A Hierarchical Framework for Movie Content Analysis: Let Computers Watch Films like Humans

Anan Liu[1,2,3], Sheng Tang[1,2], Yongdong Zhang[1,2], Yan Song[1,2], Jintao Li[1,2], Zhaoxuan Yang[3]
1. Virtual Reality Laboratory, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, 100190, China;
2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, 100190, China;
3. Department of Electronic Engineering, Tianjin University, Tianjin, 300072, China;
{liuanan,ts,zhyd,songyan,jtli}@ict.ac.cn, yangzhx@tju.edu.cn

Abstract

In this paper, we specially propose a hierarchical framework for movie content analysis. The purpose of our work is trying to realize computers' understanding for movie content, especially "Who, What, Where, How" which occur in the storyline by imitating human perception and cognition. The framework consists of two hierarchies. As for the low level part, we originally construct the human attention model with temporal information motivated by the Weber-Fechner Law to depict the variation of human perception in multiple modalities. As for the high level part, we focus on semantic understanding of different granularities of videos and simulate human cognition for movie content. Based on this hierarchical framework, we present its applications on semantic retrieval, video summarization and content filter. The promising results of users' subjective assessment indicate that the proposed framework is applicable for automatic analysis of movie content by computers.

1. Introduction

The movie industry is an active producer of video. Every year about 4,500 movies are released around the world spanning over approximately 9,000 hours of video [1]. Therefore, with such a massive amount of information, there is a great need of the way in which we can browse the movies conveniently and retrieve video clips with special semantic concepts.

Early researches in content-based video analysis mainly focus on video structurization and retrieval. The video data are news video, sports video and so on because they generally have some structure characteristics and the domain-based knowledge facilitates the video analysis. However, the retrieval results using low-level visual features are far from satisfaction. This fact has motivated the effort to perform semantic analysis. With the great success of semantic analysis in the video mentioned above, the research objective has extended to movies. The research on movie content analysis has lasted for about 10 years. Researchers mainly focus on two kinds of problems. The early work specializes in movie highlights extraction. In this area, the most representative work is respectively done by Brett Adams and Yu-Fei Ma. Due to the complex storyline and potential film grammar, it is very difficult for researchers to analyze movies as they have done in other kinds of video. In [2-5], Brett Adams et al. originally brought in the film grammar and presented an original computational approach for extraction of movie tempo to derive video saliency. Comparatively, Yu-Fei Ma et al. were engaged in applying human attention model on video analysis. In [6], they presented the construction of attention model and the application in video summarization in detail. Besides, in [27], by integrating film-making and psychology rules Hari Sundaram and Shih-Fu Chang presented an innovative method for film segmentation which is the basis for semantic analysis for movies. The research above mainly focuses on video saliency detection from the development of storyline and the variation of human perception. However, these successes can not really map the multi-modal low level features to semantic concepts. To bridge "Semantic gap", in the recent research, much work has been done on scene analysis and event detection in movies. By pattern analysis, some important semantic concepts are defined and thus low level features are integrated for semantic modeling. Lei chen et al in [7] incorporate audio and visual cues for dialog and action scene extraction. Simon Moncrieff et al. establish a correlation between the sound energy event types and horrific thematic content within film, thus enabling an automated mechanism for genre typing and scene content labeling in film in [8]. Based on this advanced research, movie clips can be annotated with semantic concepts for potential applications, such as navigation and retrieval.

Although some research has been done on movie content analysis, there exist two important problems. On one hand, the past work spans over many aspects of multimedia

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

technique while the analysis of their relationship is little. On the other hand, most of the past work focuses on the nature of videos and ignores the process of human understanding. It is well known that the ultimate goal of artificial intelligence is to make computers to think and act like humans. Therefore, as for movies which convey high level semantics of stories portrayed in video, we suggest imitating human perception and recognition for unknown things to realize the automatic movie content analysis by computers. The main contribution of our work is to propose a hierarchical framework for movie content analysis. The framework consists of two hierarchies. As for the low level part, we originally construct the human attention model with temporal information motivated by the Weber-Fechner Law to depict the variation of human perception in multiple modalities. As for the high level part, we focus on semantic understanding of different granularities of videos and simulate human cognition for movie content. With this framework, computers can mimic human behavior from "looking" to "thinking". Moreover, they can tell that "Who, What, Where, How" happen in the movies. Based on this hierarchical framework, we introduce its application on semantic retrieval, video summarization and content filter for movies. The promising results of users' subjective assessment indicate that the proposed framework is applicable for automatic analysis of movie content by computers.

