Affective Content Analysis in Comedy and Horror Videos by Audio Emotional Event Detection

Min Xu, Liang-Tien Chia, Jesse Jin*
CeMNet, School of Computer Engineering, Nanyang Technological University, Singapore
School of Design, Communication and Information Technology, the University of Newcastle*
Email: {mxu, asltchia}@ntu.edu.sg; jesse.jin@newcastle.edu.au*
Phone: +65 67906579, +61 (2) 49217912*

Abstract

We study the problem of affective content analysis. In this paper we think of affective contents as those video/audio segments, which may cause an audience's strong reactions or special emotional experiences, such as laughing or fear. Those emotional factors are related to the users' attention, evaluation, and memories of the content. The modeling of affective effects depends on the video genres. In this work, we focus on comedy and horror films to extract the affective content by detecting a set of so-called audio emotional events (AEE) such as laughing, horror sounds, etc. Those AEE can be modeled by various audio processing techniques, and they can directly reflect an audience's emotion. We use the AEE as a clue to locate corresponding video segments. Domain knowledge is more or less employed at this stage. Our experimental dataset consists of 40-minutes comedy video and 40-minutes horror film. An average recall and precision of above 90% is achieved. It is shown that, in addition to rich visual information, an appropriate usage of special audios is an effective way to assist affective content analysis.

Keywords: Affective content, Audio emotional event

1. Introduction

With the increasing amount of multimedia data, indexing and retrieval is becoming a crowded research area. Lots of research has been conducted on video structuring, event detection, and semantics modeling. These works try to provide an effective and efficient way to manage and access multimedia databases. More recently, researchers have revealed the significance of affective analysis from a personalized media point of view [1-4]. For example, many users favor a flexible tool to quickly browse the funniest or the most sentimental segments of a movie, as well as the most exciting parts of a sports game video. Compared with traditional video indexing, affective content analysis puts much more emphasis on the audience's reactions and emotions. Like semantic analysis, the affective content analysis is also challenging due to the gap between low-level perceptual features and high-level understanding. However, affective content emphasizes the factors that influence the users' attention, valuation, and memory for video contents. It does not require a deep understanding of video/audio contents. It is possible to employ some mature indexing techniques to capture the affective content by incorporating domain knowledge. To some extent, we may think of the affective analysis as a mid-layer between low-level video structuring and high-level semantics modeling.

We briefly review some related works. In [1], Alan et. al. utilized the features of motion, color, and audio to represent *arousal* and *valence*. Kang [2] employed HMM to mapped low level visual features (i.e. motion, color, shot cut rate) to high level emotional events (i.e. fear, sadness and joy). The users' feedbacks [3] were further used to adjust weights to emphasize different features in [2]. In [4], four sound energy events were identified that convey well established meanings through their dynamics. These dynamics are expected to portray and deliver certain affects and sentiments inherent to the genre of horror film.

Recently, some research works have worked on music mood recognition [5, 6]. Sounds were shown to have emotion-related meanings. A closely related work is the successful detection of excited and non-excited segments from sports video by audio analysis [7-9]. It is a typical example of audio in affective content analysis.

In this paper we want to utilize sounds in movies to detect some audio emotional events (AEE). The AEE is employed to locate the audio/video segments with affective contents. It is well known that, in sound film, movie editors usually use some specific sounds and music to highlight emotional atmosphere and promote dramatic effects. Currently we choose comedy and horror movies, as these two genres feature strong emotional actions (e.g. cheer or fear). Our AEE includes laughing, horror etc. With the AEE and simple video shot boundary, we determine the boundary of affective content by heuristic rules. Compared with previous works [1-4], our work possesses unique features in terms of using movie audio-track production knowledge.

In Section 2, we explain the reason why we use audio emotional event to index video affective content. Section 3 briefly introduces how to identify audio emotional events. Affective content detection is presented in Section 4. Section 5 discusses experimental details and results. Conclusions and future work is presented in Section 6.

0-7803-9332-5/05/$20.00 ©2005 IEEE

2. Video Affective Content and Audio Emotional Event

People watch entertainment videos for the unique experience of different emotional events. For example, why are horror movies continuously in demand? Watching horror films lets us counter our hidden fears, share them with other viewers, and eliminate the terror by meeting it head-on [10]. According to audience's preference, those video/audio segments, which may cause audiences' strong reactions or special emotional experiences, such as laughing or fear are regards as affective contents.

