

Attention Guided Football Video Content Recommendation on Mobile Devices

Reede Ren
Department of Computing Science
University of Glasgow
17 Lilybank Gardens
Glasgow, UK
reede@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science
University of Glasgow
17 Lilybank Gardens
Glasgow, UK
jj@dcs.gla.ac.uk

ABSTRACT

Live football video is the major content genre in 3G mobile service. In this paper, we introduce a realtime general highlight detection algorithm based on attention analysis. It combines attention-related media modalities into role-based attention curves, namely video director, spectator and commentator, to track viewers' feeling against game content from media data. A series of linear temporal predictors are generated from video data directly and employed to allocate strong attention changes, which are marked as scroll-back endpoints for mobile video skim. The advantages of our algorithm are that it avoids semantic uncertainty of game highlights and requires little training. We evaluated our approach using a test bed with five full games in FIFA World Cup 2002 and European League 2006 from different content suppliers, i.e. BBC and ITV to prove the robustness.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: audio-visual combined classification, video retrieval; H.3.1 [Content Analysis and Indexing]: highlight detection, index methods

General Terms

Algorithms, Design, Experimentation

Keywords

Attention feature, audio-visual fusion, linear prediction, and sports video retrieval

1. INTRODUCTION

According to the 3G survey in UK [9], live football video is the most welcome video genre in mobile video service. People watch football games on mobile phones, enjoy exciting short clips at half time and then are invited to place a bet as to which team will win the game. It becomes a common application pattern in everyday service. Compared to news and story film service, live football video

service is characterized by multiple program resources and loose audio-visual structure. Multiple resources here indicate that content suppliers offer different views of the same game or different interesting games simultaneously. Obviously, users will appreciate a realtime recommendation system, which highlights the most interesting video shots and offers appropriate controls on game content, such as extra information from background programs. To deal with such requirements, digital television broadcasting (DVB) domain developed two analogical techniques, window-in-window and window-matrix, which supply a swift user interface for video skimming and program switching. For example, window-in-window technique allows a small floating window to offer a brief view of secondary program and defines a smart key to speed up the screen switching. But these techniques can hardly be employed on mobile devices because of inherent limitations, especially the small size of display. To improve user satisfaction and enhance service quality, an active application content agent is necessary to rank video interest, mark highlight boundaries and offer swift control methods, such as switching program streams automatically. Such an agent can supply event-based media scroll-back instead of meaningless frame-by-frame technique, which is decisive for the playback function of video skimming and is helpful for payload management.

In the paper, we propose a content agent (Fig.1) in the OSI application layer based on football video attention analysis. Attention is a psychobiological measurement of content interest, which assumes the human focus and emotion variation. It quantitates reaction roles' excitement by computing attention curves and detects interesting event boundaries by attention peak segmentation. The rest of paper is organized as follows. Section 2 states related work in the field of psychological content analysis for football video and Section 3 explains the original motivation of our work. The role-based attention computing algorithm is explained in Section 4. It includes three parts, attention modalities in the football video, media modality attention models and the linear prediction algorithm for realtime attention computing. Experiment results are offered in Section 5 and Section 6 is devoted to conclusion and the guideline of future work.

2. RELATED WORK

The content agent faces two fundamental problems of semantic video analysis, how to measure the quality of content description and how to assume the interest of content. In

