

Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal

Wenqian Lin^(✉), Chao Li, and Shouqian Sun

College of Computer Science of Zhejiang University, Hangzhou, China
linwq@zju.edu.cn

Abstract. Emotions play an important role at our day-to-day activities such as cognitive process, communication and decision making. It is also very essential for interaction between human and machine. Emotion recognition has been receiving significant attention from various research communities and capturing user's emotional state such as facial expressions, voice and body language, all of which are emerging way to find the human emotions. In recent years, physiological signals based emotion recognition has drawn increasing attention. Most of the physiological signals based methods use well-designed classifiers with hand-crafted features to recognize human emotions. In this paper, we present an approach to perform emotional states classification by end-to-end learning of deep convolutional neural network (CNN), which is inspired by the breakthroughs in the image domain using deep convolutional neural network. The approach is tested using the database "DEAP" including electroencephalogram (EEG) and peripheral physiological signals. We transform EEG into images combine extract hand-crafted features of other peripheral physiological signals, and classify emotions into valence and arousal. The results show this approach is possible to improve classification accuracy.

1 Introduction

Emotion playing a significant role in human's daily activities. It is an psychological expression of affective reaction and mental state based on human subjective experience [1]. Emotion is critical aspect of the human interpersonal relationship and essential to the human communication and behaviors. Psychologists often used discrete and dimensional emotion classification systems. Eight basic emotion states (anger, fear, sadness, disgust, surprise, anticipation, acceptance, and joy) proposed by Plutchik [2] and six basic emotion states based on facial expressions (anger, disgust, fear, happiness, sadness and surprise) proposed by Ekman [3] both belong to discrete emotion classification system. And the most widely used valence and arousal emotion classification model, proposed by Russell [4] belongs to dimensional system. In this model, the valence axis represents the quality of an emotion and the arousal axis refers to the emotion activation level.

In order to improve human-machine interaction (HMI), emotion recognition began to attract increasing attention. In the past few decades, various sensory data have been used to identify human emotion, including facial expression, auditory signals, text, body language, peripheral physiological signals and electroencephalogram (EEG) [5,6]. The last two, especially EEG, can provide more objective and comprehensive information for emotion recognition in comparison with other sensory data, because they can detect the body dynamics in response to emotional states directly.

Existing EEG based emotion recognition methods can be roughly grouped into two main categories: hand-crafted features based methods with well-designed classifiers [7,8] and recurrent neural network (RNN) [9]. Inspired by [10,11], where deep CNN is successfully in some fields of identification, we transform processed EEG data into images based on different frequency bands which contains time and frequency domain information, and then the generated images and hand-crafted features extracted from peripheral physiological signals were fed into CNN models to perform fine-tuning and emotion recognition. Experiments were conducted for cross-subject evaluation on the DEAP dataset [12] to validate the effectiveness of our proposed method. Our purpose was to affirm if the classification results of our method could obtain better average accuracy and F1 score on valence and arousal classification compares with several studies on the same database before.

The rest of this paper is organized as follows. Section 2 presents the descriptions of DEAP database. Section 3 presents the proposed method. The results are discussed in Sect. 4 and we conclude the paper in Sect. 5.

2 DEAP Database

DEAP dataset, which we conduct our experiment is a multimodal dataset for emotion analysis, contains both electroencephalogram (EEG) (recorded over the scalp using 32 electrodes and the positions of the electrodes are according to 10–20 International System: Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2) and peripheral physiological signals (8 channels, include electromyogram (EMG) collected from zygomaticus major and trapezius muscles, horizontal and vertical electrooculograms (EOGs), skin temperature (TMP), galvanic skin response (GSR), blood volume pulse (BVP), and respiration (RSP)) of 32 subjects (aged between 19 and 37), as each subject watched 40 one-minute music video clips, which were carefully selected to evoke different emotional states according to the dimensional valence-arousal emotion model of subjects, and played in a random order.

After watching each music video, participants were required to report their emotion using Self-Assessment Manikins (SAM) questionnaire, rating their tastes level of five dimensions (valence, arousal, dominance, liking, and familiarity), the first four scales range from 1 to 9 and the fifth dimension range between 1 and 5. In this paper, identifications the dimensions of valence (ranging from negative to positive) and arousal (ranging from calm to active) are

addressed as two independent tasks according to valence-arousal emotion model proposed by Russell [4]. Both two tasks posed as binary classification problems.

