TEMPORAL ATTENTION GRAPH

Reede Ren and Joemon M. Jose

University of Glasgow
18 Lilybank Gardens, Glasgow, UK
reede,jj@dcs.gla.ac.uk

ABSTRACT

Temporal attention is a psychological measurement of human focus in a long perceptual process such as watching a sports video. This measurement facilitates the identification of the most attractive components in media documents, especially in videos. In this paper, we propose a graphic representation which visualizes attention related temporal sequences from multiple resolutions. Efficient image operations are used to analyze perceptual attention. This results in an effective fusion approach for temporal attention estimation. We evaluate the effectiveness by the application of general highlight detection in sports videos, as sports highlights are temporal attended area. The experimental collection includes six full football games from FIFA World Cup and European Champion.

1. INTRODUCTION

Attention is a psychological measurement and describes notice distribution in a perceptual process. As a trivial component incurs little attention, the analysis of attention information is able to filter out media noise and facilitates the identification of key components in a media document [1, 2, 3]. It therefore becomes an essential method in content-based media analysis to estimate attention intensity. This technique can be categorized into two groups, spatial and temporal attention, according to related perception process. Spatial attention estimates notice distribution on a static scene, *e.g* an image. Lopez *et al.* [1] compute spatial attention to find the most noticeable image objects under complex background. Temporal attention tracks attention variation in a long perceptual process by plotting attention curves. Evangelopoulos *et al.* [4] develop feature-based saliency models to compute audio-visual attention curves. Geometrical features such as local extrema of an attention curve are used to identify possible interesting moments in a movie. Wang *et al.* [5] mine out patterns in temporal attention to discriminate story film genres, *i.e.* romantic, terror and action film.

Attention computation is a complex process of accumulating psychological facts, *i.e.* gathering many salient features to estimate a unified attention intensity [3, 4, 2]. This is because attention is an implication of perceptual state to multiple modality stimuli such as vision, auditory and text understanding. For example, salient map [6] is a popular tool for the estimation of spatial attention, which simulates the retina in eye. Pixel brightness in a salient map highlights the attractiveness of the related image region [1, 2]. The salient map accumulates numerous salient features, *e.g.* contrast, color energy, texture and edge intensity [2, 4], because the retina receives all visual salient stimuli using a spatial sensor of light sensitive neuron layers [6]. In addition, this biological mechanism provides an evidence to support spatial attention estimation. Figure 1 shows an example of salient map in a football video, where players and other noticeable regions can be easily recognized as highlighted area.

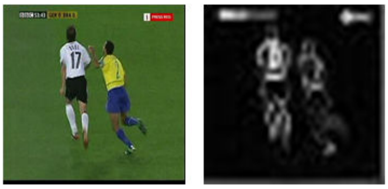

Figure 1: Static salient map

The estimation of temporal attention remains a research question [3], due to the complex psychological nature. First, temporal attention is closely associated with memory. Prior experience decides the reflection as well as attention intensity. Variant memory depth shows that temporal attention is of multi-resolution. Second, the issue of time extends salient features to time sequences or salient signals. Ma *et al.* [2] compute spatial attention on every visual frame, the amplitude of motion vector, shot boundary frequency and audio energy to estimate the perceptual importance of a sports video segment. These salient signals are physically of multiple temporal resolutions. For instances, frame-based spatial saliency is updating at 0.04 second as frame rates while shot boundary frequency refreshes at several minutes. Third, temporal attention simulates an observation sequence on content related perception, *i.e.* watching a sports video [2, 3]. The modeling of stimulus and reflection has to consider content hierarchy [3], for example, the video structure of scene and shot. To summarize, the estimation of temporal attention is a multi-resolution combination of noisy time sequences which are at multiple resolutions and updating rates.

In this paper, we originally propose a graphic representation for salient signals. This works is inspired by salient map [6] and the multi-resolution autoregressive framework (MAR) [3]. Attention can be treated as ordinal data, as we care the strength relationship rather than actual intensity. For example, a salient map is used to identify attended area, the region with the strongest attractiveness [2, 6]. It will not lose the effectiveness to change the value range of a salient signal, if the order is kept. Moreover, the MAR framework exploits a hierarchical Markovian character of content representation [3]. The operation of autoregressive analyzes sequential segments on a given temporal resolution. Note that image operations such as region smoothing mostly process neighborhood pixels and regions. These operations on a graphic representation keep Markovian constrains between attention samples and provide an efficient approach for attention analysis. Nevertheless, a graphic representation displays salient signals at multiple resolutions and leads to simultaneous processing at multiple temporal resolutions (Section 3). This advantage is noted by the psychological fact that attention perception is a parallel process on multiple memory depths [6]. In addition, we also present an efficient graphic based approach for temporal attention fusion and for highlight detection in sport videos.

