VISUAL-AURAL ATTENTION MODELING FOR TALK SHOW VIDEO HIGHLIGHT DETECTION

Yijia Zheng[1], Guangyu Zhu[2], Shuqiang Jiang[3], Qingming Huang[1,3] and Wen Gao[4]

[1]Graduate University of Chinese Academy of Sciences, Beijing, China
[2]School of Computer Science, Harbin Institute of Technology, Harbin, China
[3]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing, China
[4]School of Electronics Engineering and Computer Science, Peking University
{yjzheng,gyzhu,sqjiang,qmhuang,wgao}@jdl.ac.cn

ABSTRACT

In this paper, we propose a visual-aural attention modeling based video content analysis approach, which can be used to automatically detect the highlights of the popular TV program—talk show video. First, the visual and aural affective features are extracted to represent and model the human attention of highlight. For efficiency consideration, the adopted affective features are kept as few as possible. Then, a specific fusion strategy called ordinal-decision is used to combine the visual, aural attention models and form the attention curve for a video. This curve can reflect the change of human attention while watching TV. Finally, highlight segments are located at the peaks of the attention curve. Moreover, sentence boundary detection is used to refine the highlight boundaries in order to keep the segments' integrality and fluency. This framework is extensible and flexible in integrating more affective features with a variety of fusion schemes. Experimental results demonstrate our proposed visual-aural attention analysis approach is effective for talk show video highlight detection.

Index Terms— attention modeling, highlight extraction, ordinal-decision, affective feature, attention curve

1. INTRODUCTION

Highlight detection is one of the key technologies of content based video analysis. It offers a concise representation of the original video by showing its most representative synopsis. Previous typical work in this field is mostly done on sports video [1][2][3], the methods they used always need much domain knowledge and are hard to extend to other kinds of broadcast videos such as movie, talk show and so on. In this paper we investigate the visual-aural attention modeling based highlight detection approach on talk show video which has less computational complexity and can be extended to other TV programs easily.

Talk show video is a popular TV genre enjoyed by lots of people, and little work has been done on the analysis of this program. It has the advantage of less background noise and additive edits than other kinds of TV genres. For example, sports videos usually have lots of background noise, while movies usually contain some manual edits such as montage. Therefore, talk show video is suitable for proving the feasibility and validity of our proposed attention modeling based highlight detection approach.

Here we give the video structure definition of the talk show program employed in our work, which has the form of a few hosts, several guests, and a number of audiences，who are discussing about one topic. Usually, the hosts bring forward a topic (Fig1.a), then, the guests launch the discussion (Fig1.b) and the audiences participate in the topic and ask questions (Fig1.c).

(a) **(b)** **(c)**

Fig 1 Talk Show Video Examples

Attention as a neurobiological concept is regarded as the representation of mental power upon objects gained by observing or listening, which is the ability to concentrate mentally [4]. Modeling the primary factors that affect human attention in the process of watching TV allows us to decompose the problem of video analysis into a series of computationally less demanding and localized analyzable problems.

The pioneer work of attention modeling based video analysis is [4]. A user attention model, which estimates the attention that viewer paid to the video contents is proposed in this paper. You *et al.* inherited the ideology in [4] to propose a human perception analysis framework for video understanding based on multiple visual cues [5]. In [6], [7] the authors constructed the visual attention models and applied their models to the application of user focus detection in video frames. The attention modeling based video content analysis work is proved more consistent with human understanding and has less computational complexity.

Although they have the above advantages compared with traditional work, the current techniques on attention modeling based highlight extraction are mainly focused on the analysis of visual aspect but neglect the aural modality [5][6][7], which is another important intrinsic information source of video. Besides, the highlights are usually simply determined as the local maximums of the linear fused attention curve [4], which didn't consider the highlight asynchronous attention influence factors such as applaud and cheer in their work. To overcome the aforementioned limitations, we propose an attention analysis approach which aims to visual, aural modeling and integrates both the highlight synchronous and asynchronous attention influence factors in the fusion strategy in our work.

