KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

AFFECTIVE VIDEO ANALYSIS BY USING USERS' EEG AND SUBJECTIVE EVALUATION

Shangfei Wang*, **Yongjie Hu**

Key Lab of Computing and Communicating Software of Anhui Province,

School of Computer Science and Technology, University of Science and Technology of China

ABSTRACT

This paper describes a research project conducted to study the relationship between videos and users' induced physiological and psychological responses. Firstly, a set of 43 film clips are carefully chosen, and 20 subjects are invited to participate in our experiment. They watch several of chosen clips while their EEG signals are recorded synchronously. After each clip, the subject is required to report his real induced emotion using emotional valence, arousal, basic emotion category and intensity. Secondly, several classical movie features and EEG features are extracted, and feature selections are conducted by computing the correlation between each feature and the arousal or valence. Thirdly, selected movie features and EEG features are used to simulate the arousal and valence respectively by employing the linear relevance vector machine. Fourthly, selected movie features are used to simulate the EEG feature values, and vice verse. The results show that arousal/valence can be well estimated by either video features or EEG features. Apart from that, they also indicate that there exist certain relationship between the videos and induced EEG signals, and some relation models are acquired. Finally, clustering is conducted to map the emotion dimensions to emotion categories. Thus, the gap between videos and emotion categories, as well as the gap between the EEG and emotion categories, has been bridged to some extent. This result could provide a reference to applications in brain-computer interaction field.

Keywords: *affective video analysis, EEG, valence, arousal, emotion categories*

1. INTRODUCTION

With the enormous quantity of digital videos arising in our life, efficient and powerful video labeling and retrieval methods have become more and more important and essential.

* **Corresponding author**: School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, P.R.China, 230027, sfwang@ustc.edu.cn.

Emotion contents, as a kind of natural rules to classify and retrieve information for human, have been studied in video domain [1], [2], [3]. Most of these studies extract low-level features from videos, and adopt some machine learning methods or combine related domain knowledge, to construct emotion models with emotion categories or dimensions, which have been labeled artificially in advance. But there are some limitations in subjective assessment, such as fuzziness of the description, indeterminacy of the opinion scale and the response bias to true feeling. Apart from the assessment based on emotion, only a few of them also considered the physiological reactions[4], [5]. Physiological signals, as a kind of objective data, do not possessed above mentioned limitations, and can be taken as useful supplements to emotion subjective assessment.

In this paper, we aim at study the relationships between videos and users' induced physiological and psychological responses. The work mainly included the following four parts. First, psychophysical experiments are conducted to acquire subjects' induced physiological signals and self-assessed emotion, when they are watching video clips. Because the peripheral physiological signals have been studied [4], [5], we try to adopt electroencephalogram (EEG) signals as our physiological signals. Second, several common used EEG features and video features are extracted, and sensitive features are selected through computing the correlation between these features and self-assessed emotion dimensions (arousal/valence). Third, selected video features and EEG features are used to construct arousal/valence evaluation models by applying linear relevance vector machine (RVM [6]), and mapping models between the video contents and physiological responses are also build by the same method. Fourth, 129 evaluated video clips, which contain 3 categories of emotion (21 sad, 66 happy, and 42 fear), are used to cluster and build an emotion model mapping from emotion dimensions to specific emotion categories. The results show that emotion arousal and valence could be evaluated by both the videos and EEG signals. As a kind of physiological signals, EEG signals are hard to conceal and can't be wrongly expressed. Therefore, it provides a reference to utilize EEG as an objective emotion evaluation standard for videos. In addition, some direct relations between the videos and induced EEG signals are acquired, and the emotion clustering model also performs well based on these three basic emotion categories here.

2. PSYCHOPHYSICAL EXPERIMENTS AND DATA COLLECTION

2.1. Subjects

Twenty healthy students (18 male and 2 female, from 18 to 28) attended our experiments. All of them have a normal or corrected to normal vision and auditory acuity without mental disease. In addition, they were ensured in neutral emotion before participating in the experiment. Each of them signed an informed consent before his/her participation and was paid after the experiment.

