Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Hierarchical affective content analysis in arousal and valence dimensions

Min Xu [a,b,*], Changsheng Xu [b], Xiangjian He [a,**], Jesse S. Jin [c], Suhuai Luo [c], Yong Rui [d]

[a] School of Computing and Communications, University of Technology Sydney, Australia
[b] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[c] Faculty of Science and I.T., University of Newcastle, Australia
[d] Microsoft Research, China

ARTICLE INFO

Article history:
Received 7 March 2012
Received in revised form
14 May 2012
Accepted 19 June 2012

Keywords:
Affective content detection
Multiple modalities
Mid-level representation

ABSTRACT

Different from the existing work focusing on emotion type detection, the proposed approach in this paper provides flexibility for users to pick up their favorite affective content by choosing either emotion intensity levels or emotion types. Specifically, we propose a hierarchical structure for movie emotions and analyze emotion intensity and emotion type by using arousal and valence related features hierarchically. Firstly, three emotion intensity levels are detected by using fuzzy c-mean clustering on arousal features. Fuzzy clustering provides a mathematical model to represent vagueness, which is close to human perception. Then, valence related features are used to detect five emotion types. Considering video is continuous time series data and the occurrence of a certain emotion is affected by recent emotional history, conditional random fields (CRFs) are used to capture the context information. Outperforming Hidden Markov Model, CRF relaxes the independence assumption for states required by HMM and avoids bias problem. Experimental results show that CRF-based hierarchical method outperforms the one-step method on emotion type detection. User study shows that majority of the viewers prefer to have option of accessing movie content by emotion intensity levels. Majority of the users are satisfied with the proposed emotion detection.

1. Introduction

With the explosive growth of video production and the development of personalized multimedia services, users have become accustomed to accessing videos according to their preferences. However, an overload of video data makes users spend a great deal of time in finding the exact data they are interested in. Stored video data need to be annotated and indexed so that a compact representation of the original data can be generated to accommodate viewers' demands.

Early research focused on annotations based on low-level features, such as color features [1], motion intensity [2], and shape features [3]. Besides low-level features, video syntactic structure, such as video shot and video scene, was used for annotation [4]. Although, several promising modeling methods were proposed [5,6], low-level feature based video labeling and video structuring has little correlation with user understanding of video content. Recently, event detection and semantics modeling, therefore, becomes a key concern for video annotation and indexing [7]. Although these work tried to provide feasible ways to manage and access movie databases, there is still a gap between semantic content and user's preferences. From users' point of view,

* Corresponding author at: School of Computing and Communications, University of Technology Sydney, Australia. Tel.: +61 2 9514 4543; fax: +61 2 9514 4535.
** Corresponding author. Tel.: +61 2 9514 1816; fax: +61 2 9514 4535.
E-mail address: min.xu25@gmail.com (M. Xu).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.026

they may prefer "emotional decision" to find video segments of their interests because emotion factors directly reflect audiences' attention, evaluation and memory.

Affective content containing the amount and types of emotion and is expected to evoke audience's emotions. Affective content analysis has been conducted in many video domains, such as sports [8], music videos [9] and so on. Movies constitute a large portion of the entertainment industry. Developing effective and efficient methods to analyze, index and organize movies has great commercial potentials. Moreover, emotions are carefully presented in movies. Gathering in movie theater, most of audiences are expected to experience feelings. Therefore, movie affective content analysis occupies a dominant role in video affective computing [10–20].

Most of the existing methods tried to map low-level audio/visual features directly to emotion types by either heuristic or learning ways. Different from the existing method, we propose a hierarchical structure for emotion categories and analyze emotion intensity and emotion type hierarchically because of the following reasons.

1. The motivation of affective content analysis is to let users directly access video segments of their interests. Sometimes, users might want to watch content with high emotion intensity, where they cannot name the detailed emotion type.
2. From films' point of view, for some film genres, content with high emotion intensity directly links to video highlights, especially in horror movies and action movies. However, those highlights cannot be represented by only a detailed emotion type. For example, the action sequences in action movies cannot be simply represented by angry, fear or any particular emotions.

2. Related work

Affective analysis in Human Computer Interaction (HCI) had been researched for many years. Research on affective theories showed that valence (pleasure), arousal (agitation) and control (Dominance) are three basic underlying dimensions of affect [21–23]. In order to achieve human multimedia interaction, affective computing attracts ever-increasing research efforts. A memory-based emotion model was developed and combined with animation for the virtual character [24]. Therefore, the character could adjust the current interaction based on the existing relationship with the users. In [25], emotional sensitivity was considered for avatar to respond to users emotional nuance. To achieve HCI and data-driven animation applications, automatic facial expression recognition became an interesting and challenging problem [26,27]. Emotion factors were also considered for smart home applications [28].

With the exponential growth in the production of video and the development of personalized multimedia services, emotional decisions are more and more widely utilized in identifying user preferred content. Existing video affective content analysis had been conducted in several video domains, such as sports [8], music videos [9] and similar media outlets. Successful detection of segments with excited moments from sports video by audio analysis had been achieved in [8]. Music video affective analysis, visualization and retrieval had been implemented by using arousal and valence features [9]. Movie affective content analysis is still a challenging task due to the inflexible video syntax structure and many emotions aroused in a film, though movie affective content can provide entry points for users to access their selected content directly. Recently, affective content analysis for movies attracts more and more research efforts. Kang [10] employed HMM on motion, color, shot cut rate to detect emotional events. Hanjalic and Xu [11] utilized the features of motion, color, and audio to represent arousal and valence. Rasheed [12] 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 [13]. Hanjalic [14] 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 [15], modules were developed for detecting video tempo and music mood. A holistic method of extracting affective information from the multifaceted stream was introduced in [16]. Arifin [17], 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 [18]. In [19], Irie et al. 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 [20]. Although the existing work is able to detect some emotion types, there still exists some limitations. For example, users might not be able to name a certain emotion type of their interests. Sometime, users might want to access movies by emotion intensity instead of emotion type. Moreover, few work considers movie genres for emotion category definition. In this paper, we propose a hierarchical emotion structure for movies and analyze movie emotions by emotion intensity and emotion types hierarchically.

