Multimed Tools Appl (2014) 70:757–779
DOI 10.1007/s11042-012-1046-8

A three-level framework for affective content analysis and its case studies

Min Xu · Jinqiao Wang · Xiangjian He · Jesse S. Jin · Suhuai Luo · Hanqing Lu

Published online: 28 March 2012

Abstract Emotional factors directly reflect audiences' attention, evaluation and memory. Recently, video affective content analysis attracts more and more research efforts. Most of the existing methods map low-level affective features directly to emotions by applying machine learning. Compared to human perception process, there is actually a gap between low-level features and high-level human perception of emotion. In order to bridge the gap, we propose a three-level affective content analysis framework by introducing mid-level representation to indicate dialog, audio emotional events (e.g., horror sounds and laughters) and textual concepts (e.g., informative keywords). Mid-level representation is obtained from machine learning on low-level features and used to infer high-level affective content. We further apply the proposed framework and focus on a number of case studies. Audio emotional

M. Xu (✉) · X. He
Center for Innovation in IT Services and Applications, University of Technology, Sydney, Australia
e-mail: min.xu25@gmail.com

X. He
e-mail: xiangjian.he@uts.edu.au

M. Xu · J. Wang (✉) · H. Lu
National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: jinqiao@nlpr.ia.ac.cn

H. Lu
e-mail: luhq@nlpr.ia.ac.cn

J. S. Jin · S. Luo
School of Design, Communication and IT, University of Newcastle, Callaghan NSW 2308, Australia

J. S. Jin
e-mail: jesse.jin@newcastle.edu.au

S. Luo
e-mail: suhuai.luo@newcastle.edu.au

event, dialog and subtitle are studied to assist affective content detection in different video domains/genres. Multiple modalities are considered for affective analysis, since different modality has its own merit to evoke emotions. Experimental results shows the proposed framework is effective and efficient for affective content analysis. Audio emotional event, dialog and subtitle are promising mid-level representations.

Keywords Affective content analysis · Mid-level representation · Multiple modality

1 Introduction

With the exponential growth in the production of videos and the development of personalized multimedia services, increasing number of users have their focuses in personal accessing videos. Developing efficient methods to analyze, index and organize videos is an active research area. Recently, video affective content analysis attracts more and more research efforts. Affective content is defined as those video/audio segments which are able to cause viewers' strong reactions or special emotional experiences, such as cheer or fear. In most cases, viewers might prefer to watch affective video content since these content evoke viewers' emotions and reflect their attention, evaluation and memory. Video highlight usually overlaps with affective content.

Movies, sitcoms and TV shows constitute a large portion of entertainment industry. Affective content analysis for entertainment videos attracts increasing research efforts [1, 2, 5, 7, 11–13, 15, 19, 23, 27, 28, 34]. In [19], sound energy was used for film affective computing. Kang [15] employed HMM on motion, color, shot cut rate to detect emotional events. Hanjalic and Xu [12] utilized the features of motion, color, and audio to represent arousal and valence. Rasheed et al. [23] presented a framework to classify films into genres based on visual cues. Audio affective features were mapped onto a set of keywords with predetermined emotional interpretations. These keywords were used to demonstrate affect-based retrieval on a range of feature films [5]. Hanjalic [11] discussed the potential of the affective video content analysis for enhancing the content recommendation functionalities of the future PVR (personal video recorder) and VOD (video on demand) systems. In [7], modules were developed for detecting video tempo and music mood. A holistic method of extracting affective information from the multifaceted stream was introduced in [28]. Arifin and Cheung [1], presented a FPGA-based system for modeling the arousal content based on user saliency and film grammar. Further study detected affective content based on the pleasure-arousal-dominance emotion model [2]. A Support vector regression model was designed based on visual and auditory features for music video affective analysis, visualization, and retrieval [34]. In Irie et al. [13] introduced a latent topic driving model to classify movie affective scenes. They considered temporal transition characteristics of human emotion referring to Plutchik's emotion wheel. Considering contextual information, a Bayesian classification framework was presented for affective movie tagging [27]. Most recently, research on affective image classification proposed efficient affective features inspired by psychology and art theory and mapped these features to affective classes by using Support Vector Machine [18].

As reviewed above, most of the existing methods of video affective analysis map affective features directly to emotions by applying machine learning. These works either developed affective features or investigated feasible machine learning algorithms. Low-level features describe simple video/audio characteristics, such as color, textual, energy and so on. Emotion is a high-level (perception-level) concept to which we have cognitive access. There is a gap between low-level features and high-level human perception of emotions. Most recently, user's physiological responses have been taken into account for multimedia affective content analysis [8, 16, 20, 26]. In [20], five physiological response measures, including electro-dermal response (EDR), heart rate (HR), blood volume pulse (BVP), respiration rate (RR) and respiration amplitude (RA) were considered to produce entertainment-led video summaries. In [8], the authors indicated the fundamental role of emotion in the maintenance of physical and mental health. Neurophysiological biosignals of both electroencephalogram (EEG) and Electronic design automation (EDA) were used to measure viewers' valence and arousal discrimination while viewing pictures selected from International Affective Picture System (IAPS). Considering emotional biosignals, affective content analysis can be achieved from user's perspective. However, it might be hard for physiological response based methods to achieve real-time on-line affective analysis since it expects too much involvement from users and requires specific devices/tools for biosignal measurements.

In order to bridge the gap between low-level features and high-level emotions, a three-level framework is proposed in this paper for affective content analysis. Mid-level representations had been proposed for music genre classification in [10] where mid-level representations were pitch distribution, geometric pitch spaces, chroma length and so on. These mid-level representations were statistically computed from low-level audio features (e.g. MFCC, pitch) and might not have any direct relationships with human understanding of music genres. Different from [10], the mid-level representations in our three-level framework have their own semantic meanings and provide significant hints for users to understand affective content. Moreover, since video uses three channels that are video, audio and text, to convey video story and represent video emotions, we consider mid-level representations for multiple modalities. The definition and the details of our proposed framework will be introduced in Section 2. The main contributions of this paper are summarized as follows.

- A clear definition of mid-level representation is placed in a three-level framework for affective content analysis and its applications are explored through three case studies. Different from the exiting mid-level representations for music genre classification, our mid-level representations are motivated by simulating human perception. Therefore, our mid-level representations have their own semantic meanings and direct links to user understanding of affective content.
- Multiple modalities are considered in order to complement each other and work together towards the success of affective content analysis. Every individual case study focuses on one modality and also seeks help from other modalities. Three case studies prove the feasibility and effectiveness of our proposed three-level framework.
- By applying three-level framework, the gap between low-level features and high-level affective content can be bridged. Mid-level representation is adapted and successfully applied for affective content analysis.

