General Highlight Detection In Sport Videos

Reede Ren and Joemon M. Jose

Department of Computing Science
University of Glasgow
17 Lilybank Gardens, Glasgow, UK
{reede,jj}@dcs.gla.ac.uk

Abstract. *Attention* is a psychological measurement of human reflection against stimulus. We propose a general framework of highlight detection by comparing *attention* intensity during the watching of sports videos. Three steps are involved: adaptive selection on salient features, unified *attention* estimation and highlight identification. Adaptive selection computes feature correlation to decide an optimal set of salient features. Unified estimation combines these features by the technique of multi-resolution auto-regressive (MAR) and thus creates a temporal curve of *attention* intensity. We rank the intensity of *attention* to discriminate boundaries of highlights. Such a framework alleviates semantic uncertainty around sport highlights and leads to an efficient and effective highlight detection. The advantages are as follows: (1) the capability of using data at coarse temporal resolutions; (2) the robustness against noise caused by modality asynchronism, perception uncertainty and feature mismatch; (3) the employment of Markovian constrains on content presentation, and (4) multi-resolution estimation on *attention* intensity, which enables the precise allocation of event boundaries.

Key words:highlight detection, attention computation, sports video analysis

1 Introduction

As one of the most popular video genres in the video-on-demand service, sports video has shown its commercial value in the media industry [11]. Many value-add services, *e.g.* adaptive video skimming and content sensitive video encoding, are proposed to improve service quality. Therefore, sports highlight detection attracts great interests from both industry and academics [6], as a key function to above services.

A highlight is "something (as an event or detail) that is of major significance or special interest" (Merriam-Webster Online Dictionary 2008). This linguistic definition shows that highlights are contingent on sports contents as well as video context. A predefined collection of video events could hardly cover all possible highlights. On the other hand, a highlight may be an interesting detail rather than an event. Therefore, event-based approaches are ineffective for the identification of sports highlights. Given that all highlights incur strong reflections

among viewers, *i.e.* happiness or surprise, *attention*, the psychological measurement of human reflection, is proposed in [3] [10] as an efficient method to identify general highlights. Moreover, the estimation of *attention* intensity concerns few sports semantics. This indicates that attention-based approaches avoid semantic uncertainty caused by various video contents.

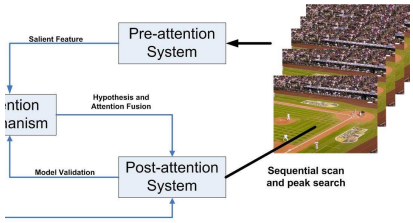

Fig. 1. Attention Perception System

An *attention* perception system [13] consists of three components (Figure 1): pre-attentive, attention combination and post-attentive system. The pre-attentive system is also called as feature-attention modelling [10], which calculates stimulus strength as well as extracts salient features. However, such an extraction of salient features is usually incomplete [8]. These features may be ineffective for the discrimination of actual attention peaks [13], because of strong perceptual noise and variant stimulus types. *Attention* combination simulates the mechanism of *attention* perception in human minds, which fuses stimuli from vision, auditory and text understanding to create a unified *attention*. The post-attentive system justifies conclusions got in the prior steps by domain knowledge.

In our mind, an *attention*-based system should answer the following research questions: (1) how to identify a set of effective salient features in a given sports video; (2) how to combine noisy salient features robustly; (3) how to estimate an unified *attention* to reflect interesting contents; and (4) how to analyse the unified attention to allocate highlights. We here take video segments which incur the strongest reflections, as highlights [10] [12]. This provides a post-attentive explanation to question 4.

