

# A Novel Affective Visualization System for Videos Based on Acoustic and Visual Features

Jianwei Niu<sup>1(✉)</sup>, Yiming Su<sup>1</sup>, Shasha Mo<sup>1</sup>, and Zeyu Zhu<sup>2</sup>

<sup>1</sup> State Key Laboratory of Virtual Reality Technology and Systems,  
School of Computer Science and Engineering, Beihang University,  
Beijing 100191, China

{niujianwei, sy1406126}@buaa.edu.cn, mo06211225@163.com

<sup>2</sup> School of Electronics and Information,  
Xi'an Jiaotong University, Xi'an 710049, China  
asxxzzy@qq.com

**Abstract.** With the fast development of social media in recent years, affective video content analysis has become a hot research topic and the relevant techniques are adopted by quite a few popular applications. In this paper, we firstly propose a novel set of audiovisual movie features to improve the accuracy of affective video content analysis, including seven audio features, eight visual features and two movie grammar features. Then, we propose an iterative method with low time complexity to select a set of more significant features for analyzing a specific emotion. And then, we adopt the BP (Back Propagation) network and circumplex model to map the low-level audiovisual features onto high-level emotions. To validate our approach, a novel video player with affective visualization is designed and implemented, which makes emotion visible and accessible to audience. Finally, we built a video dataset including 2000 video clips with manual affective annotations, and conducted extensive experiments to evaluate our proposed features, algorithms and models. The experimental results reveals that our approach outperforms state-of-the-art methods.

**Keywords:** Affective analysis · Novel features · Feature selection · Emotion visualization

## 1 Introduction

In the last decades, thousands of videos are produced every day. Since the number of videos is enormous, various applications based on affective video content analysis have become more and more popular in recent years, such as video segmentation [1], video recommendation [2,3], highlights extraction [4], and video retrieval [5]. In affective video content analysis, audiovisual features extracted from video play an important role. To some extent, significant features can bridge the gap between a video stream and the emotions the video stream may elicit.

In [5], Zhang et al. extracted music related features like tempo, beat strength, and rhythm regularity to analyze music videos. Several low-level features like saturation, lighting, pitch, and zero crossing rate were also utilized in their work. In [6], Hanjalic et al. extracted four features (motion, shot change rate, sound energy and audio pitch) to compute “arousal” and “valence”. Besides the features mentioned in the above two papers, abundant features have been introduced in previous work. Most of these features can be classified into several categories: (1) Features extracted directly from raw data, like sound energy, saturation, lighting, etc. [2, 7, 8]. (2) Classical features frequently used in signal processing like MFCC (Mel Frequency Cepstrum Coefficient), ZCR (Zero Crossing Rate), LPCC (Linear Prediction Cepstrum Coefficient), spectral rolloff, etc. [9–13]. (3) Features used for special analysis, for instance, rhythm strength and tempo for music analysis [5], and music emotion and face features for highlight extraction [4]. (4) Shot features, such as shot types, shot change rate, and motion [6, 14]. In this paper, a novel feature is proposed to describe the “harmony” in audio signal. This feature measures the amount of harmonious elements in a piece of audio. Besides, four “color emotions” in the color psychology field are also utilized [15]. These four color emotion features contain more affective information since they are the result of psychology research, and therefore, the four features are supposed to be more useful than other features. Together with another 12 common features introduced in previous work, 17 features make up the feature set used in our work.