Similar work was presented in [11]. In [11], Hanjalic et al. firstly proposed the amount and type of emotion and further proposed a computational framework for affective video content representation and modeling. A set of curves were used to depict the expected transitions from one feeling to another along a video. The paper then represented video emotion by affect curves, including arousal and valence curves. Later in [29], Zhang et al. utilized curve-based method and further worked on affective visualization. Curve-based representation significantly describes the arousal/valence changes of video. However, it might be difficult to evaluate the accuracy of generated affect curves. In [11,29], the affect curves are generated directly based on low-level features without consideration of user reaction. Different from

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

curved-based method, we map emotions in the arousal/valence space explicitly and label affective content with clear emotion categories by treating affective content analysis as a classification task. Since user reaction is vital for analyzing movie emotions, users reaction is considered by applying user labeled data as ground-truth to train the emotion identifier.

3. The framework of proposed hierarchical affective content analysis

Fig. 1 shows the proposed framework of hierarchical affective content analysis. Video shot is segmented from movies and used as a basic unit for affective content analysis. Arousal and valence features are extracted from video and audio respectively. At first step, emotion intensity which describes the degree of agitation is analyzed by using arousal related features. Fuzzy c-mean clustering (FCM) is used to identify three emotion intensity levels. At the second step, for each emotion intensity level, detailed emotion types are detected by creating conditional random fields (CRFs) model with valence related features and intensity levels. Five categories are pre-defined for emotion types: *fear*, *anger*, *happiness*, *sadness* and *neutral*.

Compared to most of the recent existing work, the proposed approach fills the gap between two classes of approaches proposed so far, namely the purely classification-based ones which map low-level features directly onto pre-specified emotion categories [10,19] and unsupervised ones, like affect curve based method [11,9]. The novel definition of emotion categories considers film genres and covers typical emotions in the popular film genres. The proposed definition has a hierarchical structure which provides users' flexibility to access movie data by either emotion intensity or emotion type. Moreover, defined categories are distinguishable in arousal–valence emotion space. Fuzzy c-mean clustering (FCM) is used to deal with uncertain intensity boundaries. FCM is very close to human understanding [30]. For different film genres, the benchmark for selecting intensity level could be different. Outperforming Hidden Markov Models (HMMs), conditional random fields (CRFs) are used to capture the context information. By using CRF, the current emotion detection not only takes the previous emotion states into account but also considers arbitrary attributes of the observation data. As a continuous research of our initial investigation [31], this study improves the following two points and performs comprehensive experiments: (1) CRF model is developed in this study to relax the independence assumption for states required by HMM and avoids bias problem. (2) User study is performed to evaluate the proposed approach since emotion depends on individual's understanding and perception.

4. Hierarchical emotion categories

Psychological studies represent emotions in three dimensional space, which are arousal, valance and dominance [23,22]. Arousal describes the intensity of emotion. Valance is related to the type of emotion, which describes the pleasure–displeasure. Dominance is related to submissiveness and control, which lacks understanding and its variance does not account for a significant factor of emotions represented by multimedia. Therefore, movie affective analysis is carried on the first two emotion dimensions, i.e. arousal and valence. A hierarchical structure of emotion categories (see Fig. 1) are defined by considering the following issues. (1) The definition for emotion categories should be based on arousal and valence. (2) The pre-defined emotion categories should include the typical emotions in the popular film genres. (3) The emotion categories should provide possible inference for movie highlights and be defined for users to access their interested content easily. (4) The emotion categories should be distinctive in two dimension arousal–valence emotion space which makes the categories distinguishable by using valence and arousal related features.

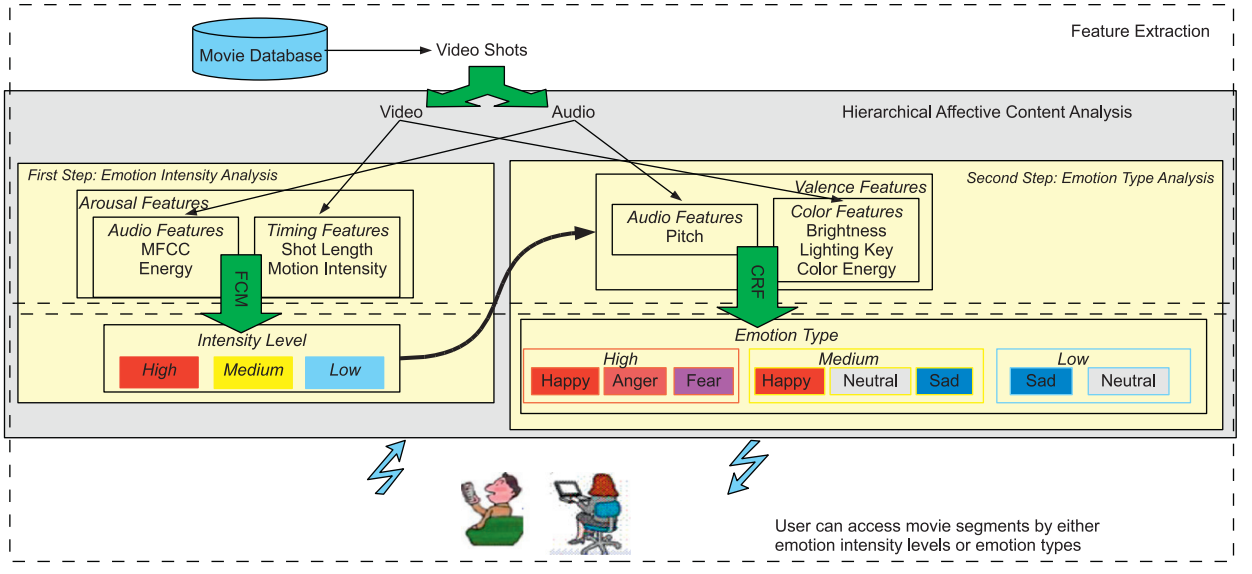

Fig. 1. The framework of hierarchical affective content analysis.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

4.1. Emotion intensity levels

At the first level of the hierarchical structure, intensity levels are defined to represent the degree of arousal. Three levels of intensity are defined: (1) *High*: expresses emotions which make persons high or agitated, such as angry. Normally, movie highlights overlap with the content with *high* because highlights are always the content trying to make audience engaged, especially for some film genres, such as action and horror. (2) *Low*: describes those emotions which make persons low or calm, such as sad or happy (romantic). (3) *Medium*: is in between of *high* and *low*. Normally, emotions at this level are relatively peaceful or not very significant.