In this paper, we model the perception process of sports video watching to estimate the intensity of viewer reflection. This leads to two improvements in comparison with prior works [10][3][12]: adaptive selection on salient features and the framework of attention fusion, *i.e.* multi-resolution autoregressive (MAR). Adaptive selection extends the pre-attentive system, which identifies the most effective

salient features to improve the robustness of *attention* estimation. The technique of MAR is equivalent to a Markovian process on graph [15], the general temporal model of video content presentation [16]. Such a combination framework therefore imposes the Markovian constraint on video presentation to *attention* perception. This is a significant improvement on *attention* based video analysis. Moreover, a video contains multiple modalities, *i.e.* audio and visual streams. These modalities are independent representation of video contents at different temporal resolutions. By sampling and matching these modalities gradually, the MAR alleviates the problem of modality over-sampling and media asynchronism. This results in a precise and robust estimation of *attention* intensity.

The paper is structured as follows. Section 2 provides a brief overview on sports highlight detection, especially *attention*-based approaches. A twofold model of *attention* perception is proposed in Section 3 to simulate the process of sports video watching. Based on this model, Section 4 presents the selection algorithm on salient features. Section 5 describes the MAR framework for *attention* fusion. The experiments on real football game videos are stated in Section 6. Conclusion is found in Section 7.

2 Related Work

The literature of highlight detection could be roughly categorised into two groups, event-based and attention-based. Event-based approaches regard some specific events as so-called highlights, although such an event collection can hardly cover all possible aspects. The detection of sports highlights is therefore specified into a sequence of event discrimination, *e.g.* goal, corner and free-kick [2]. Various Markov models have been proposed to identify these events. Lenardi *et al.* [7] model shot transmissions around game events with a controlled Markov chain. The authors take embedded audio energy as the controlling token and rank highlight candidates by the loudness. Their experimental results are evaluated by the coverage of goal events among the top 5 of candidate lists. Kang *et al.* [5] propose a bidirectional Markov model to alleviate the problem of modality asynchronism. The authors identify excited speech whilst search video objects such as goalposts in nearby shots. Xu *et al.* [17] create a group of middle-level content modalities by coupling low-level features, such as dominant colour and caption text. By these content modality, the authors build a hierarchical hidden Markov model for event detection.

Attention-based approach is an exploration from computing psychology to content analysis [10]. This methodology is relatively new in sports video analysis [3]. Ma *et al.* [10] employ a series of psychological models on pre-attention, *i.e.* motion attention model, static attention model and audio salient model, to describe the process of video watching. A set of temporal curves are created to display feature related *attention* such as motion attention, and are linearly combined to estimate the joint intensity of *"viewer attention"*. However, this massive feature

extraction introduces too much noise and challenges the later attention combination. With the increase of feature number, noise overwhelms actual attention peaks and thus fails highlight detection. Hanjalic *et al.* [3] carefully choose three features to estimate the intensity of viewer reflection, including block motion vector, shot cut density and audio energy. The authors furthermore employ a 1-minute long low-pass Kaiser window filter to smooth these features as well as enhance the signal noise rate (SNR) of feature related *attention* [4]. A robust method of attention combination is also developed. A sliding window is introduced to limit the range of observation and the authors count *attention* peaks inside to guess the appearance probability of a highlight. However, the sliding window makes constant the temporal resolution of event detection. It is difficult to allocate event boundaries precisely as well as segment video events. Such an ability is essential in many applications, *e.g.* adaptive video encoding.

3 Temporal Attention Perception Modelling

In this section, we address the temporal modelling of *attention* perception and show how to develop a MAR framework to simulate such a process.

Attention perception is a discreet temporal process in psychology: "*people notice something at this moment and other things later*". A general stimulus-attention model is proposed in [8], which consists of two differential equations to quantify the relationship among *interest*, *attention* and *reflection* of a human being in an unknown environment. Some complex issues are considered in this model, *i.e.* cultural background, personal experience and possible activity. However, the context of sports video watching indicates a predictable viewer behaviour and leads to a direct model of *attention* perception.