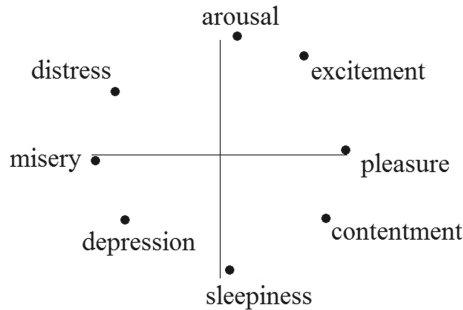


Fig. 1. Russell’s circumplex model

In order to achieve affective visualization, we adopt a categorical affective model since dimensional affective models are usually obscure to people. As shown in Fig. 1, the categorical affective model (Russell’s circumplex model) contains eight emotions (i.e., arousal, excitement, pleasure, contentment, sleepiness, depression, misery, and distress) [16]. The intensity of these eight intuitive emotions will be computed using the extracted features and BP (back propagation) network. In order to obtain significant features for computing certain emotion in movies, a novel feature selection algorithm with quadratic time complexity is proposed. Based on the affective content analysis of movies, a video

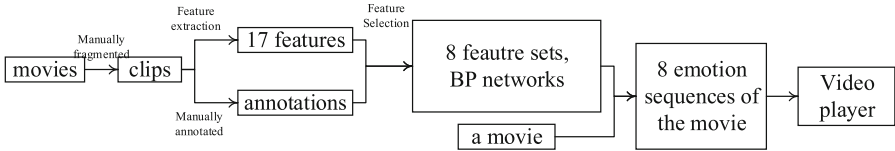
player with affective visualization is implemented, which shows the emotions of videos while playing videos. Extensive experiments are conducted to prove the effect of the novel features we used and the validity of the proposed feature selection method. The major contributions of this work include:

1. We propose a novel feature set including an audio feature depicting the “harmony” of the sound in videos and four color emotion features from the color psychology field. The features are proved to be helpful for affective video content analysis.
2. We propose a quadratic-time method to select significant features for computing a certain emotion in a movie.
3. We design and implement a player that can display video emotions while playing videos. Extensive experiments are conducted to validate our methods and player.

The rest of this paper is organized as follows. Section 2 introduces the overall methodology of this work. Section 3 explains the specific methods of extracting the novel features proposed in this work and also discusses the proposed feature selection method. Section 4 presents the data collection in this work and the experiment results. Finally, conclusions are drawn in Sect. 5.

## 2 Overall Methodology

In this paper, we aim to improve the accuracy of affective video content analysis. Figure 2 describes the framework of our work.



**Fig. 2.** The framework of our work

**Data collection and annotation:** We segment movies into short clips, and each movie clips is annotated with the intensity indices of eight emotions shown in Fig. 1. The movie clips are collected as the training or testing sets for experiments.

**Feature extraction:** Seventeen features are extracted from each movie clip and used for predicting emotions.

**Feature selection:** Apparently, it is not appropriate to use all the seventeen features to predict an emotion, and we proposed a novel method to select significant features for predicting each emotion.

**Emotion prediction:** BP networks are adopted to map features to emotions. The BP network used in this work contains one hidden layer. The training function



**Fig. 3.** User Interface of AffePlayer

is “trainlm”, and the transfer function used in the hidden and output layers is “tansig”. Eight BP networks are trained for predicting eight emotions respectively.

Affective visualization: Based on the affective content analysis, a video player, we term it *AffePlayer*<sup>1</sup>, is developed to exhibit the emotions in the playing video.

Emotion annotations of movie clips and the corresponding features are used to train eight BP networks, for the aim of computing the eight emotions of a movie. Since a movie is long, the emotions of the movie may vary greatly over time. We extract the emotions of a movie every 10 s (using windows of ten seconds and 9 s overlap between neighboring windows) and thus obtain eight emotions sequences of the movie. The UI (User Interface) of AffePlayer is shown in Fig. 3. As shown in Fig. 3, it displays video in the left window, and the emotions (bottom) and audio signals (top) in the right window. The 8 light bars represent arousal, excitement, pleasure, contentment, sleepiness, depression, misery, and distress from left to right, respectively. The dim bars between two light bars represent the transition between two emotions, which can be interpreted as the mixture of the two emotions. The height of these bars represents the intensity of emotions. When the player is playing a video, the height of these bars will change with the intensity fluctuations of emotions.

### 3 Feature Extraction and Selection

Totally 17 features are extracted, including seven acoustic features, eight visual features, and two movie grammar features. A novel method is proposed to select significant features from these 17 features for analysing emotions in movies.