To select affective contents is a challenge task. According to affective theory, scientists bicker about a definition of emotion, but they agree that emotion is not logic, and that sometimes emotions can affect a rational decision-making process [11]. In multimedia field, users may prefer "emotional decisions" to find affective content because emotional factors directly reflect an audience's attention and evaluation.

In this paper, we detect video/audio segments which make audience laugh in comedy and the scary segments in horror films as affective contents. When watching some comedy soap operas, after laughable scenario, we can hear some canned laughter. In horror films, with a horrific scenario, horror sounds are used to emphasis on the scary atmosphere and increase the dramatic effects. Therefore, laughing and horror sounds are significant to locate probable position of affective contents.

3. Audio Emotional Event Identification

Audio, which includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of video. Audio emotional events are defined as some specific audio sounds which have strong hints to video affective content. Especially in some video domain, such as sports video, comedy and horror movies, some audio sounds (e.g. excited audience sounds, audience's laughing or horror sounds) have strong relationships to the emotions of audiences.

In comedy video, except audio emotional event (canned laughter), there are non-emotional audio events, such as dialogues, silence, music and other environmental sounds. These audio events cover most audio sounds in comedy. Similarly in horror films, besides horror sounds, there are dialogue, silence and others. Subsequently, the laughing and horror sounds identification from other audio sounds can be regard as the task of audio classification. We classify audio sounds into 5 pre-defined classes (canned laughter, dialogue, silence, music and others) for comedy and 4 pre-defined classes (horror sounds, dialogue, silence and others) for horror films respectively.

3.1 HMM-Based Audio Emotional Event Identification

Audio signal exhibits the consecutive changes in values over a period of time, where variables may be predicted from earlier values. That means strong correlation exists. In consideration of the success of HMM in speech recognition, we propose our HMM based audio classification method. The proposed method includes three stages, which are feature extraction, data preparation and HMM learning, as shown in Fig.1.

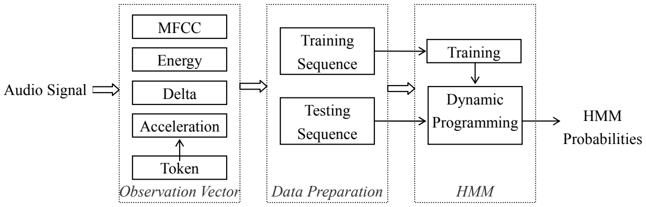

Fig. 1. Proposed audio keywords generation system

As illustrated in Fig.1, selected low-level features are firstly extracted from audio streams and tokens are added to create observation vectors. These data are then separated into two sets for training and testing. After that, HMM is trained then re-estimated by using dynamic programming. Finally, according to the maximum posterior probability, the audio keyword with the largest probability is selected to label the corresponding testing data. More details can be found in [9].

3.2 Post-Processing

Since sound itself has a continuous existence, it may be impossible to have sudden changes in the occurrence of audio events. Moreover, dominant audio events do sometimes mix with other classes. For example, there may be one or two dis-continuous samples of silence detected in continuous dialogue. Considering sequencing order of audio, we regard any audio event change within 1 or 2 seconds in the audio events sequence as an error and will be eliminated by sliding window majority-voting.

However, some of the horror sounds are sudden and short. According to our experience of watching horror movie, we are always jolted by a sudden blare which takes place in a relative silent audio track. Since most of these blares are shorter than 1 second, it is near impossible to detect by HMM-based identifier. Figure 2 shows blares examples. Furthermore, by our sliding window majority-voting, some detected sharp horror sounds are wrongly corrected as errors. In horror films, compared to other audio sounds, the amplitude of blares are large. Moreover, to enhance the scared effect, blares always happened within sounds whose amplitude is relative small. Therefore, by calculating the amplitude change of audio signal, the blares are easily detected.