3 Proposed Method

Figure 1 shows the overall architecture of our proposed method. We first perform data preprocessing to normalize all the physiological signals. Secondly, every sample of electroencephalogram (EEG) signals is transformed into six gray images according to different frequency bands. Thirdly, we extract the 81-dimensional hand-crafted features of other peripheral physiological signals. Finally, the generated images and hand-crafted features are fed into four pre-trained AlexNet models [15] to perform fine-tuning and emotion recognition.

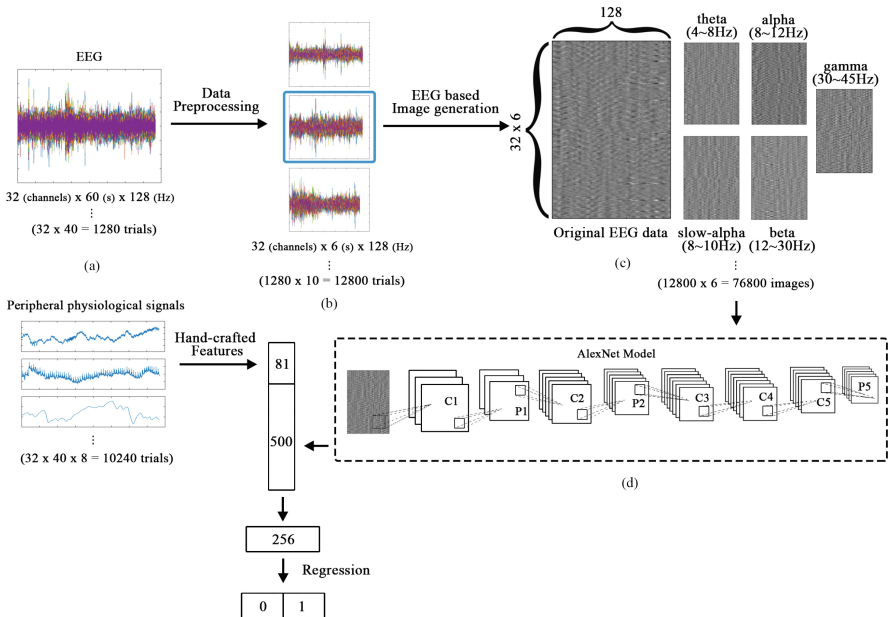


Fig. 1. An illustration of the proposed emotion recognition process. (C: convolution, P: max-pooling.)

3.1 Data Preprocessing

For all 40 channels of physiological signals in DEAP dataset, the preprocessing included down sampling the data from 512 Hz to 128 Hz. Especially for EEG data, a band-pass filter with cutoff frequencies of 4.0 and 45.0 Hz is first used to remove the unwanted noises and averaged to the common reference, as in [13, 14]. The data was segmented into 60 s trials and a 3 s pre-trial baseline removed, then

we divide the 60 s data of each trial into 10 clips to perform emotion recognition as in Sect. 3. Additionally, we perform data normalization for each channel of physiological signals as follows:

$$I_{i,j} = (I_{i,j} - \min_i) / (\max_i - \min_i) \quad (1)$$

where $I_{i,j}$ is the value of channel i at time j , \max_i and \min_i are, respectively, the maximum value and the minimum value of the channel i during $T = 60s$.

3.2 Electroencephalogram Signals Based Image Generation

After data preprocessing, the size of original EEG data for one sample is $32 \times 6 \times 128$, where 32 stands for the channels, 6 stands for the time of one clip, and 128 stands for the sampling rate and we rearrange the EEG data into a fixed size (192×128) gray image. On the other hand, according to five different frequency bands (theta (4~8 Hz), slow-alpha (8~10 Hz), alpha (8~12 Hz), beta (12~30 Hz), and gamma (30~45 Hz)), another five gray images are generated. Finally, for each sample of EEG data, we obtain 6 images. Figure 1(a) shows the details.

3.3 Peripheral Physiological Signals Based Feature Extraction

81-dimensional hand-crafted features are extracted from other eight channels of peripheral physiological signals: GSR, electrooculogram (EOG), respiration amplitude, electrocardiogram, skin temperature, blood volume by plethysmograph and electromyograms of Zygomaticus and Trapezius muscles as shown in Table 1. Before feature extraction, all the peripheral physiological signals are separately normalized to mean = 0 and s.d. = 1.

3.4 Multimodal Deep Convolutional Neural Network for Emotion Recognition

Network Structure. In this paper, the partial structure of famous deep convolutional model (AlexNet) [15] which consists of 8 parameterized layers (5 convolutional layers, 1 fully connected layer and 1 softmax layer) is adopted. We make some changes with AlexNet: (1) we encode the 81-dimensional hand-crafted feature into our CNN model by concatenating it with the hidden fully connected layer. (2) The number of hidden units in the fully connected layer is 500 and the output layer has only two neurons, which represents the two classes of the problem. Figure 1(d) shows our model's structure.

