Implicit Affective Video Tagging Using Pupillary Response

Dongdong Gui, Sheng-hua Zhong(✉), and Zhong Ming

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
ddgui@email.szu.edu.cn, {csshzhong,mingz}@szu.edu.cn

Abstract. The psychological research found that human eyes could serve as a sensitive indicator of emotional response. Pupillary response has been used to analyze the affective video content in previous studies, but the performance is not good enough. In this paper, we propose a novel method for implicit affective video tagging using pupillary response. The issue of pupil size difference between subjects has not been effectively solved, which seriously affected the performance of the implicit affective video tagging. In our method, we first define the pupil diameter baseline of each subject to diminish individual difference on pupil size. Besides, the probabilistic support vector machine (SVM) and long short term memory (LSTM) network are used to extract valuable information and output the probability estimates based on the proposed global features and sequence features obtained from the pupil dilation ratio time-series data, respectively. The final decision is made by combining the probability estimates from these two models based on the sum rule. In empirical validation, we evaluate the proposed method on a standard dataset MAHNOB-HCI. The experimental results show that the proposed method achieves better classification accuracy compared with the existing method.

Keywords: Implicit affective video tagging · Pupillary response
Emotion recognition

1 Introduction

With the explosive growth of video data, automatic video content analysis plays an increasingly significant role in various video-based applications, such as video retrieval [31], video summarization [28], video saliency detection [9], and so on. Most of traditional content-based video analysis concentrates on the main event happened in a video. The affective level is an especially important measure of the viewers' attitude toward video content. Recently, video affective content analysis has been an active research area. Video affective content analysis seeks to automatically identify the emotions elicited by videos [32]. The development of it is benefit to both the users and businesses. Users could utilize the emotional

K. Schoeffmann et al. (Eds.): MMM 2018, Part II, LNCS 10705, pp. 165–176, 2018.
https://doi.org/10.1007/978-3-319-73600-6_15

information to retrieve certain videos, and filmmakers may change their editing to make a more stimulating movie that meet the emotional flow of audiences through the help of video affective content analysis [27].

The success of this task will crucially hinge on how the effective features are defined and extracted. The direct approaches mainly rely on the automatic feature extraction from video data. The extracted audio-visual features include color histogram, the number of scene cuts, Mel- Frequency Cepstrum Coefficients (MFCC) [2], and spatio-temporal features [23]. As we known, affective information in videos is closely related to the viewer's feelings and emotions. Thus, the emerging research field of video affective analysis aims at exploiting human emotions based on the analysis of the spontaneous reactions for viewers while watching the video content, which can be called as the implicit video affective content analysis.

Since most of the psychological theories agree that physiological activity is an important component of emotional experience, and facial expression is the primary channel for emotion communication, implicit video affective content analysis mainly adopts viewers' physiological signals and spontaneous visual behaviors, especially the facial behaviors [21]. Some important physiological signals, such as: electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), skin temperature (ST), heart rate (HR), and blood volume pulse (BVP), which are controlled by the sympathetic nervous system, are capable to reflect unconscious body changes [11,12,22,25]. Compared with facial behaviors, these signals are more evident and reliable [17,21]. Unfortunately, it is difficult to obtain these physiological signals. Users are required to wear complex apparatuses to obtain these physiological signals. Moreover, the physiological signals are also sensitive to many artifacts, such as involuntary eye-movements and irregular muscle movements [15,18]. On the contrary, the visual behaviors of participants are much easier to be obtained and recorded. Only one remote visible camera is required to record the visual behaviors of viewers. Furthermore, the recorded visual behaviors of participants are undisturbed by the movements or other body conditions. Compared with the physiological signals, spontaneous visual behaviors are more convenient to measure, although it is susceptible to some environmental noises, such as lighting conditions [4], occlusion [20], and etc.

Several studies have demonstrated the feasibility of using spontaneous visual behaviors for video affective content analysis. Zhao et al. [30] extracted viewers' facial expressions to predict the affective curve to describe the process of affect changes. Their experimental results verified that both eye movements and facial expressions could be helpful in video affective content analysis and other related applications. Yeasin et al. [29] employed the recognized facial expressions to detect the emotional levels for different videos. Rather than focusing on subjects' whole facial activity, Ong and Kameyama [19] analyzed the affective video content by calculating the viewers' pupil sizes and gazing points. Katti et al. [16] employed users' pupillary dilation response for the task of the affective video summarization and the storyboard generation.