<sup>1</sup> [www.ldmc.buaa.edu.cn/AffePlayer/AffePlayer.html](http://www.ldmc.buaa.edu.cn/AffePlayer/AffePlayer.html).

### 3.1 Features

One of the novel features is proposed in this work to measure the “harmony” in audio signal. Figure 4 shows the spectrogram of noise sound and a piece of piano music, i.e., the changes of frequency over time. The lighter is the color in the spectrogram, the stronger is the amplitude of the corresponding frequency. Harmony sound show continuity in spectrogram while noisy sound show only disorder. As is shown in Fig. 4, there are continuous peak in the spectrogram of piano music but none in the spectrogram of noise sound. Based on this observation, we propose a method to compute “harmony” in audio signal by measuring the peak continuity in the spectrogram.

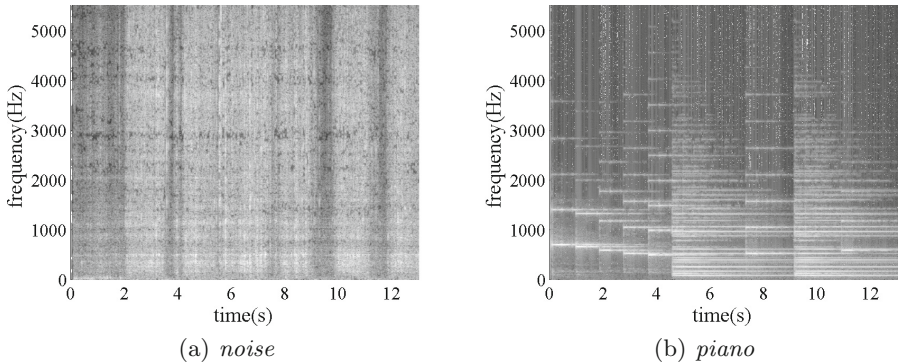


Fig. 4. Spectrogram of noise and a piece of piano music

Firstly, the Fourier transform of the audio segment is obtained every 1024 samples, and the spectrogram of the audio segment can be obtained. Assume the spectrogram of an audio segment is denoted by  $F_{ij}$ , where  $i$  is the index of frequency and  $j$  is the index of time. Then the local maximum of each  $j$  is found, and a matrix  $M$  is used to record the local maximum. In matrix  $M$ , the element  $M_{ij}$  equals 1 if  $F_{ij}$  is bigger than  $F_{i-2j}$ ,  $F_{i-1j}$ ,  $F_{i+1j}$ , and  $F_{i+2j}$ ; otherwise  $M_{ij}$  equals 0. Matrix  $M$  is used to compute the harmony of the audio. “Harmony sequence” is defined as  $(M_{ia}, M_{ib})$ , where  $a \leq b$ ,  $M_{ia-1} = 0$ ,  $M_{ib+1} = 0$ , and  $M_{ik} = 1$  ( $a \leq k \leq b$ ). It is assumed that there are  $n$  “harmony sequences” in total:  $(M_{ia}, M_{ib})_s$  ( $s = 1, 2, \dots, n$ ), then the harmony of the audio segment can be denoted as

$$H = \frac{1}{Size(M)} \sum_{1 \leq s \leq n} \|(M_{ia}, M_{ib})_s\| \quad (1)$$

where  $\|(M_{ia}, M_{ib})_s\| = (b - a)^2$ , and  $Size(M)$  denotes the size of matrix  $M$  which equals the product of the height and the width of  $M$ .