The remainder of the paper is organized as follows. In Section 2, we present the hierarchical framework for movie analysis in details. Then, the experimental results and subjective assessment are illustrated in Section 3. In Section 4, we introduce its potential applications. At last, the conclusions and future work are stated in Section 5.

2. The hierarchical framework

From the viewpoint of psychology and physiology, it is intuitive that humans usually experience a process from "looking" to "thinking" for unknown things. Therefore, the proposed framework has two levels which imitate human understanding from perception to cognition. With the hierarchical framework, we try to realize automatic movie content analysis just as computers watch films with human brain.

2.1. Low level hierarchy for human perception

As a neurobiological concept, perception means the awareness of the elements of environment through physical sensation in [9]. Human beings receive and process certain types of external stimuli, including ocular and aural stimuli, and change the concentration of mental powers on the storyline. Therefore, the low level hierarchy of the framework focuses on depicting the variation of human perception based on human attention model.

2.1.1 Human attention model

The human attention model presented in the paper is comprised of three sub-models, namely visual, audio and textual sub-model. 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 the continuous change of "*t*" 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, $\triangle I$. Therefore, for constructing more proper attention model, we must consider the information in temporal domain. Weber-Fechner Law [10] is a rule in psychology and physiology which reflects the relationship between external stimuli and human perception. With this law, we improve the human attention model for videos (*M*) as follows:

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

where k and C are experiential values.

2.1.2 Methodology

In this section, we firstly introduce the low level features in multiple modalities which directly influence human perception. Then we focus on the computational method of features in different modalities in the uniform metric of information quantity.

(a) Selection of low level features

In visual modality, drastic and complex motion of objects will have great impact on viewers' emotional feelings and attract them greatly. Therefore, as visual features, motion intensity (MI) and complexity (MC) in [11] can be used for calculation of visual information.

In aural modality, the accompanying sound is also very

important to express the semantic meanings. Especially, large volume and high pace of audio influence viewers with strength and haste. Therefore, audio energy (AE) and audio pace (AP) in [5] are calculated for the constitution of auditory information.

In text modality, the captions with more words usually convey more semantic information. To make sure the correspondence between captions and video content in temporal domain, we use video caption extraction method in [12] and Hangwang OCR SDK, a commercial software for OCR recognition, instead of movie scripts. 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 sentence information, SI, by the ratio between the total number of noun, verb and adjective and the total number of words.

(b) 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 [13]-[14], Wang proposes that information is a more proper measure for human perception. By calculating the information of visual, auditory and textual modalities, we can represent the human perception descriptor quantitatively.

For the features mentioned above, we can see that *MI*, *AE* and *TE* represent energy and *MC*, *AP* and *SI* reflects frequency. 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 convert frequency into information and integrate them as follows to formulate the features in three modalities:

$$\begin{cases} VAM(t) = MI(t) * [-\log MC(t)] \\ AAM(t) = AE(t) * [-\log AP(t)] \\ TAM(t) = TE(t) * [-\log SI(t)] \end{cases} \quad (3)$$

where we consider the features representing energy as the weights of those representing information. Then linear fusion method can be used for *HAM(t)* formulation.

2.2. High level hierarchy for human cognition

Cognition is a mental process or ability of reasoning and judgment based on awareness and perception [13]. As for a movie, viewers are usually impressed by some important semantics, such as "Who, What, Where and How" in the highlights by "thinking" as well as "looking". Therefore, the high level hierarchy of the framework focuses on semantic analysis in movie content from the four aspects mentioned above.

2.2.1 Who

"*Who*" means the main characters in one movie. To judge the main characters, we propose the voting-based method. Firstly, we implement face detection and recognition on key frames and cluster the results. Each category corresponds to one character. By majority voting, the most important characters can be found. Secondly, by face recognition [15], the occurrence of main characters in each shot or scene can be detected.