A sports video records a combination of reflections. There are three major reaction roles, spectators, commentators and video directors. These observers watch the game at the same time. They understand game content and keep video context. In psychological terminology, these observers are ready to accept stimulus. Their reflection therefore follows the stimulus-reflection model [8] (Equation 1).

$$A(t) = pX(t-\tau) + \alpha + w(t) \tag{1}$$

where $A(t)$ denotes attention intensity at the moment t; $X(t-\tau)$ refers to stimulus strength with a reflection delay τ; α stands for the threshold triggering a response; p is a reflection parameter and $w(t)$ is perceptual noise. τ is a constant related to the modality, *e.g.* 0.384 sec for vision [14]. Individual understandings from these observers affect video viewer's feeling. Directors watch camera videos, decide shot styles such as field view and close-up, and insert video editing effects, *i.e.* replay, to present the story. Spectators and commentators dominate audio tracks. As a group, stadium audience cheer at exciting moments and remain relatively silent in the rest of a game. They attract video viewers by loud plaudits. Commentator's behaviour is a little complex. On one hand, commentators

reiterate game contents and their professional jargons are detected for events annotation. On the other hand, commentators are ad-hoc spectators. Hence, the *attention* model for a viewer to watch a sports video is a combination of above observer reflections (Equation 4).

$$A_{viewer}(t) = A_{director} + aA_{spectator} + bA_{commentator} \tag{2}$$

$$= \sum_{x \in X} ((1 + a + b)p_x x(t - \tau_x) + \alpha_x + w_x(t)) \tag{3}$$

$$= \sum_{x \in X} (kx(t - \tau_x) + \alpha_x + w_x(t)) \tag{4}$$

where a,b are combination parameter for spectator and commentator reflections, respectively; x denotes a stimulus from the collection of salient features X; α refers to the response threshold; w(t) is perceptual noise and $k = (1 + a + b)p_x$.

Furthermore, a sports video is a smooth Markovian process on both time and content presentation [17]. Game contents can be described by a directed semantic graph $G = (\nu, \epsilon)$, in which vertices set ν denotes game semantics and edge set ϵ links pairs of vertices $(s, t), s, t \in \nu$, a possible event sequence. A game is therefore presented by a discrete-time Markov process $x(.)$ on G with finite states. Such a process on graph can always be extended to a state chain without loops by intuitive labelling and dynamic programming. Moreover, Hammersly-Clifford theorem [1] proves the equality between a Markov chain and an auto-regressive (AR) by comparing clique potential. In addition, a video is always with a definite start point (the root of a graph). Such a Markov process can be expressed by a first order AR model,

$$x[n] = a[n - 1]x[n - 1] + w[n] \tag{5}$$

where $w[n]$ is a set of independent Gaussian noise, $x[0]$ is the root and $x[1 \dots n]$ are a sequence of Markov states. Given that *attention* intensity reflects the importance of a game content, Equation 5 is transformed as follows.

$$A[n] = k[n - 1]A[n - 1] + w[n] \tag{6}$$

where $A[n]$ is *attention* intensity of a viewer at an event n, $k(n - 1)$ is the parameter for reflection combination, which is also regarded as the impact of a past event $n - 1$. This indicates that some efficient methods of signal processing, such as moving average, can be used for the analysis of *attention* perception.

In summary, we build a twofold model of *attention* perception for sports video watching: Equation 4 describes the transient *attention* reflection against a stimulus; and Equation 6 denotes an accumulation of *attention* in a long period.

4 Adaptive Salient Feature Selection

Here we propose an adaptive selection on possible salient features to improve system robustness. A large collection of salient features are listed in Section 4.1

for *attention* estimation in sports videos. Section 4.2 presents the algorithm for feature selection, which decides a subset of salient features for later *attention* combination, according to given video data. In another words, the set of salient features for *attention* estimation is adaptive to videos.