2.2. Stimuli

In our experiment, video clips were taken as stimulus to induce subjects' emotion. All used 43 video clips were picked from our constructed affective video dataset, which contained 21 sad, 66 happy, 5 disgust and 42 fear clips. These total 134 video clips were extracted from 4 different genre (horror, comedy, action and drama) films, and each of them contained a full emotional event (judged by the authors), which had a duration of approximately 2 to 7 minutes. These clips were pre-tested by 3 students, and all the pre-set emotion categories were identified by them. Finally, 43 video clips, which were composed of 15 happy, 12 fear, 11 sad and 5 disgust clips, were chosen from our constructed database.

2.3. Experiment procedure

Every subject watches 6 film clips selected from our database, and his/her EEG signals are recorded synchronously. After watching each film clip, the subject is required to report his real induced emotion using emotional valence, arousal, basic emotion category and intensity. We carefully design the experiments to confirm that every film clip is watched by at least two subjects. To reduce the interaction of different emotions induced by different clips, neutral clips were shown to subjects at the intervals between two different clips.

In addition, All EEG signals were acquired via a Quik-cap (Neuro Inc., El Paso, TX) with 64 Ag-AgCl electrodes arranged in an extended 10-20 system montage. Neuroscan Synamps2 bioamplifiers were used, and EEG signals were recorded using Neurscan Scan software (v 4.3.1) with the set of continuously digitize (250Hz sample rate), amplify (gain of 1000), filter (70Hz low-pass filter, without a notch filter), and mode (AC).

The details of the procedure are listed as follows. First, the subjects were given an introduction about the arrangement of the experiment, the meaning of arousal and valence, and how to self-assess their emotions. Second, he/she is asked to sit on a sofa in a special set isolating room, and wear the Quick-cap. A 29 inches LCD is placed in front of the sofa with 1 meter distance to play the videos, and an earphone is also given. Third, experiments are conducted according to the experimental arrangement, which is represented in Figure 1. After watching an eliciting video clip, the subjects filled his/her self-assessment emotion in a form which popped up automatically. And then a neutral clip was played to recover the subject' s fluctuation of emotion. To reduce the order effects for elicited emotion, movie clips were allocated randomly to each subject. During the self-assessment step, the arousal and valence are evaluated in 5-point scale, and then the category of emotion is notched while its intensity is also evaluated in 5-point scale.

Finally, 120 EEG signal recordings and 120 emotion self-assessed recordings correspond to 43 video clips are acquired. Emotion self-assessed recordings are averaged according to the corresponding video clips.