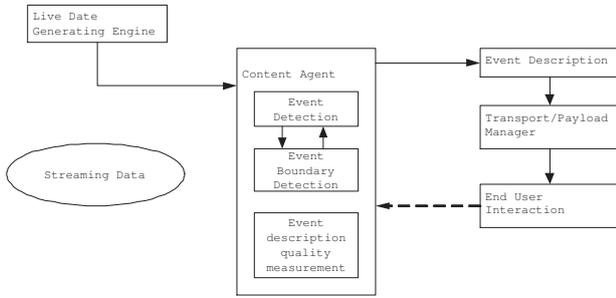


Figure 1: Content Agent

the specific football video domain, both of them are resemblance to relatively important event or highlight detection in a given temporal interval. Since highlight is the spirit of sports video, content interest weighting turns into highlight identification and the best style of content description is that catch and replay highlights at the right moment. However, event-based highlight detection methods in literature can hardly be used in such a realtime application because highlight here can not be defined by any game event set. From the semantic definition of highlight¹, Ma et al. [7] and Hanjalic et al. [3][4] proposed a psychobiological approach for general highlight detection by tracking viewer attention or excitement variation. They focus on attention capture from raw video data, namely how to map media features into the psychobiological attention space. Ma et al. [7] isolated media feature influence on perception and designed a series of feature-attention models, i.e. motion attention model, static attention model, and audio salient model. They linearly combined these feature-based attention curves to assume the intensity of ‘viewer attention’. But the isolation of features brings too much noise and makes the later fusion process fragile. With the increasing of feature number, noise will overwhelm ‘real’ attention peaks suddenly in experiments. Hanjalic et al. [3] selected a small feature set, only including block motion vector, shot cut density and audio energy. These single-feature attention(emotion) curves are smoothed by a 1-minute long Kaiser window, which significantly improves algorithm robustness. Later, [4] presented an adaptive filter to enhance curve peak and depress asynchronous noise. However, they do not address the problem of attention curve fusion, either.

3. MOTIVATION

The application of psychobiological approach is an exploration from computing science towards psychology. The mapping process from low level features to ‘attention’ intensity faces the problem of quantitative uncertainty, though their qualitative relation is ensured by computable psychology, i.e. salient map and active vision. For example, motion will attract more attention than static area, but we do not know how much the gain will be. Psychobiological experiments discover a linear reflection function between *stimulus* and *response* till *saturation* [2]. Note that *stimulus* is a combined affection from audio-visual signal. It

¹The Merriam-Webster dictionary defines highlight as something (an event or detail) that is of major significance or special interest.

partially explains why the isolated attention feature model [7] is so sensitive to asynchronous noise. Commercial encoding techniques, i.e. MPEG-1/2/4 and H. 263, deal with media data independently to save bitrate. They suppose that a 3sec-5sec misalignment will be ignored because of perception residence(MPEG-1 ISO11476-1). Such an asynchronism allowance in football video is furthermore enlarged by observer reaction bias during the production. In the view of psychology, game audio and video come from totally different reaction roles. Directors compose camera videos following personal understanding on game content, while microphones record cheers and noise automatically. In some sense, audio-based and visual-based attention peaks rarely appear on the same time. How to deal with such an asynchronism becomes the major problem in psychobiological highlight detection. Moreover, Hanjalic et al. [3] proposed an ‘average’ viewer to present so-called ‘standard’ response. There are two pitfalls in the assumption. The individual perception process is independent [2] [13] and video audience’s response can’t be collected from broadcasting video data directly. Such an assumption not only breaks the affection-reflection measurement circle, but introduces something visual in the psychological experiment. In this work, we analyze the attention/perception structure of football video so as to identify observer/reactor and their reflection in audio-visual media stream. Related salient media features are grouped according to their reactioner so as to remove observer bias and build up role-based attention state space. To solve the problem of attention asynchronism, we process these attention features on a coarse but effective resolution. Given the fact that the period of psychobiological attention reflection exceeds 0.7sec, audio-visual salient data is down-sampled to every 0.3sec by local mean. Such a low-pass filtering not only saves computing cost, but also significantly restrains random noise. Based on the multiresolution autoregressive attention model [16], we design a linear predictors array, to detect great attention variation, which marks possible start points of highlights instead of their duration. These moments define unreeling positions for interactive video skimming. Note that any operations requiring whole data can not be employed in the live video application, such as global normalization, which is widely used in [7] and [3] to increase signal noise ratio (SNR). Comparing with original MAR model, the linear predictor solution is a strict Markov and only relies on prior knowledge, though its scale is assumed statistically by the median sub-tree span of MAR tree. In addition, such an attention-based predictor can be easily realized by digital signal processor(DSP).

4. ATTENTION COMPUTING

Three major reaction roles in broadcasting football video can be easily identified, namely spectators, commentators and video directors behind visual frames (Fig.2). Their individual understanding of game content and reflections affects video viewer’s feeling and decides so-called ‘highlight’. Directors watch camera videos, edit them, decide shot style, such as field view and close-up, and insert video editing shots, i.e replay, to present the story in their eyes. Some conventions have been developed as ‘visual art’ and partially utilized in automatic game content analysis [11] [8] [12]. For example, dominant color ratio [17] and zoom depth [11] are calculated to assume shot importance, because a closer view brings more details inside the play field and will assign a lit-

tle more prominence to the shot. Replay shots are extracted to build video summaries [10], since they reiterate important moments. Response from spectator and commentator dominates audio stream. As a group, stadium audience cheer at exciting moments and remain relatively silent mostly in rest of the game. They attract video viewers by their loud plaudits and hypnotize them with silence. Commentators' behavior is complex. As a business, commentators reiterate game contents with personal style. Their specific jargons are detected as keywords to label game events [18]. On the other hand, commentators are a group of professional stadium audience. Their excitement intensity varies with the crowd. In



Figure 2: Observers in sports broadcasting

the following sections, we present a set of temporal-spatial media features in literature, which touch attention variation.