Although the pupillary response has been used to recognize the video emotion in previous studies, the performance is not good enough. The first reason is that the issue of the pupil size difference between subjects has not been solved so far. Pupillary changes during watching affective pictures could reflect human's emotion state [5]. However, some subjects have larger pupil sizes than others [26], thus absolute pupil diameter couldn't accurately reflect the emotional state of different subjects. To address this issue, we first define the pupil diameter baseline (PDB) of each subject. Then, the absolute pupil diameter of each subject can be converted to pupil dilation ratio based on his own PDB value. Besides, no studies have used recurrent neural network to learn temporal relations from pupillary response. Long short-term memory (LSTM) network is a variant of recurrent neural network capable of learning temporal representations from sequence data, but it requires enough training samples. Support vector machine (SVM) has good generalization ability, even if the samples are limited. But compared with LSTM, SVM couldn't capture the temporal information of the sequence data well. Therefore, under the condition of limited samples, we try to combine the SVM and LSTM to improve the classification accuracy. In our method, the probabilistic SVM and LSTM are used to extract valuable information and output the probability estimates based on the proposed global features and sequence features, respectively. Finally, the final decision is made by combining the output probabilities from these two models based on the sum rule. The results suggest that the probability estimates between SVM and LSTM are complementary, and the classification accuracy could be further improved after we combine the probability estimates of SVM and LSTM.

2 Method

2.1 Preprocessing

In the preprocessing part, we first exclude the missing samples of the pupil diameter due to the eye blinking. Then, the average pupil diameter of the left and the right pupil is defined as the pupil diameter time-series data. Then, we remove the pupil diameter time-series data from 0 to 2 s, which contains the initial light reflex [5]. The initial light reflex generally appears within 2 s after the video starts playing. As shown in Fig. 1, the red part of the pupillary response is the initial light reflex. We can find the pupil diameter decreases significantly due to the brightness of video clip. To eliminate the interference from the initial light reflex, we remove this part from the pupil diameter time-series data. At last, a local regression smoothing method named LOESS (Locally Weighted Scatterplot Smoothing) [8] is applied to smooth the pupil diameter time-series data. Figure 2 shows an example of the original and the smoothed pupil diameter time-series data.

2.2 Pupil Diameter Baseline Specifying

The issue of pupil size difference between subjects has never been effectively solved. To address this problem, we define the pupil diameter baseline of each

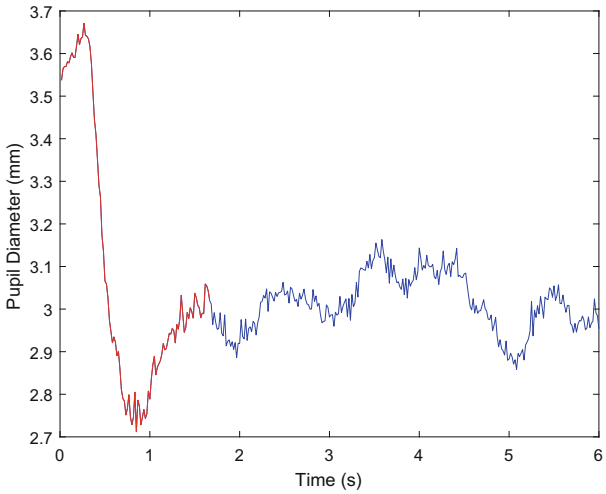


Fig. 1. An example of the pupil diameter time-series data in the first 6 s. The red part is the initial light reflex. We can find the pupil diameter decreases significantly due to the high brightness of the video.