The rest of the paper is organized as follows. In Section 2, the proposed framework will be introduced. And the contributions of our work will be discussed. Three case studies of using audio emotional events, using dialogs and using subtitles for affective content analysis will be illustrated in Sections 3, 4 and 5 respectively. Finally, conclusions will be drawn in Section 6.

2 The three-level framework for affective content analysis

Human perception is the process of attaining awareness or understanding of sensory information. While watching videos, human sensory system, including visual system and auditory system, generate sensory information as inputs for perception. To simulate human perception system and bridge the above gap, we propose a three-level affective content analysis framework in this paper. As shown in Fig. 1, we borrow the mid-level representation to indicate dialog, audio emotional events (e.g., horror sounds and laughters) and textual concepts (e.g., informative keywords). In the proposed framework, the process is delineated in three tiers comprising low-level multimodal feature representation, mid-level representation and high-level affective content analysis. Different from traditional approaches of direct mapping from low-level features to high-level emotions, mid-level feature representation bridges the perceptual gap between low-level features and high-level content. The mid-level feature representation is a concept involving video, audio and text domains and includes semantic video shots, audio emotional events, dialog, textual concepts, and so forth. The mid-level representation is generated by performing machine learning on low-level multi-modal features and is further used for high-level affective content

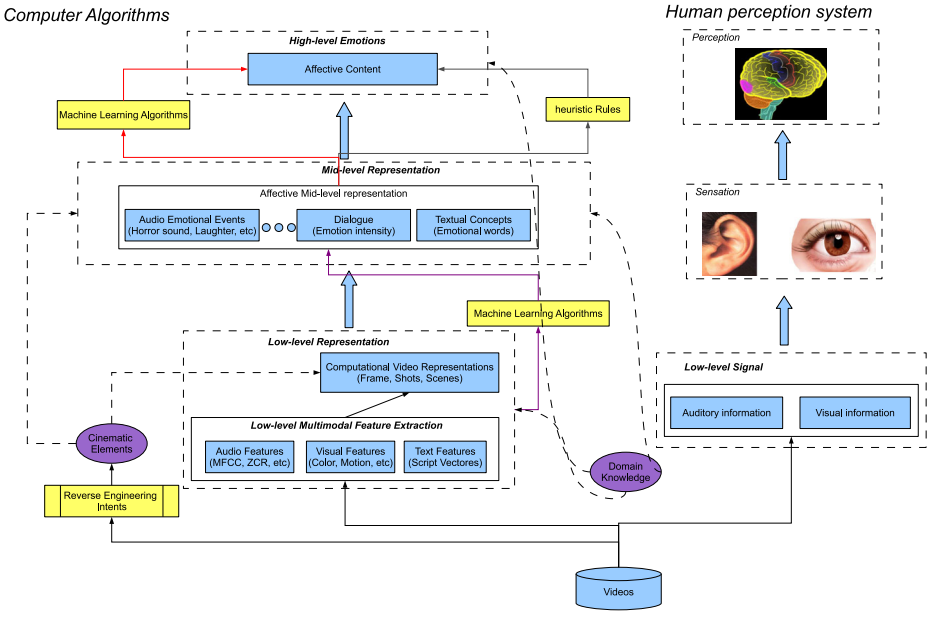


Fig. 1 Three-level emotion detection framework

analysis. From video production points of view, multiple modalities in movies were used to represent emotions and evoke emotional atmosphere. Different modality has its own merit to affect user feelings. Therefore, multimedia research leans towards analysis based on multiple modalities. Modalities compensate each other towards successes on most of the purpose of multimedia researches. In our proposed framework, domain-specific knowledge (DSK) plays a significant role at each level. In low-level representation, DSK is conducive to low-level feature selection and video shot detection. At mid-level, DSK helps to appoint mid-level representations which are connected to video contents. Furthermore, DSK takes part in high-level content analysis and supplies the justification for heuristic-rule definition. Both machine learning algorithms and heuristic-rule methods can be used for mapping mid-level representation to high-level content.

Affective content is constrained by video domains/genres, production rules and user understanding. Video genre/domain has its dominant emotional theme, which provides determinate clue for affective content detection. Affective content analysis has to consider video genres/domains. Moreover, different video domains/genres have various production rules. In order to construct specific movie structures and create vivid atmosphere, these production rules regulate the way of video shooting and utilize different modalities existing in movies. Therefore, affective content analysis has to consider different modalities together with video genres/domains and their production rules. In this paper, multiple modality mid-level representation is carefully defined, selected, and fused for different video domain, and is further applied to detect affective content. We apply the proposed framework for three case studies. In these cases, audio event, dialog and subtitle are studied to assist affective content detection in different video domains/genres based on the following three findings [21].

1. 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.
2. Dialog is the most important part of movie, which not only conveys movie story but also expresses speaker's emotions.
3. The scripts of movies provide direct access to the content of human dialogs, which also implicates human emotions.

Due to different theme topics and moods, videos are categorized into different genres. Normally, one or two dominant emotions are presented in a certain video domain/genre. Moreover, different modalities have their own advantages and disadvantages for conveying video stories and emotions. Although one modality plays dominant role in every case study, getting helps from other modalities are still necessary in some cases. The main purpose of three case studies is to demonstrate the feasibility of our proposed three-level framework and prove the effectiveness of every mid-level representation proposed in this paper. Therefore, different emotion categories are applied for individual case study by considering the video genres and advantages of the modality dominantly used in that case. In Case 1, by using audio emotional events, horror segments are detected in horror movies and laughable segments are detected in sitcoms. In order to identify affective content, Case 2 detects three intensity levels for dialog emotions. In case 3, five emotion categories, i.e., anger, sadness, fear, joy, and love are detected by analyzing scripts of the movies.

3 Case 1: using audio emotional events to detect affective content in sitcom and horror movies

AEEs are defined as some specific audio sounds which have strong clues to inference affective content. Especially in some video domains, such as sports video, comedy and horror movies, some audio sounds (e.g. excited audience sounds, audience's laughings and horror sounds) strongly represent viewer's emotions. In this case study, audio emotional events (AEEs) including laughter and horror sounds are detected and further employed with video shot boundary to locate the audio/video segments with affective contents. Sitcoms and horror movies are used in this case study.

3.1 Audio emotional event identification

In horror movies, within a horrific scenario, horror sounds are utilized to emphasise the scary atmosphere and increase the dramatic effects. Containing many laughable segments, sitcom is produced to amuse audiences and make them feel cheerful. In sitcom, canned laughters can be heard after laughable scenarios. In this case, laughters and horror sounds are significant to locate video/audio segments with laughable content and horror content. In sitcoms, besides audio emotional events (canned laughter), there are non-emotional audio events, such as dialog, silence, music and other environmental sounds. These audio events cover most audio tracks in sitcoms. In horror films, besides horror sounds, there are dialog, silence and others. Subsequently, identifying the laughters and horror sounds from other audio sounds can be regard as a task of audio classification. In this study, audio tracks are classified into five pre-defined classes (canned laughter, dialog, silence, music and others) for sitcoms and four pre-defined classes (horror sounds, dialog, silence and others) for horror movies respectively.