4.2. Emotion types

According to recent report of 'The Numbers' [32], the most popular four film genres are Action (Adventure), Horror (Thriller) , Comedy (Romance) and Drama (Melodrama). In horror movies, the dominant emotion is *fear (nervous)*. Happiness (joy) is the main emotion in comedy. It is difficult to find a dominant emotion for drama which is always full of emotions. Action movie is a film genre where action sequences, such as fights, shootouts, stunts, car chases and so on happen regularly. Those action sequences attract a lot attentions from audiences and make them stay with emotions of high intensity. The emotion could be anger, fear, happiness and so on. Ekman's basic emotions were identified by studying human facial expression and proved to be universal among humans [33]. In [16], Wang et al. adapted Ekman's basic emotions to six categories in order to have significant relevance in the film context and describe nearly all emotions in films. Based on the emotion list in [16], we elide surprise since surprise covers a big range of valence. Surprise can be pleasant, or unpleasant. Moreover, surprise is not dominant emotion of any popular film genres. Five emotion types at the second level of the hierarchical structure. *Fear* is an emotional response to a perceived threat. *Anger* is related to one's psychological interpretation of having been offended, wronged or denied and a tendency to undo that by retaliation. *Happiness* is an emotion characterized by contentment, love, satisfaction, pleasure, or joy. *Sadness* is an emotion characterized by feelings of disadvantage, loss, and helplessness. *Neutral* is used to represent no significant emotions.

According to [23,34], especially [16], the distribution of pre-defined emotion types in AV space can be roughly visualized in Fig. 3. Each emotion is distinguishable from the other as clearly shown in Fig. 3.

5. Emotion intensity detection

Emotion intensity analysis needs to provide enough candidates for further emotion type detection. Enlarging the cluster is better than losing some candidates.

5.1. Arousal feature extraction

Arousal features are extracted from both audio and video streams and further used as input to FCM for affective intensity detection.

5.1.1. Timing features

According to movie theory, timing is an important feature of films and has significant power to attract viewers' attention and to affect viewers' emotion intensity [35]. In particular, timing is about the duration and duration relationships. We defined duration and duration relationship in our previous work [36].

Duration is the length of time of similar images. A video shot is created of a series of frames (similar images), which runs for an uninterrupted period of time. Therefore, shot-length is an efficient representation for duration of similar images. We have the experience that the shot duration in a film can directly affect our emotional responses to the movie. Fast-paced montage and rapid cross-cutting often work directly to create feelings of excitement. On the other hand, if the duration of an image is longer than what we expect, we might think about why the shot continues for so long. Shot-length is represented by the number of frames in one shot.

Duration relationship indicates the relationship (difference) among images. We consider representing duration relationship by average motion intensity within one shot because motion is estimated from the difference between two frames. A sudden, explosive motion produces a startle response. Motion intensity roughly estimates the gross motion in the whole frame, including object and camera motion, which is computed as the average magnitude of motion vectors in a frame:

$$MV = \frac{1}{|\Phi|}\sum_{\phi\in\Phi}\sqrt{v_x(\phi)^2+v_y(\phi)^2} \tag{1}$$

where Φ is the set of inter-coded macro-blocks, and $\vec{v}=[v_x(\phi),v_y(\phi)]$ is the motion vector for macro-block ϕ. Then the average motion intensity is calculated for the whole shot.

5.1.2. Arousal related audio features

Sounds were shown to have emotion-related meanings [37].

Short-time energy (*STE*) is proved to be an effective audio feature and wildly used for vocal emotion detection [38]. According to the findings in [37], energy is related to evoked arousal. STE is measured by the amplitude of the audio signal varying with time. The energy of discrete-time signal is defined as

$$E_n = \sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2 \tag{2}$$

where x is the audio signal, and $w(n-m)$ is a windowing function. $w(n)=1$, if $0\le n\le N-1$ (N is the length of the window in samples). Otherwise, $w(n)=0$. Generally speaking, high energy corresponds to high emotional level while the audio segments with low emotional level have lower energy.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

Mel frequency cepstral coefficients (*MFCC*) works well for excited and non-excited audio detection [8]. Mel scale is defined as

$$F_{mel} = \frac{1000\log(1+f/1000)}{\log(2)}, \tag{3}$$

where F_{mel} is the logarithmic scale of f normal frequency scale. The mel-cepstral features can be illustrated by the mel-frequency cepstral coefficients (MFCCs), which are computed from the FFT power coefficients. The power coefficients are filtered by a triangular band pass filter banks. The filter bank consists of 19 triangular filters. They have a constant mel-frequency interval, and cover the frequency range of 0–20050 Hz. The MFCCs are calculated as

$$C_k = \sqrt{\frac{2}{k}}\sum_{k=1}^{K}(\log S_k)\cos[n(k-0.5)\pi/k], \quad n = 1,..,L, \tag{4}$$

where S_k $(k = 1,2,\ldots,K)$ is the output of the filter bank. In this research, the first four coefficients are used to generate feature vectors since they were proved to be effective for audio classification in our previous work [31].