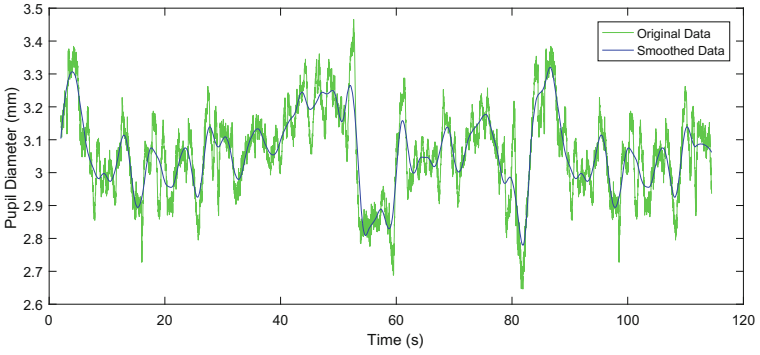


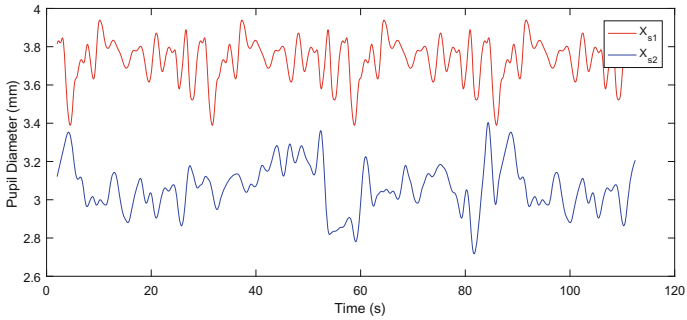
Fig. 2. An example of the original and the smoothed pupil diameter time-series data.

subject before feature extraction. In the standard emotional experiments, a collection of neutral video clips was usually prepared in advance in each trial. And a neutral video clip randomly selected from the collection was shown to the human subject before each emotional video clip. Thus, the pupil diameter baseline of the subject j is simply defined as the average pupil diameter of this subject in response to the neutral video clip:

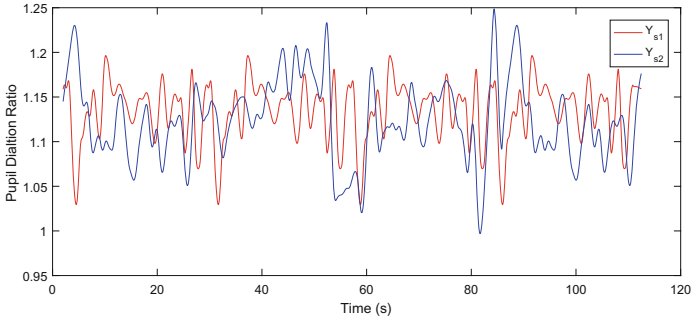
$$PDB_j = \frac{1}{n} \sum_{v \in C} d_j^v \tag{1}$$

where PDB_j denotes the pupil diameter baseline of the subject j, n is the number of video clips in the neutral video clip collection C, d_j^v is the average pupil diameter of the subject j in response to the video clip v in the neutral video clip collection C.

After we construct the pupil diameter baseline of each subject, we could use these baseline values to diminish the individual difference on the pupil size. For each subject, we transform the absolute pupil diameter time-series data (see Fig. 3a) to the pupil dilation ratio time-series data based on his own pupil diameter baseline. (see Fig. 3b). Here, the pupil dilation ratio is the ratio of the absolute pupil diameter to the pupil diameter baseline.



(a) Pupil diameter time-series signals from two subjects



(b) Pupil dilation ratio time-series signals from two subjects

Fig. 3. (a) X_{s1} and X_{s2} represent the pupil diameter time-series from subject $s1$ and $s2$ in response to the same video clip, respectively. Due to the pupil size difference between subjects, the absolute pupil diameters of $s1$ and $s2$ fluctuate around two very different values, even though $s1$ and $s2$ have the same emotional state. (b) Y_{s1} and Y_{s2} represent the corresponding pupil dilation ratio time-series of X_{s1} and X_{s2}, respectively. The PDB of $s1$ is 3.29 mm, and the PDB of $s2$ is 2.73 mm. Intuitively, after the absolute pupil diameters of $s1$ and $s2$ are divided by their own PDB, the obtained pupil dilation ratios of $s1$ and $s2$ fluctuate around two very close values.

2.3 Feature Extraction

A pupillary response signal can be represented as two different feature types: global feature and sequence feature. The global feature is used to train an SVM classifier, the sequence feature is used to train an LSTM classifier (see Sect. 2.4).

