Active Feedback Framework with Scan-Path Clustering for Deep Affective Models

Li-Ming Zhao[1], Xin-Wei Li[1], Wei-Long Zheng[1], and Bao-Liang Lu[1,2,3(✉)]

[1] Center for Brain-like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, China
{lm_zhao,college_lxw,weilong,bllu}@sjtu.edu.cn

Abstract. The attention of subjects to EEG-based emotion recognition experiments could seriously affect their emotion induction level and annotation quality of EEG data. Therefore, it is important to evaluate the raw EEG data before training the classification model. In this paper, we propose a framework to filter out low quality EEG data from participants with low attention using eye tracking data and boost the performance of deep affective models with CNN and LSTM. We introduce a novel attention-deprived experiment with dual tasks, in which the dominant task is auditory continuous performance test, identical pairs version (CPT-IP) and the subtask is emotion eliciting experiment. Motivated by the idea that subjects with attention share similar scan-path patterns under the same clips, we adopt the cosine distance based spatial-temporal scan-path analysis with eye tracking data to cluster these similar scan-paths. The average accuracy of emotion recognition using the selected EEG data with attention is about 3% higher than that of original training dataset without filtering. We also found that with the increasing distance of scan-paths between outliers and cluster center, the performance of corresponding EEG data tends to decrease.

Keywords: Eye tracking · Scan-path · Attention evaluation
EEG data filtering

1 Introduction

Recently, multimodal emotion recognition based on EEG and eye movement data has attracted increasing attention. Combining eye movements and EEG can considerably improve the performance of emotion recognition systems because eye movements and EEG are complementary to emotion recognition [8]. Eye movement data has the advantages that the device is wearable and the data is

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11302, pp. 330–340, 2018.
https://doi.org/10.1007/978-3-030-04179-3_29

easy to handle. Meanwhile, the eye movement data can be used as a multi-modal data to complement the EEG data in emotion recognition, and can also be used as a measure of whether the subject is seriously involved in the experiment.

In emotion recognition from EEG signals, using emotional film clips as stimuli to elicit emotions is one of the most popular and effective methods [12,14,16]. However, it is hard to know whether the participants have been elicited corresponding emotions through watching these clips. Traditionally, questionnaires were sent to participants for self-assessment [7], which is rather subjective. In our previous work [13], we proposed that eye tracking data could be an effective reference for evaluating EEG data quality in emotion recognition experiments. However, this work failed to explain the following points: (a) What is the relationship among the participants' attention, the scan-path pattern and the emotion recognition performance. (b) How to guide the participants to watch each clip without attention. (c) How to better measure the trend variation of different scan-paths. In this study, we propose a modified framework which is capable of evaluating the quality of data and filtering EEG data for emotion recognition.

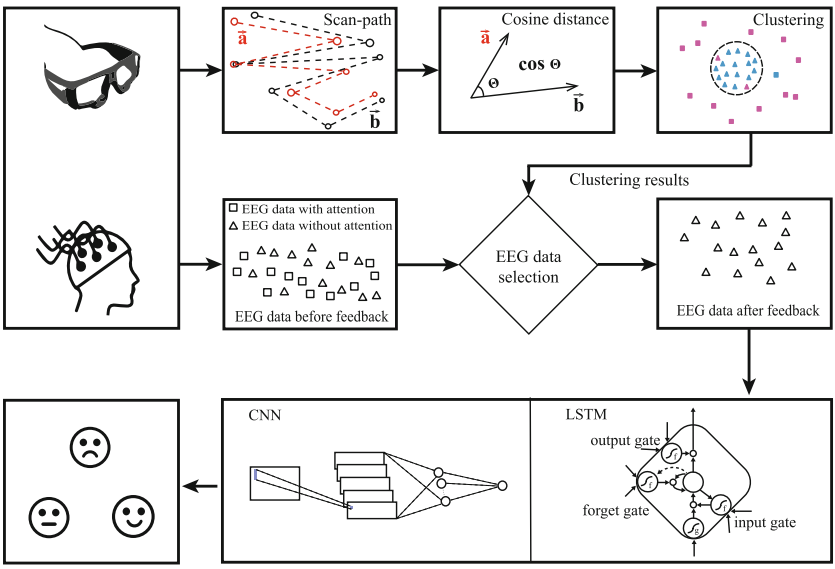

Fig. 1. The proposed framework for attention evaluation and feedback.

The flow chart of our framework is shown in Fig. 1. The EEG data and eye tracking data are collected when the participants watch the emotion film clips. The distance matrix is calculated based on cosine distances between any two scan-paths and is clustered with DBSCAN method [5]. In the feedback step, EEG data with high attention is selected by using the scan-path clustering results. The training dataset of the classifiers consists of the selected high-quality data after feedback.