Audio signal exhibits the consecutive changes in values over a period of time, where variables may be predicted from earlier values. This means strong correlation exists. In consideration of the success of HMM in speech recognition, we propose our HMM based audio classification method as show in Fig. 2. Selected low-level features are firstly extracted from audio streams and tokens are added to create observation vectors. These data then separated into two sets for training and testing. After that, HMM is trained then re-estimated by using dynamic programming. Finally, according to maximum posterior probability, the audio event with the largest probability is selected to label the corresponding testing data. Left-to-right HMM with four states is used. Mel-Frequency Cepstral Coefficient (MFCC) and Energy

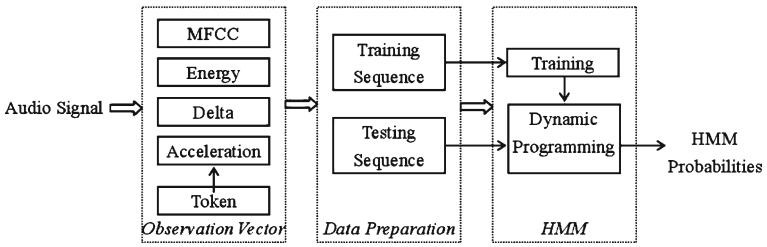


Fig. 2 HMM-based audio classification

are selected as the low-level audio features as they are successfully used in speech recognition and further proved to be efficient for audio keyword generation in [29]. Besides traditional audio features, Delta and Acceleration are further used to accentuate signal temporal characters for HMM. Delta and Acceleration effectively increase the state definition by including first and second order memory of past states. More details can be found in our already published paper [31].

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 mix with other events sometimes. For example, there may be one or two discontinuous samples of silence detected in continuous dialog. Considering sequencing order of audio, we regard any audio events changing within 1 or 2 s in the audio events sequence as an error. These errors are 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 s, it is near impossible to detect by HMM-based identifier. 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 happen within sounds whose amplitudes are relative small. Therefore, by calculating the amplitude change of audio signal, the blares can be easily detected.

3.2 Video affective content detection by audio emotional events

Video shot is used as a unit for affective content detection. In horror films, since horror sounds take place synchronously with horror scenarios, it is simple to select those video shots in which horror sounds have been identified as horrific contents. However, canned laughter always appear after laughable segments. How to decide the boundaries of laughable segments is challenging.

In a sitcom, dialog is the most popular scenario. Camera may switch among the persons in a dialog. After several shots, audiences may be amused by words or actions during dialog. Laughable segments can be detected by checking dialogs locating before canned laughters. By checking the duration and location of other audio events, we determine the starting points of laughable audio segments (LAS).

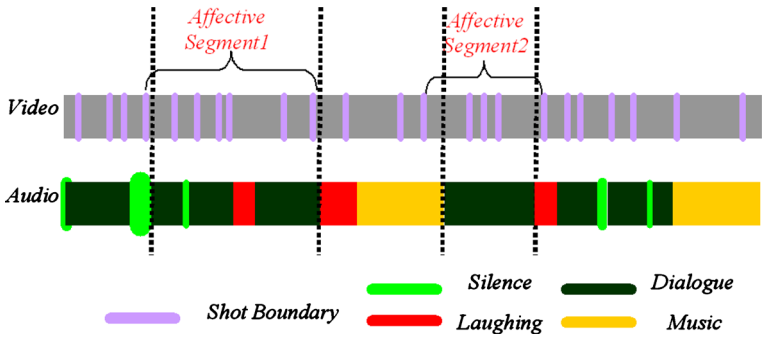


Fig. 3 Video/Audio structure from comedy videos

Table 1 Audio emotional event identification for comedy

	Dialog	Music	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)

After that, those video shots whose more than half of the length overlaps with the LAS are selected as affective content (laughable segments). A possible video/audio structure from sitcom is shown in Fig. 3.

The process of Laughable Segment (LS) selection is listed step by step as follows (t is the starting point of laughing):

1. Check the duration between two continuous laughers. If the duration < Threshold 1, go to (3). Otherwise do (2) and skip (3).
2. If there is no any end of LS set for the current sequence, set the beginning of t to be the end of LS. Search forward music or silence from t. Otherwise, directly search forward music or silence from t.
3. If there is no any end of LS set for current sequence, set the beginning of t to be the end of LS. Search forward from $t-1$. $t = t-1$, go back to (1).
4. If detected silence or music duration > Threshold 2, set the end of silence or music to be the beginning of LS. $t = t-1$. Go back to (1).

3.3 Experiments

3.3.1 Audio emotional event identification

The audio samples were from a 40 min sitcom (Friends) and 40 min Korea horror film (Face). They were collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. We used half of the data for training and the half for testing. The audio signal was segmented into 20 ms per frame which was the basic unit for feature extraction. The feature extraction was implemented in Matlab. In our study, one second was selected for HMM sample length since most audio emotional events last longer than 1 s. The Hidden Markov Model Toolkit (HTK) [9] was used for audio emotional event identification. Tables 1 and 2 show the audio event classification results. The results in parenthesis are before post-processing.

Compared with horror movies, sitcoms are mainly indoor scenes with simple environmental sounds. And for most of the cases, environmental sounds bring noises to audio classification. This may be a reason why the audio classification results of sitcom are much better than that of horror movies. The performance of HMM-based identifier is not satisfactory for horror sounds because some horror sound durations are less than the HMM sample length (1 s). After applying sliding

Table 2 Audio emotional event identification for horror

	Dialog	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)

Table 3 Affective content detection results

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

window majority-voting elimination and blares detection, results of horror movie are evidently improved.

3.3.2 Video affective content analysis

Videos implemented in video affective content analysis were same as those used in audio emotional event identification. The results of affective content detection in both sitcoms and horror movies are promising as listed in Table 3. Ground truth was labeled manually by five students from engineering department. They were required to watch the videos and label horror segments for horror movies and laughable segments for sitcoms. Each segment was labeled by a starting frame and an ending frame. Since the labels from different student might be different, the final labels were decided by majority voting. As shown in Table 3, most of the affective content are detected. However, compared to recalls, the precisions are not very satisfactory. For comedy video, it may be because some segments brought up canned laughters were not labeled in our ground truth. In horror movies, some horror sounds may only be used to highlight the overall horrific atmosphere instead of taking place synchronously with scary scenarios.

In Case 1, audio emotional events indicate possible positions of video affective content, which thus avoid blind searching for whole length of video. The detection of affective content mainly rely on audio event detection, i.e. the detection of horror sound and canned laughter. Therefore, the detection of affective content can be further improved by increasing the accuracy of audio event detection. Moreover, combining with other mid-level representations with audio events might be a promising solution to improve affective content detection.