5.2. Fuzzy c-mean clustering

Fuzzy c-mean clustering (FCM) is one of the most widely used fuzzy clustering algorithms [30]. We choose fuzzy clustering because of three reasons.

1. Fuzzy logic is based on natural language which is close to human understanding. Meanwhile, intensity level is a subjective concept heavily depending on human perception, which has unclear boundary. Fuzzy interpretations of data structures are a very natural and intuitively plausible way to formulate and solve various uncertain problems in pattern recognition.
2. Different from other classification algorithm, in fuzzy clustering, the data points can belong to more than one cluster in order to provide enough candidates for emotion type detection.
3. For different film genre, the benchmark for selecting intensity level might be different. In Fuzzy clustering, the membership grades are associated with each of the points, which indicate the degree of the data points belonging to the different clusters. We can set threshold on membership grades for each film genre to select intensity levels.

FCM is based on minimization of the objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \|x_i - c_j\|^2, \quad 1 \le m < \infty, \tag{5}$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{2/(m-1)}}, \tag{6}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{7}$$

This iteration will stop when $\max_{ij}\{|u_{ij}^{(k+1)} - u_{ij}(k)|\} < \epsilon$, where ϵ is a termination criterion between 0 and 1, whereas k is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

In our study, three clusters are set to indicate three emotion intensity levels. Considering different film genres has its own benchmark for emotion intensity levels, clustering is performed for each film genre separately. The outputs include center matrix, fuzzy partition matrix and the values of the objective function during iterations. Center matrix of final cluster centers indicates each feature dimension's coordinates. Fuzzy partition matrix is composed of membership functions. By checking the sample points close to each cluster center, we can identify the corresponding intensity level to the cluster. We enlarge or diminish the area for each cluster on data space by setting thresholds for the partition degree experimentally to achieve the best results.

6. Emotion type detection

Emotion types are further detected by performing CRFs on valence features and detected emotion intensity levels (as shown in Figure).

6.1. Valence feature extraction

Valence features are also extracted from both video and audio streams.

6.1.1. Valence related visual features

We firstly describe visual valence features. *Brightness* is dramatically exploited to evoke emotions. According to [34], valence is strongly correlated to brightness. In movies, the abundance of brightness is to highlight the pleasant atmosphere. In Contrast, low valence emotions, such as fear or sad, are romanced by dim scenes. We simply use the brightness value in HSB (Hue, Saturation, Brightness) model, which is also known as HSV (Hue, Saturation, Value).

Lighting key measures the contrast between dark and light. Besides brightness, light and shade are used together in movie scenes to create affective effects. High-key lighting with high brightness and small light/dark contrast is normally used for joyous scenes, whereas low-key lighting represents heavy light/dark contrast to emphasize unpleased feelings. The lighting key is calculated for each frame by the mean and the deviation of brightness for each pixel [12].

Color energy is used to measure the color contrast. Colorful scenes are used to evoke joyous feelings. Color energy is calculated by the product of the raw energy and color contrast [16].

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

6.1.2. Valence related audio features

Pitch is successfully used as one of the valence features in [11]. Pitch is significant for emotion detection, especially for the emotions in speech and music. According to the findings in [37], average pitch in speech signals is related to valence. We use the same method of [11] to calculate pitch.

6.2. Conditional random fields (CRFs)

For emotion type detection, a statistical model of sequential data should be applied to capture the context information because of the following two reasons:

- Video/audio signal exhibits the consecutive changes in values over a period of time, where variables may be predicted from earlier values.
- The occurrence of a certain emotion depends not only on the current status, but also on the recent emotional history.

In our previous work, Hidden Markov Models (HMMs) were used for emotion type detection [31]. For HMMs, generative models, the parameters are typically trained to maximize a joint probability of paired observation and label sequences. HMMs need to enumerate all possible observation sequences, which are represented as independent from the other elements in an observation sequence. However, for the task of emotion type detection, observation sequences are best represented in terms of multiple interacting features and long-range dependencies between observation elements. Therefore, conditional random field (CRF) is used in this paper to support tractable inferences of emotion type. CRF model defines a conditional probability $p(y|x)$ specifying the probabilities of possible label sequences y given a particular observation sequence x, rather than a joint distribution over both label and observation sequences defined by HMM. The conditional nature of CRF results in the relaxation of the independence assumptions required by HMMs and ensure tractable inference. Therefore, the selected features can represent attributes at different levels of granularity of the same observations or aggregate properties of the observation sequence. Moreover, CRF solves the label bias problem of other conditional Markov models by applying a single exponential model for the joint probability of entire sequence of labels given the observation sequence. The weights of different features at different states can be traded off against each other.

As shown in Fig. 2, the conditional probability of label sequence y given observation sequences x is defined as

$$p(y|x,\wedge)=\frac{1}{Z(x)}\exp\left(\sum_{j=1}^{J}(\lambda_j F_j(y,x)\right) \tag{8}$$

where $\wedge=\lambda_1,\ldots\lambda_J$ is a set of parameters of CRF, J is the number of feature functions. And, $Z(x)$ is a normalization factor:

$$F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x_1,x_2,x_3,x_4,x_5,i) \tag{9}$$

n is the length of label and observation sequences. In our study, the observation x has five observation sequences $x_1,x_2,\ldots,x_5$ which are Brightness, Lighting key, Color energy, Pitch and Emotion intensity. In the training process, parameters are trained by $\mathrm{Argmax}(\log(p(y|x_1,x_2,x_3,x_4,x_5,\wedge)))$. L-BFGS quasi-Newton method is used for training [39]. After training, the CRF can be used for emotion type detection. Emotion type Y will be labeled if $p(Y|X_1,X_2,X_3,X_4,X_5,\wedge)$ is maximized.

