

Temporal Attention Fusion For Sports Event Detection

Reede Ren, Yue Feng, Joemon Jose

MIR Group, University of Glasgow,
17 Lilybank Gardens, Glasgow, UK, G12 8QQ
{reede,yuefeng,jj}@dcs.gla.ac.uk

Keywords:attention fusion, event detection, video retrieval

Abstract

The employment of psychological measurement, *attention*, alleviates the semantic uncertainty around video events and leads to an effective general event detection approach. This paper proposes a multi-resolution autoregressive framework to estimate a unified *attention* curve from multi-modality salient features at different temporal resolutions. The highlights of this work are: (1) the capability of using data at very coarse temporal resolutions, e.g. three minutes; (2) the robustness against noise caused by modality asynchronism and feature collection size; and (3) the utilisation of Markovian temporal constrains on content presentation. This approach achieved 100% goal event coverage in the football video collection of the FIFA World Cup 2002, 2006 and UEFA League 2006.

1 Introduction

Event detection is one of the most important research topics in video retrieval [7][1][3]. As stream-like data, video documents lack clear content structures, e.g. paragraph, section and chapter. This absence results in many problems during content management and video indexing. It is difficult to create an index directly from a video, because a randomly selected video clip is semantically incomplete and can hardly explain itself. A content sensitive segmentation or event detection is therefore essential from many aspects. Such a technique should (1) allocate meaningful video intervals or video *sentences*; (2) decide possible content topics so as to facilitate future queries; and (3) remove trivial or duplicated video clips from video index. However, there is no clear definition what a video event is, given the variety of video contents. This semantic ambiguity indicates the difficulty during developing a general event detection approach by modelling video contents.

A video event of importance attracts attention from views. This indicates that *attention*, the psychological measurement of attractiveness, can be used to find video events without involving too many semantic details [7]. A new event detection approach, attention analysis, is therefore introduced, which treats a video stream as a sequence of multiple modality stimuli and regards video segments which incur the strongest reflections as video events.

An attention perception system [11] consists of three components (Figure 1): pre-attentive, attention combination and post-attentive system. The pre-attentive system quickly calculates stimulus strength or extracts salient features. Additionally, such a process is also mentioned as feature-attention modelling [7]. However, we do not assume this extraction is complete or even that these features can distinguish possible attention peaks from the background noise by themselves. This is because of strong noise in perception. The attention combination is to simulate attention mechanism in human minds. In this stage, salient features are fused as stimuli from vision, auditory and text understanding, to estimate a unified attention reflection. The post-attentive system justifies conclusions of attention combination by prior knowledge of video contents and thus completes event segmentation.

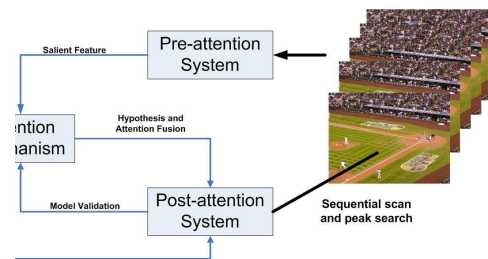


Figure 1: Attention Mechanism

This paper improves a multi-resolution autoregressive framework (MAR) for attention combination [10]. As a common hypothesis, a video document is a smooth Markovian process on both time and content presentation. Attention perception is an observation of this process and therefore is a smooth Markovian random process, too. A multi-resolution autoregressive framework is equivalent to a Markovian process on graph [15]. This indicates that our framework can meet such a Markovian temporal constraint on attention perception well. Moreover, a video document contains multi-resolution data, i.e. audio and visual streams. These modalities are independent observations or stimulus-reflection processes from different temporal and content resolutions. The new MAR framework samples and matches these modalities gradually from multiple resolutions, which avoids the problem of over-sampling as well as alleviates media asynchronism.

The rest of paper is organised as follows. A brief literature review in the domain of attention analysis for sports highlight detection is offered in Section 2. Section 3 states our pre-attentive system, including audio-visual salient features and the self-entropy project function, which maps signal values to attention intensity. The MAR fusion model is presented in Section 4. Experimental results and conclusion will be found in Section 5 and Section 6, respectively.