In order to evaluate our framework, we design a novel attention-deprived experiment. The dominant task named as Auditory CPT-IP, derived from CPT-IP [2] which is used to activate particular regions of brain. We conduct both the attention-deprived experiment as well as the normal emotion eliciting experiment, and the data go through the process shown in Fig. 1. We compare the selected training dataset to original full data, and test the affective models on the same test dataset. The models include SVM, CNN and LSTM, which are popular in emotion recognition areas [15]. Finally, we analyze the relationship between the dissimilarity of scan-path and the performance of corresponding EEG data.

2 Experiment Settings

Previous studies have explored the reliability of using movie clips to induce emotions [12]. In our work, 15 Chinese movie clips with highly emotional contents are selected to induce three corresponding emotional states, i.e., positive, negative and neutral [16]. The subjects are instructed to sit comfortably, facing a large screen and move as least as possible to reduce the interference of artifact. There is a 5 s hint for starting, a 60 s rest for subjects to recover from elicited emotions in each trial. During these experiments both eye movement signal and EEG signal are collected simultaneously. Eye movement signals are recorded using SMI 30 Hz ETG eye tracking glasses. EEG signals are recorded by an ESI NeuroScan System with sampling rate 1000 Hz from a 62-channel electrode cap according to the international 10–20 system. A set of control experiments are performed under these common conditions.

2.1 Experiment with Attention

Sixteen subjects (7 males) aged between 20 and 24 years old, with normal or corrected-to-normal vision and normal hearing, participate in this experiment. All the participants are required to watch the clips with intently and elicit their own corresponding emotions. One loudspeakers plays the audio of movie clips and the volume is adjusted to the appropriate size. Both eye tracking data and EEG data are collected as shown in Fig. 2(a).

2.2 Attention-Deprived Experiment

In this experiment, according to the theory of cognitive psychology, attention-deprived tasks are added to video-based emotion experiments. Study shows that multitasking will cause switching costs and mixing costs [10], which could seriously affect the task-handle ability of the brain. In the video-based emotional stimulation experiment, participants have to understand both the auditory and visual information. Therefore, we design an auditory CPT-IP experiment to fight for the attention of the subjects in video task.

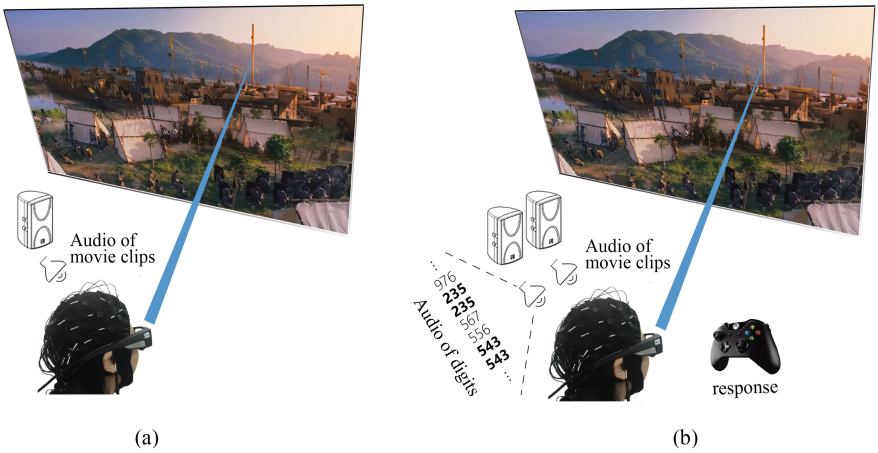

Fig. 2. Setting of experimental scene. (a) In the experiment with whole attention, the participants watched the movie clips normally. (b) In the attention-deprived experiment, two loudspeakers are used to play the audio of movies and digits. Gamepad is used to response and the response key needs to be pressed once a number was repeated.

As shown in Fig. 2(b), in the CPT-IP experiment, participants work through several conditions of a continuous performance task with the task to identify identical pairs of 3-digit numbers. Participants are presented a continuous stream of 3-digit numbers per second. The response key needs to be pressed once a number was repeated (Go trials). For non-repeating stimuli, participants are instructed to wait for the next repetition (NoGo trials).

Two separate loudspeakers are used in this attention-deprived experiment, one for the audio of movie clips and the other for CPT-IP auditory digits. Both speakers are set in the same place, with the same volume. Participants are required to complete the dominant auditory CPT-IP task as well as possible while watching the movie clips at the same time. We use the same equipment to collect the eye tracking data and the EEG data. Fourteen subjects, 7 males and 7 females, aged from 19 to 24, with self-reported normal or corrected-to-normal vision and normal hearing, participate in this experiment.

