AFFECTIVE SPORTS HIGHLIGHT DETECTION

Reede Ren[*], *Jeomon Jose*[*], *He Yin*[+]

[*]omputing Science Dept., Glasgow University; [+]IBM, China
[*]17 Lilybank Gardens, G12 8RZ, Glasgow, UK
[*]phone: + (44) 01413306292, email: reede,jj@dcs.gla.ac.uk; [+]email:heyinhy@cn.ibm.com

ABSTRACT

This paper explores a psychological attention approach for sports highlight detection. A multiresolution autoregressive algorithm is proposed to fuse misaligned audio-visual time sequences and estimate an unified attention curve. Game highlights are found by ranking attention intensity; content-based events are filtered out by allocating local attention peaks. The test bed includes six complete football games from World Cup 2002, 2006 and Champion League 2006, and two content suppliers, BBC and ITV. Two evaluations are presented, the comparison on average attention and event attention, and the ranking of goal events. Experiments show this fusion framework is robust on different data collections.

1. INTRODUCTION

As a combination of audio-visual stimulus, video conveys various emotion stories, i.e. happy and sad. Affective analysis deals with these affect aspects and helps the content mining in the domain of sports videos and story films. Endowing a system with the capability of affective understanding has introduced many interesting applications, e.g. comedy and action film discrimination [19], sports structure analysis [6], event detection [11], sports highlight detection [15] [3], and audio emotion indexing [16]. The advantage of affective approaches comes from psychological observations. In the neural situation, people's feeling is mostly similar against a given stimulus [12]. A large set of affective features have been discovered in the literature of computing psychology and classified by their qualitative effects (Table 1). Moreover, interesting video contents always attract attention and incur strong emotion variation. To some extent, video highlights co-occurs with attention peaks at a high probability [3] [4]. Affective analysis offers a reliable generic highlight identification method, while concerning few content details.

Current works in the literature of affective analysis focus on the projection from modality features into psychological spaces. Ma et al. [10] isolated feature influences on perception by a series of independent feature-attention models, i.e. motion attention model, static attention model, and audio salient model. These feature-based attention curves are linearly combined to calculate the intensity of so-called "viewer attention" in generic videos. But the isolation of feature modalities makes the later fusion fragile. Hanjalic et al. [3] selected a small feature set, including block motion vector, shot cut density and audio energy. They assume the possibility of highlights by counting the peak number of feature-attention curves in a floating window. This fusion method relies on the signal noise ratio (SNR). An 1-minute low-pass Kaiser window filter is used to smooth feature-attention (emotion) curves. Later, an adaptive filter is proposed in [4] to enhance curve peaks. However, many questions still remain. These approaches focus on the filtering of attention signals from background noise, but discard the nature of modalities and the content of videos. The duration of content events vary with modality. For example, an excited shout of "goal" in a football game will cost several minutes of video to iterate the same content. A single temporal resolution over all modalities can hardly afford this difference. Moreover, the video content is usually regarded as a Markov process on graph. But the local maximum detection is just an application of the second order Markov chain. There is a complexity gap. Nevertheless, though feature affective models are supported by psychological observations, how to combine multiple affective models is unsure in psychology till now. Decision from different affective models may conflict with each other, but human cannot be in two affective states, i.e. sad and happy, at the same time.

In this paper, we present an optimised algorithm for misaligned affective feature fusion in the application of sports video highlight detection. A multiresolution autoregressive model(MAR) is proposed to combine and distribute modality information from different modalities and multiple resolutions. The MAR is equivalent to a Markov model on graph [21]. Then affective features are ranked according to their standard deviations, in which the top 3 are combined to assume the overall attention. The rest of paper is organised as follows. Section 2 introduces some background knowledge in psychology and the computing of affective features. Section 3 is devoted to the MAR fusion model. The experiment results and conclusion will be found in Section 4 and Section 5, respectively.

2. ATTENTION COMPUTING

Attention is a fundamental concept in psychology, which has been studied from the dawn of modern psychology. Attention is regarded as the gateway for cognition processes, such as decision making, memory and emotion. A widely accepted hypothesis is that the sum of attention keeps constant. The perception is to distribute attention onto different stimulus and the ratio of attention reflects the observer's interest or reaction intensity towards stimulus. In [8], a set of differential attention-interest equations is developed to describe the relationship among interest, attention, and human activity before stimulus in an unknown environment. It is possible to find a solution to attention intensity in a given context. Generally, attention is proportional to the strength of stimulus or the amount of pan-out information as far as information theory concerns. In sports videos, attention is roughly proportional to the interest of contents, where the stimulus comes from.