2 Related Work

As an exploration from computing psychology to content analysis, attention analysis is a relatively new approach in event detection. There are very few works in this field. Ma et al. [7] isolated media feature influences on perception by a series of feature-attention models, i.e. motion attention model, static attention model, and audio salient model. The authors linearly combined these feature-based attention curves to estimate the intensity of *viewer attention*. But this isolation introduces too much noise and leads to a fragile fusion step. With increasing feature counts, noise usually overwhelms *real* attention peaks in experiments. Hanjalic et al. [3] hence selected a small feature set, including block motion vector, shot cut density and audio energy. The authors counted the peak number of feature-attention curves inside a predefined floating window to estimate the probability of a game highlight. Apparently, such an approach strongly relies on signal noise ratio (SNR) of selected attention signals and the width of observation window. Later, these authors applied an adaptive filter and a 1-minute long low-pass Kaiser window filter to improve the SNR in [4]. However, neither of these works addresses the problem of signal multi-resolution and temporal constrains between different modalities.

3 Pre-Attentive Computing

Temporal variation, spatial contrast and stimuli strength are major factors attracting attention [6]. Moreover, the watching of sports videos is not a plain stimulus-reaction process but a content understanding with rich domain knowledge. We select seven audio and visual salient features: three from audio, the base band energy [5], speech pitch frequency [14], and the first order derivatives of Mel-frequency Cepstral Coefficients (MFCC) [2]. The left four visual features are visual harmony, shot boundary frequency [12], shot zoom depth [9] and game pitch ratio [9]. These salient features are supposed to reflect the strength of modality stimuli. For example, audio base band energy measures the loudness of background noise. The higher is audio based band energy, the louder the audio stream is, and the stronger audio stimulus would be [7] [5].

Visual harmony (Vh) is proposed to measure static spatial contrast in a visual frame. Since most of experimental video

documents are of MPEG-1 format, we use block mean hue (Equation 1) and block hue covariance (Equation 2) to compute visual harmony (Equation 3) for $n \times n$ image block with the centre at (i, j) ,

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

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

where C is the pixel colour. We use an 256-bin histogram to count the block covariance distribution. Therefore, visual harmony (Vh) is,

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

where P_n is the portion of bin n over all histogram.

In [3], [4], and [7], salient features are normalised as *attention* intensity. However, we suppose it is physically unjustified to combine different modality features, i.e. audio and colour, directly. In psychology experiments [8], both weak and strong stimuli can incur strong reflections if and only if such a stimulus change is big enough. Moreover, there are many psychobiological hypotheses about stimulus-reflection functions, such as linear and log-like reflection [8]. The normalisation of salient features is a simplification of linear reflection which is not so plausible in auditory perception [6]. To avoid these dilemmas, we favour the explanation from information theory: (1) *attention* is the ability of consuming information; and (2) the pan-out speed of message decides the distribution of *attention*. Therefore, the overall attention is,

$$I_{attention} = \vec{A} \vec{E} \quad (4)$$

where \vec{A} is the *attention* distribution vector over all modalities; \vec{E} is stimulus intensity vector, each element of which stands for the information provided by a given modality. Self-information (Equation 5) is used to assume modality contribution, since it is the amount of information which a certain event contributes to the overall knowledge.

$$Entropy = -\log_2(P_i) \quad (5)$$

where P_i is the appearance probability of a feature at the given value i . Hence self-information can be easily computed by feature histogram. Moreover, some features have a known prior distribution which can improve the overall self entropy estimation. For instance, shot boundary frequency follows a Weibull distribution [12]. We use the EM algorithm to find the best fit and thus compute the self information entropy of these features. Nevertheless, the self-information entropy can be accumulated with time and by different modality stimuli. Figure 2 compares the attention intensity distribution of audio energy by the normalisation project function [7] and self-information in the final game of the FIFA World Cup 2006, France vs. Italy.