4 Case 2: using dialog to detect affective content in movies

Movie dialog is affected by participants' emotions and therefore represents movie emotions. In this case study, the efficiency of using dialog to detect affective content was researched since dialog is an important element in movie to convey movie story and represent movie emotions. Dialogs are firstly identified by SVM learning. Secondly, dialog emotions are detected by HMM learning based on energy, MFCC and pitch features. Finally, dialog emotion is used as a significant clue for affective content detection. Three emotion intensity levels are detected. Figure 4 shows the system architecture.

4.1 Dialog categorization

Due to different background of dialog, movie dialogs are classified into three categories i.e., plain dialog, dialog with music and dialog with other sounds.

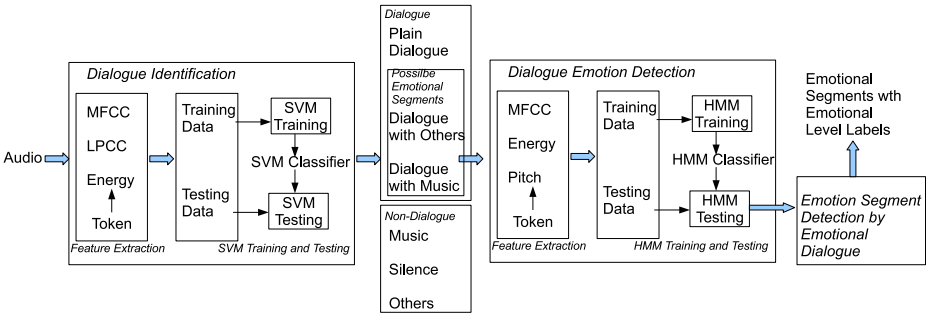


Fig. 4 The framework for movie dialog emotion detection

- Plain dialog: Sometimes, when we watch a movie, we can only hear dialog without any other sounds or other sounds can not be noticed while we listen to dialog. Those dialogs are defined as plain dialog. With the happening of plain dialogs, scenes, images or objects do not attract audience's attention. Most probably, the camera switches between two or more persons attending the dialog. The director may want audiences to focus on the dialog and well understand dialog content. Normally, those dialogs convey movie story or narrative smoothly without strong emotions.
- Dialog with music: Dialog with music is referred to those movie dialogs which interlace with movie music or has music background. Music used in movie romances emotional atmosphere and expresses a certain emotion. According to [25], film music has its power to evoke specific scenes, images, and characters when heard apart from the film it accompanies, thus move audiences to laughter or tears. Those movie dialogs with music actually need music to promote dialogs' contagious effects on emotions. Therefore, dialogs with music are very significant for movie emotion detection. Sometimes, dialog with music appears after plain dialog and is followed by music. For example, two persons start talking for a while, and later it comes some music to enhance their emotions. After the dialog, the music is still on for a while to look after audiences' remaining feelings. The audio sounds take place according to some potential orders. These orders are called as audio transition patterns in the rest of this section.
- Dialog with other sounds: Other sounds can be any sounds other than music, dialog and silence. Typical examples of other sounds are the sounds of explosion, gun shooting, bird singing, water flowing and the environment sounds of certain stages. Generally speaking, dialog with other sounds take place along the happening of other events. For example, two guys are talking while fighting. From the emotion detection point of view, this kind of dialog is more complicated than previous two since it is hard to tell whether the dialog plays the dominant role or the events happening along dialog is dominant. Sometimes, the contents of this kind of dialog are not significant, which is just to emphasise the emotions of happening events, whereas sometimes, the happening events enhance dialog.

Besides plain dialog, dialog with music and dialog with other sounds, music, silence and others take place regularly in movie. Those audio sounds almost cover all the movie audio signals. Distinguishing dialog from other audio sounds is treated as a

classification problem. Firstly, we segment audio signal into frames of 20 ms, which is a basic unit for feature extraction and future classification. Non-linear support vector classifier is applied on Energy, MFCC and LPCC features of audio signals to identify three categories of dialogs [29]. Post-processing is performed by exploiting a sliding window to eliminate those sudden changes by majority-voting on the sound type from a sequence of frame-based classification results. Our related publications can be found [30, 32].

4.2 Dialog emotion detection

As discussed in Section 4.1, compared to plain dialog, dialog with music and dialog with other sounds are significant for emotion detection. Dialog emotions are detected for dialog with music and dialog with other sounds.

To detect dialog emotion, we face two difficulties. Firstly, most of the existing algorithms are not robust to achieve speaker-independent task. During a conversation, every sentence has its own emotion. It is hardly to detect moods for the individual sentence. Secondly, psychology research shows various definitions for human emotion categories. It is hard to define emotion categories for dialog detection. Moreover, the technical constrains make it difficult to detect detailed emotion categories from dialog.

Due to the above reasons, three emotion intensity levels were detected with the following considerations.

- The emotion intensity levels are labeled for the whole coherent dialog segments instead of every individual sentence.
- Emotion is a complex psychophysical process which is a continuous psychological response and physical expressions. The current emotion status is affected by the previous one. HMM-based classifiers treat audio signal as a continuous time-series data and employ hidden states transitions to capture context information.
- Dialog with music and dialog with other sounds should be carried on respectively.

The right part of Fig. 4 shows the detailed processing of dialog emotion detection. Energy and pitch were proven to be effective audio feature and wildly used for vocal emotion detection [24, 33]. Moreover, MFCC worked well for excited and non-excited detection [32]. Therefore, we extract these features to get feature vectors for HMM training and testing. The same as those shown in Section 3.1, the left-right HMM structure with five states was selected for HMM topology. Three HMMs are trained for emotion intensity levels. For each coming audio sample sequence, the likelihood of every HMM was computed. The intensity level with biggest HMM likelihood was set.

4.3 Movie affective content detection by dialog emotion intensities

Movie emotional segments should include dialog with emotions but not should be limited by the length of dialog. Sometimes, movie audio occurrence presents certain transition patterns which provide reference for locating affective content. We discovered three transition patterns in terms of dialog, as shown in Fig. 5.

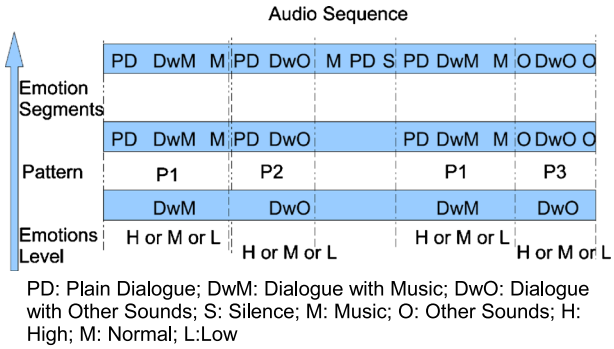


Fig. 5 Movie affective content locating

Affective content detection can be regarded as identify these three transition patterns with dialog emotions. In Fig. 5, the structure of affective content detection by using movie dialog is shown from bottom to up. Firstly, Emotion intensity levels are detected for dialogs. Secondly, patterns with dialog are located in the movie. Finally, every movie segment with affective content is labeled with the emotion intensity level of the dialog in that segment. Only those segments containing dialog with High and Low emotion intensity level are finally selected as affective content.