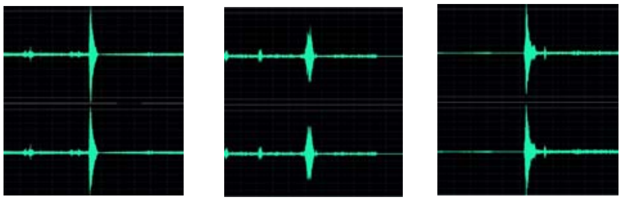

Fig. 2. Two channels of blares in horror movies

4. Video Affective Content Located by Audio Emotional Events

Video affective content selection by audio emotional events is a challenging research topic, since video/audio segments' boundary selection has to be considered. In horror films, because the horror sounds take place synchronously with horror scenario, it is simple to select those shots in which horror sounds has been identified as horrific affective contents. However, canned laughter appearing after funny scenario makes affective segments selection a little complicated for comedy videos. We make use of other audio events such as silence and music and combine video shot boundary information to make some heuristic decision rules for laughable affective content.

In comedy video, especially some comedy soap opera, dialogue is the most familiar scenario. The camera may switch among the persons in a dialogue. After several shots, audiences may be amused by words or actions during dialogue. Therefore, the laughable segments may be detected from the dialogues before audience's laughing. By checking duration of other audio events we determine the starting points of laughable audio segments (LAS). Later on, we select those video shots, which have more than half of the shot length overlapping with the LAS as comedy affective content. A possible video/audio structure from comedy videos is shown in Figure 3.

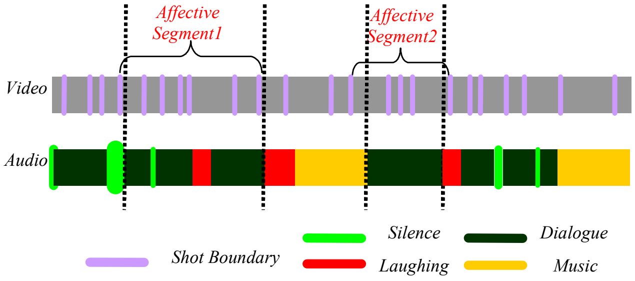

Fig. 3. Video/Audio structure from comedy videos

The process of LAS selection is listed step by step as following: (*t* is the starting point of laughing)

1) Check the duration between two continuous laughing. If (*L(t)-L(t-1)<Threshold1*), go to 3) otherwise do 2) and skip 3)
2) Search forward music or silence from *t*.
3) If (no any end of LAC detected for current sequence), set t as end of LAC. Search forward from *t-1*.
4) If (*detected silence or music duration> Threshold2*), set the end of silence or music as beginning of AAS. *t=t-1*. Go back to 1).

5. Experiments

In this section, we list our experimental details and results on both audio emotional event identification and affective content detection. The ground truth is manually labeled by 5 students. They label audio emotional events by listening to the audio stream without watching the video. The audio stream labeled with audio events are used for training and testing. By watching the video and listening to audio at same time, affective contents are labeled based on students' understanding.

5.1 Audio emotional event identification

The audio samples come from a 40 minutes comedy soap opera (Friends) and 40 minutes of horror films (Korea, Face). They are collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. We used half the dataset for training and half for testing.

The audio signal is segmented into 20 ms per frame which is the basic unit for feature extraction. Mel-Frequency Cepstral Coefficient (MFCC) and Energy are selected as the low-level audio features as they are successfully used in speech recognition and further proved to be efficient for audio event identification in [9]. Delta and Acceleration are further used to accentuate signal temporal characters for HMM [9].

For the HMM learning, different HMM structure may model different states transition process, which could influence results. Moreover, as each kind of audio events have their own durations, we need to choose appropriate sample length for training. Since HMM is not the focus of this paper, we simply use left-to-right HMM with 4 states. 1 second is selected for sample length since most audio emotional events last longer than 1 second. Table 1 and 2 show the audio event classification results. The results in parenthesis are before post-processing.