2.2.2 What

"*What*" denotes events conveying semantics, namely, what happens in the story. Dialogue event and exciting event usually account for most of a film. Therefore, we implement dialogue event detector, action event detector and warfare event detector (two kinds of exciting events) on the movie.

(a) Dialogue event detection

As for dialogue event, there usually appears speech in aural modality and captions in textual modality. Different from the former methods mainly depending on audio information, we integrate aural and textual modalities for dialogue detection because audio in movies is usually complicated and is difficult for classification. Firstly, we implement the audio classification on videos, the method of which will be introduced in Section 2.2.3. Then, we will use textual attention sub-model to extract dialogue events by edge analysis in [5] for candidates of speech scenes.

(b) Action event detection

As for action event, there usually exists drastic and complicated object motion and large volume and high pace of audio. Therefore, we integrate the film grammar and human perception for action event detection with the algorithm in [16].

(c) Warfare event detection

As for warfare event, there exists gunfire as well as the similar characteristics with action event in visual and aural modalities. Therefore, on the basis of action event detector, we model the SVM-based gunfire descriptor with the method in [17]. If both detectors give the positive decisions at the same time, warfare event can be detected.

2.2.3 Where

"*Where*" means the surrounding of events. Here we not only analyze the spatial place of the occurrence of events but also the audio classification which is the background of the happening of events and is an important cue for movie analysis.

(a) Spatial place

We use SVM-based concepts modeling framework in [18] to model the related concepts proposed in LSCOM [19], including: indoor, outdoor, river, ocean, hill, forest. These classifiers can be implemented on keyframes of movies to judge the spatial place where the event happens.

(b) Audio classification

The audio stream is firstly segmented into

non-overlapped 20-ms short time frame. Then five frame-level audio feature: Short-Time Energy Function, Short-Time Zero-Crossing Rate, Frequency energy, Sub-band energy ratio, Mel-frequency cepstral coefficients, are extracted to represent the character of each short time frame. Finally, each audio clip is classified into four kinds: silence, speech, music and others with the SVM-based approach in [20].

2.2.4 How

"*How*" is related with affective understanding for movie content. Due to the seemingly inscrutable nature of emotions and the difficulty of bridging the affective gap [21], affective analysis for videos become a challenging research area. Our ongoing research mainly consults to [22]. By audiovisual features extraction, we realize the scenes classification grounded upon psychology and cinematography, including: Neutral, Fear, Joyous, Surprise, Anger and Sad.

By associating the four aspects presented above with temporal constrains in [23], the relationship between "Who, What, Where and How" is founded. Then, computers can "think" like humans and deeply analyze the movie content with the basic ability of human cognition.

3. Experiments and evaluation

3.1. Experiments

To demonstrate the effectiveness of the proposed hierarchical framework for movie analysis, we select ten movies for experiments. The details of each movie will be presented in Table 1. As for the input movies, video structurization, including: shot boundary detection & keyframe extraction in [24], and scene boundary detection in [25], is implemented on them. Then movie content is analyzed depending on the hierarchical framework. Figure 1 shows the experimental results for the movie, "Crouching Tiger, Hidden Dragon". In Figure 1. (a), for convenient navigation of movies, we put the video into a two-dimensional Cartesian Coordinates in the hierarchical browsing subwindow. Along the vertical dimension are the key frames of the same scene while along the horizontal dimension is the linear temporal dimension of scene sequences. Figure 1. (b-d) show human attention changes with the development of storyline in three modalities. In Figure 1. (e-h), "Who, What, Where and How" are separately annotated for movie understanding. From Figure 1, we can see that complicated movie content can be automatically analyzed and displayed by computers under the hierarchical framework. Therefore the proposed framework can be widely used for potential application, such as video retrieval, summarization and filter.

3.2. Evaluation

Although much work has been done on movie content analysis, there is no standard method to evaluate performance of the framework. The assessment is a strong subjective task. To evaluate the proposed hierarchical framework, we have carried out a user study experiment and invited 10 persons to give their subjective scores to each step of the framework. Besides, we quantify the subjective assessment of each step into five levels, from 5 to 1 which means changes from good to bad, respectively. Then the average score of each step for the movie is calculated. The statistical results are listed in Table 2.