4.4 Experiments

The experiments involved three parts: dialog identification, dialog emotion detection and movie affective content detection.

4.4.1 Dialog identification

Around 4 hours' movie segments collected from 3 movies (Cold Mountain, Billy Elliot and Love Actually) were used to test our dialog identification module. The same as method used in Case 1, audio signals were collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. The audio signal was segmented into 20 ms per frame which was the basic unit for feature extraction. Features were extracted by using Matlab. For SVM to identify dialog, we used radial basis function (RBF) as kernel function, $K(x_i, x_j) = exp(-r||x_i - x_j||^2)$, $r > 0$, for SVM classification. LIBSVM [6] was applied for dialog detection. One-against-all multi-class approach was used.

Table 4 shows dialog identification results.

Table 4 Performance of dialog identification

	Dialog			Music	Silence	Others
	PD	DwM	DwO			
Recall (%)	92.45	93.28	91.27	93.52	91.34	93.15
Precision (%)	90.41	93.78	92.25	94.17	91.58	85.14

PD Plain Dialog; *DwM* Dialog with Music; *DwO* Dialog with Others

Table 5 Performance of dialog emotion detection

	Dialog with music			Dialog with others		
	High	Normal	Low	High	Normal	Low
Recall (%)	87.25	82.96	72.13	82.79	73.35	65.14
Precision (%)	88.35	80.26	80.74	77.15	71.10	69.21

4.4.2 Dialog emotion detection

The original data was labeled with emotion intensity levels by 10 students (5 females, 5 males) from IT departments. The detection results were compared with the ground truth labeled by students, as shown in Table 5.

One second was selected for HMM sample length since most audio emotional events were longer than 1 s. Again, HTK [9] was used for dialog type detection.From Table 5, we find that detection on dialog with music outperforms detection on dialog with others at least 5% on average. A possible reason is that other background sounds become noise during emotion detection. Another finding is that emotions with level 'high' are detected more accurate than emotions with other levels. It may be because 'high' is easy to be distinguished from 'normal' and 'low'. The reason may be that high frequency and high energy signal of 'high', while 'normal' and 'low' are relative hard to be differentiated.

4.4.3 Movie affective content detection

Affective content detection results are directly affected by dialog emotion analysis. By using the proposed detection methods (as shown in Fig. 5), more than 85% of the affective contents are detected. However, about 30% of the detected affective segments are not exactly what viewers expect to view. Most of the unexpected emotional segments are from the segments which contains dialog with other sounds. Other sounds together with dialog depend on the environment where the dialog takes place. Sometimes, the environment sounds might be very loud (or even more dominant than dialog) which actually affect the identification of dialog emotions. As a significant part of movie, dialog has been used and proven to be feasible as an emotional clue for movie emotion segments detection in this study. Although dialog emotions are used straightforward to indicate affective content, detecting dialog boundaries, identifying dialog types and audio transmission patterns are still challenging. Also how to determinate the duration of affective content is a challenging task.

5 Case 3: using subtitle and audio to detect affective content in sitcom

This case study attempts to extract affective content by analyzing the subtitle files of DVD/DivX videos and utilizing audio event to assist affective content detection. As discussed in Section 4, the affective aspect of the video content is significantly represented by humans' dialogs. The scripts of sitcoms provide direct access to the content of human dialogs. The scripts extracted from DVD/DivX only include time and content of dialog. Compared to complex video and audio processing, it is relatively convenient and easy to get useful affective information through analyzing

scripts. In this study, we chose sitcom videos as test bed. The system flow is shown in Fig. 6. Scripts with time stamps and audio stream are firstly extracted from the subtitle file associated with video. Then, videos are segmented by script partition instead of traditional video shot segmentation. Informative keywords in scripts are detected for each partition to locate possible affective video content. Finally, audio events are detected to complete affective content detection.

The unique features of this work are listed in two points:

1. Using subtitle can directly access video content and avoid complex video/audio analysis process.
2. Compared to traditional video shots, video segmentation by scripts partition is not affected by camera changes and shooting angles and effectively extracts meaningful video segments with compact content.

5.1 Video segmentation by script partition

The ability to segment video into meaningful segments is an important aspect of video indexing. Video shot relies on camera changes and shooting angles. Sometimes, it is hard to include video segments with compact contents in one shot. For example, the camera switches between two persons in dialog produce several shots. Recently, audio information has been considered for video segmentation [14]. However, segmenting video by audio information is limited by audio analysis technique. The objective of video segmentation is to group together those video frames that have a close semantic thread running through them. From the scripts' point of view, the temporally adjacent *ScriptElements* tend to convey a semantic notion together. Therefore, a video segmentation method by script partition is proposed to effectively extract meaningful video segments with compact affective content.

DVD/DivX videos come with separate subtitle or script files for each frame in the video sequence. This study focus on scripts recorded as strings in text files. Each script in the file consists of an index for the script, the time for appearance and disappearance of the script with respect to the beginning of the video and the text of the script. The subtitle file is parsed into *ScriptElements*, where each

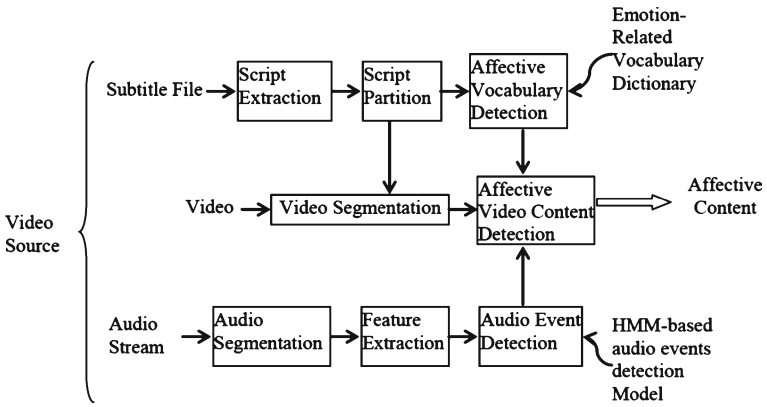


Fig. 6 The system flow of affective content detection

ScriptElement has the following three attributes: 'Start Time', 'End Time' and 'Text'. We use the information in the *ScriptElements* to partition the video. *ScriptElements* constitute a dialog or an extended narration having a high 'semantic correlation' among themselves. In videos, when there is a dialog or a long narration that extends to several frames, the *ScriptElement* gap is very small. We utilize the time gap between *ScriptElements* as the clue for script partition. This time gap, which we call *ScriptElement* gap, is defined as the time gap between the 'EndTime' of the previous *ScriptElement* and the 'StartTime' of the current *ScriptElement*. Hence, *ScriptElement* gap is a useful parameter by which we group together semantically relevant *ScriptElements*, thereby creating a partition of the scripts. In the proposed method, the *ScriptElements* are partitioned by thresholding the *ScriptElement* gap. We call each partition a *ScriptSegment* which corresponds to a video segment by time correspondence. Figure 7 shows the distribution of the time gap between *ScriptElements* for a total of 450 *ScriptElements*. In our experiments, we segment video by script partition with 2 s as the threshold for *ScriptElement* gap.