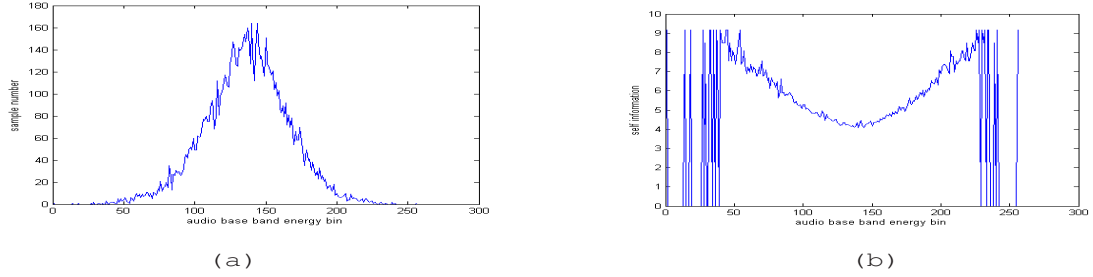


Figure 2: (a) 256-bin Histogram of Normalised Audio Based Band Energy (b) 256-bin Self-information Histogram of Audio Base Band Energy

4 Audio-visual MAR Fusion Model

The complexity of modality fusion rises not only from media asynchronism and the difference in event discrimination, but also from the contents that a video stream iterates. It is difficult to map audio and visual segments onto the semantics. Attention analysis introduces a unified psychological measurement as the middle layer between low level features and video semantics, and therefore provides a bridge across this semantic gap. We propose a multi-modality fusion algorithm based on the multi-resolution autoregressive framework (MAR). Willsky et al. [15] argued that the MAR is equivalent to a Markov process on graph. Each node in the MAR tree denotes an optimised Markov state. The construction of a MAR tree will experience all possible Markov states from a given time point, e.g. the start of game, without pre-defined Markov states. Model parameters are automatically learnt by a three-step recursive algorithm without complex training processes. Moreover, the estimation of unified *attention* is generated by a large set of multi-modality signals from multiple resolutions. Such an estimation hence is robust against noise. Note that the employment of data at coarse resolutions, i.e. the shot frequency over five minutes, is a significant improvement in this framework. This is not only because prior work [7] [3] [4] in attention analysis can hardly use them, but also because these coarse resolution data are able to increase the precision of event detection significantly.

Since the attention distribution vector \vec{A} (Equation 4) is unknown in most perception cases [8], it is unrealistic to directly combine attention intensities from different modalities. Note that audio and visual streams in a video can be regarded as two independent observations. We therefore use one attention curve, i.e. audio, as a measurement to the others and thus combine all attention curves by the MAR model. This hypothesis can be described as follows. Denote the set of resolution by $\mathbb{R} = \{1, \dots, R\}$, with $r = R$ being the finest resolution. We set the finest combination resolution as 1.4 times of the longest shot duration, which will be about 50 sec in experiments. Node N at scale r is $N_n^{(r)} = \{1 : 2^r\}$. Let $x(s)$ be the observation vector of visual attention on the node

s , $y(s)$ for the audio, the combination process can be described as,

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

where H is set as a vector of $\{1, \dots, 1\}^1$; N is the normalisation parameter; and $v(s)$ denotes a Gaussian noise on the tree. We use the binary tree in this application. Therefore, the projection from finer resolution to coarse resolution will be

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

where $x(s|s-)$ is the sub-tree under node (s), $w(s)$ is the Gaussian noise. For convenience, we here use the binary tree.

We follow the general algorithm in [15] to estimate MAR parameters. Two steps are included, fine-to-coarse filtering and coarse-to-fine smoothing. The fine-to-coarse step is a generalisation of the Kalman filter for tree models, which contains a three-step recursion: prediction, measure updating and merge when moving up to a coarse resolution. Different from the prior work [10], we impose modality features gradually, which will be introduced as a new measurement in the merge step. The coarse-to-fine step smoothes these estimations at fine resolutions by spreading information gained at coarse resolutions. The details of these algorithms will be found in the following sections.

4.1 Fine-to-coarse Filtering

Let $\hat{x}(s|s)$ be the optimal estimation of attention intensity $x(s)$ 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.