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	$+$
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	$*$
scene affect vector	valance	$*$

Table 1: Director-based Attention Feature, $+$ stands for the positive qualitative relation between feature and attention, while $-$ is for negative and $*$ for unsure. If the feature induces increase of attention intensity, it will be defined as positive feature, and the negative otherwise. If the feature can bring both positive and negative affection in different contexts, it will be titled as qualitative unsure.

2.1 Attention and Markov

The video content is a Markov process. If we regard game events as states, the content process is an experience of state transitions, in which game contents have changed from the state it was in the moment before. The sports video is a record of this Markov; the audio and visual stream are observation sequences from different sensors. The extracted temporal sequence of modality attention is a discrete temporal sequence with Markov character. However, though the game content is a Markov, the semantic uncertainty and the context complexity decide that the content process can hardly be described by any Markov model with given states. There is a trade-off between the number of Markov states and model generality. More states will increase detection precision at the cost of model extensibility. Considering artefacts in video production, a Markov model with too many states will be fragile. Additionally, traditional Markov Chain Monte Carlo (MCMC) techniques will be troubled by the strong temporal correlation among hidden states in sports videos. Nevertheless, audio and visual sensors are with different resolution and sampling rate. Multiple resolution and asynchronism are essential aspects of the content-based audio-visual fusion process. For example, audio is a generally brief media style; visual stream carries rich but sometimes helpless details. A loud shout ”goal” from commentator costs several minutes of visual data to reiterate the same story. All these mismatches from resolution, date sampling and media alignments, hint that the multi-modality fusion has to be carried out on a coarse temporal resolution. To some extent, audio and visual stream are only synchronous on semantic events. In short, Markov states in the audio and visual stream are asynchronous in most cases except the semantic level.

2.2 Attention Features

Temporal variation, spatial contrast and stimuli strength, i.e. pure red colour, are major facts attracting attention [9][17]. However, a sports video is not a plain stimulus-reaction experiment, but a content understanding with plenty of domain knowledge. The semantics of video object is an important issue in the affection assumption. For example, the ball is always a focus in football games [2] and the goalpost will predict a possible shot event [5]. Their appearance attracts great attention in the context of football games.

Six visual features are selected, including visual harmony, shot frequency, shot type, play field ratio, the mean amplitude of motion vector, and uniform size. Visual harmony is proposed for the measurement of static spatial contrast. Given the block-based approach in commercial encoder standards, i.e. MPEG-1(8×8 blocks), block mean hue (Eq.1) and block hue covariance (Eq.2) for $n \times n$ image block with the centre at (i, j),

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

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

where C is the pixel colour. We use an 256-bin histogram to count the block covariance distribution. Then the visual harmony of a frame is,

$$Vh = \arg\max_{N} \sum_{n=0}^{N} (-P_n \log(P_n)) \tag{3}$$

where P_n is the portion of bin n over all histogram. The visual harmony Vh is the block covariance value at the bin N. The uniform size is assumed by a FST detector on pyramid [13]. In sports videos, most shot transitions are of the cut type. Shot boundaries are identified by the two-threshold algorithm in [18]. We classified visual shots into four categories, field-away, close-up, replay and field view by algorithms in [13]. The play field ratio is the play field area over the whole image, which is detected by the Gaussian mixed model of grass hue [13].

The audio in sports videos is mixed by comments and noise from spectators. In [20], speech pitch and cross-zero-ratio are used to detect excited commentators, while the excitement of spectators is assumed by base band audio energy [7]. The audio salient feature bag includes base band energy, spectral centroid, spectral flux and octave energy.

2.3 Entropy Measurement

Instead of the widely accepted normalisation [3] [4] [19], we propose an self information measure to estimate the intensity of attention and alleviate the dependency on data collection. As far as cognition psychology concerns, 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 message will decide the distribution of attention. This hypothesis bring two helpful conclusions. First, it is intuitive to introduce self-information (Eq.4), which is defined as the amount of information that knowledge about a certain event, adds to overall knowledge.

$$Entropy = -\log_2(P_i) \tag{4}$$

where P_i is the appearance probability of a feature at the given value i. The self-information can be robustly computed by the histogram estimation of feature distribution. Second, the unified attention state can be regard as,

$$I_{attention} = \overrightarrow{A}\overrightarrow{E} \tag{5}$$

where $\overrightarrow{A}$ is the vector of attention ratio over modalities; E is the modality contribution, which stands for the information supplied by a given modality. This equation (Eq.5) hints a Kalman filter like technique to estimate the unified attention intensity with the temporal smooth constraint. However, $\overrightarrow{A}$ is unknown in most cases of perception processes.

3. AUDIO-VISUAL FUSION MODEL

Multi-resolution autoregressive model(MAR) is a multi-scale recursive linear dynamic model [1], which simulates a random process by a serial of AR models on multiple scales. It combines heterogeneous data in different spectral bands and at different resolution following given criteria, such as fractal smoothness. A general algorithm for MAR parameter estimation is proposed in [21]. Here we specialise this algorithm and extend a later feature ranking step to assume the unified attention curve at different temporal resolutions.