End-to-End Fine-Tuning. In order to fine-tune the pre-trained AlexNet model, the size of input skeleton sequence based image must be compatible with AlexNet's input size which is known as 227×227 pixel size. We first rescale each image to 256×256 pixel size and then randomly cropping and mean-subtracting

Table 1. Description for the hand-crafted features of peripheral physiological signals

Feature name	Feature index	Feature details
EOG & EMG frequency-domain features	1~5	Eye blink rate Average PSD of vertical Horizontal EOG Trapezius EMG Zygomaticus EMG
EOG and EMG time-domain features	6~21	Mean, Variance Zero-crossing rate The approximate entropy of 4 EOG and EMG channels
Skin temperature features	22~27	Average PSD in the frequency bands (0~0.1 Hz) and (0.1~0.2 Hz) Mean, Variance Approximate entropy Mean of derivative
GSR features	28~52	Mean Mean of derivative Mean of negative derivative values Proportion of negative values in all Derivative values Number of local minima Mean of rising time 15 PSD values in frequency band (0~2.4 Hz) Zero-crossing rates and means for the bands of (0~0.2 Hz) and (0~0.8 Hz)
Blood volume pressure features	53~59	Power ratio between the frequency bands of (0.04~0.15 Hz) and (0.15~0.5 Hz) Average PSD in the frequency bands of (0.1~0.2 Hz), (0.2~0.3 Hz), (0.3~0.4 Hz), (0.01~0.08 Hz), (0.08~0.15 Hz) and (0.15~0.5 Hz)
Respiration features	60~81	Power ratio between frequency bands of (0.05~0.25 Hz) and (0.25~0.5 Hz) Mean Mean of derivative Centroid of PSD Respiration rate 15 values of PSD in frequency band (0~2.4 Hz) peak-to-peak time

are adopted. For the last softmax layer (i.e., the output layer), the number of the unit is the same as the number of the emotion classes. Each sample of EEG signals is transformed into 6 images, which are separately fed into the proposed multimodal CNN with the same label for fine-tuning. Regarding the

hyper-parameter setting, we empirically selected the size of mini-batches for the SGD as 200. Moreover, we set the initial learning rate to 0.001, which is decreased by multiplying it by 0.1 at every 500th iteration. The fine-tuning is stopped after 50 epochs.

Emotion Recognition. In this study, we separately focus on valence and arousal scales because Koelstra et al. [12] finds that there is a significant difference between low and high among these two emotions. For each trial, two labels were generated. The affective level in valence space described HV (high valence) or LV (low valence), and the affective level in valence space described HA (high arousal) or LA (low arousal). The label 1 indicates high valence/arousal and the label 0 indicates low valence/arousal. Considering subject-specificity of the subjective ratings, the binary emotional classes could be much proper generated based on personal threshold, which determine the target classes by clustering subjective rating data for each subject using classical k -means clustering algorithm [14]. The threshold is computed by the midpoint of two cluster centers as examples shown in Fig. 2 of subject 1, and all threshold values for 32 subjects summarized in Table 2. The classification performances based on them.