Global Feature. Given a pupil dilation ratio time-series, some emotion related features can be extracted from the whole time-series. Table 1 shows the features extracted from the pupillary response signal. The time domain features include the average value and the standard deviation of the pupillary response. The average value can be used to distinguish different arousal levels [5]. The frequency domain features include the spectral power from four bands ($0\,\text{Hz} < f \leq 0.2\,\text{Hz}$, $0.2\,\text{Hz} < f \leq 0.4\,\text{Hz}$, $0.4\,\text{Hz} < f \leq 0.6\,\text{Hz}$, $0.6\,\text{Hz} < f \leq 1\,\text{Hz}$), these spectral features are related to the mental activity [24].

Sequence Feature. For the same pupil dilation ratio time-series, it can also be represented as a sequence feature. A time window divides the time-series data into segments with the length λ, and the segments are with 50% overlap. The average shot length (ASL) of most films is less than 10 s [3]. Hence, we set the length λ to 10 s. Then, the features listed in Table 1 can be extracted from each segment and combined as a sequence. Finally, the time-series data is represented as a $m \times n$ sequence feature, where m is the sequence length, and n is the feature dimension of each segment.

Table 1. The emotion-related features extracted from the pupillary response.

Domain	Extracted features
Time	Average, standard deviation
Frequency	Spectral power in the following bands: (0, 0.2]Hz, (0.2, 0.4]Hz, (0.4, 0.6]Hz, (0.6, 1]Hz

2.4 Recognition Method

This section describes the proposed recognition method, which includes three parts: (1) probabilistic support vector machine; (2) long short-term memory network; (3) decision level fusion.

Probabilistic Support Vector Machine. Support vector machine is a powerful algorithm, even if the training samples are limited. However, the output of SVM cannot be directly used as a probability estimate of multi-class. For this reason, Huang et al. [14] proposed a probabilistic SVM model, which could output the probability estimate of multi-class.

Long Short-Term Memory Network. Long short-term memory network is a type of recurrent neural network capable of learning order dependence in sequence prediction problems [13]. To obtain the probability estimate of each category, the softmax function is applied to the output layer.

Decision Level Fusion. In our method, the probabilistic SVM outputs the probability estimate based on the global feature, and the LSTM network outputs the probability estimate based on the sequence feature. The final probability estimate is then made by combining the output probabilities from the two models with the sum rule [6]. For a given trial, the sum rule is defined as follows:

$$g_i = \frac{\sum_{m \in M} P_m(c_i|\mathbf{f})}{\sum_{i=1}^{K} \sum_{m \in M} P_m(c_i|\mathbf{f})} = \frac{1}{|M|} \sum_{m \in M} P_m(c_i|\mathbf{f}) \tag{2}$$

where M is the ensemble of the classification models chosen for fusion, $|M|$ is the number of these models in M, and K is the number of classes. $P_m(c_i|\mathbf{f})$ is the posterior probability of feature $\mathbf{f}$ belongs to class c_i obtained by model m. The final decision is made by selecting the class c_i with the highest g_i.

In our work, (2) can be simplified as:

$$g_i = \frac{1}{2}(P_{svm}(c_i|\mathbf{f_v}) + P_{lstm}(c_i|\mathbf{f_s})) \tag{3}$$

In (3), $P_{svm}(c_i|\mathbf{f_v})$ is the posterior probability of the vector feature $\mathbf{f_v}$ belongs to class c_i obtained by a learned probabilistic SVM, $P_{lstm}(c_i|\mathbf{f_s})$ is the posterior probability of the sequence feature $\mathbf{f_s}$ belongs to class c_i by a learned LSTM network. The class c_i with the highest g_i is the final decision.

3 Experiments

3.1 Experimental Setting

In this paper, we evaluate the performance of the proposed method on the MAHNOB-HCI database [24]. The MAHNOB-HCI database is a multimodal database recorded in response to affective stimuli with the purpose of implicit video affective tagging. The eye gaze data recorded from four participants (P3, P16, P25, and P26) of the total 27 participants are not used in our experiment due to the incompleteness of the data collection. After watching a video clip, all participants reported their felt emotions as keywords feedback. In our experiments, we use the responded keyword with the highest proportion of all participants as the label of each corresponding video clip. As shown in Table 2, these emotional labels can also be mapped into the commonly used two-dimensional arousal-valence space, containing arousal and valence dimensions [10].