4.1 Attention Feature

Psychobiological research on visual attention has shown that strong variation, stimuli strength and spatial contrast are major facts attracting attention [6][13]. Since game videos focus on the close environment, play field, directors mainly rely on fast variation and contrast to stimulate viewers' attention. They zoom-in what they consider interesting, replay what they assume important, fast cut and change shots to offer different view points towards game events [19]. However, replay and field-away shots interrupt the continuous perception process and might trouble viewer's understanding. They are rarely employed unless carrying essential game aspects. So the length of replay and field-away shots monotonically increases with viewer attention level. Moreover, replay shots are sandwiched with special video edit effects, which increase attention intensity, too. Another perception issue is zoom depth of shot, which is proportional to the area of rectangle of interest (ROI), an important measure of static salience. The attention feature for audience and commentators is relatively simple. Loud and greatly varying sound always catch attention. Table.1 lists attention modalities in literature.

4.2 Role-based Attention

Fusing media modalities from the same observer, such as director, diminish reaction bias and will ease the later latency assumption. Three role-based attention curves are computed, namely video director, spectator and commentator curve. Video directors' attention modalities reflects static salience and visual temporal variation. The static salience is assumed by zoom depth and ROI area, while visual temporal variation is calculated by shot frequency and shot length. According to Table.1, the four tuple set (Max

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	*
penalty box	rect of interest	*
shot duration	temporal variance	-
shot cut frequency	temporal variance	+
motion vector	temporal variance	*
zoom-in sequence	temporal variance	+
replay	temporal contrast	*
off-field shot	temporal contrast	*
baseband energy	loudness	+
cross zero ratio	sound variation	+
speech band energy	sound variation	+
keyword	semantic	*

Table 1: Director-based Attention Feature, + stands for the proportional qualitative relationship between feature and attention, while - is for inverse proportional and * for unsure.

object size, Mean color contrast, Average motion, Shot frequency) is designed for video director attention. Max object size measures given video object in football video, i.e. uniform, face and goalpost and is calculated by the video object contour, a MPEG-4 feature. Shot frequency is the shot number in a given temporal interval, such as 1 minutes in experiments, while average motion counts the average motion block number per frame during the same period.

The attention intensity of spectator is proportional to the background noise [5]. We use average audio baseband energy in 1sec long window and its absolute difference from an given interval mean to describe audience attention and its variation trend. Four scales, 5 sec, 10 sec, 30 sec and 1 minute are selected. The audience attention is a five element vector, $(E_0, D_5, D_{10}, D_{30}, D_{60})$, where E_0 is baseband energy, and $D_5, D_{10}, D_{30}, D_{60}$ for the absolute difference from 5 sec, 10 sec, 30 sec, 60 sec mean audio energy, respectively. Speech speed and loudness of commentators hint their excitement. [18] computed a low band of LPCC parameters from 0 to 3 and the cross zero ratio to assume speaker attention. Given the noisy situation, we just take the sum of LPCC coefficients and cross zero ratio in 1.5 sec to assume commentator attention.

4.3 Attention Fusion and Highlight Detection

The complexity of audio-visual information fusion comes from not only the media asynchronism and different event resolution, but what they observed. The audio-visual data stream reflects the semantic story behind video. No matter what kind of middle presentation layer is used, i.e. text description [14] [15], it is hard to match audio and visual segments onto their semantics. The advantage of psychobiological approach is that it avoids such a gap and combines audio-visual information according to their affection on measurable attention signal. Note that computable psychological methods guarantee such an extraction process from multimedia

data with a high confidence. Attention-based approach will be useful in the content analysis of passion-lead videos, i.e. sports video and music video.