4.1.1 Initialisation

Start 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

¹This is a widely accepted hypothesis in perception research that all modalities are of the same importance [3] [4] and [13].

the sub-tree are

$$\hat{x}(s|s-) = 0 \quad (8)$$

$$P(s|s-) = P_x(s) \quad (9)$$

4.1.2 Measure Updating

The measurement updating is identical to the analogous equations in Kalman filter, although here it changes observation values from other modality, i.e. audio.

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

where $v(s)$ is the measurement innovations,

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

which is zero-mean with covariance,

$$V(s) = HP(s|s-)H^T \quad (12)$$

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

$$K(s) = P(s|s-)H^TV^{-1}(s) \quad (13)$$

$$P(s|s) = [I - K(s)H]P(s|s-) \quad (14)$$

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

4.1.3 Sub-tree fusion

The second step is to merge estimations 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,

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

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

The error covariance matrix $P(s|sa_i)$ indicates the attention weight distribution on modality features at a fine resolution. We keep this matrix for the later coarse-to-fine smoothing. To avoid the extra noise caused by the assumption of coarse resolution features such as shot frequency at fine resolutions [10], each layer of the MAR tree is regarded as an individual Markov process. Therefore, an extra round of Kalman filtering is carried out when a new feature is available at a coarse resolution otherwise these covariance and estimations would be calculated by Equation 7.

4.1.4 Fine-to-Coarse Prediction

An one-step fine-to-coarse prediction is proposed to estimate $\hat{x}(s|sa_i)$ and the error covariance for each child of s .

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

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

where

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

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

4.2 Coarse-to-Fine Smoothing

When this fine-to-coarse filtering reaches the root, all possible time delays between modality events are experienced; the error covariance and optimised estimations at all nodes are calculated. Attention from different modalities, i.e. audio and visual, are therefore synchronous if and only if the temporal resolution is coarse enough. In particular, the coarse-to-fine step fuses a node s with its optimal smoothed estimations and covariance from the 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

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

4.3 Unified Attention Estimation

Prior estimation-updating steps, including fine-to-coarse and coarse-to-fine, fuse information from attention sampling at different temporal resolution and modality features. We estimate an unified attention as a mean of all available visual attentions on a given resolution (Equation 24) to avoid resolution difference among modalities.

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

where $x_i(s)$ denotes the visual attention vector at a resolution i , H a unit vector and N the normalisation parameter. The algorithm for event detection hence is a tree search of attention peaks.

5 Experiment

The evaluation set includes six whole MPEG-1 game videos from the collection of 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. Figure 3 and 4 displays the estimated attention curve in both halves of the game Arsenal vs Barcelona, UEFA League 2006.

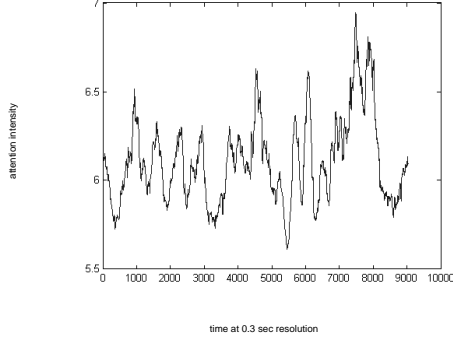


Figure 3: Attention Curve in the First Halve of Arsenal vs. Barcelona

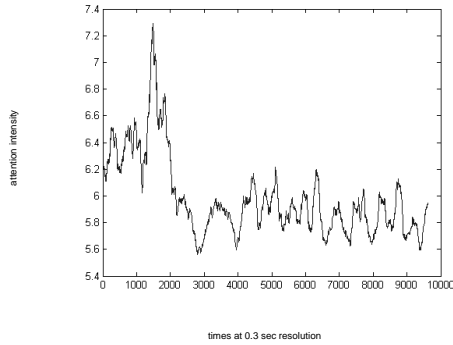


Figure 4: Attention Curve in the Second Halve of Arsenal vs. Barcelona

Game records are collected from the FIFA and BBC Sports website to define the ground truth of video event list. All games are divided into halves, e.g. Brazil-Germany I for the first half of the final game in World Cup 2002 and II for the second half, to remove the middle break but keep other broadcasting aspects, such as player entering, triumph, and coach information boards.