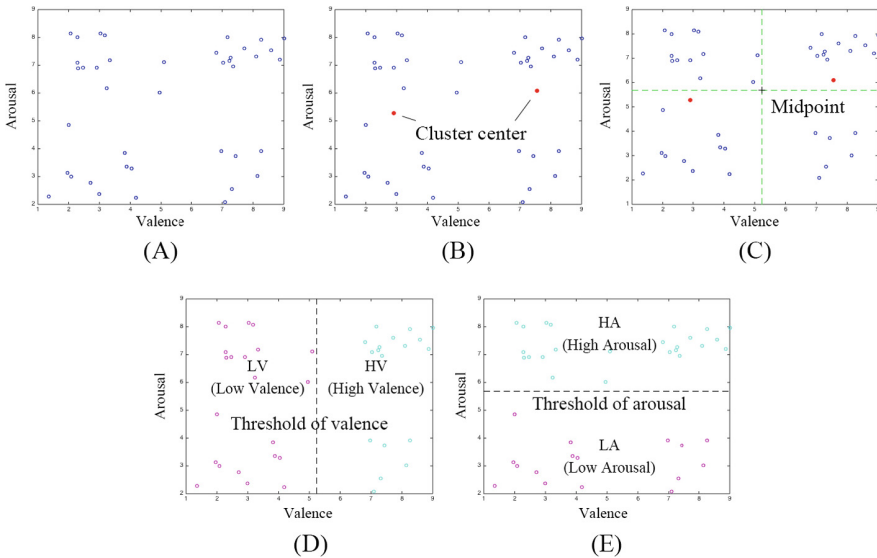


Fig. 2. The target emotion classes generated based on personal threshold. (A) Ratings of all 40 one-minute music videos by subject 1. (B) Two cluster centers based on results of k -means clustering. (C) Midpoint of two cluster centers. (D) The low and high valence states discretized by the threshold. (E) The low and high arousal states discretized by the threshold.

Table 2. All threshold values for 32 subjects.

Subject	Arousal	Valence	Subject	Arousal	Valence
1	5.6803	5.2342	17	5.1932	5.0815
2	5.6126	6.0166	18	5.5781	5.5596
3	3.7776	5.5513	19	5.4990	5.3685
4	4.5916	4.6503	20	5.6172	5.8185
5	5.1736	4.9791	21	6.0432	5.6618
6	4.6612	5.7579	22	5.3251	4.2624
7	5.0705	4.8358	23	3.6487	6.1354
8	5.6286	5.8466	24	5.8675	4.9634
9	5.6759	5.4592	25	5.9870	5.3552
10	5.6759	5.0015	26	3.8795	4.8234
11	5.1886	4.0322	27	4.6934	5.8161
12	6.3644	4.9731	28	4.7856	5.3817
13	6.6635	4.8578	29	4.3479	4.5732
14	5.4360	4.9597	30	5.1283	5.5714
15	4.7245	5.8538	31	5.6703	4.6661
16	4.7233	4.2413	32	5.6419	5.1586

Because each sample of EEG signals is transformed into 6 images, during testing, the class scores of all images are averaged to form the final prediction of the action class i as follows:

$$P = \left(\sum_{k=1}^6 O_k \right) / 6 \quad (2)$$

and

$$i = \arg \max_{i \in [0,1]} P_i \quad (3)$$

where O represents the output vector of our proposed CNN model and P is the final probability of high/low emotion.

3.5 Evaluation Criteria

As our analysis of emotion is a binary classification, we adopt mean classification accuracy and F1-score as the final evaluation criteria. The mean classification accuracy is calculated as follows:

$$Mean_Acc = (n_{TP} + n_{TN}) / (n_{TP} + n_{TN} + n_{FP} + n_{FN}) \quad (4)$$

where n_{TP} , n_{TN} , n_{FP} and n_{FN} denote the numbers of correct classified $label : 1$ instance, the numbers of correctively classified $label : 1$ instance, the numbers

of correctly classified *label* : 0 instance, the numbers of incorrectly classified *label* : 1 instance and the numbers of incorrectly classified *label* : 0 instance, respectively.

The precision for recognizing the high-class (1) instances is defined as p_1 ,

$$p_1 = (n_{TP}) / (n_{TP} + n_{FP}) \quad (5)$$

The precision for recognizing the low-class (0) instances is defined as p_0 ,

$$p_0 = (n_{TN}) / (n_{TN} + n_{FN}) \quad (6)$$

Finally, the F1-score is calculated:

$$p_f = 2p_0p_1 / (p_0 + p_1) \quad (7)$$

4 Results

Experiments of valence and arousal classification were conducted. We divide the last 60s data of each trial into 10 clips to perform emotion recognition and the 10-fold cross validation technique was carried out to evaluate the performance. Which partitioned samples into 10 disjoint subsets. One of the subsets was used as test sample at each fold, rest subsets were used as training samples and repeated 10 times till all the subsets were used as test sample ones. The average classification accuracy and F1-score values for all subjects were computed at each fold and were averaged at the end of the experiment. The classification results of the method we proposed was further compared with results obtained by other methods in Table 3.

Table 3. Experimental results on DEAP Dataset

Methods	Arousal		Valence	
	Accuracy [%]	F1-score [%]	Accuracy [%]	F1-score [%]
Koelstra et al. [12]	62.00	63.10	62.70	65.20
Liu and Sourina [16]	76.51	–	50.80	–
Naser and Saha [17]	66.20	–	64.30	–
Yoon and Chung [18]	70.10	–	70.90	–
Wang and Shang [19]	51.20	–	60.90	–
Chen et al. [20]	69.09	68.96	67.89	67.83
Li et al. [21]	64.20	–	58.40	–
Atkinson and Campos [22]	73.06	–	73.14	–
Yin et al. [13]	84.18	77.98	83.04	79.50
Our proposed method	87.30	78.24	85.50	80.06

It shows our proposed method could obtain better average accuracy and F1 score on valence and arousal classification compares with several studies on the same database before. More specifically, separately with respect to arousal and valence, the performance is improved by 3.12% and 2.46%, and F1 score also enhanced 0.26% and 0.56%.