The rest of the paper is organized as follows. Section 2 introduces the overview of proposed approach. Visual-aural attention model construction is presented in Section 3. Section 4 describes the details of the ordinal-decision fusion strategy in talk show video highlight detection scheme. Experimental results are reported in

Section 5. Conclusions and acknowledgement are drawn in Section 6 and 7.

2. ARCHITECTURE OF VISUAL-AURAL ATTENTION MODELING BASED VIDEO HIGHLIGHT DETECTION

The framework of our proposed approach is shown in Fig.2, which consists of three major modules: 1) visual-aural attention model construction, 2) talk show video attention curve generation and 3) talk show video highlight extraction.

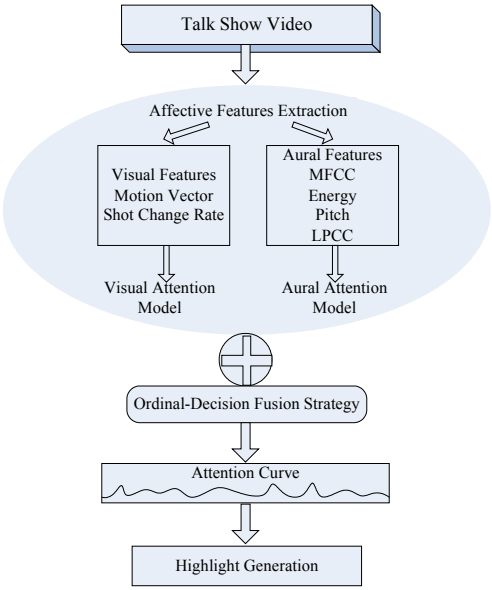

Fig. 2 Attention Analysis for Video Highlight Extraction

As video highlights are strongly subjective, therefore, it is important to generate them based on the general understanding of audiences. We consider the dominant reasons that affect human attention when watching videos and construct the corresponding models to obtain the more acceptable results.

Video is a compound of image sequences and audio tracks, which deliver information with their own primary elements. The formation of visual and aural attention can be considered as the combination of multiple factors. Firstly, we divide the video into visual and aural channels. Affective features are extracted for representing the visual and aural attention. An Ordinal-Fusion strategy is exploited to construct the final visual-aural model.

Then, the attention curve is obtained from the fused visual-aural model. Highlight segments are localized at the peaks of this curve.

Finally, the speech boundaries are detected to locate the final boundaries of the highlights and guarantee the segments natural and fluent. With the confidence that the sentence can not be interrupted, we select the segments satisfy this qualification as the final video highlights.

Compared with previous work, the main contributions of our work are 1)propose an aural attention modeling approach which commits to simulate the human aural perception changes while watching TV, and 2) propose a feasible Ordinal-Fusion strategy for combining the highlight synchronous and the highlight asynchronous attention models. Besides, for efficiency consideration, the number of adopted affective features is kept as few as possible. So in the visual attention modeling approach we use the more video-highlight-related "video-based" affective features and leave the less video-highlight-related "image-based" features out.

3. VISUAL-AURAL ATTENTION MODELING

In this section, we present an approach of video attention modeling in terms of the visual and aural perceptive sources.

3.1 Visual Attention Modeling

In visual modality, the features such as color, texture, orientation etc. which can be calculated from a single video frame are called "image-based" features. As a video highlight is always consists of a series of frames, a single frame can't affect much to the whole series. These "image-based" features are not only expensive to compute but also less related with the highlight generation from the view of the whole video. So we use the affective features that calculated between video frames called "video-based" features to model the visual attention.