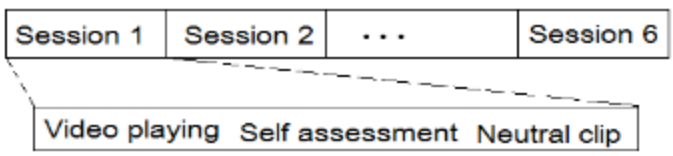

Figure 1: Experimental arrangement

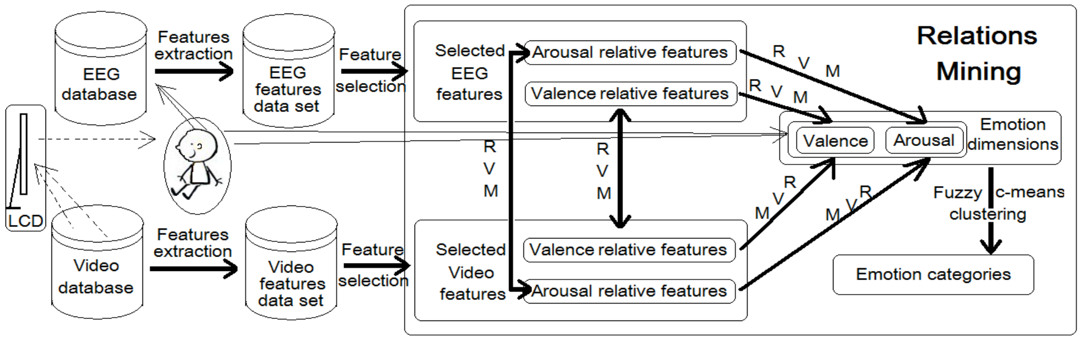

Figure 2: Framework of our approach

3. FEATURE EXTRACTION AND RELATIONS MINING

In this section, we study the relationships among the videos, EEG signals and self-report emotion evaluations from the above data. The framework is showed in Figure. 2, which composes of feature extractions, feature selections and relations mining.

3.1. Feature extraction

3.1.1. Audio and visual low-level features

Audio information does affect human's emotion [7], and its features had been studied in audio retrieval and audio classification [8], [9], and have obtained a pretty good effect. Therefore, the features, which are widely used in audio and speech processing, are chosen to construct our audio feature vector. The audio feature vector used here contained *74* low-level audio features, which were listed in Table 1.

Table 1: Low-level audio features

Feature set	Extracted features
Energy	Average energy of the audio features
Formants	Average and variance of the First 3 formants (6 features)
Fundamental frequency(FF)	Average and variance of the Fundamental frequency (2 features)
LPCC	LPCC (16 features), derivative of LPCC (16 features)
MFCC	MFCC coefficients (13 features), derivative of MFCC (16 features)
Silence ratio	Proportion of silence in a time
ZCR	Zero crossing rate

In these features above, Energy, Fundamental Frequency, LPCC and MFCC were all extracted by using the PRAAT software package [10]. It should be noted that the audio channels of the video clips were extracted as WAV format with a sampling rate of 44 kHz.

From the cinematographic perspective, lighting key [11] and color energy [11] are very important to elicit emotions. Therefore, we transformed the film frames in the HSV space, and then extracted the lighting key. Color energy, which measures the joint valence-arousal quality of a scene arising from the color composition alone, was extracted in HLS space. Visual Excitement [11] presents the motion in a video sequence, and was usual thought related to the arousal. So it was also adopted as a visual feature. Finally, a total of 3 visual features have been extracted for the following processes.

3.1.2. EEG features

EEG signals are always contaminated by various noises, such as electromagnetic noises, eye blink artifacts and so on. Therefore, noise mitigation should be conducted first. As a common approach to remove various noises in EEG, Independent components analysis (ICA) [12], [13] is adopted here to mitigate noises. After the noise mitigation, all the EEG signals corresponding to the same film clip are averaged for the following feature extraction.

EEG features mainly conclude temporal and frequency features. In the frequency domain, the power spectrum [14] is computed for the delta (0-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-18 Hz), mid(18-30) and gamma (30-40 Hz) frequency bands (according to the Schwab frequency classification), which means the ratio of the power in one frequency band to the power in all six frequency bands. For signals from each channel, all the above features are extracted. At last, 448 temporal features and 384 frequency features are acquired. While in temporal domain, zero-crossing rate (ZCR) and 6 statistics μ_X, σ_X, δ_X, $\overline{\delta_X}$, γ_X, $\overline{\gamma_X}$ [15] are adopted.(In following definitions, t is the sampling number and T is the total number of sample .)

$$\mu_X = \frac{1}{T}\sum_{t=1}^{T} X(t), \quad \sigma_X = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t)-\mu_x)^2}, \quad \delta_X = \frac{1}{T-1}\sum_{t=1}^{T-1}|X(t+1)-X(t)|,$$

$$\overline{\delta_X} = \frac{1}{T-1}\sum_{t=1}^{T-1}|\overline{X}(t+1)-\overline{X}(t)| = \frac{\delta_X}{\sigma_X}, \quad \gamma_X = \frac{1}{T-2}\sum_{t=1}^{T-2}|X(t+2)-X(t)|,$$

$$\overline{\gamma_X} = \frac{1}{T-2}\sum_{t=1}^{T-2}|\overline{X}(t+2)-\overline{X}(t)| = \frac{\gamma_X}{\sigma_X};$$