We compare the proposed method with the only existing method [24] which used this standard dataset, and we use the classification accuracy and F1 score as the evaluation metrics. Leave-one-participant-out cross-validation is applied

to validate the performance. The LibSVM [7] with RBF kernel is used as an implementation of probabilistic SVM. A parameter selection tool included in LibSVM automatically select the parameters C and γ for the RBF kernel. The LSTM implementation of an open source neural network library Keras [1] is used in our experiments. We use one hidden layer of 128 LSTM units. In order to prevent overfitting of the network, the dropout fraction of the input is set to 0.3. The batch size for LSTM is set to 8, and the number of epochs is set to 100.

Table 2. The emotional labels and the arousal-valence dimension labels for the video clips in MAHNOB-HCI database.

Emotional label	Arousal	Valence	Video clips sources
Neutral	Calm	Neutral	AccuWeather New York, Dallas, Detroit weather report
Sadness	Calm	Unpleasant	Gangs of New York, the thin red line, American history X
Disgust	Calm	Unpleasant	Hannibal, ear worm
Joy	Medium	Pleasant	Mr. Bean's holiday, love actually, the thin red line
Amusement	Medium	Pleasant	Kill Bill VOL I, Mr. Bean's holiday, love actually, funny cats, funny
Fear	Activated	Unpleasant	The shining (2), silent hill
Anger	Activated	Unpleasant	Hannibal, ear worm

3.2 The Effect of Individual Difference Reduction

Table 3 presents the average classification accuracies of different models with or without individual difference reduction. Without individual difference reduction, all features are extracted from the absolute pupil diameter time-series. With individual difference reduction, all features are extracted from the pupil dilation ratio time-series. From this table, we could find that all models could obtain better accuracies with the features extracted from the pupil dilation ratio time-series data. Notably, the average accuracies with SVM are improved more than 9%, but the average accuracies with LSTM are only improved about 4%. The possible reason for this result is that LSTM is not as sensitive as SVM to the individual difference. Although the pupil size difference exists between subjects, LSTM is still capable of learning the temporal representations from pupillary responses. On the whole, the results demonstrate that the proposed method can effectively reduce the pupil size difference between subjects.

3.3 The Comparison of Classification Results

In this section, we compare the proposed method with the only existing method [24] which used this standard dataset. Table 4 presents the average classification

Table 3. The average classification accuracies of different models with or without individual difference reduction. Without individual difference reduction, all features are extracted from the original pupil diameter time-series. With individual difference reduction, all features are extracted from the pupil dilation ratio time-series.

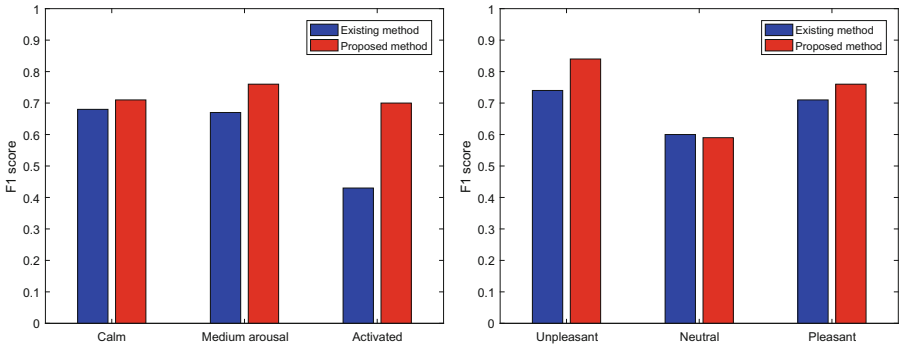
Model	Without individual difference reduction		With individual difference reduction	
	Arousal	Valence	Arousal	Valence
SVM	60.0%	66.3%	**69.3%**	**76.3%**
LSTM	62.4%	66.1%	**66.1%**	**70.4%**
Decision level fusion	63.9%	69.1%	**73.0%**	**78.5%**

accuracies and F1 scores of different methods. The results demonstrate that our method could achieve better performance. In Fig. 4a, the average F1 scores of different arousal classes are shown. The proposed method obtains higher F1 scores for all the arousal classes. Besides, we can find that the proposed method significantly improves the F1 score of the activated class. In general, most of the participants have significant pupil dilation when watching the climax of a high arousal video. The proposed method is capable of capturing the temporal information to recognize the activated emotions.