5.2 Affective script detection

Words expressing emotions are defined as informative keywords which are important clues for affective content detection. Scripts containing those informative keywords are detected in this study.

5.2.1 Script vector representation

After partitioning the scripts into segments, we build an index for each script segment. We adopt the term-frequency inverse document frequency (*tfidf*) vector space model [4], which has been widely used for information retrieval. The first step involves removal of stop words, e.g. 'about', 'I'. The Potter Stemming algorithm [22] is used to obtain the stem of each word, e.g. the stem for the word 'families' is 'family'. The stems are collected into a dictionary, and then used to construct the script vector

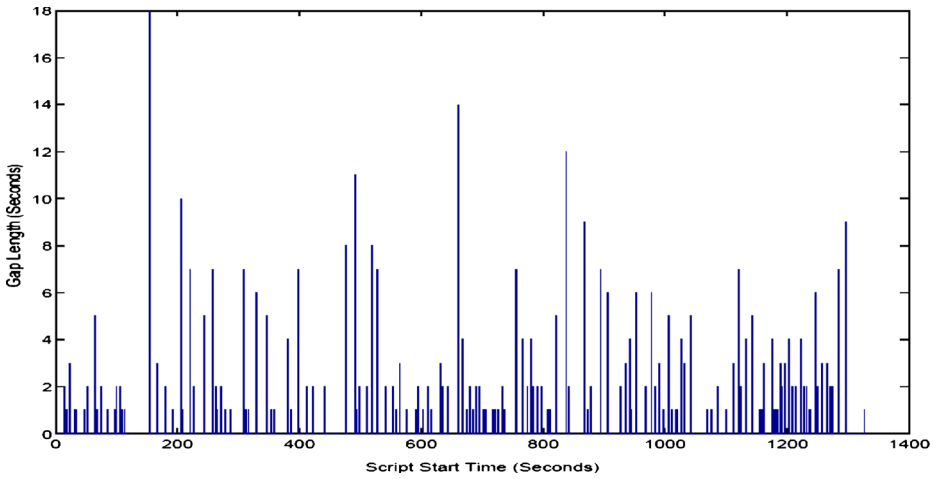


Fig. 7 Time gap between the script element of sitcom video

for each segment. Just as the vector space model represented a document with a single column vector, we represent the script segment using the *tfidf* function [3] given by

$$tfidf\left(t_k, d_j\right) = \#\left(t_k, d_j\right) \log \frac{|S_s|}{\#S_s(t_k)} \tag{1}$$

where $\#(t_k, d_j)$ denotes the number of times that a word t_k occurs in segment d_j, $|S_s|$ is the cardinality of the set S_s of all segments, and $\#S_s(t_k)$ denotes the number of segments in which the word t_k occurred. This function states that (a) the more often a term occurs in a segment, the more it represents its contents, and (b) the more segments a term occurs in, the less discriminating it is. The *tfidf* function for a particular segment is converted to a set of normalized weights for each word belonging to the segment according to

$$(W_{k,j}) = \frac{tfidf\left(t_k, d_j\right)}{\sqrt{\sum_{i=1}^{T}\left(tfidf\left(t_i, d_j\right)\right)^2}} \tag{2}$$

Here, $W_{k,j}$ is the weight of the word t_k in segment d_j and T is the total number of words in the dictionary. This is done to ensure that every segment extracted from the subtitle file had equal length and that the weights are in [0, 1]. These weights are collected together into a vector for a particular segment such that the vector acts as a semantic index to that segment. We call this vector as the '*tfidf* vector' in the following discussion.

Based on script vector representation, we collected all the column script vectors together into a matrix of order $T \times |S_s|$, called the script matrix.

5.2.2 The dictionary of affective vocabularies

To find some relationship between scripts and affective content, we build up dictionary of affective vocabularies. affective vocabularies are categorized into five basic categories: anger, sadness, fear, joy, and love [17]. To experimentally explore the possible usage of script, the five emotion categories are detected for video content according to the five basic categories. Examples are shown as follows.

1. Anger: shit, angry, rage, wrath, damnit, ...
2. Sadness: sad, depressed, upset, sorry, ...
3. Fear: horrible, scared, fear, frighten, terrify, ...
4. Joy: happy, cheer, joy, pleased, glad, ...
5. Love: love, romantic, affection, ...

5.2.3 Affective script detection by informative keywords query

We further detect the script segments which include informative keywords vocabularies. The detection can be regarded as a query task. If we want to detect the affective script of joy, we query informative keywords in category of joy. The query can be in the form of a single word in which case the query vector (which has the same dimensions as the *tfidf* vector) will consist of a single non-zero element. For example, a query with the word 'happy' will result in a query vector like [0...1...0], where only the entry of the vector corresponding to the word 'happy' is set to one. The query can also take into account of the n emotional words in the category, where the query

vector will look like $[0...1/\sqrt{n}...1/\sqrt{n}...]$. In our study, the words that are presented in the query have higher values in the query vector. The result of the querying process is the return of script segments which are geometrically close to the query vector. Here, we will use the cosine of the angle between the query vector and the columns of the script matrix as a measure,

$$\cos\theta_j = \frac{{a_j}^T q}{||a_j||^2||q||^2} = \frac{\sum_{i=1}^{T} a_{ij}q_i}{\sqrt{\sum_{i=1}^{T} a_{ij}^2}\sqrt{\sum_{i=1}^{T} {q_i}^2}} \tag{3}$$

For $j = 1...|S_s|$, where a_j is a column vector from the script matrix, q is the query vector and T is the number of words. Those script vectors for which (3) exceeds a certain threshold are considered relevant. Alternatively, we could sort the values of $\cos\theta_j$ to present the top n results.

5.3 Audio event detection

In this section, five kinds of audio events are detected as: *silence*, *dialogue*, *laughing*, *music* and *other environment sounds* which appear in audio stream extracted from video source. Same work has been introduced in Section 3.1.