7. Experiments

Experiments include two main parts: emotion intensity detection and emotion type detection. We use 720 min videos from eight movies, including 6201 video shots, to test the proposed approach. Experimental data cover four movie genres (Action, Horror, Drama and Comedy). Ten volunteers label experimental data manually. Since the duration of video shot is short, a video segment which contains a number of video shots is used as a unit for labeling. The volunteers are asked to watch the movie and label video segments with both emotion intensity level and emotion type. The duration of a video segment is decided by each volunteer when labeling. One video segment should have same emotion labels consistently. The shots within one segment have a same intensity label and a same emotion type label. Each movie is labeled by at least three volunteer. The final labels (i.e. intensity level and emotion type) for each shot are assigned by majority voting. For each movie genre, video

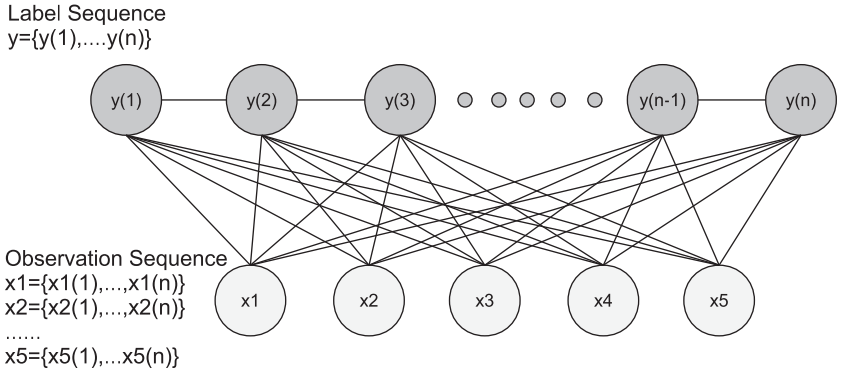

Fig. 2. The proposed CRF model.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

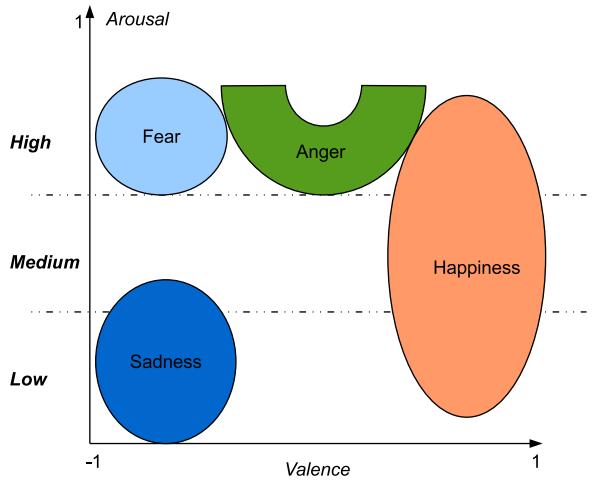

Fig. 3. The distribution of emotion types in AV.

of 180 min are collected from two movies of that genre. The shot distribution of emotion types in different film genres can be found in Table 1. For emotion type detection, parameters of CRF need to be decided by supervised learning. To make the training data of each emotion type almost even, 720 min of the training data are picked up and composed from other 16 movies which are different from the data mentioned above. Manually labeled emotion intensity level are used for training. The emotion intensity level outputted from fuzzy clustering is used as one dimension of the input for CRF testing.

7.1. Emotion intensity detection

Fig. 4 plots the experimental results for emotion intensity analysis. For display purpose, only two dimensions out of eight are used to illustrate the feature distribution and clustering results for 259 video shots from horror movies. From the upper-left plot, upper-right plot and lower-right plot, we can find that the high emotional level shots mostly take place along with the high motion intensity and fast shot changing, i.e. short shot-length. The plot of upper-right, lower-left, and lower-right shows that high emotional level shots take place along with the high energy. Upper-left and lower-left plots show that the low emotional level shots take place when the MFCC is low. By comparing with manfully labeled data, we find that the content with high intensity covers over 80% of the horror and action movie highlights.

7.2. Emotion type detection

Emotions have some persistence in time. A sudden change in the shot sequences can be considered as an error. A sliding window is exploited on the sequence of detected emotion type with window length of 4 and step-size of 1 to eliminate those sudden changes by majority-voting. The numbers of correctly detected shots in each emotion type are listed in Table 1. Accuracies in Tables 2–4 are further calculated from Table 1.

Table 1
Testing data (shots) distribution and emotion type detection results (eight movies, 720 min).

Genre	Action (180 min)					Horror (180 min)				
Movie	Star wars and speed					The ring and face				
Type	F	A	H	S	N	F	A	H	S	N
Manual	471	823	84	72	580	963	67	18	99	606
One-step	342	691	64	54	431	819	50	12	67	461
Hierarchical(HMM)	359	718	60	62	424	843	56	10	67	458
Hierarchical(CRF)	368	724	62	64	430	828	56	11	69	442
Genre	Drama (180 min)					Comedy (180 min)				
Movie	Cold mountain and Billy Elliot					Love actually and Mr. Bean's holiday				
Type	F	A	H	S	N	F	A	H	S	N
Manual	54	306	198	216	414	16	41	904	44	319
One-step	33	218	163	180	320	11	25	760	25	219
Hierarchical(HMM)	36	224	157	178	315	12	28	754	25	215
Hierarchical(CRF)	40	231	163	179	318	11	31	756	27	225

Note: F: Fear; A: Anger; H: Happiness; S: Sadness; N: Neutral.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