4.1 Salient Feature

As the theory of psycho-biology asserts, temporal variation, stimuli strength and spatial contrast are major facts in visual attention [9]. Video directors mainly rely on fast shot variation to excite viewers, such as replay and quick switching camera viewpoints [18]. Loud and greatly varying noise from spectators always catches notice. Moreover, the watching of sports videos requires rich domain knowledge. The semantics of video objects and audio key words plays an important role in *attention* computation. For example, a goalpost attracts great interest as the forecast of a goal [2].

Table.1 lists most salient features reported in literature [18][10][2][3]. Psychological explanations and possible affection on *attention* intensity are also annotated. In addition, related algorithms for feature extraction are found in [12].

4.2 Feature Selection

Equation 6 shows that effective salient features should reach local extremes at important game events. Signal correlation (Equation 7) therefore becomes an effectiveness measurement for salient features in *attention* estimation, if perceptual noise is a Gaussian white noise with zero mean.

$$r_{XY} = \| \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y} \| \tag{7}$$

where X,Y are two salient features with n samples; $\bar{X}, \bar{Y}$ denote the average and s_X, s_Y refer to standard deviations of X and Y, respectively. $r_{XY} \in (0, 1]$ and $r_{XY} = 1$ iff the strength of X and Y are of the same linear direction. However, this measurement is not so robust in computation. This is because: (1) salient signals, e.g. shot frequency and audio energy, are of various sequence length due to the difference in sampling rate [1]; (2) random perceptual delay mismatches salient signals (Equation 4).

There are two facts in video watching which can alleviate the above problem: (1) the duration of most events is less than 5 minutes [2]; (2) the average reflection delay is less than 15 sec for ready viewers [18]. We therefore use a 5-minute moving average to smooth salient signals. This could reduce perceptual noise effectively. We collect maximum and minimum every five minutes and compute the correlation between respective maximum/minimum sequences. We assume

[1] We compute shot frequency every 50 sec and audio energy every 0.3 sec.

Salient Feature	Psychological Facts	Qualitative Affection on Attention
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 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	*
MFCC and delta	sound variation	*
spectral roll-off	sound variation	+
spectral centroid	loudness	+
spectral flux	loudness	+
octave energy	loudness	+
music scale	sound variation	*
audio type proportion	valance classification	*
scene affect vector	valance classification	*

Table 1. Attention Related Salient Feature, + stands for the positive qualitative relation between feature strength and *attention*, where the feature induces an increase of *attention* intensity. − denotes negative qualitative relation, which decreases *attention* intensity, and $*$ for unsure.

the correlation distribution is a Gaussian with the mean of one (Equation 7). Therefore, a score which suggests feature effectiveness, is decided as the probability of a correlation value belonging to the given Gaussian distribution. Salient signals with the largest N scores are kept for *attention* combination.

5 Multi-resolution Auto-regressive Fusion

In this section, we address the problem of *attention* combination. Equation 6 shows that attention perception can be described by an autoregressive (AR) process. Given that salient features are sampled at different temporal resolutions, it is reasonable to employ a MAR for *attention* combination.

A MAR is a scale-recursive linear dynamic model, which simulates a random

process by a set of AR models on multiple scales. A general two-pass parameter estimation algorithm is proposed in [15], which includes a fine-to-coarse filtering followed by a coarse-to-fine smoothing. The fine-to-coarse step is a three-step recursion of measurement updating, fine-to-coarse prediction and information fusion when moving to a coarse resolution. The coarse-to-fine step combines smoothed estimations and covariances at coarse resolutions with the statistics computed in the first fine-to-coarse sweep. We extend this general algorithm for *attention* combination. Different from the prior work [12], we start from the salient signal with the finest temporal resolution, e.g. zoom depth and game pitch ratio; and gradually impose other salient features as an updated measurement in the merge step. The details of our algorithm are presented as follows.

5.1 Fine-to-coarse Filtering

Let $\hat{x}(s|s)$ be the optimal estimation of attention intensity $x(s)$ at a node s, together with $P(s|s)$, the error covariance.