Since the attention ratio distribution $\overrightarrow{A}$ is unknown, the direct modality combination is unrealistic. However, audio and visual stream are two independent observations on the same message production process, i.e. a football video. We can employ one media attention curve as a measurement to the other. In another word, audio and visual attention curves are independent but rough observations; a better estimation can be drawn by a MAR tree. Denote the set of resolution by $R = \{1,...,R\}$, with $r = R$ being the finest resolution. Since extracted visual salient features are of different resolutions, for example, the shot frequency is meaningless if the width of an observation window is less than the shot length, we set the finest combination resolution as 1.4 times of the longest shot duration. Additionally, the window in experiments is about 50 sec, very close to Hanjalic's 1-minute window [3]. The node N at scale r is $N_n^{(r)} = \{1 : 2^r\}$ in the bitree. Let $x(s)$ be the observation vector of visual attention at a node s, $y(s)$ for the audio, the discrete-time process can be described by a linear stochastic difference equation,

$$y(s) = \frac{1}{N}Hx(s) + v(s) \tag{6}$$

We assume the contribution of visual salient features are of the same importance. H is a vector of $\{1,...,1\}$ and N is the normalisation parameter. $v(s)$ is a Gaussian noise on the tree. For simplicity, we use the binary tree in the MAR model, the projection from finer resolution to coarse resolution will be

$$x(s) = [0.5, 0.5]^T x(s|s-) + w(s) \tag{7}$$

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. The three-step algorithm is presented in the following sections.

3.1 Fine-to-coarse Sweep

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

3.1.1 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

$$\begin{aligned} \hat{x}(s|s-) &= 0 \qquad (8) \\ P(s|s-) &= P_x(s) \qquad (9) \end{aligned}$$

3.1.2 Measure Updating

The measurement updating is identical to the analogous equations in Kalman filter.

$$\hat{x}(s|s) = \hat{x}(s|s-) + K(s)v(s) \tag{10}$$

where v(s) is the measurement innovations,

$$v(s) = y(s) - H\hat{x}(s|s-) \tag{11}$$

which is zero-mean with covariance,

$$V(s) = HP(s|s-)H^T \tag{12}$$

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

$$\begin{aligned} K(s) &= P(s|s-)H^T V^{-1}(s) \qquad (13) \\ P(s|s) &= [I - K(s)H]P(s|s-) \qquad (14) \end{aligned}$$

3.1.3 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,

$$\begin{aligned} \hat{x}(s|s-) &= P(s|s-)\sum_{i=1}^{K_s} P^{-1}(s|sa_i)\hat{x}(s|sa_i) \qquad (15) \\ P^{-1}(s|s-) &= P_x^{-1}(s) + \sum_{i=1}^{K_s}[P^{-1}(s|sa_i) - P_x^{-1}(s)] \, (16) \end{aligned}$$

3.1.4 Fine-to-Coarse Prediction

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

$$\hat{x}(s|sa_i) = F(sa_i)\hat{x}(sa_i|sa_i) \quad (17)$$

$$P(s|sa_i) = F(sa_i)P(sa_i|sa_i)F^T(sa_i) + U(sa_i) \quad (18)$$

where

$$F(s) = P_x(s\bar{r})A^T(s)P_x^{-1}(s) \quad (19)$$

$$U(s) = P_x(s\bar{r}) - F(s)A(s)P_x(s\bar{r}) \quad (20)$$

3.2 Coarse-to-Fine Sweep

When the fine-to-coarse sweep reaches the root, the covariance and estimation at all nodes are ready. 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, because they 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 $s\bar{r}$.

$$\hat{x}_s(s) = x(\hat{s}|s) + J(s)[\hat{x}_s(s\bar{r}) - \hat{x}(s\bar{r}|s)] \quad (21)$$

$$\hat{P}_e(s) = P(s|s) + J(s)[P_e(s\bar{r}) - P(s\bar{r}|s)] \quad (22)$$

where

$$J(s) = P(s|s)F^T(s)P^{-1}(s\bar{r}|s) \quad (23)$$

3.3 Unified Attention Estimation

In the prior two steps, the knowledge from the audio stream is distributed into visual attention sequences. If a salient feature element is coherent with the unified attention process, the covariance of that feature will be small. At each node, we rank all features according the value of their covariance and estimate unified attention as a mean of top three visual attention (Eq.24). The algorithm for highlight allocation is a tree search process and is carried out at multi-scales.

$$A_i(s) = \frac{1}{N}H_i x_i(s) \quad (24)$$

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.