From the columns of Table 2, we can see that the integral results for the framework are satisfied and the average score for each step is over the neutral score, 2.5. We will analyze the results in detail as following: (a) In the low level hierarchy of the framework, by mapping the low level features into three attention sub-models, we implement quantitatively representing the variation of human attention. Because the human attention in different modalities well matches three attention sub models, the scores in the first three columns are very high. (b) In the high level hierarchy of the framework, by modeling specific semantic descriptors, we realize the detection of characters, events, surroundings and emotion. Due to the successful technique in face detection and recognition and clear definition of events, the fourth and fifth columns show that the results are still satisfied although they are relatively lower than those in the first three columns. However, due to the difficulties in modeling the surrounding descriptor with limited data and the complicated nature of human emotion, the results in the sixth and seventh columns are a little low and need to be improved in the future.

From the rows in Table 2, we can see that the results of movie content analysis under the hierarchical framework are also satisfied and the lowest average score is 3.70. We will analyze the results in detail as following: (a) As for the first six movies corresponding to the first six rows, the subjective evaluation is high due to the relatively simple storyline. The framework presents a good structure with semantic annotation for the content of each movie. (b) As for the last four movies, the scores are a little low because of the following two reasons: the scenarios of the movies are very complicated and lead to the difficulties in the discriminable description of semantics with low level features; the advanced techniques, such as ceaseless changes in illumination, color, motion, make it difficult to detect specific events, especially "Where and How" which get the low scores.