Table 4. The average classification accuracies and the F1 scores of different methods.

Method	Classification accuracy		Average F1	
	Arousal	Valence	Arousal	Valence
Existing method [24]	63.5%	68.8%	0.60	0.68
Proposed method	**73.0%**	**78.5%**	**0.72**	**0.77**

For the valence classes, our method could obtain higher F1 scores on the unpleasant and pleasant classes. But we can also find our method obtains a slightly lower F1 score on the neutral category (see Fig. 4b). The bad performance on the neutral class of the proposed method is primarily due to the sample imbalance. The number of the neutral samples is only 15% of the total. Table 5 presents the average F1 scores of all valence classes obtained by different models. Compared with the classical shallow model SVM, as a deep learning model, LSTM has more model parameters, which makes the training of it requires a large number of labeled training samples. Therefore, the lack of the neutral training samples leads to a low F1 score 0.19 of the neutral category. In future, we will integrate the imbalance techniques into our method to address the imbalance problem in implicit affective video tagging. Despite LSTM has this limitation, it is still good at extracting effective information from a sequence. Hence, from Table 5, we can find the final fusion result in most of the cases is better than the respective output of SVM and LSTM.



(a) F1 scores of arousal classes (b) F1 scores of valence classes

Fig. 4. The average F1 scores of different arousal and valence classes.

Table 5. The average F1 scores of different valence classes obtained by different models in the proposed method.

Model	Unpleasant	Neutral	Pleasant
SVM	0.80	**0.79**	0.69
LSTM	0.79	0.19	0.71
Decision level fusion	**0.84**	0.59	**0.76**

4 Conclusion

In this paper, we propose a novel method for implicit affective video tagging using pupillary response. Our method includes three parts. First, we construct the pupil diameter baseline of each subject to reduce the individual difference on pupil size. Then, the probabilistic SVM and LSTM network are used to obtain the probability estimates based on the proposed global features and the sequence features, respectively. Finally, the final classification decision is made by combining the probability estimates from these two models. We evaluate our method on a standard dataset MAHNOB-HCI. The experimental results show that the proposed method is effective to reduce the pupil size difference between subjects. Compared with the existing method, our method could achieve better classification accuracy. Moreover, from our results, we can also find the probability estimates between SVM and LSTM are complementary. By combining the probability estimates of SVM and LSTM, the classification accuracy could be further improved.

In the future, we will investigate the computational emotion recognition using multimodal physiological signals, such as EEG (electroencephalography).

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61502311, No. 61672358), the (Key) Project of Department of Education of Guangdong Province (No. 2014GKCG031, No. 12JGXM-MS29,

No. 2015SQXX0), the Natural Science Foundation of Guangdong Province (No. 2016A030310053), the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) (No. U1501501), the Shenzhen high-level overseas talents program, and the Tencent "Rhinoceros Birds" Scientific Research Foundation for Young Teachers of Shenzhen University (2015, 2016).