The remainder of this paper is organized as follows. Section 2 surveys relate work in saliency based sports highlight detection and saliency fusion. The graphic representation for salient signals is defined in Section 3. Algorithms for temporal attention estimation and highlight detection are addressed in Section 4. Experimental results of sports highlight detection are found in Section 5 for six football game videos from FIFA World Cup 2002/6 and European Championship 2006. Section 6 provides a short conclusion and discussion.

2. RELATED WORK

Attention based video analysis is an exploration from computing psychology [2, 7]. Ma *et al.* [2] propose a series of psychological models on pre-attention, *i.e.* motion attention model, static attention model and audio salient model, to describe the perceptual process of video watching. This results in a set of temporal attention curves, *e.g.* motion attention curve, static attention curve and audio attention curve. Ma *et al.* linearly combine these curves to estimate a joint intensity of *"viewer attention"*. Too much noise is however introduced due to the massive extraction of salient features [3]. This makes the late highlight detection fragile. Real video events may be suppressed or vanished with the aggregation of salient signals. Evangelopoulos *et al.* [4] develop an energy function to remove noisy visual salient feature as well as to improve the robustness of linear combination. Hanjalic *et al.* [7] carefully choose a three-feature collection, including block motion vector, shot cut density and audio energy. An 1-minute long low-pass Kaiser window filter is furthermore employed to smooth feature-based attention curves and to improve signal noise ratio [7]. Hanjalic *et al.* count curve peaks inside a sliding window and regard this measurement as a probability estimation of highlight appearance. Although this approach of salient combination and event detection is robust against perceptual noise, the usage of sliding windows makes constant the temporal resolution of event detection. Ren *et al.* [3] exploit content structure of sports videos and propose a hierarchical Markovian constrain on feature-based attention curves. This results in a robust fusion framework of multi-resolution autoregressive but the computational complexity is high ($O(N^3 logN)$, where N is the length of salient signals).

3. GRAPHIC REPRESENTATION

In this section, we address the graphic representation for salient signals. As the MAR framework [3] is also a popular model in texture analysis [8], we think it is reasonable to propose a graphic representation for salient signals. This may reduce the computational complexity of attention fusion by using efficient image operators. In addition, the graphic representation visualizes attention data and facilitates the late attention-based content analysis. We normalize salient signals into the scale $[0, 255]$. Signal intensities are treated as gray intensities; signal samples are aligned according to time stamp; a texture image is thus created to represent a time sequence, *i.e.* salient signal. An example is shown in Figure 2, where the upper part is a salient signal and the bottom is the related texture graphic representation.

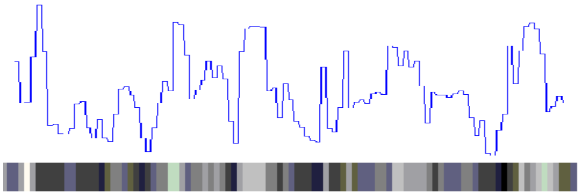

Figure 2: Graphic representation for salient signal

The graphic representation is convenient to present multi-resolution data. Figure 3 is a multi-resolution graphic representation for 15-min audio energy between the 27th to 42nd minute in the FIFA World Cup 2002 final game. Eight temporal resolutions are displayed, including 1 sec, 5 sec, 10 sec, 20 sec, 50 sec, 100 sec, 200 sec and 500 sec, from up to down. The multi-resolution representation provides a direct method for the signal combination on multiple temporal resolutions.

4. TEMPORAL ATTENTION ESTIMATION

In this section, we describe salient features, fusion algorithm and the evaluation system of sports highlight detection.

4.1 Saliency Feature Computation

Three salient features are used in this work, including block motion vector, audio energy and shot cut density. This feature collection is the same as that in [3, 7], which facilitates the evaluation on attention fusion and sports highlight detection.