3 Method

3.1 Attention Evaluation

Generating Gaze Sequence. When subjects watching videos, their gaze position in each video frame would form a scan-path. Therefore, the eye movements can be regarded as a spatial-temporal sequence. The gaze sequence generation is a manual process using the BeGaze software from SMI. Since SMI ETG has a sampling rate of 30 Hz, we can acquire 30 raw sample points per second. Each

sample point includes the information of time, gaze position and pupil diameter, which is labeled with three different kinds of event type, including fixation, saccade and blink. In particular, the gaze position will be recorded as 0 for sample points whose event type is blink. Therefore, we fix these gaze positions by using linear interpolation method. Since the gaze sequence we extract from the raw data usually have strong fluctuations, we apply the moving average approach with the window of 6s to filter out the local jitter for sequence similarity comparison. Finally, the eye movement sequence is divided into 15 segments, according to the start and end time of each movie clips. The gaze sequence can be encoded into the following vector space both horizontally and vertically:

$$\begin{cases} \boldsymbol{S}_{i_c}^{\boldsymbol{x}} = [x_{c_1}^i,\ x_{c_2}^i,\ \ldots,\ x_{c_m}^i] \\ \boldsymbol{S}_{i_c}^{\boldsymbol{y}} = [y_{c_1}^i,\ y_{c_2}^i,\ \ldots,\ y_{c_m}^i] \end{cases} \quad i = 1,\ 2, \ldots,\ 30; c = 1,\ 2, \ldots,\ 15, \tag{1}$$

where i and c represent the participant number and clip number, respectively, m is the dimension of the gaze vector associated with the length of each movie clip, and x and y stand for the horizontal gaze position and the vertical gaze position, respectively.

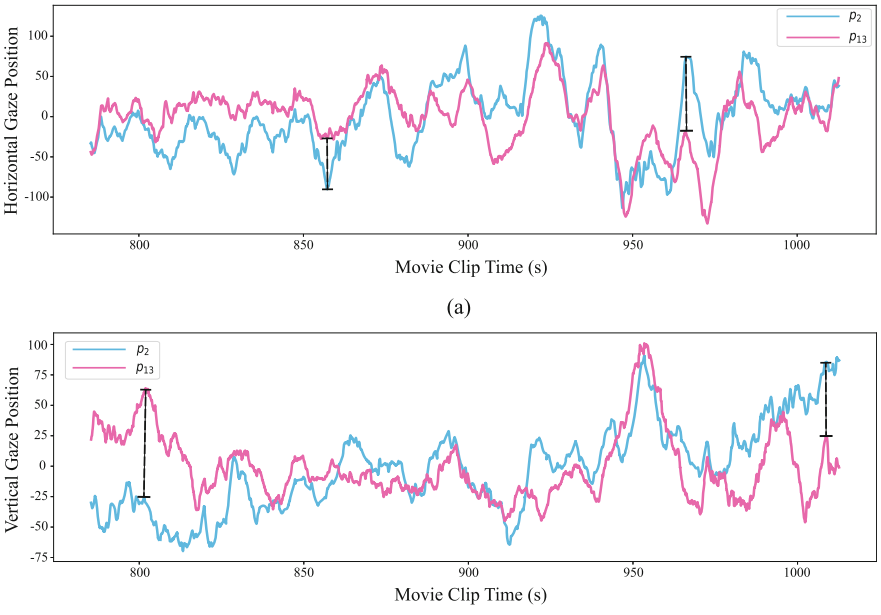

Fig. 3. The scan-path of two subjects with attention in movie clip 3. The black dotted line indicates a large difference in amplitude. (a) The map of horizontal gaze vector. (b) The map of vertical gaze vector.

Similarity Measures. The gaze sequence, as described above, should share high similarity among data collected from subjects with high attention, whereas there will be large differences between the subjects involved in the attention-deprived experiment. Before clustering, we must determine how to measure the similarity between different gaze sequence. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis. A wide variety of similarity measures can be used for clustering, such as Euclidean distance and cosine similarity which are very popular similarity measures for clustering [6]. The Euclidean distance of two vectors is defined as:

$$D_E(\boldsymbol{a}, \boldsymbol{b}) \ = (\sum_{t=1}^{m} |a_t - b_t|^2)^{1/2}. \tag{2}$$

Cosine similarity is one of the most popular similarity measures applied to clustering. Given two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, their cosine similarity is

$$D_C(\boldsymbol{a}, \boldsymbol{b}) \ = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}| \times |\boldsymbol{b}|}. \tag{3}$$

When gaze sequence is represented as gaze vectors in two directions, the similarity of two sequences corresponds to the correlation between the vectors. In general, the gaze vector has more than 7000 dimensions. Euclidean distance will not be a good metric in such a high dimensional space. Figure 3 depicts the horizontal and vertical gaze vector of movie clip 3 from two subjects with high attention. From Fig. 3, we can see that although the red and blue line have similar trend, they can sometimes be very different in amplitude, as illustrated in black dotted line. This difference is caused by head movements, which is hard to eliminate. In other words, although two participants looked at the same place, when they moved their heads, the SMI ETG would move along with it, thus the recorded gaze position would also be different. In this case, Euclidean measurement calculates the absolute distance between the two vectors, which is not suitable for our data set.