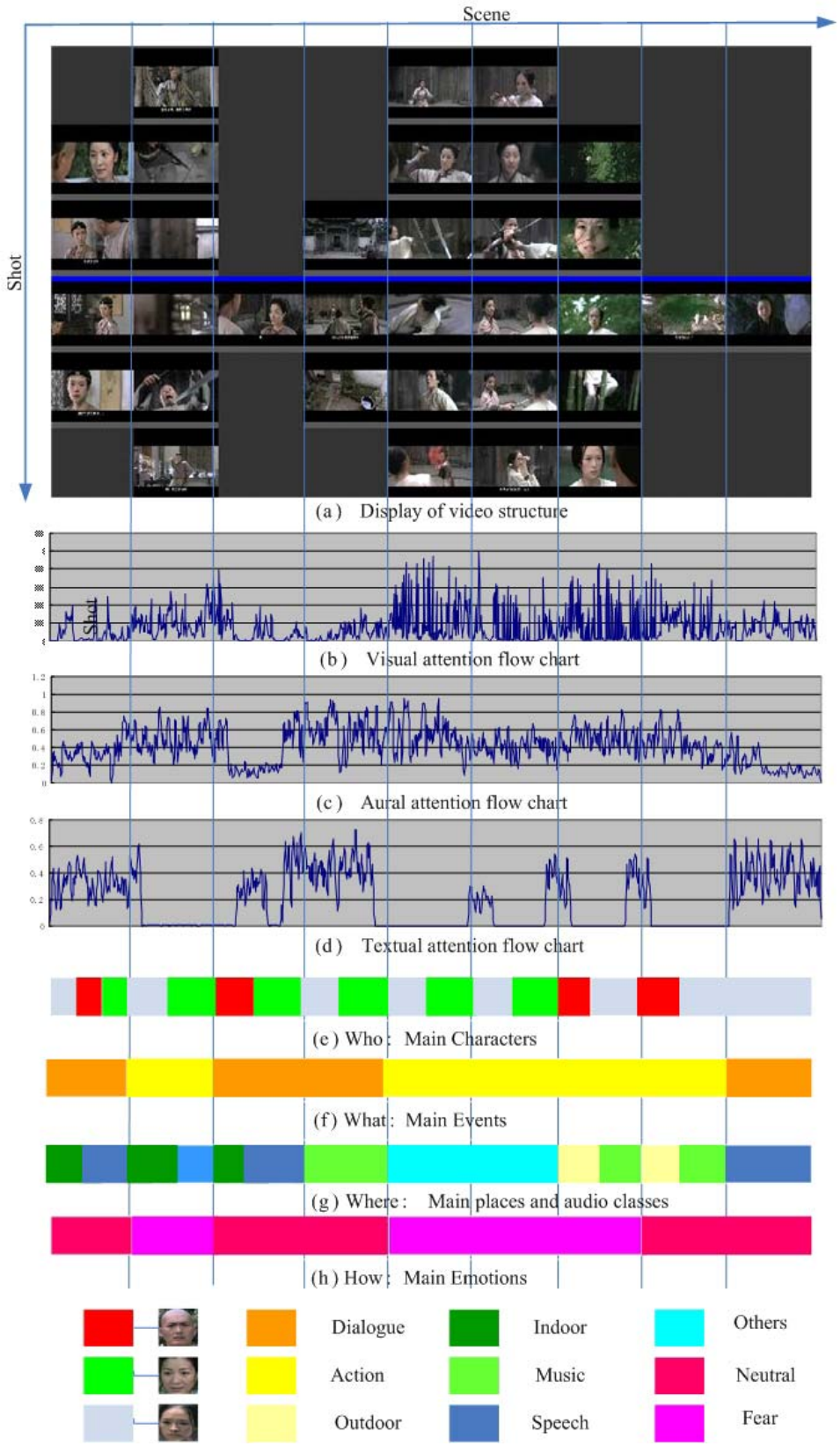


Figure 1: The experiment results of the hierarchical framework for the movie, “Crouching Tiger, Hidden Dragon”.

Table 1. Detailed information about each movie

Movie Title	Fear-less	Crouching Tiger, Hidden Dragon	Fist of Legend	Gladiator	The Matrix1	Minority Report	Enemy At The Gates	Wind talkers	Pearl Harbor	Thin Red Line
Runtime (min)	110	120	103	155	113	145	131	134	183	170
Movie Genre	Action+Drama				Action+Sci-fi		Action+War			
File Format	MPEG-1									
Audio Format	16 bits/sample, mono, 22kHz									
Delivery:(f/s)	25	30	25	30	30	30	30	25	30	25

Table 2. Experimental results

Movie Title	Visual Attention	Aural Attention	Textual Attention	Who	What	Where	How	Average per film
Fearless	4.6	4.3	5.0	3.9	4.1	3.0	3.4	4.04
Crouching Tiger, Hidden Dragon	4.8	4.6	5.0	4	4.2	3.5	3.6	4.24
Fist of Legend	4.5	4.2	4.8	3.7	4.0	3.2	3.2	3.94
Gladiator	4.0	4.2	5.0	3.5	3.8	3.4	3.1	3.86
The Matrix	4.2	4.5	5.0	3.8	3.6	3.0	3.3	3.91
Enemy At The Gates	4.5	4.6	4.6	3.4	3.9	3.3	3.1	3.91
Wind talkers	4.1	4.3	5.0	3.7	3.6	2.7	2.8	3.74
Pearl Harbor	4.6	4.1	4.5	3.6	3.4	3.2	3.0	3.77
Thin Red Line	3.8	4.2	5.0	3.5	3.7	2.9	3.1	3.74
Minority Report	4.3	4.1	5.0	3.5	3.3	2.8	2.9	3.70
Average per step	4.34	4.31	4.89	3.66	3.76	3.10	3.15	——

4. Application

Based on this hierarchical framework for movie content analysis, we present its potential applications on semantic retrieval, video summarization and content filter.

4.1. Semantic retrieval

Although researchers have been engaged in semantic retrieval for many years, the query is still restricted to the phrase which corresponds to one semantic. However, humans used to query with a sentence containing "Who, What, Where and How". Therefore, we not only need to detect the semantics in different classes but also need to associate them to realize friendly interaction between humans and computers.

In our framework, there is an important step to associate the detected semantics with time constrains after hierarchical analysis. Thus, we found the relationship for multiple classes of semantics to support the complicated humans' queries. As for the movie "Crouching Tiger, Hidden Dragon", the most famous scene is that Mubai Li and Jiaolong Yu fight in the bamboo cluster. For this query, we map it as following: "Who" to "Mubai Li and Jiaolong Yu", "What" to "Action event", "Where" to "bamboo cluster". Due to the limited concepts for "where" which can be detected, we classify "bamboo cluster" into the class of "outdoor". Then the feedback is shown in Figure 2.