Ma *et al.* [2] regard motion vector as perceptual response of optic nerves. Block motion vector from the compressed field however is a rough estimation of this salient character. In addition, there are massive motion prediction errors in sports videos, as play field is with uniform color and texture. To improve the precision, we exclude play field from motion intensity computation. The motion intensity I at macro block (i, j) is computed as the magnitude of motion vectors besides play field.

$$I(i,j) = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{MaxMag} \qquad (1)$$

where $dx_{i,j}$ and $dy_{i,j}$ denote components of motion vector; MaxMag is the maximum magnitude. The average of motion vector intensity is computed for every visual frame.

Audio energy measures the loudness. Auditory perception is complex, as many physical and perceptual aspects are concerned, *e.g.* frequency, audio type and speech contents. In sports video, the intensity of auditory stimulus can be roughly estimated by short period energy [2, 7]. In addition, audio encoder in MPEG has already considered the bandwidth issue of auditory perception. We therefore compute audio energy as the normalized sum of all audio tracks in *500ms*.

Shot is a relatively static view in a video. Shot cut density reflects the temporal deviation on vision. We detect shot boundary by a two-thread algorithm [9]. Shot cut density is the number of shot boundaries in every 100 seconds.

4.2 Multi-resolution Temporal Attention Estimation

The estimation of temporal attention usually involves many salient signals, for example, more than six salient signals are employed in [2] and [3]. We use a moving average algorithm to generate multi-resolution data [3]. This operation alleviates the problem of signal asynchronism due to the difference in updating rate and temporal resolution (Section 1). A multi-resolution representation is therefore created for every salient signal from the finest resolution, *i.e.* 1 second for audio energy, up to 500 seconds, as Figure 3 shows. This means that a salient signal is transformed into a texture image. We aggregate these texture images to estimate temporal attention the same as a static salient map does on salient features. The result of temporal attention estimation is also a texture image or a multi-resolution graphic representation for unified temporal attention. In addition, as a texture image is a multi-resolution representation, our approach is a parallel processing on salient signals at multiple temporal resolutions. In psychological terminology, we investigate many memory depths at the same time.

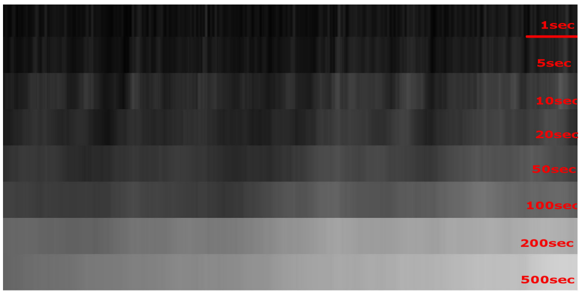

Figure 3: Multi-resolution graphic representation for audio energy in FIFA World Cup 2002 final game

4.3 Sports Video Highlight Detection

In this section, we demonstrate an efficient method which employs the graphic representation for sports highlight detection. Sports highlights are attended area in temporal attention, as the most interesting and attractive duration in a game video [2]. This indicates highlights are of high temporal attention intensity or bright areas in the graphic representation. Moreover, highlights can be observed from most resolutions due to their long duration, *e.g.* more than 84 seconds [3]. Sports highlight detection therefore turns into an image segmentation which allocates regions with a high brightness. A gray histogram is computed and we exploit the criterion of maximum entropy to select the threshold for image segmentation. For example, Figure 4 shows a 30-bin histogram extracted from the graphic representation of temporal attention in the second half of FIFA World Cup 2002 final game. After image segmentation, a morphological open operation is carried out to remove small regions as detection noise. We calculate the average gray scale in detected regions and take the top five with the highest average brightness as highlights in every temporal resolution. The boundaries of these regions are projected to time stamps in order to decide highlight moment as well as compute highlight duration.

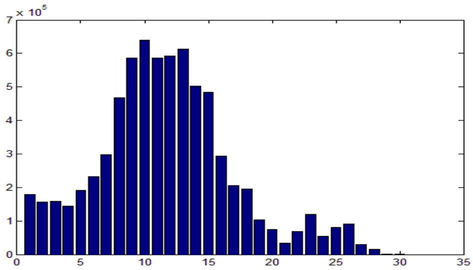

Figure 4: 30-bin gray histogram of temporal attention in the second half of FIFA World Cup 2002 final game

5. EXPERIMENT

We take the approach of multi-resolution autoregressive [3] as the baseline. The same salient signal set {average block motion, shot cut density, base band audio energy} is used as [7, 3] for temporal attention estimation.

