HUMAN ATTENTION MODEL FOR SEMANTIC SCENE ANALYSIS IN MOVIES

Anan Liu[1,2,3], Yongdong Zhang[2,3], Yan Song[2,3], Dongming Zhang[2,3], Jintao Li[2,3], Zhaoxuan Yang[1]

[1]School of Electronic Engineering, Tianjin University, Tianjin, 300072, China
[2] Virtual Reality Laboratory, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, 100080, China
[3] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, 100080, China
liuanan@ict.ac.cn

ABSTRACT

In this paper, we specifically propose the *Weber-Fechner Law*-based human attention model for semantic scene analysis in movies. Different from traditional video processing techniques, we pay more attention on bringing in the related subjects, such as psychology, physiology and cognitive informatics, for content-based video analysis. The innovation of our work has two aspects. Firstly, we originally construct the human attention model with temporal information instructed by the Weber-Fechner Law. Secondly, motivated by cognitive informatics, we formulate the computational methodology of features in visual, audio and textual modalities in the uniform metric of information quantity. With human attention analysis and semantic scene detection, we build a system for hierarchical browse and edit with semantics annotation. Large-scale experiments demonstrate the effectiveness and generality of the proposed human attention model for movie analysis.

Index Terms— Human Attention Model, Cognitive Informatics, Weber-Fechner Law, Movie

1. INTRODUCTION

Researchers have been engaged in content-based video analysis and retrieval for a long time. However, the past focus is mainly on the video itself. Therefore, many difficulties have emerged as for the semantic analysis of video, especially movies that convey high level semantics of stories. Therefore, it is time that we brought in the related subjects, such as cognitive informatics and psychology, for movie analysis from the viewpoint of human.

The most representative work for movie content analysis is done by Brett Adams et al. In [1], they bring in the film grammar and present an original computational approach for movie tempo to derive story sections and events. Due to the limitation of using only a single modality, Hsuan-Wei Chen et al in [2] use the features of audio energy to modify tempo proposed by Brett Adams et al. Although some research has been done on tempo model for movie analysis, there exists one important problem. The complex storyline and potential film grammar make it quite difficult to map the low level features into special descriptors of semantic events. Therefore, little advanced work has been done on movie content analysis with the characteristics of films in the recent years.

Motivated by [3], we bring in human attention modal to simulate human response to stimuli, including ceaselessly changed visual, audio and textual signal. Although the computational attention methodologies have been studied for many years, researchers mainly focus on the construction of topographical saliency map and the application of the attention modal is restricted to the processing of the single frame. Yu-Fei Ma et al. are those of few researchers who are engaging in applying human attention model on video analysis. In [4], they detailedly present the attention model and the application in video summarization. However, they do not deeply analyze the relationship between user attention and the external stimuli and neglect the temporal information for the attention model. Therefore, we propose Weber-Fechner Law based human attention model for semantic scene analysis in movies. The main contributions of our work are twofold: Firstly, we improve the model with temporal variation depending on psychology and physiology; Secondly, we bring in the cognitive informatics, a newly emerging subject, for video analysis and formulate the computational methodology of features in the uniform metric of information quantity. Therefore, the human attention model proposed in the paper is more proper to describe the change of user attention with the external stimuli.

The remainder of the paper is organized as follows. In Section 2, we introduce the framework of Weber-Fechner Law based human attention model. Then, we specifically illustrate the computational methodology of the features in the uniform metric of information quantity in Section 3. In Section 4, the application of the model is introduced. The

experimental results are presented in Section 5. At last, the conclusions and future work are stated in Section 6.

2. WEBER-FECHNER LAW BASED HUMAN ATTETION MODEL

As for human perception, people usually focus on the conspicuous changes both in visual, audio and textual modalities. Therefore, for effectively analyzing the movie content, we need mining human attention descriptors from the viewers' standpoint.