4.3.1 Attention signal sampling

As a psychobiological measurement, attention describes human behavior before stimulus. Such a reaction period will exceed 0.384 sec against strong and simple stimulus, such as a flash inside a dark room and the transmission between attention states will cost the similar duration [13]. The interval between two attention peaks will be 0.7 sec at least. Moreover, [4] utilized an 1-minute long low-pass filter to smooth attention (excitement) curve in experiments, which indicates that the bandwidth of attention signal is more narrow than we imaged. Audio and visual stream are obviously over-sampled for attention signal assumption. Decreasing sample rate will not only save computing cost but also avoid noise introduction. The finest data resolution in current system is set at 0.3 sec to fulfil Nyquist-Shannon sampling theorem. Fig.3 and Fig.4 show the spectator and director attention in the second half of final game in FIFA World Cup 2002, respectively. In both of figures, we mark the moment of goal event according to official game record from FIFA website. The attention peak displacement partially proves the existence of reflection bias between observers.

4.3.2 Multi-resolution autoregressive attention model

As a reflection of game content, attention signal keeps the similar embedded temporal structure as video story's. A widely accepted assumption on such a structure is that it is a markov process on graph. [18] and [17] proposed some very complex hidden markov models, such as hierarchical hidden markov model and coupled hidden markov model, to simulate content movement in football video. Without losing generality, these temporal structures are simplifications from a markov process on a graph. All these models face similar problems, how to define the number of markov state and how to train such a model. Given the variation of game content and artifact, few successful works have been reported in literature.

[16] proved that a multi-resolution autoregressive tree (MAR) is equivalent and with the same ability to a markov process on graph in the temporal sequence analysis. MAR (Fig.5) is a scale-recursive linear dynamic model [1], which employ a tree to combine heterogenous data, i.e. visual, audio, caption text and other complement media sources, on different bands and resolution under some given requirements, i.e. $1/f$ smoothness. Such a model assumes that each node in the coarse resolution is a linear combination of nodes in fine resolution. The optimization process includes a fine-to-coarse filtering sweep and a coarse-to-fine smoothing sweep. The fine-to-coarse recursion corresponds to the multiresolution analysis of signals and is a variation of Kalman filter for multi-scale models on tree. In the step, the number of nodes in coarse layer is decided by recursive measurement updating and their parameters are computed by minimizing prediction error. The coarse-to-fine sweep is the multiresolution synthesis of signals, in which higher resolution details is added at each scale and random noise is suppressed. MAR model organizes attention segmentation according to their temporal coherence into subtrees and treats its subtrees independent from each other. In our application, each

a subtree is regarded as a game content element. Moreover, the multi-resolution character of MAR structure meets the multi-scale nature of semantical content, which has no counterpart in the literature of game event detection.

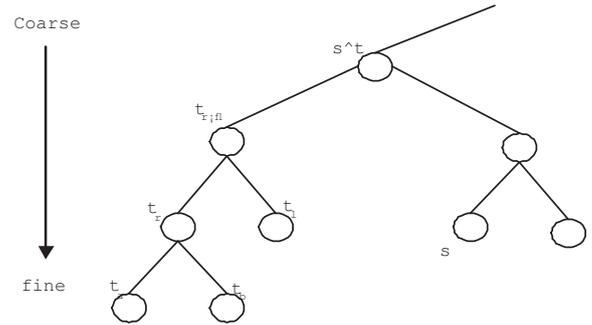


Figure 5: Dyadic tree for 1D signal autoregressive analysis and some notation used in paper

In our system, each node of MAR tree is a triple state holding role-based attention (spectator, commentator, director). But sometimes attention state of commentator is ignored in experiment data set to avoid too much noise, because we can hardly compute commentator attention intensity with a high confidence from the mixture of commentator and spectator audio in broadcasting video. The same to the general MAR framework, the MAR tree building-up process includes two steps, fine-to-coarse and coarse-to-fine, to minimize the residual prediction error under the smooth measurement. Such a model can be even simplified into an internal MAR, in which each node is a linear combination of prior knowledge, if we remove replay shots inside to ensure the non-loop topological structure. Though MAR model is strong enough for abnormal event detection in football video, there are still a lot of pitfalls. It is obvious that MAR model is a model built directly on a given data. Different from markov model, MAR model needs few training and all its parameters is calculated during the optimization process at the cost that such a model can hardly be extended to adopt other game videos. In some sense, a MAR model contains the whole knowledge of the observed process to decide the best topological structure. So a MAR-based algorithm will not fulfill markov condition we have mentioned in Section.3. In addition, the heavy computing cost necessary for MAR tree building decides that the MAR model can hardly be employed in real time application. But the model offers a theoretical explanation for our linear predictor system and brings a statistical assumption of predictor scale in application. Tab.2 lists the length of MAR subtree span in our evaluation data set. We use the mean of median span degree as the predictor scale in experiment. The MAR tree offers a overlook on game content and can be used in related applications of temporal structure mining, i.e. automatical replay production, content-based video indexing and decomposition.