Initialisation Start with salient features at the finest temporal resolution. For each leaf s, the estimation of $\hat{x}(s|s-)$ and the covariance $P(s|s-)$ from the sub-tree are as follows.

$$\hat{x}(s|s-) = 0 \tag{8}$$
$$P(s|s-) = P_x(s) \tag{9}$$

Measure Updating is identical to the analogous step in a Kalman filter, although only estimations are changed here. If there is no measure available, go to sub-tree fusion directly.

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

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,

$$K(s) = P(s|s-)H^T V^{-1}(s) \tag{13}$$
$$P(s|s) = [I - K(s)H]P(s|s-) \tag{14}$$

Repeat the above steps until $\|P(s|s)\|$ is smaller than a given threshold.

Sub-tree fusion merges estimations from immediate children at s. 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.

$$\hat{x}(s|s-) = P(s|s-)\sum_{i=1}^{K_s} P^{-1}(s|sa_i)\hat{x}(s|sa_i) \tag{15}$$

$$P^{-1}(s|s-) = P_x^{-1}(s) + \sum_{i=1}^{K_s}[P^{-1}(s|sa_i) - P_x^{-1}(s)] \tag{16}$$

Error covariance matrix $P(s|sa_i)$ indicates the distribution of *attention* weight on salient features at the given resolution. This matrix is kept for the later coarse-to-fine smoothing. To avoid noise incurred by signal interpolation [12], we regard every layer in the MAR tree as an individual Markov process and limit the scope of recursive smoothing.

Fine-to-Coarse Prediction estimates $\hat{x}(s|sa_i)$ and error covariance matrix $P(s|sa_i)$ of the parent s from its children sa_i.

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

$$P(s|sa_i) = F(sa_i)P(sa_i|sa_i)F^T(sa_i) + U(sa_i) \tag{18}$$

where

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

$$U(s) = P_x(s\overline{r}) - F(s)A(s)P_x(s\overline{r}) \tag{20}$$

5.2 Coarse-to-Fine Smoothing

When the fine-to-coarse filtering reaches a predefined coarse resolution or the root, the MAR has experienced all possible reflection delays and completed parameter estimation. The error covariance and optimised estimations are calculated at all nodes. Then the coarse-to-fine smoothing spreads optimal estimations and covariance from parents $s\overline{r}$ and improves the estimation at finer resolutions s.

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

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

where

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

6 Experiment

The evaluation collection includes six entire game videos in MPEG-1 format from FIFA World Cup 2002, World Cup 2006, and UEFA Champions League

2006: 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. We gathered game records from the FIFA and BBC Sports website as the ground truth of video event list. All videos are divided into halves, *e.g.* Brazil-Germany I for the first half of the final game in World Cup 2002. The middle break is removed but we keep other broadcasting aspects such as player entering, triumph, and coach information board.

We use the ratio of *attention* intensity on events and other general video clips (Equation 24) to evaluate system robustness. A high ratio is preferred.

$$R_{attention} = \frac{E(A_{events})}{E(A)} \sim \frac{E(A_{goal})}{E(A)} \tag{24}$$

where E is the expectation function, and A_{events}, A_{goal}, A denote estimated attention intensity on events, goals and the entire game, respectively. Table 2 compares the average of attention intensity over different temporal resolutions. Many interesting conclusions are reached: (1) the maximum of average attention appears at the temporal resolution of 76 sec; (2) the delta maximum is at the resolution of about 5 min (304sec). The observation window with 5-minute width is the best choice for event detection whilst 1-minute for event segmentation.