The human attention model presented in the paper is comprised of visual, audio and textual sub-models. Any extractable features in the three modalities can be integrated into this framework because of its extendable ability. Then, the sub-models can be integrated with fusion methods to generate one attention model for shots. Therefore, the human attention model for the single shot, *HAM(t)*, can be represented as follows:

$$\begin{cases} VAM(t) = \sum_{i=1}^{l} w_i * VF(t)_i \\ AAM(t) = \sum_{j=1}^{m} w_j * AF_j(t) \\ TAM(t) = \sum_{k=1}^{n} w_k * TF_k(t) \\ HAM(t) = Fusion[VAM(t), AAM(t), TAM(t)] \end{cases} \quad (1)$$

where *VAM(t)*, *AAM(t)* and *TAM(t)* respectively represent visual attention sub-model, audio attention sub-model and textual sub-model; w_i, w_j and w_k respectively denote the weights for visual, audio and textual sub-models; *VF(t)*, *AF(t)* and *TF(t)* respectively mean the features in visual, audio and textual modalities; *Fusion[ ]* is the operation of fusion scheme. Here, the unit of "*t*" is "*shot*" and its continuous change means the concessive shots in one video.

It is perceptive that the human perception is not straightforwardly related with the external stimuli, I, but with the increase, ΔI . Therefore, for constructing attention model for the entire movie, we must consider the information in temporal domain. Weber-Fechner Law [5] is a rule in psychology and physiology which reflects the relationship between external stimuli and human perception. With this law, we formulate the Weber-Fechner Law based human attention model, *M*, for videos as follows:

$$M = k * log[HAM(t)] + C \quad (2)$$

where k and C are experienced constant.

3. COMPUTATION METHODOLOGY

In this section, we firstly introduce the methods of feature extraction and then present the formulation of them in the uniform metric of information quantity

3.1. Features Extraction

It is commonly considered that the features for video analysis are classified into two kinds, low level features and high level features. In this section we introduce the related features for movie analysis.

(a) Low level features:

The low level features in visual and auditory modalities, such as motion energy (*ME*), phase distribution (*PD*), audio energy (*AE*) and audio change frequency (*AF*), have been detailedly introduced in [4]. Here, we focus on introducing the novel textual features.

As for movies, the captions with more words usually attract more attention of viewers. Therefore, we use the length of the sentence, *TE*, to represent its importance. Besides, we use lexical analysis system *ICTCLAS*, a free software, to implement the word segment on each sentence and to extract noun, verb and adjective which contain more information for viewers. Then we formulate the word frequency, *WF*, as follows:

$$WF = -1/Log(\frac{N}{N_{word}}) \quad (3)$$

where, in one sentence, *N* denotes the total number of noun, verb and adjective and N_{word} means the total number of words.

(b) High level features:

High level features usually indicate the special semantic meanings. Therefore, the special detectors in different modalities and their combination can be used for movie analysis. In visual modality, we adopt the face detector (*FD*) in [6], sex detector (*SD*) in [7] for semantic scene detection. In auditory modality, the SVM-based gunfire descriptor (*SGD*) in [8] and SVM-based Audio classifiers (*SAC*) in [9] are constructed.

3.2. Features Formulation

Quantitative representation of human perception with scientific basis has been a problem unsolved ideally for a long time. However, cognitive informatics, the newly emerging subject, has founded the relative theoretical foundation and proposed an advisable method. In [10], Wang proposes that information is a more proper measure for human perception. With calculating the information of visual, auditory and textual modalities, we can represent the human perception descriptor quantitatively. Therefore, the paper focuses on the computational method of some features in different modalities in the uniform metric of information quantity.

For the features mentioned above, we can see that *ME*, *AE* and *TE* represent energy and *PD*, *AF* and *WF* reflects information. Therefore they belong to two categories and are not additive. However, the high values of them are all positive to represent the large visual information and indicate the strong stimuli. Therefore, we integrate them as follows to formulate the features in three modalities:

$$\begin{cases} VI = ME * PD \\ AI = AE * AF \\ TI = TE * WF \end{cases} \quad (4)$$

where *VI*, *AI* and *TI* respectively correspond to visual, audio and textual information. We consider the features representing energy as the weights of those representing information. As for the other four features, the detectors (*FD*, *SD*, *SGD* and *SAC*) can calculate the confidence score of their existence. Then, we can convert the probability (*P*) into information (*I*) with the following equation:

$$I = -Log(1/P) \quad (5)$$

Consequently, we formulate the features in three modalities in the uniform metric of information quantity. Then they are additive and comparable.