The experimental collection includes six entire game videos in MPEG-1 format from FIFA World Cup 2002, World Cup 2006, and UEFA Champion 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. Game records are gathered from the FIFA and BBC Sports web site 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.

Figure 5 is the temporal attention aggregation image for the second half of the final game in the FIFA World Cup 2002. The light area denotes sports highlights and gray scale refers to the content attractiveness.

We take the temporal resolution of 200 seconds to evaluate the performance of highlight detection. This is because 200-second is the closest to the suggestion temporal resolution of 304-second for event detection [3]. If a ground truth event is overlapped more than 40% by detected regions, we label this event as detected. In Table 1, we count the

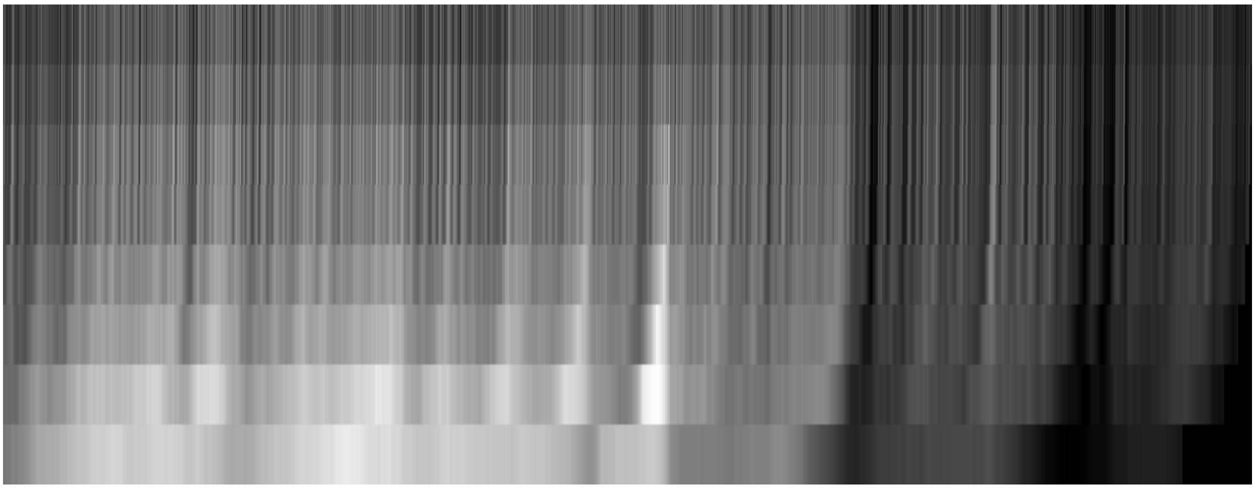

Figure 5: Temporal attention aggregation image for the second half of FIFA World Cup 2002 final game

coverage of goal events in the top five brightest regions as a direct performance measurement on highlight detection [10]. This is because goal events are widely accepted as domain highlights. Experimental results show that graphic based highlight detection (Section 4.3) is effective in goal event identification.

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

Table 1: Performance of Goal Detection (*goal events are replayed for several times)

Both MAR and graphic based highlight detection achieve 100% precision for goal events, although some rank changes exist. We suppose there are two causes. First, graphic-based method lacks the step of knowledge propagation from coarse temporal resolutions to fine resolutions [3]. Some long highlights may be divided into two segments due to a relatively plain moment in the middle. This will change the order of a highlight sequence. Second, a texture image is an efficient representation for salient signals at cost of precision, which may introduce noise and hide signal difference. However, such a rank variation does not affect the final result of goal detection, as most media applications, *e.g.* video skimming and video retrieval, show multiple game events simultaneously.

We use OpenCV C++ library to implement both approaches of temporal attention estimation, graphic-based and MAR. The experimental platform is a workstation with a Intel Core2 CPU at 1.8GHZ and 2GB memory. Table 2 shows time costs, involving attention estimation and highlight detection. We find the graphic representation results in a significant improvement on computational efficiency. This is due to two computational advantages of graphic representation: (1) image like data are efficiently managed by current CPU structures; and (2) frequent matrix multiplication which is necessary in MAR, is avoided in the graphic-based approach.