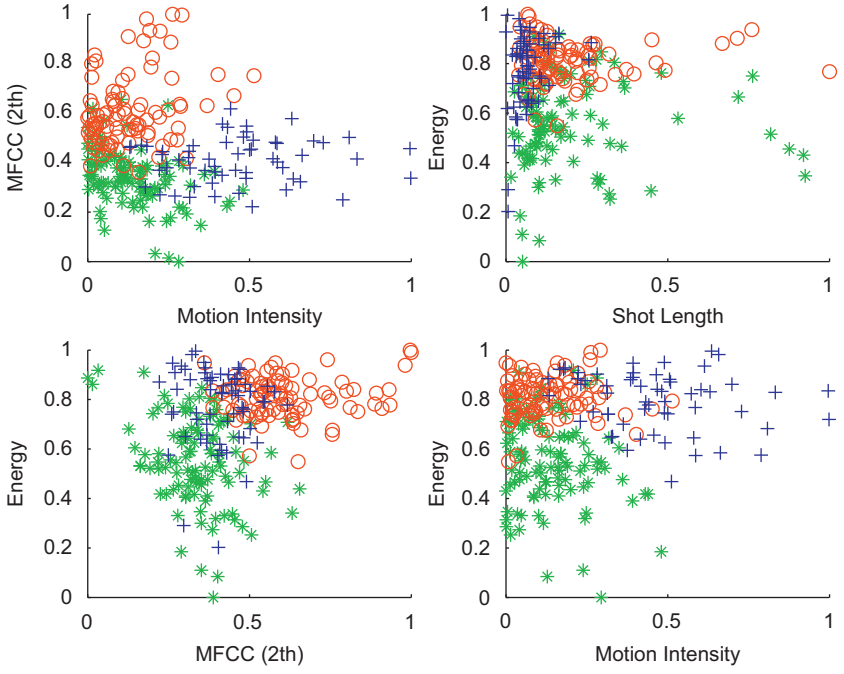

Fig. 4. Results of Fuzzy Clustering (Blue '+': High; Green '*': Low; Red '○': Medium). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

To justify the efficiency of the proposed hierarchical method, we implement one-step method which is widely used in affective content analysis for comparison on the same data set. In the experiments of one-step method, both arousal and valence features are mixed to generate observation vectors as the inputs to HMMs. We also implement HMM-based hierarchical approach to compare with CRF-based hierarchical approach. Tables 2, 3 and 4 show the detection results by using one-step method, HMM-based hierarchical approach and CRF-based hierarchical approach.

From Table 2, we find that CRF-based hierarchical approach outperforms one-step approach for all movie genres used in experiments. Compared to one-step method, the hierarchical method increases the accuracy of 2.3% on average for Action and Horror movies. One possible reason is that emotion intensity analysis reduces the detection range for the dominant emotion types (*fear*, *anger*, etc.) in action and horror movies. The accuracy for emotion type detection in Action (83.8%) and Horror (81.8%) movies are better than that in Drama (76.6%) and Comedy (78.0%). It might be due to the less emotion contrast in Drama and Comedy than that in Action and Horror.

By using CRF-based hierarchical approach, the improvement of accuracy for *fear*, *anger* and *sadness* are 4.85%, 7.96% and 5% respectively. For some emotion categories, such as *happiness* and *Neutral*, the detection accuracy by using hierarchical approach is not significantly increased. It might be because that *happy* crosses three emotion intensity levels (as shown in Fig. 3). The Hierarchical method needs to identify the samples of *happiness* from three different intensity levels. Accumulated error from emotion intensity levels detection might affect the accuracy of emotion type detection. Even though, the accuracy of *happiness* detection using hierarchical method is still satisfactory. Meanwhile, the wrong classification of *happiness* affects classification result of *Neutral* more or less.

As a significant improvement to one-step approach, hierarchical approach provides viewers flexibility to access movie emotions by both emotion intensity and emotion type. In order to prove the above claim, we further perform a user study in the following section. Compared to our previous work of HMM-based hierarchical method [31], it is easy to find that CRF-based method outperforms HMM-based method in most of the categories.

7.3. User evaluation

To further evaluate the performance of our approach, user studies are carried out among 30 movie viewers. Considering the age effects on evaluation, the viewers cover a age gap from 22 to 60. There are two purposes of user study. Firstly, we want to know whether viewers prefer to have emotion intensity as an option for them to select content of interests. Secondly, we want to investigate users' satisfaction on movie emotion query based on our hierarchical emotion analysis. For the first purpose, each viewer is firstly requested to watch a whole movie. Then, detected movie segments with consistent emotion label are shown to them according to their queries of either a emotion intensity or a certain emotion type. After watching, they are requested to answer a question: 'compared to accessing video segments by selecting only emotion types, do you think accessing movie segments by emotion intensity can rich your watching experience?'.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

Table 2
Emotion type detection accuracy (%) I (eight movies, 720 min).

Genre	Action (180 min)					Horror (180 min)				
Movie	Star wars and speed					The ring and face				
Type	F	A	H	S	N	F	A	H	S	N
One-step	72.61	83.96	76.19	75.00	74.31	85.05	74.63	66.67	67.68	76.07
HMM	76.22	87.24	71.43	86.11	73.10	87.54	83.58	55.56	67.68	75.58
CRF	78.13	87.97	73.81	88.89	74.14	85.98	83.58	61.11	69.70	72.94
Genre	Drama (180 min)					Comedy (180 min)				
Movie	Cold mountain and Billy Elliot					Love actually and Mr. Bean's holiday				
Type	F	A	H	S	N	F	A	H	S	N
One-step	61.11	71.24	82.32	83.33	77.29	68.75	60.98	84.07	56.82	68.65
HMM	66.67	73.20	79.29	82.41	76.09	75.00	68.29	83.41	56.82	67.40
CRF	74.07	75.49	82.32	82.87	76.81	68.75	75.61	83.63	61.36	70.53

Note: F: Fear; A: Anger; H: Happiness; S: Sadness; N: Neutral.

Table 3
Emotion type detection accuracy (corresponding to emotion type) (eight movies, 720 min).

	F	A	H	S	N
One-step	71.88	72.70	77.31	70.71	74.08
HMM	76.36	78.08	72.42	73.26	73.04
CRF	76.73	80.66	76.11	75.71	73.61

Note: F: Fear; A: Anger; H: Happiness; S: Sadness; N: Neutral.

Table 4
Emotion type detection accuracy (corresponding to movie genre) (eight movies, 720 min).