4. APPLICATION OF HUMAN ATTENTION MODEL

We build an interactive system, "SmartMoviePlayer", for movie browse and edit. The system consists of three parts: Structuralization, Semantic scene detection and Interactive interface. We will illustrate the three sections as follows.

4.1. Structuralization

Structuralization for movie content analysis includes two steps: shot boundary detection & keyframe extraction, and scene boundary detection. We implement our previous method presented in [11] for them.

4.2. Semantic scene detection

In our system, we realize the detection of action, war, dialogue, sex and music scenes which are useful for movie edit and viewers' navigation. The features used for them are listed in Table 1. Then we use the linear fusion of the features for the formulation of the specific human attention models and we get the attention flow plot to depict the change of human attention with the development of the storyline. Moreover, the attention flow plot is smoothed with a Gaussian filter since human perception gradually changes because of memory retention. At last, different semantic scenes are detected based on the edge analysis method with Deriche's recursive filtering algorithm in [1].

4.3. Interactive interface

As shown in Fig.1, our interface is composed of five parts: a folder browsing subwindow, a media file subwindow; a playing-back subwindow (left-bottom); a hierarchical browsing subwindow; and finally a storyboard subwindow (right-bottom). Besides, different scenes are annotated for interest-oriented navigation of viewers and convenient edit of moviemaker. The detailed introduction is in [11].

Table 1. Features for semantic scene detection

Concept	Action	War	Dialogue	Sex	Audio
Visual	ME,PD	ME,PD	FD	SD	——
Audio	AE,AF	SGD	SAC	——	SAC
Textual	——	——	TE,WF	——	——

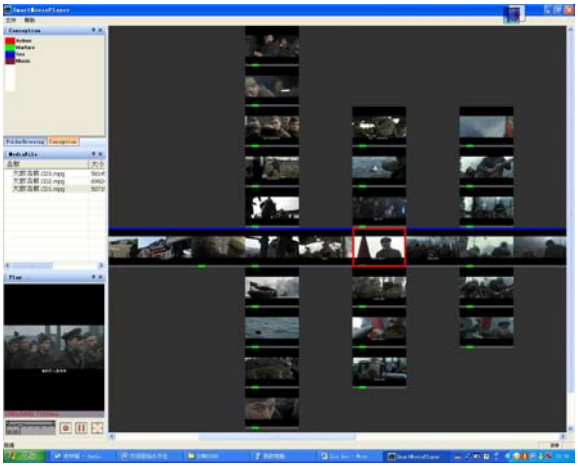

Fig.1 SmartMoviePlayer: hierarchical browsing of War movie, "Enemy at the Gates".

5. EXPERIMENTAL RESULTS

To demonstrate the effectiveness and generality of human attention model for semantic scenes detection, we select twelve movies for our experiments. By human judgment, we found the ground truth for experimental result analysis. Precision and Recall are used to measure the results, which are well known rules in information retrieval field. The detailed experimental results are shown in Table 2-6.

From Table 2-6, we can see that recall value is satisfactory, which means that human attention model can well depict the characteristics of the semantic events. Comparatively, precision value is a little lower. The main reasons are that some other concepts may have the similar characteristics and thereby give the false positives.

6. CONCLUSION AND FUTURE WORK

In this paper, we specifically propose the *Weber-Fechner Law*-based human attention model for semantic scene analysis in movies. Besides, we formulate the computational methodology of features in multiple modalities in the uniform metric of information quantity motivated by cognitive informatics. With human attention analysis and semantic scene detection, we build a system for hierarchical browse and edit with semantics annotation. Large-scale experiments demonstrate its effectiveness and generality of human attention model for movie analysis.

In the future, we will advance our research on content-based video analysis with the related subjects for human perception modeling and semantics representation in movie content analysis.