	MAR (sec)	Graphic based (sec)
Ger-Bra I	274.0	33.0
Ger-Bra II	321.0	41.0
Bra-Tur I	244.0	28.0
Bra-Tur II	230.0	30.0
Ger-Kor I	361.0	27.0
Ger-Kor II	437.0	39.0
Mil-Bar I	330.0	26.0
Mil-Bar II	332.0	28.0
Ars-Bar I	289.0	29.0
Ars-Bar II	294.0	26.0
Ita-Fra I	291.0	32.0
Ita-Fra II	324.0	35.0

Table 2: Time cost for highlight detection

Graphic representation based highlight detection keeps the merits from temporal attention analysis. As [3], we compare professionally marked highlights from BBC Sports and FIFA web site in Table 3 for the game of Italy vs. France, World Cup 2006. Most of manually selected highlights are covered. This indicates the effectiveness of graphic-based approach in the application of real-time video content filtering such as computer assisted video editing.

FIFA	BBC Sports	Rank
Players enter the field	-	5(I)
Penalty	Zidane Penalty	1(I)
Goal	Goal	2,4(I)
-	Zidane expulsion	4(II)
Italian Triumph	-	1(II)

Table 3: Game highlights and related rank in France vs Italy (I,II game halve)

6. CONCLUSION AND DISCUSSION

In this paper, we originally propose a graphic representation for temporal attention sequences, *i.e.* salient signals. This representation results in an efficient fusion algorithm for temporal attention estimation, as well as an effective approach for temporal attended area detection, such as the allocation of sports video highlights. Experimental results show that this graphic representation greatly enhances efficiency while not reducing the effectiveness of temporal attention analysis. We conclude that this new approach meets the requirement of online media service, such as real-time video content filtering, video skimming and summarisation.

Many aspects remains research in the graphic representation. The algorithm of temporal attention fusion is a direct simulation of static salient map. This sounds reasonable but ignores the issue of time which plays an important role in temporal attention. For example, it is difficult to identify the optimal resolution for attended area detection directly from the aggregation image. We also observe that the set and the rank of attended areas vary with temporal resolutions (Figure 5), although this phenomenon can be explained by the psychological fact that the memory and the prior experience decide on possible reflections. In this paper, we rely on external knowledge from [3] to choose a proper time resolution for sports event detection. A further study on temporal resolution selection is necessary for graphic based temporal attention analysis. Moreover, the approach of graphic representation facilitates temporal attention aggregation at multiple temporal resolutions. This is welcomed by psychological facts. However, a formal psychological explanation on this representation remains a research question. Although we regard the aggregation image as a parallel memory which keeps stimuli and decides on reactions according to prior experience, a careful psychological experiment is necessary to support this conclusion.

7. ACKNOWLEDGEMENT

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

REFERENCES

[1] María T. López, Miguel A. Fernández, Antonio Fernández-Caballero, José Mira, and Ana E. Delgado, “Dynamic visual attention model in image sequences,” *Image Vision Comput.*, vol. 25, no. 5, pp. 597–613, 2007.

[2] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[3] Reede Ren and Joemon M. Jose, “General highlight detection in sport videos,” in *ACM Multimedia Modeling 2009*, 2009, pp. 27–38.

[4] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, “Movie summarization based on audiovisual saliency detection,” in *ICIP 2008*, Oct. 2008, pp. 2528–2531.

[5] Hee Lin Wang and Loong Fah Cheong, “Affective understanding in film,” *IEEE Trans. Circuits Syst. Video Techn.*, vol. 16, no. 6, pp. 689–704, 2006.

[6] M.J. Lesser and D.K.C. Murray, “Mind as a dynamical system: Implications for autism,” in *Durham conference Psychobiology of autism*, 1998.

[7] A. Hanjalic and L.Q. Xu, “Affective video content repression and model,” *IEEE Trans on Multimedia*, vol. 7, no. 1, pp. 143–155, Feb 2005.

[8] J. Mao and A.K. Jain, “Texture classification and segmentation using multiresolution simultaneous autoregressive models,” *Pattern Recognition*, vol. 25, no. 2, pp. 173–188, 1992.

[9] Reede Ren and J.M Jose, “Football video segmentation based on video production strategy,” in *ECIR 2005*, 2005.

[10] 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*, vol. 14, pp. 634–643, May 2004.