4.3.3 Realtime linear predictor system

Given the limitation of MAR model, we develop a series of linear prediction models based on autoregressive tree for live football highlights detection. As the cost of simplification,

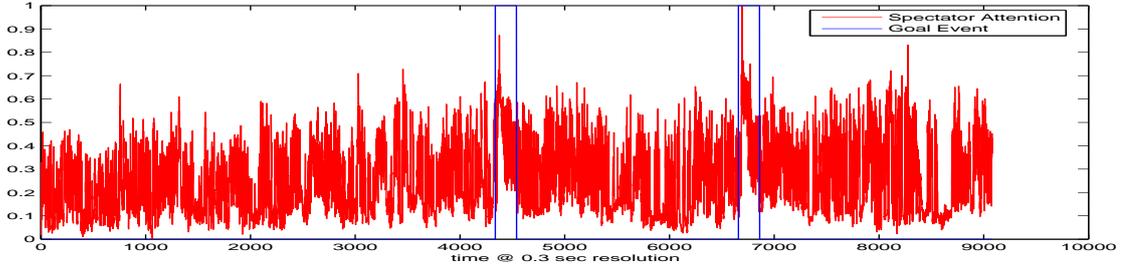


Figure 3: Normalized spectator attention curve @ 0.3 sec resolution in the second half of World Cup 2002 final game

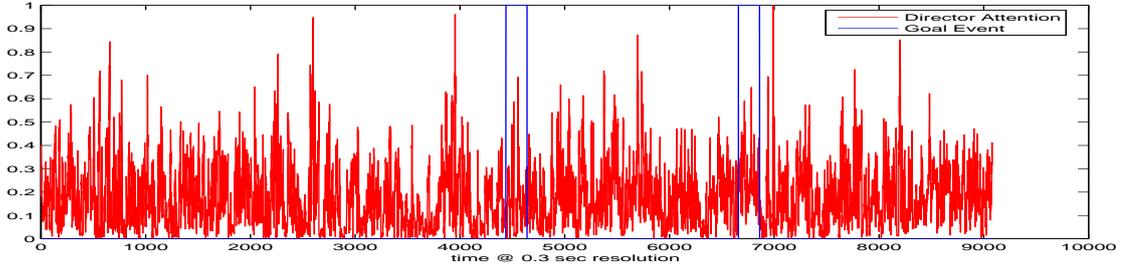


Figure 4: Normalized director attention curve @ 0.3 sec resolution in the second half of World Cup 2002 final game

	Min		Median		Max	
	Aud	Dir	Aud	Dir	Aud	Dir
Ger-Bra I	73	27	147	59	1406	1511
Ger-Bra II	82	21	151	64	1322	973
Bra-Tur I	126	41	153	42	1587	1470
Bra-Tur II	153	62	171	52	1210	1107
Ger-Kor I	41	23	214	74	970	760
Ger-Kor II	53	35	162	69	981	873
Mil-Bar I	70	34	167	127	1344	1406
Mil-Bar II	76	46	184	104	1947	1512
Ars-Bar I	67	38	174	62	1327	1006
Ars-Bar II	73	36	220	71	1211	1241

Table 2: Sub-tree Span @ 0.3 sec resolution, where aud and dir stands for audience and director, respectively

such a prediction array loses the ability of attention-based video decomposition and can not define the duration of game highlights. But it marks the abnormal moment in the long video sequence with few computing cost and become a more general model than MAR. The most important of all, these moments marked are computed according to the assumption of perception or attention intensity variation. They provide a better pathway to understand game content and meet the requirement of content-based video skimming. Moreover, the prediction array can be easily simulated by hardware, if necessary. As we have mentioned, the predictor scale is assumed by the median of MAR's subtree span. The following is the formal description of our linear prediction system.