3.2. Feature selection and relations mining

3.2.1. Feature selection

Not all the extracted features are relevant to emotion space. To employ the effective information and reduce the noise caused by superfluous features for the following regression modeling, feature selection should be conducted. In this study, we take the Bivariate Correlation (a correlation calculating method in SPSS) to compute the correlations between each feature and valence or arousal, and the features are selected according to the correlation threshold. Only the absolute correlation coefficient surpasses 0.25($|\rho| > 0.25$) with p-value below 0.05, the significant correlation between two vectors is supposed to exist. The p-value means the probability of obtaining the correlation coefficient at least as the actually observed.

3.2.2. Relations mining

After the features are selected, an arousal estimate model and a valence estimate model are built by using the arousal-related features and the valence-related features respectively. This procedure is performed separately for video features and EEG features.

$$y = \sum_{i=1}^{N} w(i)x(i) + w(0) \qquad \textbf{Eq. 1}$$

The estimate models are constructed by a regression method, RVM. And the models finally acquired can be represented in Eq.1, where y means the simulation value, x(i) represents the i-th feature, N denotes the total number of the participant features, and w(i) is the weight corresponding to the i-th feature.

Taking into account of the small sample size, leave-one-out method is used to verify the validity of our estimate models. First, one sample is chosen as the test sample, and the other samples are all used to train. Second, the acquired weights are then applied to the test sample, and the mean squared error (MSE) between the computed estimate value and the original actual value is calculated. Each sample is chosen as the test sample for only once, and the above two steps are repeated until all samples have been chosen. The results of these regressions will be shown in next section.

The RVM is also used to mine the direct relations between the selected video features and EEG features. The results are listed in Table 5 and Table 6.

Finally, the relationship between emotion categories and dimensions is also analyzed by using Fuzzy C-Means clustering. Disgust samples are abandoned for its small size of 5, and the other three kinds of samples (21 sad, 66 happy, and 42 fear) are applied in clustering method. The results are showed in next section.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Results of feature selection

For each of the EEG features and video features, the correlation coefficients between them and self-assessed emotion are calculated. After this process, 59 EEG features and 2 video features are selected for arousal, while 32 EEG features and 26 video features are selected for valence. For the large number of the selected features, 14 EEG features for arousal with absolute correlations over 0.35, 11 EEG features and 14 video features for valence with absolute correlations over 0.40 are selected and showed in Table 2.

Table 2: EEG features with high correlation with self assessment

	Arousal	Correlation	Valence	Correlation
1	FP2,Delta	0.35	T8,Delta	-0.4070
2	HEOG, Delta	-0.36	FC2,Delta	-0.4810
3	AF3, Delta	0.3690	FC2, Theta	0.4650
4	T7,Alpha	-0.3610	FC2, Alpha	0.5150
5	F5,Alpha	0.3630	T8, Beta	0.4270
6	HEOG, Beta	0.3750	FC3,Beta	0.4520
7	T7, Mid	-0.3680	T8,Mid	0.4310
8	F5, Mid	0.3580	FC2,Mid	0.4740
9	AF3, σ_X	0.4340	T8, Gama	0.4050
10	FC4, ZCR	0.4180	FC2,Gama	0.4710
11	TP7, ZCR	0.4710	HEOG, ZCR	0.4090
12	T7, ZCR	0.4240		
13	P5, ZCR	0.4150		
14	AF7, ZCR	0.5060		

From the Table 2, for the arousal, EEG signals' ZCR is an important feature, because there are 5 channel signals' ZCR are chosen. While for the valence, 5 frequency features of the channel FC2 and 4 frequency features of the channel T8 are selected, and it means that the positions of channel FC2 and T8 are closely related to the valence. In addition, none of the channel is related to both arousal and valence except the HEOG, which represent the horizontal electrooculogram signal.