References

1. Keras: an open source neural network library. http://keras.io
2. 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, Cham (2014). https://doi.org/10.1007/978-3-319-04114-8_26
3. Baxter, M.: Notes on cinemetric data analysis (2014)
4. Belhumeur, P.N., Kriegman, D.J.: What is the set of images of an object under all possible illumination conditions? Int. J. Comput. Vis. **28**(3), 245–260 (1998)
5. Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J.: The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology **45**(4), 602–607 (2008)
6. Chanel, G., Kierkels, J.J., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum.-Comput. Stud. **67**(8), 607–627 (2009)
7. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 1–27 (2011)
8. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **74**(368), 829–836 (1979)
9. Fang, Y., Lin, W., Chen, Z., Tsai, C., Lin, C.: A video saliency detection model in compressed domain. IEEE Trans. Circ. Syst. Video Technol. **24**(1), 27–38 (2014)
10. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.C.: The world of emotions is not two-dimensional. Psychol. Sci. **18**(12), 1050–1057 (2007)
11. Gajraj, R., Doi, M., Mantzaridis, H., Kenny, G.: Analysis of the EEG bispectrum, auditory evoked potentials and the EEG power spectrum during repeated transitions from consciousness to unconsciousness. Br. J. Anaesth. **80**(1), 46–52 (1998)
12. Guggisberg, A.G., Hess, C.W., Mathis, J.: The significance of the sympathetic nervous system in the pathophysiology of periodic leg movements in sleep. Sleep **30**(6), 755–766 (2007)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Huang, T., Weng, R.C., Lin, C.J.: Generalized Bradley-Terry models and multi-class probability estimates. J. Mach. Learn. Res. **7**(Jan), 85–115 (2006)
15. Iwasaki, M., Kellinghaus, C., Alexopoulos, A.V., Burgess, R.C., Kumar, A.N., Han, Y.H., Lüders, H.O., Leigh, R.J.: Effects of eyelid closure, blinks, and eye movements on the electroencephalogram. Clin. Neurophysiol. **116**(4), 878–885 (2005)
16. Katti, H., Yadati, K., Kankanhalli, M., Tat-Seng, C.: Affective video summarization and story board generation using pupillary dilation and eye gaze. In: Proceedings of the International Symposium on Multimedia, pp. 319–326. IEEE (2011)
17. Kreibig, S.D.: Autonomic nervous system activity in emotion: a review. Biol. Psychol. **84**(3), 394–421 (2010)
18. Lins, O.G., Picton, T.W., Berg, P., Scherg, M.: Ocular artifacts in EEG and event-related potentials I: scalp topography. Brain Topogr. **6**(1), 51–63 (1993)
19. Ong, K., Kameyama, W.: Classification of video shots based on human affect. J. Inst. Image Inf. Telev. Eng. **63**(6), 847–856 (2009)

20. Poursaberi, A., Araabi, B.N.: Iris recognition for partially occluded images: methodology and sensitivity analysis. EURASIP J. Appl. Sig. Process. **2007**(1), 20 (2007)
21. Rainville, P., Bechara, A., Naqvi, N., Damasio, A.R.: Basic emotions are associated with distinct patterns of cardiorespiratory activity. Int. J. Psychophysiol. **61**(1), 5–18 (2006)
22. Robinson, B.F., Epstein, S.E., Beiser, G.D., Braunwald, E.: Control of heart rate by the autonomic nervous system. Circ. Res. **19**(2), 400–411 (1966)
23. Shao, L., Zhen, X., Tao, D., Li, X.: Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Trans. Cybern. **44**(6), 817–827 (2014)
24. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**(1), 42–55 (2012)
25. Tang, Y.Y., Ma, Y., Fan, Y., Feng, H., Wang, J., Feng, S., Lu, Q., Hu, B., Lin, Y., Li, J., et al.: Central and autonomic nervous system interaction is altered by short-term meditation. Proc. Natl. Acad. Sci. **106**(22), 8865–8870 (2009)
26. Tsukahara, J.S., Harrison, T.L., Engle, R.W.: The relationship between baseline pupil size and intelligence. Cogn. Psychol. **91**, 109–123 (2016)
27. Wang, S., Ji, Q.: Video affective content analysis: a survey of state-of-the-art methods. IEEE Trans. Affect. Comput. **6**(4), 410–430 (2015)
28. Wu, J., Zhong, S.H., Jiang, J., Yang, Y.: A novel clustering method for static video summarization. Multimedia Tools Appl., 1–17 (2016)
29. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimedia **8**(3), 500–508 (2006)
30. Zhao, S., Yao, H., Sun, X., Xu, P., Liu, X., Ji, R.: Video indexing and recommendation based on affective analysis of viewers. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 1473–1476. ACM (2011)
31. Zhu, Y., Huang, X., Huang, Q., Tian, Q.: Large-scale video copy retrieval with temporal-concentration sift. Neurocomputing **187**, 83–91 (2016)
32. Zhu, Y., Jiang, Z., Peng, J., Zhong, S.: Video affective content analysis based on protagonist via convolutional neural network. In: Chen, E., Gong, Y., Tie, Y. (eds.) PCM 2016. LNCS, vol. 9916, pp. 170–180. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48890-5_17