5.3.1 Music mood labelling

Music is relatively complex. Incidental music is one of the clearest insights that we have. Into incidental music, the makers of the film want us to think and feel in reaction to what is happening on the screen. However, automatic music mood detection is a challenging research topic which has not been solved satisfactorily. In this section, we classify music mood manually. Eight students label music with anger, sadness, fear, joy and love by listening to the music segments individually and discuss later to get accordant affective labels for the music segments.

5.4 Locating affective content

Affective contents are located by checking emotion-related vocabularies in subtitles and later supplemented by the results of audio event.

Some emotions are easily inferred by detected informative keywords. For example, when 'sorry', 'depress' and 'lonely' are detected in the same script segment, we can tell with high probability that this segment expresses 'sadness'. However, sometimes when people talk about 'love', they may have no feelings of love. For example, if 'love' and 'divorce' are mentioned together, it is hard to tell the emotion of the segments is about love or sadness. It is possible that, within one segment, the detected informative keywords cross more than one affective categories. In this case, we check the number of informative keywords from each affective category and only select the segments which include at least more than one informative keywords from the same category as the possible segments with affective video content. As shown in Fig. 8, the first segment is ignored because the emotion-related vocabulary in one class appears only once. In Segment n, the number of vocabularies in category joy is J, in category anger is A, in category fear is F, in category sadness is S and in love is L. We set Flag=max (J, A, F, S, L) to present the affective label for the corresponding

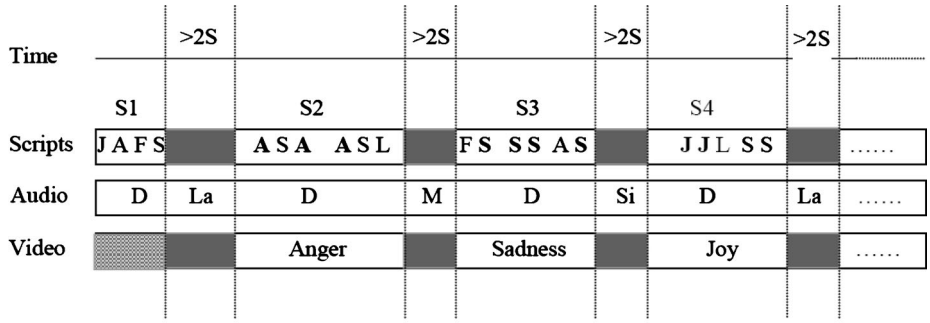


Fig. 8 Affective video content locating

video segment. If among J, A, F, S and L, there are more than one parameter having the same value (See S4 in Fig. 8. J=S=2), we have to seek help from the audio events.

There exists a relationship between audio events and affective contents. For example, it is impossible that the event of laughing takes place after the segment of sadness. Instead, *Silence* or sad music may appear after a segment of sadness. See S4 in Fig. 8, when J=S and audio event of *laughing* is detected after the segment, the affective label for this video segment should be joy. Some decision rules for the case of more than one parameter among J, A, F, S and L having same value are summarized as follows.

1. If *music* is detected after the video segment, affective label is decided by Music Mood.
2. If *laughing* detected after the video segment in which J is one of the max (J, A, F, S, L), the affective label should be joy.
3. *silence* that is at the end of one video segment is used to eliminate the label of joy.

5.5 Experiments

The audio samples were used to train HMM models come from a 300 min real-time sitcom video. The testing data were from around 260 min of TV sitcom video (Friends) with subtitle files and audio. Audio data were collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. Table 6 shows the results of audio event detection on 260 min audio samples from TV sitcoms. Both audio event detection and affective script detection were implemented in Matlab.

The ground truth was manually labeled by eight students. Table 7 shows the experimental results. Most segments of affective content were detected by informative

Table 6 Audio event identification result

	Dialogue	Music	Canned laughter	Silence	Others
Recall (%)	98.98	96.21	99.17	97.18	100
Precision (%)	99.04	98.19	99.19	96.57	86.35

Table 7 Affective content detection result

	Sadness	Anger	Fear	Love	Joy
Ground truth	17	6	9	8	19
Correctly detected	16	6	9	6	18
False alarm	3	1	0	4	4
Recall (%)	94.12	100	100	75	94.74
Precision (%)	84.21	85.71	100	60	81.82

keywords and audio events successfully. The detection of love is not very satisfactory. We need to make an assumption when using informative keywords to detect affective content. The assumption is that people mention emotion related vocabularies when they experience that emotion. However for some real cases, it does not always follow the assumption. By analyzing the wrong detected cases, we realized that when people mention love related vocabularies, it might not because they experience the emotion of love. This can be the main reason that both precision and recall of love detection are not satisfactory.

6 Conclusions

A three-level affective content analysis framework has been proposed and proved to be effective in this paper. As three vital components in videos, audio sound, dialog and subtitle are used to evoke video emotions. In this paper, we have explored how three-modality mid-level representations that are audio sound, dialog and subtitle, contribute to affective analysis. Mid-level representations are generated from low-level multi-modality features and further used for affective content detection. Experiments have proved that audio sound, dialog and subtitle are efficient to infer affective contents. Mid-level representation bridges the gap between low-level features and high-level user perception. Modalities compensate each other towards success on affective content analysis.

Video domain and movie genre constrain affective analysis. In future, a generic solution needs to be considered. Moreover, multiple modality fusion is also vital. The current fusion is based on heuristic rules. Statistical fusion should be considered.

Acknowledgements This research was supported by National Natural Science Foundation of China No. 61003161, No. 60905008 and UTS ECR Grant.

References

1. Arifin S, Cheung PYK (2006) User attention based arousal content modeling. In: Proceedings of the international conference on image processing, pp 433–436
2. Arifin S, Cheung PYK (2007) A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. In: Proceedings of the ACM multimedia conference, pp 68–77
3. Berry MW, Drmavc Z, Jessup ER (1999) Matrices, vector spaces, and information retrieval. SIAM Rev 41(2):335–362
4. Berry MW, Dumais ST, O'Brien GW (1995) Using linear algebra for intelligent information retrieval. SIAM Rev 37:301–328
5. Chan C, Jones GJF (2005) Affect-based indexing and retrieval of films. In: Proceedings of the ACM multimedia conference, pp 427–430