From the Table 3, for the arousal, only the fundamental frequency and silence ratio are selected from all video features. In the light of the value of these two correlations, the greater the fundamental frequency, the higher the arousal presents, and the lower the silence ratio, the higher the arousal shows.

For the valence, 2 visual features and 12 audio features are chosen. Among these features, MFCC can be thought as the most important features for the valence, because over half of the selected features belong to the MFCC and the greatest value of the correlation 0.6140 is also derived from the MFCC. However, it is noted that lighting key and color energy are also closely related to the valence with correlations of almost 0.6, which shows that increasing the lighting key and color energy can induce subjects' higher valence.

Table 3: Video features with high correlation with the assessed valence

	Arousal	Correlation		Valence	Correlation
1	1st FF	0.3210	7	3rd MFCC	-0.4860
2	Silence ratio	-0.3790	8	5th MFCC	-0.5650
	Valence	Correlation	9	7th MFCC	-0.4430
1	Lighting key	0.5920	10	6th MFCC derivative	0.4710
2	Color energy	0.5980	11	8th MFCC derivative	0.4630
3	1st LPCC	0.4820	12	9th MFCC derivative	0.6140
4	8th LPCC	-0.4800	13	11th MFCC derivative	0.5860
5	11th LPCC	-0.4470	14	1st ZCR	0.4360
6	1st MFCC	0.4350	-	-	-

After choosing all emotion-related features, the correlation coefficients between selected EEG features and video features are calculated. Table 4 shows the results from the arousal, and Table 5 shows the results from the valence.

From the Table 4, 10 pairs of EEG features and video features are selected. Generally, fundamental frequency can elicit users' changes of EEG's ZCR, while silence ration can induce power changes of some frequency bands in T7 channel.

Table 4: Feature pairs with high correlations between EEG and videos from the arousal aspect

	FPZ/alpha	FPZ/mid	FC4/ZCR	TP7/ZCR	T7/ZCR	C6/ZCR
1st FF	0.302	0.362	0.389	0.320	0.344	0.373
	T7/theta	T7/alpha	FT7/alpha	T7/mid	-	-
Silence ratio	0.326	0.372	0.375	0.336	-	-

Table 5: Feature pairs with high correlations between EEG and videos from the valence aspect

	5th MFCC	6th MFCC	7th MFCC	8th MFCC	10th MFCC	11th MFCC
T8,delta	0.6510	-	0.6590	-0.6520	-	0.6630
FT7,theta	-	-	-	-	-	-0.6450
T8, Beta	-0.7300	0.6260	-0.7330	0.7630	0.6800	-0.7390
T8, Mid	-0.6630	-	-0.6690	0.6520	-	-0.6630
T8,Gama	-0.6880	-	0.7020	0.7020	0.6170	-0.6990
C4,Mean	-0.6180	-	-0.6280	-	-	-

There are total 243 pairs of EEG features and video features are selected according the correlation threshold 0.25. Because of the large number of the feature pairs, it is hard to show and analyze all the pairs. At last, 22 feature pairs with correlation over 0.6 are chosen and presented in Table 5. It can be seen that MFCC is closely related to the EEG's frequency information because there are 5 frequency EEG features in all selected 6 EEG features. And there are 19 pairs in all 22 selected feature pairs are related to the EEG features derived from T8, it shows that T8 is an important channel response to the audio stimulus.

In general, the correlations between EEG features and video features based on the valence are greater than the correlations based on arousal. It may be explained that individual differences of assessing arousal are greater than those of assessing valence.

4.2. Results of the regression

Table 6: MSE between estimated aro/val grades and self assessments

	Arousal	Valence
Video features	0.0645	0.1632
EEG features	0.0858	0.1658

Table 7: MSE between estimated feature values and true feature values

Video to EEG	EEG to Video			
FP2,Delta (2.4271E-4)	Silence ratio (0.0016)	1st LPCC (0.0308)	8th LPCC (1.8223E-4)	11th LPCC (1.1473E-4)

14 EEG features and 2 video features related to the arousal are used to simulate the arousal, while 11 EEG features and 14 video features related to the valence (From Table 2 and Table 3) are applied to simulate the valence. The MSE are computed 43 times using the cross validation technique described in 3.2.2, and the results are showed in Table 6.