4. 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 included, three from World Cup 2002, Brazil vs German(final), Brazil vs Turkey(semi final), and German 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 are recorded from BBC and ITV in MPEG-1 PAL format with the visual resolution at 352×288 while audio at 224kbit/s. To set up ground truth, we collect game records from the FIFA official website to define the list of content-based game events, while the highlight collection comes from BBC Sports website and FIFA highlight videos. Note that the temporal resolution of official game records is of minute, and there is a start point misalignment between broadcasting and real games. A 30 sec allowance is set to match official records and experiment results. Each of games are 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.

We manually labelled all content-based events in the second half of the final game in World Cup 2002, according to the FIFA game record. Table 2 compares the difference between event attention and average attention intensity under multiple resolutions. Note that the maximum of average event attention appears at the temporal resolution of 76 sec, while the maximum of signal noise ratio at the resolution of 5 min(304sec). The fact shows that the observation window with 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 takes place about every 5 minutes and the replay duration is about 1 minute.

Resolution	1.2	38	76	152	304	600
event mean*	6.628	6.628	**6.807**	6.743	6.671	6.563
event mean	6.832	6.874	**7.522**	7.271	7.113	7.110
average	4.020	3.974	4.122	3.532	3.432	**3.342**

Table 2: Attention intensity under different resolution in 2^{nd} half in Brazil vs German, World Cup 2002(* without visual feature rank)

The precision of goal detection is a popular evaluation of highlight detection. Table 3 concludes the number of goal events found in the top five of the attention list and their rank. The average attention value of goal events is show in Table 4.

	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	1,3
Ita-Fra I	2	2	1,2,4*
Ita-Fra II	0	-	-

Table 3: Performance of Goal Detection (*goal events are replayed for several times)

5. CONCLUSION AND DISCUSSION

In the paper, we present an affective fusion algorithm for sports highlight detection. The system offers an attention credit as a prior for content-based game event filtering and identification. Our major contribution is the introduction

	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 4: Goal and general contents attention

of the multi-resolution fusion algorithm, which solves the problem of modality asynchronism and offers an estimation of the best temporal resolution for multiple applications, e.g. event segmentation. The performance of self-information measurement is interesting. In experiments, the overall information gain of different videos are similar, though the gain distribution varies. This fact indicates a possible statistics model for affect-content analysis in the domain of sports videos. However, different from the popular normalisation operator[10][19], self-information sometimes offers a high credit to the low intensity of stimulus. For example, silent clips in audio come with a highlight self-information value, because they are rare in videos, which is against the psychobiological assumption on attention. But, in some cases, this character is welcome, e.g. replay, when the director will switch off audio input.

The regression of attention over content-based events is a statistics pathway for event-based video analysis. It discovers interesting clips and then guess their contents. This approach is similar to the text retrieval, where document information is concluded firstly from the statistics of text terms and then from the content and semantics. Furthermore, we will try to build an attention model for content-based video analysis with more labelled data.

6. 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.

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] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization.

[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] Y. Kang, J. Lim, M. Kankanhalli, C.-S. Xu, and Q. Tian. Goal detection in soccer video using audio/visual keywords. *ICIP2004*, 3:1629 – 1632, Oct 2004.

[6] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. Hmm based structuring of tennis videos using visual and audio cues. In *ICME*, pages 309–312, 2003.

[7] 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.

[8] M. Lesser and D. Murray. Mind as a dynamical system: Implications for autism. In *Durham conference Psychobiology of autism: current research and practice*, 1998.

[9] M. S. Lew. *Principles of Visual Information Retrieval*. Springer, 1996.

[10] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *ACM Multimedia 02*, 2002.

[11] S. Moncrieff, S.Venkatesh, and C. Dorai. Horror film genre typing and scene labelling via audio analysis. In *ICME*, pages 193–196, 2003.

[12] C. Osgood, G.J.Suci, and P.H.Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

[13] R. Ren and J. Jose. Football video segmentation based on video production strategy. In *ECIR 2005*, 2005.

[14] R. Ren and J. Jose. Attention guided football video recommendation system on mobile device. In *MobiMedia 2006*, 2006.

[15] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball program. In *ACM Multimedia*, page 105115, 2000.

[16] A. Salway and M. Graham. Extracting information about emotions in films. In *ACM Multimedia*, page 299302, 2003.

[17] A. M. Treisman and N. G.Kanwisher. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8:218–226, 1988.

[18] N. Vasconcelos and A. Lippman. Bayesian video shot segmentation. In *NIPS*, pages 1009–1015, 2000.

[19] H. L. Wang and L. F. Cheong. Affective understanding in film. *IEEE Trans. Circuits Syst. Video Techn.*, 16(6):689–704, 2006.

[20] 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.

[21] A. Willsky. Multiresolution markov models for signal and image processing. In *Proceedings of the IEEE 90 (8) (2002) 1396-1458. 33*, 2002.