4.2. Video summarization

Video summarization is very useful for making video preview. Nowadays, editors usually manually select highlights to make movie trailers. Because it is laborious and time consuming, there is great need for automatic video summary generation.

Based on the movie content analysis under the proposed framework, we proposed two methods for video summary. The first one is an interactive method. Human attention model gives the attention value for different shots and then editors subjectively select the highlights as shown in Figure 3.(a). The second one is an automatic method. By inputting sentence-level requests, the summary is automatically generated. This method is based on semantic retrieval presented in Section 4.1. After detecting the scenes corresponding to the queries, the shots including the scene key frames (the frames under the blue horizontal line) are selected as the highlights as shown in Figure 3.(b). Then, the selected video clips are sent to the bottom storyboard

subwindow. The clips in the box will be connected together from left to right to form a new video clip.

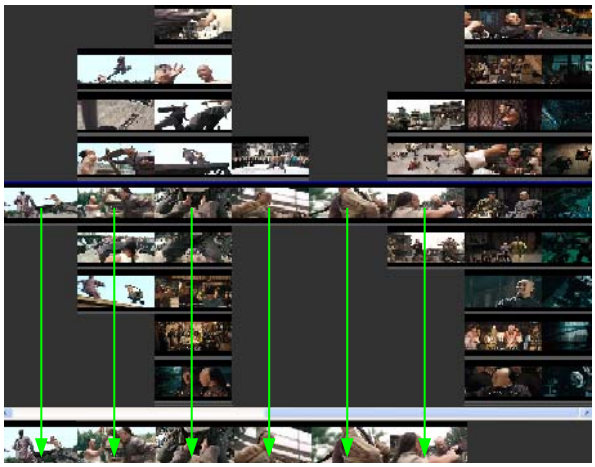
4.3. Content filter

Nowadays there exists great need for detecting and blocking harmful video content, including sex and violence and so on.

After content analysis and semantic annotation under this framework, we can directly find the harmful sections and blocking them with corresponding detectors. We model the violence detector mentioned above and sex detector depending on the method in [26]. Then the harmful clips in the movie can be detected and blocked as shown in Figure 4.

Figure 2: Feedback of sentence-level semantic retrieval for the movie, "Crouching Tiger, Hidden Dragon"

(a) Example of interactive method for video summarization.



(b) Example of automatic method for video summarization.

Figure 3: Video summary with two methods for the movie, "Fearless".

Figure 4: Blocking sex shots for the movie, "La Belle".

5. Conclusion and future work

In this paper, we propose a hierarchical framework for movie content analysis. The framework consists of two hierarchies. As for the low level part, we construct the human attention model to depict the variation of human perception in multiple modalities. As for the high level part, we focus on semantic understanding of videos and simulate human cognition for movie content. With this framework, computers can mimic human behavior from "looking" to "thinking" and watch films like humans. Based on this hierarchical framework, we introduce its application to semantic retrieval, video summarization and content filter. The promising results of users' subjective assessment indicate that the proposed framework is applicable for automatic movie content analysis by computers.

Although some work has been done on this area, the state

of the art research is not enough due to the complex storyline and the potential semantics. Therefore, more attention will be paid on this challenging research in the following aspects: (a) Integrating the related subjects to mine the proper descriptors in multimodal for semantics representation; (b) Analyzing the relationship of the descriptors for modeling semantic event; (c) Paying more attention on effective analysis to imitate high level human cognition.

Acknowledgement

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), 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).

References

[1] Howhard D. Wactlar. The Challenges of Continuous Capture, Contemporaneous Analysis and Customized Summarization of Video Content, CMU, USA.

[2] Brett Adams, Chitra Dorai, Svetha Venkatesh. Novel Apporach to Determining Tempo and Dramatic Story Sections in Motion Pictures. Proc. of IEEE ICIP, 2000.

[3] Brett Adams, Chitra Dorai, Svetha Venkatesh. Study of Shot Length and Motion as Contributing Factors to Movie Tempo. Proc. of ACM Multimedia 2000.

