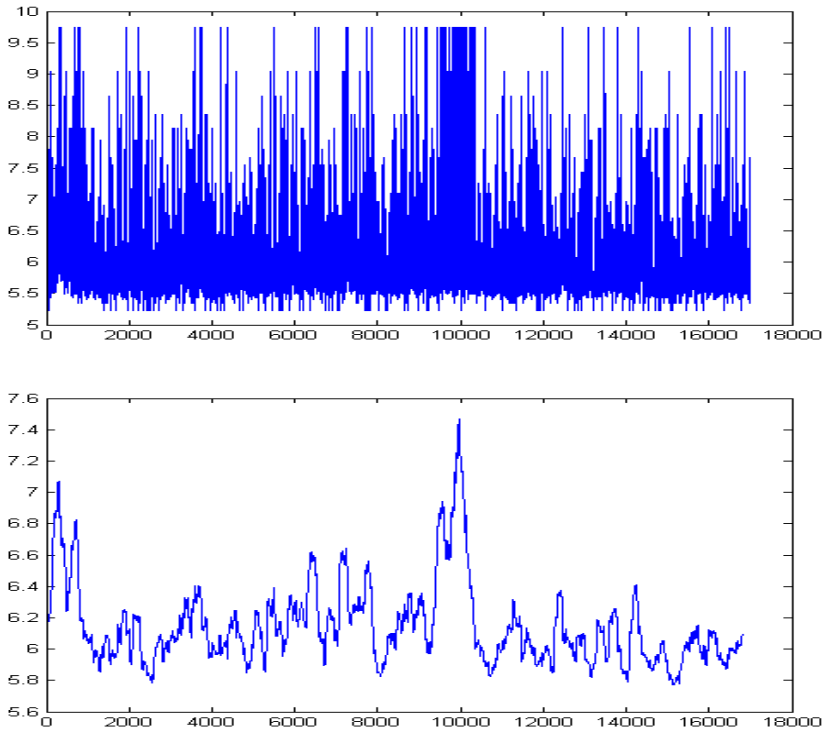


Figure C.6: 2^{nd} Half of Italy vs France in FIFA World Cup 2006

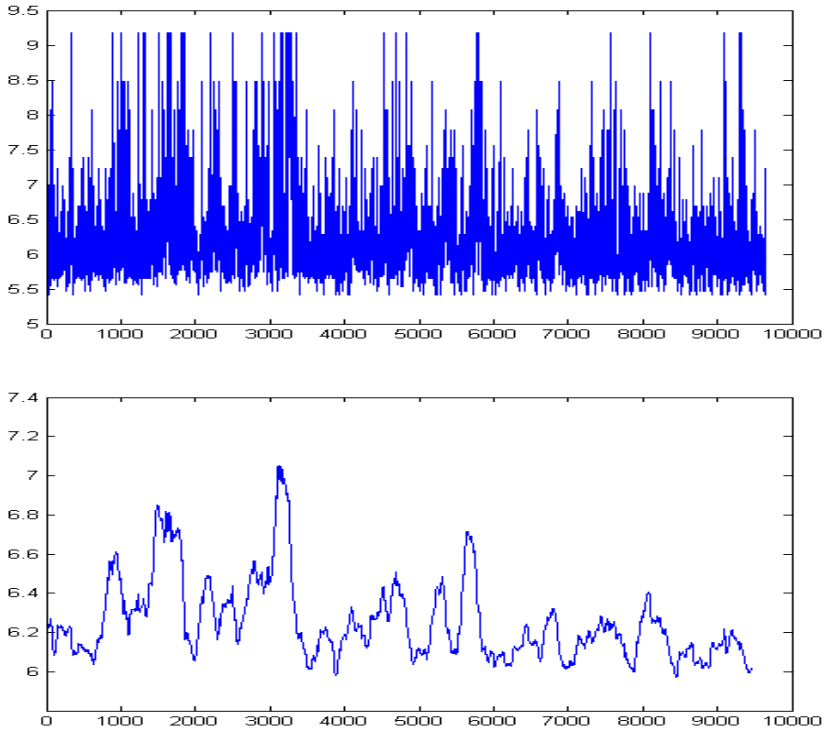


Figure C.7: 1^{st} Half of Korea vs Germany in FIFA World Cup 2002