Color is important in affective video content analysis since existing research results suggest that color directly affects human’s emotion. The research results

from color psychology reveal that different color induce different feelings of people, and based on the result, four novel color features (Warm-Cool, Transparent-Turbid, Vivid-Sombre, and Hard-Soft) are utilized in this work. Nakamura et al. investigated twelve pairs of color emotions and quantify these color emotions. As the results of their experiments, empirical formulae were obtained, and we select four typical color emotions as the features used in our work. The four color emotions are calculated in the CIELAB color space and the specific calculation methods are:

Warm-Cool:

$$WC = 3.5[\cos(h - 50) + 1]B - 80 \quad (2)$$

where  $B$  represents the brightness calculated by  $B = 50C^*(1 - \Delta h_{290}/360)/D$ , and  $D$  is the color depth calculated by  $D = (100 - L^*) + (0.1 + \Delta h_{290}/360)(1 - \Delta h_{290}/360)C^*$ .

Transparent-Turbid:

$$TT = [\{5(L^* - 40)\}^2 + \{5.8\{1 + 0.35\cos(\Delta h_{220})\}(1 - \Delta h_{290}/360)C^*\}^2]^{1/2} - 180 \quad (3)$$

Vivid-Sombre:

$$VS = [\{2.5(L^* - 50)\}^2 + \{5.4(1 - \Delta h_{290}/360)\}^2]^{1/2} - 130 \quad (4)$$

Hard-Soft:

$$HS = [(3.2L^*)^2 + \{2.4(1 - \Delta h_{290}/360)C^*\}^2]^{1/2} - 180 \quad (5)$$

In the above equations,  $L^*$  is the CIELAB lightness,  $C^*$  is the CIELAB chroma,  $h$  is the CIELAB hue-angle, and  $\Delta h_{290}$  is the CIELAB hue-angle difference from  $h = 290$  [15].

Besides the novel features mentioned before, other twelve common features are also extracted. Six acoustic features including: (1) Sound energy (the average amplitude of audio signals), (2) Sound centroid (the average frequency of the fourier transform of audio signals), (3) Spectral contrast of an audio segment (extracted in three sub-bands resembling that in Lu's work) [17], (4) Silence ratio (the proportion of the time duration of silent sound to the entire time length). Four visual features including: (1) Darkness ratio and brightness ratio extracted in the HSV (Hue-Saturation-Value) color space according to the value of V (Value) dimension, (2) Saturation, which has strong effects on emotions [18], is the average saturation of all frames. (3) Color energy is extracted following that in [19]. Two movie grammar features are extracted in this paper: motion and shot change rate.

### 3.2 Feature Selection

It is a simple way to feed all seventeen features into a BP network and get the emotions of a video, but is not rational. Since some features are not relevant to

a certain emotion, that is, some features may be “detrimental” for calculating a certain emotion. Thus, it is necessary to find a way to select significant features and exclude redundant features.

If we are going to obtain the optimal combination of features for analysing emotions, the computational complexity will be  $O(2^n)$  (where  $n$  is the number of features extracted from videos), which is unpractical. Therefore, we propose a method to select significant features with quadratic time complexity. Though it is hard to get the optimal feature combination, we can obtain a near-optimal solution with low time complexity. For emotion  $e_1$ , assume that  $S_1$  is the set of selected features to predict  $e_1$  and  $S_0$  is the set of features that are not selected. Initially,  $S_0$  contains all the 17 extracted features, while  $S_1$  contains none. Then one feature  $f$  in  $S_0$  is chosen and added into set  $S_1$ , where  $f$  meets the condition:

$$mse(f, D) \leq mse(f', D) |_{f' \in S_0 \& f' \neq f} \quad (6)$$

$D$  denotes the training set,  $f'$  denotes the feature in  $S_0$ .  $f'$  is different from the chosen feature  $f$ .  $mse(f, D)$  means the Mean Square Error (MSE) while using feature  $f$  to predict emotion  $e_1$  on data set  $D$ . Equation 6 means that feature  $f$  performs better than any other feature in  $S_0$  on predicting  $e_1$ . The Mean Square Error is recorded as  $m$ . The chosen feature  $f$  is moved from  $S_0$  to  $S_1$ . Next, features in  $S_1$  (now there is only one selected feature) are used to predict emotion  $e_1$ , and 80% of the movie clips, which are worse predicted than the other 20% movie clips, make up a data set named  $D'$ . Then, a feature  $f$  in  $S_0$  is chosen and added into set  $S_1$  if  $f$  meet the condition:

$$mse(f, D') \leq mse(f', D') |_{f' \in S_0 \& f' \neq f} \quad (7)$$

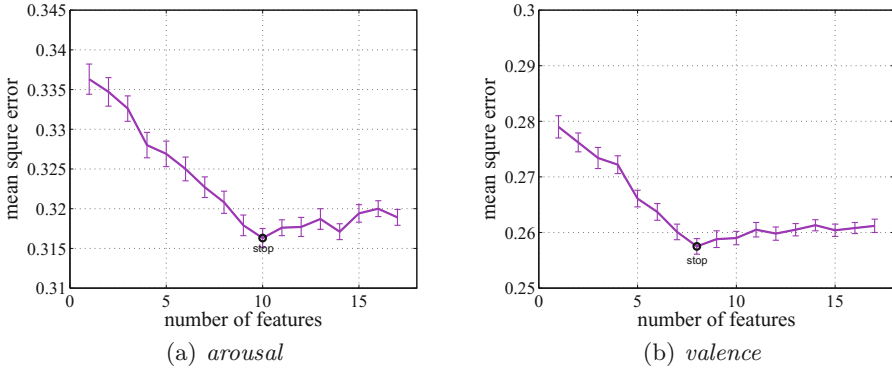
Emotion  $e_1$  in set  $D$  is predicted using features in  $S_1$  and the Mean Square Error  $m'$  is obtained. If  $m'$  is greater than  $m$ , then the selecting process is terminated and remove the last selected feature from  $S_1$ ; or else assign  $m'$  to  $m$  and continually choosing a feature and move it from  $S_0$  to  $S_1$  until there are no features in  $S_0$ . At last, the features in set  $S_1$  are the features selected to predict emotion  $e_1$ . There are eight emotions used in this paper, and therefore, eight feature sets will be obtained to predict the eight emotions of videos respectively.

## 4 Experimental Results

Experiments are conducted on both our own collected database and the LIRIS-ACCEDE database [20]. In the first subsection, we introduce the process of building our own database. In the second subsection, experiments conducted on the LIRIS-ACCEDE database are presented. The significant features for valence and arousal are obtained and the results of our method are compared to that in [20]. In the third subsection, experiments are conducted on our own database. The most relevant single feature of each emotion is found from the 17 extracted features, and the feature combinations for predicting eight emotions are obtained using the method we proposed in Sect. 3.

#### 4.1 Data Collection and Annotation

Our database consists of 2000 video clips manually segmented from 112 movies which can be download from our website<sup>2</sup>. The movies are various in genres, including action, comedy, thriller, romance, etc. Languages are English, Chinese, Spanish, Japanese, and Korean. Each movie clip lasts about 10s and all the movie clips are annotated with the intensity of eight emotions from the Russell’s circumplex model [16]. The intensity of eight emotions are divided into 5 grades: 1, 2, 3, 4, and 5 correspond to very weak, weak, normal, strong, and very strong, respectively. The annotators are fifty students from our university, including 13 females and 37 males. A web site is built for students to complete their annotating work. After the students watch a movie clip, they are requested to estimate the intensity of eight emotions and annotate the movie clip with eight values denoting the intensity of eight emotions respectively. Thus, for each emotion of a movie clip, there are 50 values annotated by 50 different persons. A video clip is excluded from our dataset if the Mean Square Error of any emotion annotations is bigger than 1. Fifty values are averaged and rounded to an integer, and the average value is taken as the final annotation of the emotion.