Better results are acquired for arousal than for valence, and all MSE values are considerable small.

After simulating the arousal and valence, all selected EEG features are also used to simulate each single selected video feature. And the same process are conducted for utilizing all selected video features to simulate each selected EEG features. The results are presented in Table 7.

With regard to simulating EEG features, only one feature, Delta power ratio in FP2 channel, performs well with 2.4271E-4 MSE. While on simulation video features, silence ratio and 3 LPCC features performs well with small MSE value. The results mean that relations between videos and the EEG signals induced by them do exist, and some mapping models can be built between the features from these two domains. For example, we can apply these selected video features to predict the Delta power ratio in FP2 channel.

4.3. Results of the clustering

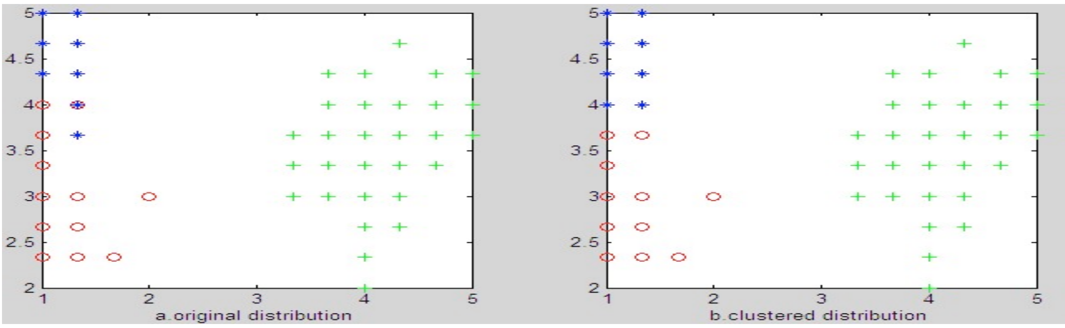

Figure 3: The distributions of these three categories before and after the clustering

Finally, emotion clustering is conducted, and the results are showed in Figure 3. The accuracy rate of 94.57% means that these three emotions can be classified and recognized easily from the dimensional arousal/valence. In fact, a stimulus video clip can induce several emotion experiences simultaneously for subjects, which is not convenience for processing. So we only record the subjects' primary emotion. However, after the cluster emotion model is constructed, the probabilities for each emotion categories of the subjects can be estimated.

5. CONCLUSION

In this paper, affective video analyses have been conducted from the point of induced EEG signals and self-assessed emotions. EEG signals of subjects are recorded at the same time when subjects are watching video clips. Through calculating the correlations between extracted EEG features, video features, and the subjects' self-evaluated arousal/valence, features closely related to arousal/valence are determined. And then, three section of analysis are performed as follows.

First, the selected EEG features and video features are used to conduct arousal/valence estimation respectively. The results of the cross validation perform well, which means arousal/valence of the videos can be estimated by EEG signals or videos themselves instead of artificial assessing. This is because EEG signals is a kind of physiological signals, which is objective and never express wrongly, our results may provide a reference for the future affect estimation with EEG features. Furthermore, the results of the arousal/valence estimation with video features show that movie recommenders could propose some films to satisfy customers' emotional needs. And there is a promising future to apply the results in especially personalized film recommendation systems.

Second, after the above arousal/valence models are validated, closely related feature pairs between EEG and video are selected by computing the correlations between selected EEG features and video features. The results confirm the relations between video clips and the EEG signals induced by them. To find the detailed information, these two kinds of features are used to simulating each feature of the other kind respectively with a regress method RVM, and 5 features that can be simulated well are acquired finally. The result of this part opens the doors to study the affective videos by using EEG signals directly.