5 Conclusion

In this paper, we proposed to transform different frequency band of EEG signals into six gray images which contains time and frequency domain information, and extracted hand-crafted features of other peripheral physiological signals. These images and features were then fed into a AlexNet model to perform end-to-end fine-tuning. To achieve better performances, data preprocessing of the original signal was also adopted. The provided experimental results prove the effectiveness and validate the proposed contributions of our method by achieving superior performance over the existing methods on DEAP Dataset.

References

1. Mauss, I.B., Robinson, M.D.: Measures of emotion: a review. *Cogn. Emot.* **23**(2), 209–237 (2009)
2. Plutchik, R.: *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*, 1st edn. American Psychological Association, Washington (2003)
3. Ekman, P.: *Basic Emotions*. Handbook of Cognition and Emotion. Wiley, New York (1999)
4. Russell, J.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)
5. Nie, D., Wang, X.W., Shi, L.C., et al.: EEG-based emotion recognition during watching movies. In: International IEEE/EMBS Conference on Neural Engineering, pp. 667–670. IEEE Xplore (2011)
6. Liu, Y., Sourina, O., Nguyen, M.K.: Real-time EEG-based human emotion recognition and visualization. In: International Conference on Cyberworlds, pp. 262–269. IEEE Computer Society (2010)
7. Heraz, A., Razaki, R., Frasson, C.: Using machine learning to predict learner emotional state from brainwaves. In: IEEE International Conference on Advanced Learning Technologies, ICALT 2007, July 18–20 2007, Niigata, Japan, DBLP, pp. 853–857 (2007)
8. Wang, X.-W., Nie, D., Lu, B.-L.: EEG-based emotion recognition using frequency domain features and support vector machines. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) *ICONIP 2011*. LNCS, vol. 7062, pp. 734–743. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24955-6_87
9. Ishino, K., Hagiwara, M.: A feeling estimation system using a simple electroencephalograph. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 5, pp. 4204–4209. IEEE Xplore (2003)
10. Shiraga, K., Makihara, Y., Muramatsu, D., et al.: GEINet: view-invariant gait recognition using a convolutional neural network. In: International Conference on Biometrics, pp. 1–8. IEEE (2016)

11. Li, C., Min, X., Sun, S., Lin, W., Tang, Z.: DeepGait: a learning deep convolutional representation for view-invariant gait recognition using joint Bayesian. *Appl. Sci.* **7**(3), 210 (2017)
12. Koelstra, S., Muhl, C., Soleymani, M., et al.: DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **3**(1), 18–31 (2012)
13. Yin, Z., Zhao, M., Wang, Y., et al.: Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Comput. Methods Programs Biomed.* **140**, 93–110 (2017)
14. Yin, Z., Wang, Y., Liu, L., Zhang, W., Zhang, J.: Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Front. Neurobot.* **11** (2017)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *International Conference on Neural Information Processing Systems*, pp. 1097–1105. Curran Associates Inc. (2012)
16. Liu, Y., Sourina, O.: EEG-based valence level recognition for real-time applications. In: *International Conference on Cyberworlds*, pp. 53–60 (2012)
17. Naser, D.S., Saha, G.: Recognition of emotions induced by music videos using DT-CWPT. In: *Medical Informatics and Telemedicine*, pp. 53–57. IEEE (2013)
18. Yoon, H.J., Chung, S.Y.: EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm. *Comput. Biol. Med.* **43**(12), 2230–2237 (2013)
19. Wang, D., Shang, Y.: Modeling physiological data with deep belief networks. *Int. J. Inf. Educ. Technol.* **3**(5), 505–511 (2013)
20. Chen, J., Hu, B., Moore, P., et al.: Electroencephalogram-based emotion assessment system using ontology and data mining techniques. *Appl. Soft Comput.* **30**, 663–674 (2015)
21. Li, X., Zhang, P., Song, D., Yu, G., Hou, Y., Hu, B.: EEG based emotion identification using unsupervised deep feature learning. In: *SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research*, 13 August 2015
22. Atkinson, J., Campos, D.: Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Syst. Appl.* **47**, 35–41 (2015)