To evaluate the effectiveness of attention analysis, we propose the measurement of *attention* intensity ratio on events and other general video clips (Equation 25). According to the hypothesis of attention analysis, a large ratio reflects the efficiency of proposed measurements and related fusion algorithm.

$$R_{attention} = \frac{E(A_{events})}{E(A)} \sim \frac{E(A_{goal})}{E(A)} \quad (25)$$

where E is the expectation function, and A_{events}, A_{goal}, A denotes estimated attention intensity on events, goals and the

whole game, respectively.

We use the feature set {average block motion, shot cut density, base band audio energy} in [3] and [4] to compare attention fusion frameworks. The linear combination in [7] is taken as the baseline. Table 1 states six fusion approaches: Linear I [7] directly adds up normalised feature values; Linear II linearly combines normalised features with the optimised weights from the fine-to-coarse step of MAR; MAR I compares attention intensity on the leaves (0.3 sec) with the self-information projection function; MAR II on 1-minute resolution; Linear III and MAR III are similar to Linear I and MAR II respectively, but work on our seven feature set. The performance of MAR framework is better than linear combination at all resolution whist MAR III gets the highest attention ratio. Additionally, the performance of Linear III is worse than Linear I. This shows that linear combination is not robust in attention fusion.

As [4], we counted the coverage of goal events in the top five of *attention* peaks (Table 2). It is interesting to find that many replay segments of goal events are detected without using any temporal sequence models. This is a welcome result since replay is a special video editing effect to highlight important game events. Note that such a phenomena has not been reported in prior works [7] [3]. We owe this advantage to the self-information projection function, because the manual editing brings many rare feature values, i.e. the silence in audio and the jump of shot frequency, which will gain a high information increment.

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

6 Conclusion

Attention analysis is an application of computing psychology in content-based video analysis. This approach shows its

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

efficiency in the detection of content-based video highlights and the weighting of content importance. In this paper, we developed a gradual MAR fusion framework to simulate the multi-resolution attention perception process under the Markovian temporal constraint. Feature-based *attention* curves are combined to find an optimised estimation of video affection from multiple modalities and temporal resolutions. The advantages of the MAR framework are: (1) the employment of information at coarse resolutions, which can hardly be used in content-based video analysis before; (2) the multi-resolution sampling and matching framework which alleviates media asynchronism caused by media resolution gap; (3) the extensibility and robustness on a large and noisy feature space.

7 Acknowledgement

The research leading to this paper was supported by European Commission under contracts FP6-045032 (Semedia) and FP6-027122 (Salero).

References

- [1] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization.
- [2] Daniel P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. <http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat>.
- [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.Q. 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] Michael S. Lew. *Principles of Visual Information Retrieval*. Springer, 1996.
- [7] Yuefei Ma, Lie Lu, Hongjiang Zhang, and Mingjing Li. A user attention model for video summarization. In *ACM Multimedia 02*, 2002.
- [8] C.E. Osgood, G.J.Suci, and P.H.Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.
- [9] Reede Ren and J.M Jose. Football video segmentation based on video production strategy. In *ECIR 2005*, 2005.
- [10] Reede Ren, J.M. Jose, and Yin He. Affective sports highlight detection. In *the 15th European Signal Processing Conference*, pages 728–732, Poznan, Poland, Sept. 2007.
- [11] Hemant D. Tagare, Kentaro Toyama, and Jonathan 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.
- [12] Nuno Vasconcelos and Andrew Lippman. Bayesian video shot segmentation. In *NIPS*, pages 1009–1015, 2000.
- [13] Hee Lin Wang and Loong Fah Cheong. Affective understanding in film. *IEEE Trans. Circuits Syst. Video Techn.*, 16(6):689–704, 2006.
- [14] Jinjun Wang, Changsheng Xu, Engsiong Chng, Kongwah Wah, and Qi Tian. Automatic replay generation for soccer video broadcasting. In *ACM Multimedia 2004*, pages 32–39, New York, NY, USA, 2004. ACM Press.
- [15] A. Willsky. Multiresolution markov models for signal and image processing. In *Proceedings of the IEEE 90 (8) (2002) 1396-1458*. 33, 2002.