A linear predictor model forecasts the amplitude of a sig-

nal at time m , $x(m)$, using a linearly weighted combination of P past samples as,

$$x(m) = \sum_{k=1}^P a_k x(m-k) + e(m) = ax + e(m) \quad (1)$$

where $e(m)$ is the prediction error, which carries information. By minimizing least mean square error, a can be calculated directly from the autocorrelation matrix of samples.

$$a = R_{xx}^{-1} r_{xx} \quad (2)$$

where $R_{xx} = E[xx']$ is the autocorrelation matrix of the input vector, $r_{xx} = E[x(m)x]$, and E is the mean operator. A recursive algorithm is proposed to compute the coefficients of a predictor of order P ,

Algorithm: Predictor Coefficient Computing

```

Given sample  $x[0..P]$ ;
 $a[0 : P] = 1$ ;
for(int i = 0; i ≤ P; i++)
{
 $r_{xx}[i] = x[i]x$ ; }
error[0] =  $r_{xx}[0]$ 
for(int i = 1; i ≤ P; i++)
{
Delta[i - 1] =  $r_{xx}[i] - \sum_{k=1}^{i-1} a_k^{(i-1)} r_{xx}(i - k)$ 
 $k[i] = \frac{Delta[i-1]}{error[i-1]}$ 
 $a[i] = k[i]$ 
 $a[j] = a_j - k[i]a[i - j]$ 
error[i] =  $(1 - k[i]^2)error[i - 1]$ 
}

```

When the predictor error $\|e(m)\|$ exceeds a given threshold, the model will be updated by recomputing coefficients and a break is marked on the attention curve as a boundary of attention segments. For each break, the sign add-up operator on $e(m)$ elements (Eq.3) is carried out to decide its direction. If the sum exceeds zero, the boundary will be labeled as up-shift, which indicates an increase of attention, otherwise it will be a drop-down.

$$Boundary = \sum_i [sgn(e(m_i))] \quad (3)$$

The quality of a media stream can be assumed by the number of up-shift boundaries, which stands for strong variations in the attention curve. Moreover, such a measurement is media independent and can be extended to multiple-stream data, whose quality is the average up-shift boundary number over all streams. This approach offers a replacement for physical shot segmentation and takes perception variation into account instead of a change of camera view. Nevertheless, the up-shift boundary places the end for scroll-back operation in video skim. The highlight period is the temporal interval between a up-shift boundary and the close-by drop-down.

5. EXPERIMENT

Five games are collected from BBC and ITV sports to build up test bed. Three of them come from World Cup 2002, German vs Brazil, Brazil vs Turkey, and German vs Korea. The later two are from Champion League 2006, Arsenal vs Barcelona, and AC Milan vs Barcelona. They are encoded in MPEG-1 with visual resolution at 352×288 and audio at 44KHz/16bit. The total length is about 14 hours, including interview and celebration. All games are divided in halves, and each half is considered as an independent game. For example, Ger-Bra I stands for the first half of final game in World Cup 2002, German vs Brazil. Attention from video director, spectator and commentator are sampled at different speed. The video director attention is sampled every 1 sec with prediction order 17, while spectator and commentators' attention at 2 sec with prediction order 20. Table.3 shows the number of up-shift boundaries detected and relate shots.

	Shot Number	Upshift Number		
		Director	Audience	Comment
Ger-Bra I	105	61	31	29
Ger-Bra II	173	74	37	31
Bra-Tur I	160	67	33	24
Bra-Tur II	148	62	36	26
Ger-Kor I	173	71	47	30
Ger-Kor II	185	69	39	31
Mil-Bar I	212	98	41	33
Mil-Bar II	198	117	46	35
Ars-Bar I	177	79	54	37
Ars-Bar II	190	91	61	49

Table 3: Attention Upshift Boundary Detection

A 2-minute long sliding window is employed to detect highlights. These filtered video clips are ranked by the sum of

up-shift edge number inside. As most works in highlight detection, Table.4 presents goal events² found in the top 10 of the ordered list and their rank. Two results are compared from director attention and from the combination of audience and director attention, where we assume that audience and director attention are roughly synchronous because of the coarse resolution (2 minutes).