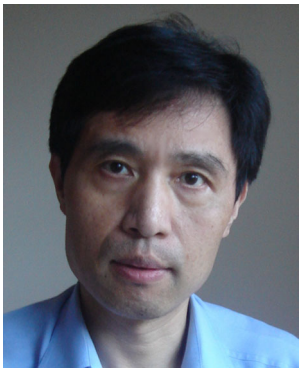
6. Chang CC, Lin CJ (2006) Libsvm – a library for support vector machines. In: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
7. Chen YH, Kuo JH, Chu WT, Wu JL (2006) Movie emotional event detection based on music mood and video tempo. In: Proceedings of IEEE international conference on consumer electronics, pp 151–152
8. Frantzidis CA, Bratsas C, Klados M, Konstantinidis E, Lithari CD, Vivas AB, Papadelis CL, Kaldoudi E, Pappas C, Bamidis PD (2010) On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. IEEE Trans Inf Technol Biomed 14(2):309–318
9. Gales M, Woodland P (2006) Recent progress in large vocabulary continuous speech recognition: an htk perspective. In: ICASSP tutorial
10. Gruhne M, Dittmar C (2009) Comparison of harmonic mid-level representations for genre recognition. In: Proceedings of third international workshop on learning semantics of audio signals, pp 91–102
11. Hanjalic A (2006) Extracting moods from pictures and sounds: towards truly personalized tv. IEEE Signal Process Mag 23(2):90–100
12. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimedia 7(1):143–154
13. Irie G, Hidaka K, Satou T, Kojima A, Yamasaki T, Aizawa K (2009) Latent topic driving model for movie affective scene classification. In: Proceedings of the ACM multimedia conference, pp 565–568
14. Jiang H, Lin T, Zhang HJ (2000) Video segmentation with the assistance of audio content analysis. In: Proceedings of IEEE international conference on multimedia & expo, vol 3, pp 1507–1510
15. Kang HB (2003) Affective content detection using hmms. In: Proceedings of the ACM multimedia conference, pp 259–262
16. Kim J, Andre E (2008) Emotion-specific dichotomous classification and feature-level fusion of multichannel biosignals for automatic emotion recognition. In: Proceedings of IEEE international conference on multisensor fusion and integration for intelligent systems, pp 114–118
17. Kovecses Z (2003) Metaphor and emotion language, culture, and body in human feeling. Cambridge University Press
18. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the ACM international conference on multimedia, pp 83–92
19. Moncrieff S, Dorai CSV (2001) Affect computing in film through sound energy dynamics. In: Proceedings of the ACM on multimedia conference, pp 525–527
20. Money AG, Agius H (2010) Elvis: entertainment-led video summaries. ACM T Multim Comput 6(3):1–30
21. Plantinga C, Smith GM (1999) Passionate views: film, cognition and emotion. The Johns Hopkins University Press
22. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
23. Rasheed Z, Sheikh Y, Shah M (2005) On the use of computable features for film classification. IEEE Trans Circuits Syst Video Technol 15(1):52–64
24. Sebe N, Cohen I, Gevers T, Huang TS (2006) Emotion recognition based on joint visual and audio cues. In: Proceedings of the 18th international conference on pattern recognition, pp 1136–1139
25. Smith J (1998) The sounds of commerce: marketing popular film music. Columbia University Press
26. Soleymani M, Chanel G, Kierkels JJM, Pun T (2008) Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Proceedings of 2008 10th IEEE international symposium on multimedia (ISM '08), pp 228–235
27. Soleymani M, Kierkels JJM, Chanel G, Pun T (2009) A bayesian framework for video affective representation. In: Proceedings of the international conference on affective computing and intelligent interaction
28. Wang HL, Cheong LF (2006) Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology 16(6):689–704
29. Xu M, Xu CS, Duan LY, Jin JS, Luo S (2008) Audio keywords generation for sports video analysis. ACM T Multim Comput 4(2):1–23
30. Xu M, Duan LY, Xu CS, Tian Q (2003) A fusion scheme of visual and auditory modalities for event detection in sports video. In: Proceedings of IEEE international conference on acoustic, speech, and signal processing, vol 3, pp 189–192

31. Xu M, Duan L, Cai J, Chia LT, Xu C, Tian Q (2004) Hmm-based audio keyword generation. In: Proceedings of IEEE pacific rim conference on multimedia (PCM), pp 566–574
32. Xu M, Maddage NC, Xu CS, Kankanhalli M, Tian Q (2003) Creating audio keywords for event detection in soccer video. In: Proceedings of international conference on multimedia & expo, vol 2, pp 143–154
33. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Trans Multimedia 9(6):424–428
34. Zhang S, Huang Q, Jiang S, Gao W, Tian Q (2010) Affective visualization and retrieval for music video. IEEE Trans Multimedia 12(6):510–522

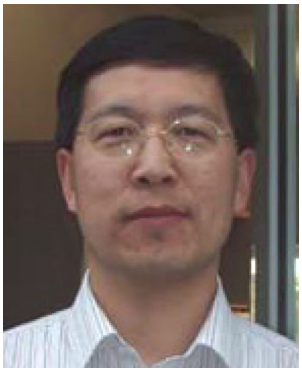
Min Xu received the B.E. degree from University of Science and Technology of China, in 2000, M.S degree from National University of Singapore in 2004 and Ph.D degree from University of Newcastle, Australia in 2010. Her research interests include multimedia content analysis, video adaptation, interactive multimedia, pattern recognition and computer vision.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently an Assistant Professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Xiangjian He received his B.S. degree in Mathematics from Xiamen University in 1982, his M.S. degree in Applied Mathematics from Fuzhou University in 1986, and his Ph.D. degree in Computing Sciences from the University of Technology, Sydney, Australia in 1999. From 1982 to 1985, he was with Fuzhou University. From 1991 to 1996, he was with the University of New England. Since 1999, he has been with the University of Technology, Sydney, Australia. He is currently a Full Professor and the Director of Computer Vision and Recognition Laboratory at the University of Technology, Sydney.

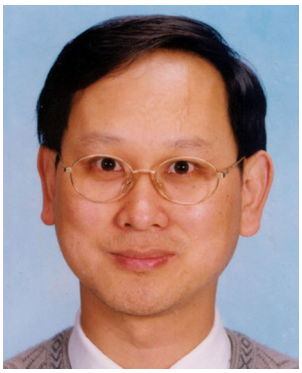


Jesse S. Jin graduated with a Ph.D. from University of Otago, New Zealand in 1992. He is the Chair Professor of IT in the School of Design, Communication and IT, University of Newcastle. He also chaired the Academic Board of the ITIC College in Sydney. His research interests include Multimedia, Medical Imaging and Computer Vision.



Suhuai Luo received B.E. and M.E. degrees in Radio Engineering from Nanjing University of Posts and Telecommunications China in 1982 and 1987, respectively, and PhD degree in Electrical Engineering from the University of Sydney Australia in 1995.

From 1995 to 2004, he worked as a senior research scientist with the Commonwealth Scientific and Industrial Research Organisation Australia and the Bioinformatics Institute Singapore. He is currently a senior lecturer with the University of Newcastle Australia. His research interest is in information technology and multimedia, including health informatics, machine learning, image processing, computer vision, and Internet-oriented IT applications.



Hanqing Lu Professor received his B.E. degree in 1982 and his M.E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree in 1992 from Huazhong University of Sciences and Technology. Currently, he is a deputy director of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences.

His research interests include image and video analysis, medical image processing, object recognition, etc. He has published more than 100 papers in these fields.