The visual modality is not only contains the spatial information but also the temporal information that both affect human attention. We employ the average motion vector (***AMV***) in 1 second to represent the spatial dimension. Usually the larger ***AMV*** means a scene with higher motion intensity which is more likely to attract audience's attention. Though sometimes motion vectors (MV) don't reflect the "true" motion field well, but utilizing MV can greatly help for reducing more computation complexity than that of the fine-grained optical flows. The spatial model M_{spa} is defined as

$$M_{spa} = (\sum_{i=1}^{k} MV_i)/k \times 100\% \quad (1)$$

where MV_i is the motion vector of the ith frame obtained by the decoding process, k is the frame rate of video (e.g. 25fps).

In temporal dimension, the shot change rate (***SCR***) is used to represent the camera motion. When the shots change more frequently in talk show video, it usually indicates a strenuous discussion scene and the audience's attention will be prominently influenced. The temporal model M_{tem} is defined as

$$M_{tem} = e^{((1-(n(k)-p(k))/\delta)} \times 100\% \quad (2)$$

where $p(k)$ and $n(k)$ are the positions (frame indexes) of the two adjacent shot boundaries to the left and right of the frame k, and the parameter δ is a constant which guarantees the M_{tem} values distribute in the scale between 0% and 100%.

3.2 Aural Attention Modeling

All the abovementioned related work on human attention modeling either ignores the aural dimension or models this dimension comparatively simple [4][5][6][7]. The audio-track usually contains immense amounts of useful information and it normally has closer link to semantic event than the visual information. In our work, we propose a feasible aural attention modeling approach which can simulate the audiences' perception to certain extent.

The speech, applaud and laugh are the typical sounds in talk show videos, and human usually pay more attention to these sounds which convey more information than the others. Highlight segments usually followed by a sequence of applaud or laugh. The louder the applaud or laugh is, the more exciting the scene is.

Based on our previous work on audio classification [8], we use the support vector machine (SVM) to classify the audio track into speech, applaud, laugh, silence, music and other specific sounds. Low level audio features: short time energy, zero crossing rate, pitch, LPCC, MFCC are used for the SVM training and test. The audio affective features average speech energy (***ASE***) and average pitch (***AP***) are used to represent speech. Applaud and laugh are described by ***ASE***. We define the speech model M_{spe}, applaud model M_{app} and laugh model M_{lau} as:

$$M_{spe} = (\sum_{i=1}^{n} Energy_i) \cdot (\sum_{i=1}^{n} Pitch_i) / n^2 \times 100\% \quad (3)$$

$$M_{app} = (\sum_{i=1}^{p} Energy_i) / p \times 100\% \quad (4)$$

$$M_{lau} = (\sum_{i=1}^{q} Energy_i) / q \times 100\% \quad (5)$$

where $Energy_i$ and $Pitch_i$ are the energy and pitch of the ith sampling signal. n, p, q are the numbers of samples in one second, respectively.

All the attention values calculated from the above visual and aural attention models are normalized into the interval of [0, 1] using Gaussian normalization. So for each talk show video clip, we can get its several visual and aural attention curves which reflect the changes of human attention.

4. TALK SHOW VIDEO HIGHLIGHT GENERATION

Along the temporal axis, the values of visual and aural attention in each second form several attention curves, which are good expressions of video content. The attention curves are shown in Fig. 3.

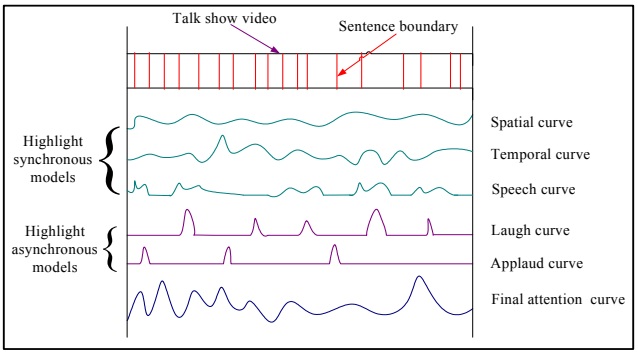

Fig. 3 Talk Show Video Attention Curves

Using formula (1) to (5) we can obtain the spatial curve C_{sc}, temporal curve C_{tc}, speech curve C_{spec}, laugh curve C_{lauc} and applaud curve C_{appc}. The ordinal-decision fusion strategy shown in Fig. 4 and highlight detection algorithm described in Table 1 are used to fuse the visual and aural attention models. We can obtain the video highlights from the fused video attention curve.