[4] Brett Adams, Chitra Dorai, Svetha Venkatesh. Role of Shot Length in Characterizing Tempo and Dramatic Story Sections in Motion Pictures. Proc. of IEEE Pacific Rim Conference on Multimedia, 2000.

[5] Brett Adams, Chitra Dorai, Svetha Venkatesh. Toward Automatic Extraction of Expressice Elements from Motion Pictures: Tempo. IEEE Transactions on Multimedia, 2002.

[6] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, et al, "A User Attention Model for Video Summarization," Proc. of the tenth ACM international conference on Multimedia, 2002.

[7] Lei Chen, Sariq J. Rizvi, M.Tamer Ozsu. Incorporating Audio Cues into Dialog and Action Scene Extraction. Proc. of SPIE Storage and Retrieval for Media Databases, 2003

[8] Moncrieff S, Venkatesh S, Dorai C. Horror film genre typing and scene labelling via audio analysis. Proc. of ICME, Baltimore, USA, 6-9 July 2003, Vol 2, p193-196.

[9] Junyong You, Guizhong Liu, Li Sun, et al, A Multiple Visual Models Based Perceptive Analysis Framework for Multilevel Video Summarization, IEEE Transactions on Circuits and Systems for Video Technology, 2007.

[10] Selig Hecht, "The Visual Discrimination of Intensity and the Weber-Fechner Law," Journal of General Physiology, Vol 7, 1924.

[11] Anan Liu, Jintao Li, Yongdong Zhang, et al, Human Attention Model for Action Movie Analysis. Proc of 2nd International Conference on Pervasive Computing and Applications, 2007, UK.

[12] Rainer Lienhart, Axel Wernicke, Localizing and Segmenting Text in Images and Videos. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 4, April 2002.

[13] Yingxu Wang, On Cognitive Informatics, Proc. of the First IEEE International Conference on Cognitive Informatics, 2002.

[14] Yingxu Wang, On a New Frontier: Cognitive Informatics, Invited Talk, Proc. of the 7th International Conference on Object-Oriented Information Systems, Canada, 2001.

[15] Zhao Ming,Chen Chun,Li S Z,et al. Subspace analysis and optimization for AAM based face alignment. In: Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition. Seoul,South Korea,2004.290-295.

[16] Anan Liu, Jintao Li, Yongdong Zhang, et al. An Innovative Model of Tempo and Its Application in Action Scene Detection for Movie Analysis. Proc. of International Workshop of Application of Computer Vision (WACV), USA, 2008.

[17] Moncrieff, S., Dorai, C., Venkatesh, S. Detecting Indexical Signs in Film Audio for Scene Interpretation. Proc. of ICME, 2001.

[18] Sheng Tang, Yong-Dong Zhang, Jin-Tao Li, et al., TRECVID 2007 High-Level Feature Extraction By MCG-ICT-CAS. Proc. of TRECVID Workshop, USA, 2006.

[19] LSCOM Lexicon Definitions and Annotations (Version 1.0). Columbia University ADVENT Technical Report #217-2006-3, March 2006.

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

[21] A. Mittal and L.F. Cheong, Framework for synthesizing semanticlevel indexes, Multimedia Tools Appl., vol. 20, no. 2, pp. 135–158, 2003.

[22] Hee Lin Wang, Loong-Fah Cheong. Affective Understanding in Film. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 16, No. 6, June 2006.

[23] Andrew Salway, Bart Lehane, Noel E. O'Connor, Associating Characters with Events in Films. Proc. of CIVR'07, July 9–11, Amsterdam, The Netherlands, 2007.

[24] Yueting Zhuang, Yong Rui, Thomas S. Huang et al. Adaptive key frame extraction using unsupervised clustering. Image Processing, ICIP 1998.

[25] Zeeshan Rasheed, Mubarak Shah. Detection and Representation of Scenes in Videos. IEEE Transaction on Multimedia, Vol7, NO.6, December, 2005.

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

[27] Hari Sundaram, Shih-Fu Chang, "Determining computable scenes in films and their structures using audio-visual memory models," ACM MM, 2000, California, USA, 2000.