**Fig. 5.** Mean Square Error of feature selection for arousal and valence (95% confidence interval)

**Table 1.** Features used to calculate arousal and valence in LIRIS-ACCEDE database

Emotion	Features
Arousal	<b>Warm-cool, vivid-sombre, harmony</b> , centroid, low contrast, mid contrast, high contrast, color energy, dark color ratio, silence ratio
Valence	<b>Warm-cool, transparent-turbid, vivid-sombre</b> , centroid, <b>hard-soft</b> , saturation, color energy, dark color ratio

<sup>2</sup> [www.ldmc.buaa.edu.cn/AffePlayer/database.rar](http://www.ldmc.buaa.edu.cn/AffePlayer/database.rar).



## 4.2 Experiments on LIRIS-ACCEDE Database

The results of experiment on the LIRIS-ACCEDE database are presented in this subsection. Figure 5 shows the variation of Mean Square Error while selecting features for predicting arousal and valence. At first, only one feature is selected for predicting emotions with relatively big error. With the increase of the number of features, the Mean Square Error become smaller. According to the algorithm introduced in Sect. 3, the process end at the black circle marked in Fig. 5. That is, ten features are selected for arousal and eight features for valence as shown in Table 1. As can be seen from Fig. 5, the prediction error will increase slightly if all 17 features are used. Although there may exist better feature combinations for predicting arousal or valence, the process of the feature selection shown in Fig. 5 proves that the results in Table 1 are at least near-optimal solutions. The proposed feature “harmony” and two color emotions are in the feature combination for arousal and four color emotions are included in the feature combination for valence, which demonstrate the effectiveness of the novel features we introduced in Sect. 3. Experiments are conducted based on protocol A and protocol B introduced in [20]. The results of experiments are shown in Fig. 6. It can be seen from Fig. 6 that the method proposed in this paper performs better on the prediction of valence and the prediction of valence is as good as that in [20].

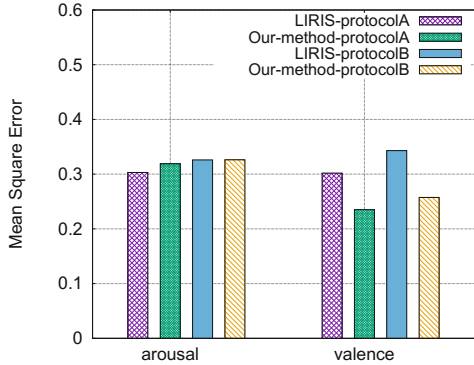


Fig. 6. Results of experiments on LIRIS-ACCEDE database.

## 4.3 Experiments on Our Own Database

In this subsection, the database collected in this work is used in several experiments. Seventy percent of the 2000 movie clips are used as the training set and the rest thirty percent are used as the testing set. We apply the feature selection method on the training set, and the variation of Mean Square Errors in the process of feature selection are shown in Fig. 7. Eight feature combinations are selected for predicting eight emotions using the method proposed in Sect. 3. The black circles marked in the figure represent the terminations of feature selection,

and the selected features are shown in Table 2. As can be seen from Table 2, the novel features we proposed and introduced from color psychology play important roles in emotion predictions.

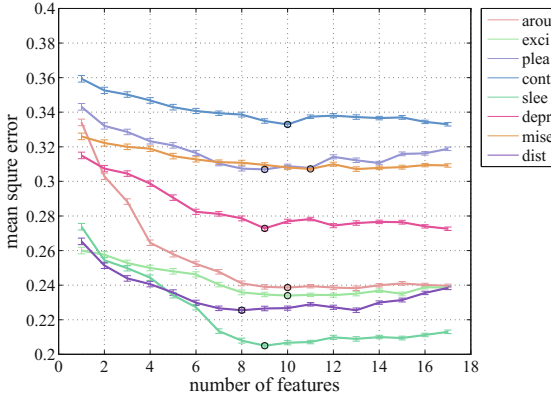


Fig. 7. MSE of feature selection for 8 emotions (95% confidence interval)