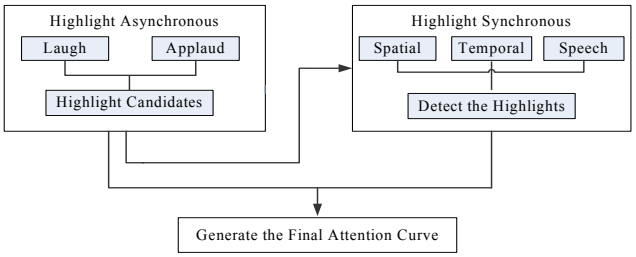

Fig. 4 Ordinal-Decision Fusion Strategy

Table 1 Talk Show Video Highlight Extraction

Step1: Use the highlight asynchronous curves C_{appc}, C_{lauc} to localize the highlight segments following the algorithm proposed in [9]. The main idea is to make use of some audio events such as silence and combine video shot boundary information for highlight candidates generation.

Step2: Use the highlight synchronous curves C_{sc}, C_{tc}, C_{spec} to further estimate the exciting degree of each highlight candidate obtained by Step 1. If one candidate's M_{spe}, M_{spa}, M_{tem} and its followed applaud and laugh are larger and longer than others, it will be paid more attention to. Each highlight candidate's attention value M_a is computed as:

$$M_a = (\lambda_{spa} \cdot M_{spa} + \lambda_{tem} \cdot M_{tem} + \lambda_{spe} \cdot M_{spe}) \cdot e^{(\sum_{i=1}^{p} M_{app})} \cdot e^{(\sum_{i=1}^{q} M_{lau})} * G(n) \quad (6)$$

λ_{spa}, λ_{tem}, λ_{spe} are each model's weight with $\lambda_{spa} + \lambda_{tem} + \lambda_{spe} = 1$, $\lambda_{spa}, \lambda_{tem}, \lambda_{spe} \geq 0$,. p, q are the length of laugh and applaud followed the highlight segments measured by second, respectively. $G(n)$ is the Gaussian smoothing window, n is the smoothing parameter.

Step3: Let S_1 and S_2 are the left and right boundaries of one highlight candidate, F_a and F_b are the frame indexes of their nearest speech boundaries detected by our audio classification algorithm in [8]. So the segment between F_a and F_b are one of the final highlight we need.

5. EXPERIMENT

There are some manually selected highlight segments chose by the editor in talk show videos. These manually edited highlights reflect audience's subjective perception and are taken as ground truth in our experiment. We took six talk show videos "Art and Life", "Dialogue" and "Tiger. Talk" recorded from the broadcasted programs of China Central Television (CCTV) and Hong Kong Phoenix Television (PHTV) as the testing data. The testing videos cover different programs and different length in order to verify the effectiveness and generality of our proposed approach. The details of the testing data are presented in Table 2.

Table 2 Testing Data

No.	Video	Shot	Length
1	Art&life1.mpg	435	96'24''
2	Art&life2.mpg	297	53'47''
3	Dialogue1.mpg	293	45'45''
4	Dialogue2.mpg	304	41'05''
5	Tiger.Talk1.mpg	326	45'27''
6	Tiger.Talk2.mpg	343	45'46''

The testing results followed our proposed visual-aural attention modeling based talk show video highlight extraction approach are shown in Table 3.