	Action	Horror	Drama	Comedy
One-step	80.18	74.02	75.06	67.85
HMM	78.82	73.99	75.53	70.18
CRF	80.59	74.66	78.31	71.98

Table 5
User evaluation of emotion detection.

	Bad	Poor	Fair	Good	Excellent
Group 1	0 (0%)	5 (16.7%)	11 (36.7%)	13 (43.3%)	1 (3%)
Group 2	1 (3%)	4 (13.3%)	10 (33.3%)	14 (46.7%)	1 (3%)

From their answers, we find over 86% viewers (26 viewers) agree that emotion intensity provides them flexibility to access movie segments, especially when they are not sure of detailed emotion types. For the second purpose, we adopt the double stimulus impairment scale (DSIS) method [40] with some modifications to evaluate users' satisfaction on the video segments response to their request. Five scales from "bad" to "Excellent" are used for users to vote for their satisfactory. For the second purpose, two groups of user study are designed as: (1) without pre-knowledge of movies; (2) with pre-knowledge of movies. Table 5 shows the number of voting from reviewers for each scale.

7.3.1. Group 1: without pre-knowledge of movies

In the first group, viewers are asked to select an intensity level or emotion type at one time. According to viewer's selection, video shots are randomly selected from viewer preferred category. Continuous video shots with same emotion categories are played together. Viewers are required to repeat the above process five times. Then, viewers are required to vote their satisfaction by selecting one from "Excellent", "Good", "Fair", "Poor" or "Bad". From Table 5, we find that most of the voting (80%) concentrate on "Fair" and "Good". There are 3% viewers vote for "Excellent". This indicates that most of the users are satisfied with the current emotion detection.

7.3.2. Group 2: with pre-knowledge of movies

In the second group, viewers are firstly asked to watch a movie from beginning to the end. After that, they are further required to repeat the experiments in Group 1 for the movie they have watched. Same as Group 1, there are

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026

still 83% viewers vote for "Fair", "Good", or "Excellent". Compared to Group 1, 3.4% more viewers feel "Good" with the detected emotional content. By comparing the voting from the same viewer between two groups of experiments, we find that 13 viewers changed their voting. Five of them think that the emotion detection become less satisfied in Group 2. The reason they told us is that the emotion detection cannot provide their expected video shots. These expected video shots are impressive to them while they are watching the movies. Eight of them feel more satisfied in Group 2 because they think some emotions are evoked in movie context. Therefore, after watching the whole movie, they can easily understand the detected emotional shots.

7.4. Discussion

User study indicates that most of the viewers prefer to use emotion intensity as a query option to access movie segments. Moreover, most of the viewers are satisfied with emotion based queries which heavily depend on our emotion detection. By comparing two group experiments, we find that viewers understanding of emotions affected by the context in movies. Comparing viewer feedback from two groups, we find that movie context affects viewers' understanding of movie emotions. Without movie context, viewers sometimes cannot efficiently understand emotions evoked in movies.

8. Conclusions

A hierarchical structure of emotion category is proposed in this study. Based on the hierarchical emotion definition, emotion intensity and emotion type are detected in two steps. Fuzzy c-mean clustering is well performed on arousal features to find out three levels of emotion intensity. Over 80% of the movie highlights are detected by finding content with high emotion intensity for horror and action movies. Valence features together with emotion intensity are used for CRF-based emotion type identification. Experimental results show that the proposed hierarchical emotion detection method not only outperforms one-step method but also provides users flexibility to access movie segments by either emotion intensity or emotion type. CRF outperforms HMM for emotion type detection. User study is carried out among 30 viewers. Most of the viewers are satisfied with emotion detection results. Pre-knowledge of the movie content helps viewers to understand detected emotional content. Over 86% viewers agree that the proposed emotion detection method provides flexibility for user to access their interested movie content. Currently, due to lacking public data, ground truth is labeled manually, which is relative subjective. Internet videos will be considered for experiments in future work, because most of the Internet videos are published under certain categories and with viewers' comments. Social data is quite significant for both ground truth labeling and affective content detection. Affective content detection is meaningful for emotion based human video interaction. Emotion-based interaction provides a platform for user to access videos according to the emotional factors. In future, the affective content analysis will be extended to other applications, especially for Internet video recommendation. On the other hand, the interaction has the capability to allow video to response to users' emotion requirements. Human emotion or preference on emotions should also be considered to achieve an automatic interaction. Social data are also very important to estimate users' preference which will be investigated to help video recommendation.

Acknowledgements

The work is supported by National Natural Science Foundation of China No. 61003161 and UTS Early Career Research Grant.