Table 2. Experimental results for action scene detection

Movie Title	Crouching TigerHidden Dragon	Fist of Legend	The Matrix	Pearl Harbor
Ground	8	8	7	12
False accepted	2	1	0	2
False rejected	0	1	1	0
Precision(%)	80	87.5	100	85.7
Recall (%)	100	87.5	86	100
Average Precision (%)	88.3	**Average Recall (%)**		93.4

Table 3. Experimental results for war scene detection

Movie Title	Enemy at the Gates	Wind talkers	The Thin Red Line
Ground	19	21	5
False accepted	5	3	1
False rejected	2	2	0
Precision(%)	77.3	86.4	83.3
Recall (%)	89.5	90.5	100
Average Precision (%)	82.3	**Average Recall (%)**	93.3

Table 4. Experimental results for dialogue scene detection

Movie Title	Crouching TigerHidden Dragon	Fist of Legend	Wind talkers	Pearl Harbor
Ground	8	8	9	12
False accepted	2	1	2	2
False rejected	0	1	1	0
Precision(%)	80	87.5	80	85.7
Recall (%)	100	87.5	88.9	100
Average Precision (%)	83.3	**Average Recall (%)**		94.1

Table 5. Experimental results for sex scene detection

Movie Title	La Belle	Sex is Zero
Ground	14	20
False accepted	0	2
False rejected	0	1
Precision(%)	100	90.5
Recall (%)	100	95
Average Precision (%)	95.25	**Average Recall (%)**	97.5

Table 6. Experimental results for music scene detection

Movie Title	Brave heart	Legends of the Fall	Barnyard
Ground	28	21	35
False accepted	3	1	3
False rejected	0	3	3
Precision(%)	90.3	94.7	91.4
Recall (%)	100	85.7	91.4
Average Precision (%)	92.1	**Average Recall (%)**	92.4

7. ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), the National Nature Science Foundation of China (60773056), the Beijing New Star Project on Science & Technology (2007B071),the Knowledge Innovation Project of The Institute of Computing Technology, Chinese Academy of Sciences (20076031) and Key project supported by Natural Science Foundation of Tianjin (No. 07JCZDJC05800).

8. REFERENCES

[1] Brett Adams, Chitra Dorai, Svetha Venkatesh. "Toward Automatic Extraction of Expressive Elements from Motion Pictures: Tempo." *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp:472-481, 2002.

[2] Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, et al. "Action Movies Segmentation and Summarization Based on Tempo Analysis." Proc of 6th ACM MIR, pp: 251-258, 2004.

[3] L.Itti, C.Koch, E.Niebur, "Computational Modaling of Visual Attention," *Nature Reviews Neuroscience*, 2001 Mar;2(3):194-203.

[4] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, et al, "A User Attention Model for Video Summarization," Proc. of ACM MM, pp:533-542, 2002.

[5] Selig Hecht, "The Visual Discrimination of Intensity and the Weber-Fechner Law," Journal of General Physiology, vol 7, pp: 235-267, 1924.

[6] Zhao Ming, Chen Chun, ,et al. "Subspace analysis and optimization for AAM based face alignment." Proc. of IEEE International Conference on Automatic Face and Gesture Recognition. Seoul,South Korea, pp: 290-295,2004.

[7] Qiang Zhu, Ching-Tung Wu, et al,"An Adaptive Skin Model and Its Application to Objectionable Image Filtering," ACM MM 2004, New York, pp:56-63, 2004.

[8] Moncrieff, S., Dorai, C., Venkatesh, S., "Detecting Indexical Signs in Film Audio for Scene Interpretation," Proc. of ICME, pp:989-992, 2001.

[9] Bai Liang, Hu Yaali, "Feature analysis and extraction for audio automatic classification," Proc. of International Conference on Systems, Man and Cybernetics, vol.1, pp: 767-772, 2005.

[10] Yingxu Wang, "On Cognitive Informatics," Proc. of International Conference on Cognitive Informatics,pp:34-42, 2002.

[11]Sheng Tang, Yong-Dong Zhang, Jin-Tao Li et al. "Rushes Exploitation 2006 By CAS MCG." In Proc. TRECVID Workshop, Gaithesburg, USA , Nov. 2006.