Table 3 Experimental Results

Video	Manual	Automatic	intersection	accuracy
Art&life1.mpg	15	17	14	82.4%
Art&life2.mpg	9	9	8	88.9%
Dialogue1.mpg	18	16	15	83.3%
Dialogue2.mpg	14	13	11	78.6%
Tiger.Talk1.mpg	12	14	11	78.6%
Tiger.Talk2.mpg	14	15	12	80.0%

The weights in formula (6) could be decided base on either user demand or the property of the particular video. In our experiment, all weights are set to 1/3, and the *n* in formula (6) is 60. In Table 3, the second column is the number of manually selected highlights, and the third column is the number of automatically generated highlights. The fourth column is the intersection of column two and column three. The last column is the accuracy of our algorithm compared with human perception. The accuracy in Table 3 is calculated as:

$$accuracy = \frac{Automatic \cap Manual}{\max\{Automatic, Manual\}} \times 100\% \quad (7)$$

From the experimental results we can see that the two main problems—how to establish an aural model which can simulate human attention and how to extract the talk show video highlights which are more consistent with audience's subjective understanding are both solved satisfactorily.

The deviation between the automatic extraction and the audience's subjective perception is mainly due to two factors. One is the error of audio classification, which influences the accuracy of speech, applaud and laugh attention modeling. The second reason is the shot boundary and speech boundary detection are not exact enough which lead to the inaccurate localization of the final highlight positions. Although it has the above limitations, our visual-aural attention modeling based perceptive analysis of talk show video highlight extraction approach is proved feasible and effective by experiments.

6. COCLUSION

In this paper, we proposed a computable visual-aural attention model by analyzing audience's perception when he/she is watching TV. The model has been designed and a set of modeling methods for audio, visual affective features have been proposed. As an important application of the attention analysis framework, we present a feasible solution of talk show video highlight extraction, which can obtain the highlights from the viewpoint of human understanding. Compared with existing methods, our proposed approach is based on the changes of human attention rather than simple content changes of videos, which is more consistent with human understanding.

7. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China under grant No. 60773136 and No.60702035, National Hi-Tech Development Program (863 Program) of China under grant No. 2006AA01Z117, "Science 100 Plan" of Chinese Academy of Sciences under grant No.99T3002T03, and the Knowledge Innovation Program of the Chinese Academy of Sciences under grant No.20076032.

8. REFERENCES

[1] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling", *IEEE Trans. Multimedia*, Dec. 2005, vol.7, pp. 1114-1122.

[2] Z. Xiong, R. Radhakrishnan, A. Divakararan, T. Huang, "Highlights extraction from sports video based on an audio-visual marker detection framework", *IEEE Int. Conf. on Multimedia and Expo*, Jul.2005.

[3] Y. Rui, A. Gupta, A. Acero, "Automatically extracting highlights for TV baseball programs", in *Proc. ACM Int. Conf. Multimedia,* 2000, pp. 105-115.

[4] Y. Ma, X. Hua, and L. Lu *et al.*, "A generic framework of user attention model and its application in video summarization", *IEEE Trans. Multimedia*, Oct. 2005, vol. 7, no. 5, pp. 907–919.

[5] J. You, G. Liu, H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization", *IEEE Trans. Circuits and System for Video Technology*, Mar. 2007, vol. 17, pp. 273-285.

[6] C. Ho, W. Cheng, T. Pan, J. Wu, "A user-attention based focus detection framework and its applications", *ICICS-PCM*, Dec. 2003, vol. 3, pp. 1315-1319.

[7] M. Guironnet, N. Guyader, D. Pellerin, P. Ladret, "Spatio-temporal attention model for video content analysis", *Int. Conf. on Image Processing*, Sep. 2005, vol. 3, pp. 1156-1159.

[8] Y. Zheng, G. Zhu, S. Jiang, Q. Huang, "Highlight Ranking for Racquet Sports Video in User Attention Subspaces Based on Relevance Feedback", *IEEE Int. Conf. on Multimedia and Expo*, Jul. 2007, pp. 104-107.

[9] M. Xu, L. Chia, J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection", *IEEE Int. Conf. on Multimedia and Expo*, Jul. 2005.