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafiner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the {QBIC} System, IEEE Computer Special Issue on Content-Based Retrieval 28 (9) (1995) 23–32.
[2] J. Wei, Z.N. Liu, I. Gertner, A novel motion-based active video indexing method, Proceedings of IEEE International Conference on Multimedia Computing and System, 2 (1999) 460–465.
[3] E. Ardizzo, M.L. Cascia, V.D. Gesu, C. Valenti, Content-based indexing of image and video databases by global and shape features, in: Proceedings of 13th International Conference on Patter Recognition, vol. 3, 1996, pp. 140–144.
[4] A. Hanjalic, Shot-boundary detection: unravelled and resolved, IEEE Transactions on Circuits and Systems for Video Technology 12 (2) (2002) 90–105.
[5] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Transactions on Multimedia 11 (3) (2009) 465–476.
[6] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multi-graph learning, in: IEE, Transactions on Circuits and Systems for Video Technology 19 (5) (2009) 733–746.
[7] L.Y. Duan, M. Xu, C. Xu, Q. Tian, A unified framework for semantic shot classification in sports video, IEEE Transactions on Multimedia 7 (6) (2005) 1066–1083.
[8] M. Xu, N.C. Maddage, C.-S. Xu, M. Kankanhalli, Q. Tian, Creating audio keywords for event detection in soccer video, Proceedings of International Conference on Multimedia & Expo, 2 (2003) 143–154.
[9] S. Zhang, Q. Huang, Q. Tian, S. Jiang, W. Gao, i.MTV: an integrated system for MTV affective analysis, in: Proceedings of ACM International Conference on Multimedia, 2008, 985–986.
[10] H.-B. Kang, Affective content detection using HMMs, in: Proceedings of the ACM Multimedia Conference, 2003, pp. 259–262.
[11] A. Hanjalic, L.Q. Xu, Affective video content representation and modeling, IEEE Transaction on Multimedia 7 (1) (2005) 143–154.
[12] Z. Rasheed, Y. Sheikh, M. Shah, On the use of computable features for film classification, IEEE Transactions on Circuits and Systems for Video Technology 15 (1) (2005) 52–64.
[13] C. Chan, G.J.F. Jones, Affect-based indexing and retrieval of films, in: Proceedings of the ACM Multimedia Conference, 2005, pp. 427–430.
[14] A. Hanjalic, Extracting moods from pictures and sounds: towards truly personalized TV, IEEE Signal Processing Magazine 23 (2) (2006) 90–100.
[15] Y.-H. Chen, J.-H. Kuo, W.-T. Chu, J.-L. Wu, Movie emotional event detection based on music mood and video tempo, in: Proceedings of IEEE International Conference on Consumer Electronics, 2006, pp. 151–152.
[16] H.L. Wang, L.F. Cheong, Affective understanding in film, IEEE Transactions on Circuits and Systems for Video Technology 16 (6) (2006) 689–704.
[17] S. Arifin, P.Y.K. Cheung, User attention based arousal content modeling, in: Proceedings of the International Conference on Image Processing, 2006, pp. 433–436.
[18] S. Arifin, P.Y.K. Cheung, A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information, in: Proceedings of the ACM Multimedia Conference, 2007, pp. 68–77.
[19] G. Irie, K. Hidaka, T.Satou, A. Kojima, T. Yamasaki, K. Aizawa, Latent topic driving model for movie affective scene classification, in: Proceedings of the ACM Multimedia Conference, 2009, 565–568.
[20] M. Soleymani, J.J.M. Kierkels, G. Chanel, T. Pun, A bayesian framework for video affective representation, in: Proceedings of the

International Conference on Affective Computing and Intelligent interaction, 2009.
[21] M. Bradley, Emotions: essays on emotion theory, in: Lawrence Erlbaum, Hillsdale, NJ, Stephanie H.M. van Goozen, N.E. van de Poll, J. A. Sergeant, 1994, 97–134.
[22] P. J. Lang, The network model of emotion: motivational connections, Perspectives on Anger and Emotion, Advances in Social Cognition, Psychology Press, in: Robert S. Wyer, Thomas K. Srull, (Eds.), 1993, pp. 109–133.
[23] J.A. Russell, A. Mehrabian, Evidence for A Three-factor theory of emotions, Journal of Research in Personality 11 (3) (1977) 273–294.
[24] Z. Kasap, M. Benmoussa, P. Chaudhuri, N. Magnenat-Thalmann, Making them remember-emotional virtual characters with memory, IEEE Computer Graphics and Applications 29 (2) (2009) 20–29.
[25] T.S. Huang, M.A. Hasegawa-Johnson, S.M. Chu, Z. Zeng, H. Tang, Sensitive talking heads, IEEE Signal Processing Magazine 26 (4) (2009) 67–72.
[26] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, T.S. Huang, Authentic facial expression analysis, Image and Vision Computing 25 (12) (2007) 1856–1863.
[27] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image and Vision Computing 27 (6) (2009) 803–816.
[28] A. Khalili, C. Wu, H. Aghajan, Autonomous learning of user's preference of music and light services in smart home, in: Proceedings of the behavior monitoring and interpretation workshop at German AI Conference, 2009.
[29] S. Zhang, Q. Tian, Q. Huang, W. Gao, S. Li, Utilizing affective analysis for efficient movie browsing, in: Proceedings of IEEE International Conference on Image Processing, 2009, 1853–1856.
[30] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algoritms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[31] M. Xu, J.S. Jin, S. Luo, L. Duan, Hierarchical movie affective content analysis based on arousal and valence features, in: Proceedings of the ACM Multimedia Conference, 2008, pp. 677–680.
[32] Nash Information Services, LLC, The Numbers — Movie Box Office Data, Film Stars, Idle Speculation <http://www.the-numbers.com/charts/weekly/2009/20091225.php>, 2009.
[33] P. Ekman, Universals and cultural differences in the judgments of facial expressions of emotion, Journal of Personality and Social Psychology 54 (4) (1987) 712–717.
[34] P. Valdez, A. Mehrabian, Effects of color on emotions, Journal of Experimental Psychology: General 123 (4) (1994) 394–409.
[35] C. Plantinga, G.M. Smith, Passionate Views: Film, Cognition, and Emotion, The Johns Hopkins University Press, 1999.
[36] M. Xu, S. Luo, J.S. Jin, Affective content detection by using timing features and fuzzy clustering, in: Proceedings of the Advances in Multimedia Information Processing — PCM 2008, Lecture Notes in Computer Science, vol. 5353, 2008, pp. 685–692.
[37] R.W. Picard, Affective Computing, The MIT Press, Cambridge, MA, 2000.
[38] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, S. Levinson, Audio–visual affective recognition, IEE, Transaction on Multimedia 9 (2007) 424–428.
[39] E. Chong, S. Zak, An Introduction to Optimization, Wiley, New York, 1996.
[40] S.-F. Chang, D. Zhong, R. Kumar, Real-time content-based adaptive streaming of sports video, Proceedings of the IEEE Workshop Content-Based Access to Video/Image Library 2001, pp. 139–146.

Please cite this article as: M. Xu, et al., Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing (2012), http://dx.doi.org/10.1016/j.sigpro.2012.06.026