In current experiments, we ignore commentator's reflection because the mixed audio brings too much noise in the computing process of commentator attention. Commentator information will be useful in the event identification and labelling. But the introduction of commentator attention decreases overall system performance. Though such a strange phenomena is caused partially by the mixture noise in the production, we assume it reflects the two-faced nature of commentator. As a job, commentators should keep clam during their explanation but can hardly hold their personal emotion in some special cases, for example, the game with England or British teams in BBC. They are biased. The attention curve of commentator should be even in most of time but with some very strong variation, which is not propositional to the interesting level of game event. In some sense, the solution of spectator and director combination may be a better choice for general attention analysis and their result is good enough for content recommendation in our experiments.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented an attention analysis framework for live football video processing. It is proposed for realtime interesting event detection and video skim generation. The algorithm consists of two-step processes. The video signal, including audio, is processed first by extracting attention-related media modalities, which are coupled into three role-based attention curves, namely video director, spectator and commentators. These curves reflect independent emotional feeling against game content from ad-hoc viewers and make it possible to identify interesting segments in the view of human perception. A series of linear predictors are proposed to assume the temporal evolution between attention states. The prediction failure indicates a strong and fast change in attention and its temporal intensity is employed to allocate game highlights. Moreover, we fused role-based attention curves by counting their state signals. Though the result promises a better performance comparing with director attention only, there are still questions on fusion method and the information entropy distribution across media in video data. Our Current research focuses on two problems, how to measure and compare system performance according to game interest instead of plain event detection precision and how to combine of multiple attention curves with confidence, so as to further reduce the number of false detections. Both of them lead to a multi-modal content model for passion-lead video analysis. Nevertheless, involving user feedback in the algorithm is an interesting topic to address in the future.

²Given the uncertainty of game highlight, most works in game highlight identification only measure the precision of goal detection.

	Goal Number	Director Attention Only		Director and Audience Attention	
		Detected Goal Events	Rank in List	Detected Goal Events	Rank in List
Ger-Bra I	0	-	-	-	-
Ger-Bra II	2	2	1,3	2	1,2
Bra-Tur I	0	-	-	-	-
Bra-Tur II	1	1	1	1	1
Ger-Kor I	0	-	-	-	-
Ger-Kor II	1	1	1	1	1
Mil-Bar I	0	-	-	-	-
Mil-Bar II	1	1	2	1	1
Ars-Bar I	1	1	4	1	2
Ars-Bar II	2	2	2,3	2	1,2

Table 4: Performance of Goal Detection

7. ACKNOWLEDGEMENT

The research leading to this paper was partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

8. REFERENCES

- [1] K. C. Chou, A. S. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion and regularization. *IEEE Trans on automatic control*, 39(3):464–478, Mar 1994.
- [2] J. Crary. *Suspensions of Perception: Attention, Spectacle and Modern Culture*. Cambridge, MA: MIT Press, 1999.
- [3] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Trans. on Multimedia*, 7(6):1114–1122, Dec 2005.
- [4] A. Hanjalic and L. Xu. Affective video content repression and model. *IEEE Trans on Multimedia*, 7(1):143–155, Feb 2005.
- [5] R. Lenardi, P. Migliorati, and M. Prandini. 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, May 2004.
- [6] M. S. Lew. *Principles of Visual Information Retrieval*. Springer, 1996.
- [7] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *ACM Multimedia02*, 2002.
- [8] N. Babaguchi, Y. Kawai, and T. Kitashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. Multimedia*, 4:68–75, Mar 2002.
- [9] G. News. 3g football best mobile service, Jan 2005.
- [10] H. Pan, P. Beek, and M.I. Sezan. Detection of slowmotion replay segments in sports video for highlights generation. In *IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2001.
- [11] R. Ren and J. Jose. Football video segmentation based on video production strategy. In *ECIR 2005*, 2005.
- [12] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. Classification of self-consumable highlights for soccer video summaries. In *ICME 2004*, 2004.
- [13] A. M. Treisman and N. G. Kanwisher. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8:218–226, 1988.
- [14] J. Wang, C. Xu, E. Chng, L. Duan, K. Wan, and Q. Tian. Automatic generation of personalized music sports video. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 735–744, New York, NY, USA, 2005. ACM Press.
- [15] J. Wang, C. Xu, E. Chng, K. Wah, and Q. Tian. Automatic replay generation for soccer video broadcasting. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 32–39, New York, NY, USA, 2004. ACM Press.
- [16] A. Willsky. Multiresolution markov models for signal and image processing. In *Proceedings of the IEEE 90 (8) (2002) 1396-1458*, 33, 2002.
- [17] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2002.
- [18] H. Xu and T. Chua. The fusion of audio-visual features and external knowledge for event detection in team sports video. In *MIR 2004*, 2004.
- [19] H. Zettl. *Sight, Sound, Motion: Applied Media Aesthetics*. Wadsworth, Belmont CA, 1990.