Table 1: Audio emotional event identification for Comedy

	Dialogue	Music	Canned Laughter	Silence	Others
Recall	98.98% (98.73%)	96.21% (90.24%)	99.17% (98.89%)	97.18% (94.60%)	100% (97.21%)
Precision	99.04% (99.05%)	98.19% 97.37%	99.19% (98.72%)	96.57% (94.60%)	86.35% (50%)

Table 2: Audio emotional event identification for Horror

	Dialogue	Silence	Horror Sounds	Others
Recall	95.29% (89.81%)	91.72% (84.85%)	96.66% (79.88%)	90.89% (77.78%)
Precision	97.89% (96.43%)	88.21% (75.68%)	92.99% (88.24%)	81.96% (64.81%)

Compared with horror movies, comedy soap operas are mainly indoor scenes with simple environmental sounds. This may be a reason why the audio classification results of comedy are much better than horror films. The performance of HMM-based identifier is not satisfactorily for horror sounds because some horror sounds' duration are less than the HMM sample length (1 second). After sliding window majority-voting elimination and blares detection by wave amplitude change detection, results of horror movie are evidently improved. From the final experimental results, we find the audio emotional events are identified exactly. This makes affective content detection possible after audio emotional event identification.

5.2 Video affective content analysis

The results of affective content detection in comedy videos and horror movies are listed in Table 3. Most of the affective content can be detected. However, the precision is not very satisfactory. For comedy video, it could be because some segments make audiences laughing but not labeled in our ground truth. In horror movie, some horror sounds may just be used to highlight the overall horrific atmosphere instead of taking place synchronously with scared scenario.

Table 3: Affective content detection results

	Comedy videos	Horror movies
Recall	97.61%	97.11%
Precision	91.3%	90.68%

5.3 Discussion

Audio emotional events indicate the probable position of video affective content, which is very significant and it provide a focus for future analysis by other cues instead of blindly analyzing video from beginning to the end.

Although the affective content detection results show that the proposed method is effective and efficient on our test data, this rule-based method may have some limitations to other video source, i.e. the rules may not be generic. We are extending the research to take advantage of video information. Especially, in horror movie, with horror scenario taking place, the visual scene turns to dark and the rate of shot change may be reduced. Moreover, according to our experience, the horror scenario may take place only when one or two persons appear in the scene instead of a crowed scene. Thereby, face detection may be added to our affective content detection system. From caption analysis point of view, the text of dialogue will provide important cues to affective analysis. In future, the affective content analysis will be broadened to multiple-modalities analysis including visual, audio and text domains.

6. Conclusions

In this paper, a comprehensive attempt of affective content detection for comedy and horror films has been completed. Affective content analysis provides probability to access multimedia database by audience's emotions and preference. Compared to semantic analysis, affective content avoids in depth video understanding such as detailed events in video or what the movie is about. At the early research stage, we mainly rely on audio emotional event to locate probable position of affective content. Initial experimental results of affective content detection show, with the help of other audio events and video shot boundary information, detected audio emotional events efficiently play remarkable roles for affective content detection. Our research is extending to analysis other modalities' emotional cues.

7. REFERENCES

[1] A. Hanjalic and L.-Q. Xu, "User-oriented Affective Video Content Analysis", In Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries' 01

[2] H.-B. Kang, "Affective Content Detection using HMMs", In Proc. of ACM Multimedia' 03.

[3] H.-B. Kang, "Emotional Event Detection Using Relevance Feedback", In Proc. of International Conference on Image Processing' 03.

[4] S. Moncrieff, C. Dorai and S. Venkatesh, "Affect Computing in Film through Sound Energy Dynamics", In Proc. of ACM Multimedia' 01.

[5] D. Liu, L. Lu and H. -J. Zhang, "Automatic Mood Detection from Acoustic Music Data", In Proc. of International Symposium on Music Information Retrieval' 03

[6] H. Katayose, M. Imai, and S. Inokuchi, "Sentiment Extraction in Music", In Proc. of International Conference on Pattern Recognition' 02.

[7] M. Xu, L-Y. Duan, C.-S. Xu, Q. Tian, "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video," In Proc. of IEEE International Conference on Acoustic, Speech, and Signal Processing' 03

[8] M. Xu, N. C. Maddage, C.-S. Xu, M. Kankanhalli, Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video," In Proc. of International Conference on Multimedia & Expo' 03

[9] M. Xu, L.-Y. Duan, J. Cai, L. –T. Chia, C. -S. Xu, Q. Tian, "HMM-Based Audio Keywords Generation for Sports Video Analysis," In Proc. of IEEE Pacific Rim Conference on Multimedia' 04

[10] Astrid Bullen, "A Short History of the Horror Film", http://www.chiff.com/a/horror-movie.htm.

[11] Rosalind W. Picard, "Affective Computing," The MIT Press, Cambridge, Massachusetts London, England.