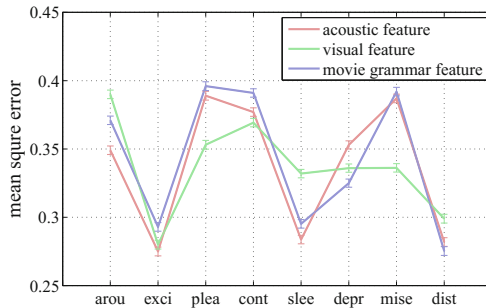
Table 2. Features used to estimate the 8 emotions in our work

Emotion	Features
Arousal	Motion, shot change rate, sound energy, <b>harmony</b> , sound centroid, low contrast, mid contrast, high contrast, color energy, silence ratio
Excitement	Motion, <b>transparent-turbid</b> , <b>vidid-sombre</b> , <b>hard-soft</b> , silence ratio, low contrast, mid contrast, high contrast, brightness ratio, darkness ratio
Pleasure	Shot change rate, <b>warm-cool</b> , <b>transparent-turbid</b> , <b>vidid-sombre</b> , color energy, <b>hard-soft</b> , sound centroid, darkness ratio, brightness ratio
Contentment	Shot change rate, <b>warm-cool</b> , <b>transparent-turbid</b> , <b>vidid-sombre</b> , <b>hard-soft</b> , sound centroid, silence ratio, darkness ratio, brightness ratio, low contrast
Sleepiness	Motion, shot change rate, sound energy, <b>harmony</b> , high contrast, sound centroid, low contrast, mid contrast, silence ratio
Depression	<b>Warm-cool</b> , <b>transparent-turbid</b> , <b>vidid-sombre</b> , <b>hard-soft</b> , brightness ratio, low contrast, mid contrast, color energy, darkness ratio
Misery	Shot change rate, <b>warm-cool</b> , <b>transparent-turbid</b> , color energy, <b>vidid-sombre</b> , <b>hard-soft</b> , sound centroid, saturation, darkness ratio, brightness ratio, silence ratio
Distress	Motion, shot change rate, <b>vidid-sombre</b> , darkness ratio, low contrast, mid contrast, color energy, silence ratio

The first feature chosen for a certain emotion can be regarded as the most relevant feature of the emotion. Color emotion vivid-sombre is most relevant to pleasure, contentment, and misery. The feature most relevant to sleepiness is silence ratio, which is the same as common sense. Silence ration is also most relevant to distress and motion, mid-band spectral contrast and bright ratio are most relevant to arousal, excitement and depression respectively. It can be seen from Fig. 7 that excitement is best predicted with single feature, which means mid-band spectral is really suitable for predicting excitement.

The feature combinations for predicting eight emotions are shown in Table 2. The feature combinations for predicting each emotion are different from each other. As shown in Fig. 7, the Mean Square Errors of the computation are between 0.205 and 0.333. It can be seen from Fig. 7 that the computation results vary with emotions, and the computations of arousal, excitement, sleepiness, and misery are more precise than that of the other four emotions, which means the features extracted are more suitable for predicting arousal, excitement, sleepiness and distress. Among these emotions, sleepiness is the best estimated emotions with MSE of 0.205 and contentment is the worst estimated with MSE of 0.333. The average MSE of computing eight emotions is 0.2654. Therefore, we could say that compared to the result in [20], the features and the framework in this work perform better.

Another experiment is conducted to find the impact of feature types on analyzing different video emotions. The 17 features contain three types of feature, that is, acoustic features, visual features, and movie grammar features. Every feature is used to compute the eight emotions respectively. The MSEs are averaged within each type of features and the results are shown in Fig. 8. Visual features perform best in computing pleasure, contentment, and misery, which means features about color and lighting have greater impact on emotions pleasure and unpleasure. Acoustic features and movie grammar features are more suitable to compute the intensity of arousal and sleepiness.