Third, Fuzzy C-Means clustering method is adopted to transform emotion space from dimensions to categories. Emotions are always evaluated with categories in daily life. Therefore, it makes the affective assessment with basic emotion categories be possible.

Compared with other related works, our contributions can be concluded in three aspects. First, most of the researchers only concern with the relations between videos and subjects' psychological responses. However, the relations between videos and subjects' physiological responses are also investigated in our work. Second, a few researchers studied the relations between the videos and the induced peripheral physiological signals, but EEG signals are not included. In our work, EEG signals, which could represented the emotion more than other physiological signals, are studied to confirm their relations with stimulus videos and induced emotions. Third, emotion categories, which are used more popular in daily life, are also considered besides the emotion dimensions (arousal/valence). This could provide our work with more applicable value. Moreover, only three basic categories are included in our emotion clustering process, which couldn't represent all primary emotion categories. Therefore, we will enlarge our video database with more kinds of videos with different emotions, and make a full-scale research work.

ACKNOWLEDGEMENT

This paper is supported by National 863 Program (2008AA01Z122), Anhui Provincial Natural Science Foundation (No.070412056) and SRF for ROCS, SEM.

REFERENCES

1. Wei, C.-Y., Dimitrova, N., & Chang, S.-F. Color-Mood Analysis of Films Based on Syntactic and Psychological Models, IEEE International Conference on Multimedia and Expo, pp. 831-834, 2004.
2. Arifin, S..Affective Level Video Segmentation by Utilizing the Pleasure-Arousal-Dominance Information [J], IEEE transactions on multimedia, Vol.10, No.7, 2008

3. Arifin, S., & Cheung, P. Y. K. A novel probabilistic approach to modeling the pleasure-arousal-dominance content of the video based on "Working memory", Paper presented at the International Conference on Semantic Computing (ICSC 2007), Irvine, CA. Sep 17-19, 2007.

4. Mohammad Soleymani, Guillaume Chanel, Joep J. M. Kierkels, Thierry Pun. Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses, Tenth IEEE International Symposium on Multimedia, pp. 228-235, 2008.

5. Arthur, G. M. and A. Harry. Feasibility of Personalized Affective Video Summaries. Affect and Emotion in Human-Computer Interaction, Theory to Applications, Springer-Verlag, pp. 194-208, 2008.

6. M. E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. of Machine learning research, Vol. 1, No. 1, pp. 211-244, 2001.

7. R.W. Picard, Affective computing, The MIT press, 1997.

8. D. G. Li, I. K. Sethi, N. Dimitrova, and T. Mcgee, Classification of general audio data for content-based retrieval, Pattern Recognition Letters, Vol. 22, No. 5, pp. 533-544, 2001

9. C. Lei, S. Gunduz, and M. T. Ozsu, “Mixed Type Audio Classification with Support Vector Machine”, IEEE Multimedia and Expo, pp.781-784, July 2006.

10. P. Boersma and D. Weenink, Praat: doing phonetics by computer, computer program, 2008.

11. Wang, H. L., & Cheong, L. F., Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 16, No. 6, pp. 689-704, 2006.

12. Makeig S, et al. Matlab Toolbox for analysis of electrophysiological data. http://www.cnl.salk.edu/～scott/ica.html, 1997.

13. Dan-hua Zhu, Ji-jun Tong, Yu-quan Chen, An ICA-based method for automatic eye blink artifact correction in multi-channel EEG, Technology and Applications in Biomedicine, pp.338-341, 30-31 May 2008.

14. Ji Zhong, Qin Shuren., Detection Of EEG Basic Rhythm Feature By Using Band Relative Intensity Ratio(BRIR)[A], The 28th IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP’2003) [C]．Hong Kong：IEEE Signal Processing Society．pp. 429—432, 2003.

15. K. Takahashi, Remarks on Emotion Recognition from Bio-Potential Signals, 2nd International Conference on Automomous Robots and Agents, 2000.