Temporal Resolution (sec)	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 2. Attention intensity under different resolution in the $2^n d$ half in Brazil vs Germany, World Cup 2002

Feature set {average block motion, shot cut density, base band audio energy}[4] is used to evaluate approaches of attention combination. We take linear combination [10] as baseline. Table 3 presents six approaches: Linear I directly adds up normalised salient features [10]; Linear II linearly combines normalised salient features but with the weight from the fine-to-coarse filtering; MAR I uses the self-information [12]; MAR II works on 1-minute resolution; Linear III and MAR III are similar to Linear I and MAR II respectively, but employ a set of seven salient features from adaptive feature selection. The MAR outperforms linear combination in most cases. Adaptive selection is effective to improve the ratio of average attention intensity (Equation 24). The performance of Linear III is worse than Linear I, because linear combination cannot afford perceptual noise.

	Linear I	Linear II	MAR I	MAR II	Linear III	MAR III
Ger-Bra II	1.522	1.874	1.802	1.997	1.333	**2.141**
Bra-Tur II	1.671	1.944	1.972	2.187	1.461	**2.245**
Ger-Kor II	1.142	1.326	1.411	1.563	1.274	**1.665**
Mil-Bar II	1.377	1.700	1.741	2.043	1.276	**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**	1.151	1.732
Ita-Fra I	1.302	1.377	1.420	**1.723**	1.044	1.658

Table 3. Attention ratio(goals vs. general contents) under different combination algorithms

The MAR based approach achieved 100% precision in the detection of goal events. As an interesting case study, we compare professionally marked highlight lists from BBC Sports and FIFA website in Table 4 for Italy vs. France, World Cup 2006. *Attention*-based detection covers most of manually selected highlights.

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 4. Game highlights and attention Rank in France vs Italy (I,II game halve)

7 Conclusion

Attention-based approach is an application of computing psychology in video analysis. Such an approach is efficient in the identification of sports highlights. We propose an abstract model of *attention* perception to simulate the process of video watching, which leads to an adaptive selection on salient features and a combination framework of multi-resolution autoregressive. Adaptive selection exploits the characters of temporal accumulation on *attention* perception. A measurement of signal correlation is therefore suggested at a coarse temporal resolution to evaluate feature effectiveness. The MAR framework is based on the multi-resolution nature of *attention* perception. The advantages of the MAR framework are as follows: (1) the employment of data at coarse temporal resolutions, which can hardly be used before in content-based video analysis; (2) the multi-resolution framework of data sampling and matching, which alleviates media asynchronism; (3) the extensibility and robustness on a large feature space.

8 Acknowledgement

The research leading to this paper was supported by European Commission under contracts FP6-045032 (Semedia).

References

1. J. Besag. Spatial interaction and statistical analysis of lattice system. *Journal of Royal Statistical Society*, 36(2):192–236, 1974.
2. A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 12(7):796–807, July 2003.
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. A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan. Browsing sports video: trends in sports-related indexing and retrieval work. *Signal Processing Magazine, IEEE*, 23(2):47–58, March 2006.
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. G. News. 3g football best mobile service, Jan 2005.
12. R. Ren, J. Jose, and Y. He. Affective sports highlight detection. In *the 15th European Signal Processing Conference*, pages 728–732, Poznan, Poland, Sept. 2007.
13. H. D. Tagare, K. Toyama, and J. G. Wang. A maximum-likelihood strategy for directing attention during visual search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(5):490–500, May 2001.
14. A. M. Treisman and N. G.Kanwisher. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8:218–226, 1988.
15. A. Willsky. Multiresolution markov models for signal and image processing. In *Proceedings of the IEEE 90 (8) (2002) 1396-1458. 33*, 2002.
16. C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan. Live sports event detection based on broadcast video and web-casting text. In *ACM Multimedia 2006*, 2006.
17. G. Xu, Y. Ma, H. Zhang, and S. Yang. An hmm-based framework for video semantic analysis. *IEEE Trans on Circuits and System for Video Technology*, 15:1422–1433, Nov 2005.
18. H. Zettl. *Sight, Sound, Motion: Applied Media Aesthetics.* Wadsworth, Belmont CA, 1990.