**Fig. 8.** MSE of computing emotions using different type of features (95% confidence interval)

## 5 Conclusion

In this work, a novel feature set is used in affective video content analysis, including an acoustic feature depicting the “harmony” in audio signal and four color emotions. The “harmony” feature performs well in predicting emotion arousal and sleepiness. And the four color emotion features which contain the psychology information are useful in predicting almost all the emotions. A feature selection method is proposed to obtain significant features, and experimental results proves the effectiveness of the feature selection method.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China (Grant Nos. 61572060, 61190125, 61472024) and CERNET Innovation Project 2015 (Grant No. NGII20151004).

## References

1. Arifin, S., Cheung, P.: Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Trans. Multimed.* **10**(7), 1325–1341 (2008)
2. Niu, J., Zhao, X., Zhu, L., Li, H.: Affivir: an affect-based Internet video recommendation system. *Neurocomputing* **120**, 422–433 (2013)
3. Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., Jose, J.M.: Enriching user profiling with affective features for the improvement of a multimodal recommender system. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*, p. 29. ACM (2009)
4. Lin, K.S., Lee, A., Yang, Y.H., Lee, C.T., Chen, H.: Automatic highlight extraction for drama video using music emotion and human face features. *Neurocomputing* **119**, 111–117 (2013)
5. Zhang, S., Huang, Q., Jiang, S., Gao, W., Tian, Q.: Affective visualization and retrieval for music video. *IEEE Trans. Multimed.* **12**(6), 510–522 (2010)
6. Hanjalic, A., Xu, L.Q.: Affective video content representation and modeling. *IEEE Trans. Multimed.* **7**(1), 143–154 (2005)
7. Chan, C.H., Jones, G.J.F.: An affect-based video retrieval system with open vocabulary querying. In: *Detyniecki, M., Knees, P., Nürnberger, A., Schedl, M., Stober, S. (eds.) AMR 2010. LNCS, vol. 6817, pp. 103–117. Springer, Heidelberg (2012). doi:10.1007/978-3-642-27169-4\_8*
8. Zhang, S., Tian, Q., Jiang, S., Gao, W.: Affective MTV analysis based on arousal and valence features. In: *2008 IEEE International Conference on Multimedia and Expo*, pp. 1369–1372. IEEE (2008)
9. Canini, L., Benini, S., Leonardi, R.: Affective recommendation of movies based on selected connotative features. *IEEE Trans. Circ. Syst. Video Technol.* **23**(4), 636–647 (2013)
10. Xu, M., Wang, J., He, X., et al.: A three-level framework for affective content analysis and its case studies. *Multimed. Tools Appl.* **70**(4), 757–779 (2014)
11. Cui, Y., Jin, J.S., Zhang, S., Tian, Q.: Music video affective understanding using feature importance analysis. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 213–219. ACM (2010)
12. Yazdani, A., Kappeler, K., Ebrahimi, T.: Affective content analysis of music video clips. In: *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, pp. 7–12. ACM (2011)

13. Acar, E., Hopfgartner, F., Albayrak, S.: Understanding affective content of music videos through learned representations. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014. LNCS, vol. 8325, pp. 303–314. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-04114-8\\_26](https://doi.org/10.1007/978-3-319-04114-8_26)
14. Canini, L., Benini, S., Leonardi, R.: Affective analysis on patterns of shot types in movies. In: International Symposium on Image and Signal Processing and Analysis, pp. 253–258 (2011)
15. Xin, J.H., Cheng, K.M., Chong, T.F.: Quantifying colour emotion-what has been achieved. *Res. J. Text. Apparel* **2**(1), 46–54 (1998)
16. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)
17. Lu, L., Liu, D., Zhang, H.: Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio Speech Lang. Process.* **14**(1), 5–18 (2006)
18. Valdez, P., Mehrabian, A.: Effects of color on emotions. *J. Exp. Psychol. Gen.* **123**(4), 394–409 (1994)
19. Wang, H.L., Cheong, L.-F.: Affective understanding in film. *IEEE Trans. Circ. Syst. Video Technol.* **16**(6), 689–704 (2006)
20. Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L.: LIRIS-ACCEDE: a video database for affective content analysis. *IEEE Trans. Affect. Comput.* **6**(1), 43–55 (2015)