Compared with Euclidean distance, cosine distance is good at capturing the similarity of patterns of feature changes, at the same time disregarding the absolute amplitude of the compared feature vectors [11]. Therefore, we choose cosine distance as the similarity measurement method finally. The average cosine distance between two gaze sequences $\boldsymbol{S_i}$ and $\boldsymbol{S_j}$ can be calculated as

$$\begin{aligned} Dis_C(\boldsymbol{S_i}, \boldsymbol{S_j}) \ &= \tfrac{1}{15} \textstyle\sum_{c=1}^{15} \left(\tfrac{1}{2} \left(D_C(\boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{i}_c}, \boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{j}_c}) + D_C(\boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{i}_c}, \boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{j}_c}) \right) \right) \\ &= \tfrac{1}{30} \textstyle\sum_{c=1}^{15} \left(\frac{\boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{i}_c} \cdot \boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{j}_c}}{|\boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{i}_c}| \times |\boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{j}_c}|} + \frac{\boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{i}_c} \cdot \boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{j}_c}}{|\boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{i}_c}| \times |\boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{j}_c}|} \right), \end{aligned} \tag{4}$$

where $\boldsymbol{S}^{\boldsymbol{x}}_{\boldsymbol{i}_c}$ and $\boldsymbol{S}^{\boldsymbol{y}}_{\boldsymbol{i}_c}$ are defined in Eq. 1.

Clustering. Through clustering the scan-path by unsupervised clustering algorithm, we can distinguish whether the subjects are involved in the video. Based on the cosine distance matrix calculated by the previous steps, we perform a density-based clustering algorithm called DBSCAN [5] to fulfil our task. We chose DBSCAN because it is density-based and allows using precalculated distance matrix as the input. The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Meanwhile, clusters found by DBSCAN can be any shape, as opposed to k-means [4] which assumes that clusters are convex shaped. There are two parameters to the algorithm, *min_samples* and *eps*, which define formally what we mean when we say dense. Higher *min_samples* or lower *eps* indicate higher density necessary to form a cluster. By performing this algorithm, the subjects are classified into outliers and non-outliers, which are considered as subjects without attention and subjects with attention, respectively. Metric Multidimensional scaling (MDS) [1] is used for visualizing the clustering results. Both DBSCAN and MDS algorithms used here are built from the scikit-learn toolkit [9].

3.2 EEG Data Filtering for Emotion Recognition

The clustering results provide us with an indicator for filtering EEG data, which means if one subject is labeled with outliers, his or her corresponding EEG data will be removed from the training data set.

EEG Feature Extraction and Smoothing. Considering the effectiveness of differential entropy (DE) in EEG-based emotion recognition [3], we choose DE as the EEG feature. The DE features are extracted in five frequency bands : $\delta(1-3\ Hz)$, $\theta(4-7\ Hz)$, $\alpha(8-13\ Hz)$, $\beta(14-30\ Hz)$ and $\gamma(31-50\ Hz)$. A 256-point Short-time Fourier transform (STFT) with 1 s non-overlapping Hanning window is used to calculate the average DE features of each channel on these bands. Since 62-channel EEG signals are collected, we obtain 310 dimensional features for each sample. A linear dynamic system (LDS) approach is used to eliminate the rapid changes of DE features, which makes the features more reliable.

Classification. We test the clustering result on a series of classification models including Linear SVM, LSTM and CNN. To explore the influence of feedback for EEG-based emotion recognition on these models, we use all 30 subjects' EEG data recorded while they watching the first 9 movie clips, including 3 positive, 3 negative and 3 neutral clips, as the training set before feedback and remove the EEG data who is labeled with outliers from the training set when doing the feedback session. To ensure fairness, the rest six sessions, including 2 positive, 2 negative and 2 neutral clips from 16 subjects in video-base emotion experiment are used as the testing data set which keep the same before and after feedback.

4 Experimental Results

4.1 Results of Attention Evaluation

Firstly, we count the performance of the 14 subjects in the CPT-IP task. The accuracies of 12 subjects are higher than 93%, while the accuracies of the rest 2 subjects are below 85%. We randomly select 6 subjects for visualization, 3 subjects with attention and 3 subjects from attention-deprived experiment, denoted by p and n respectively. Figure 4 illustrate the horizontal and vertical gaze position along with time in one movie clip. The subjects with attention share a similar scan-path, while the subjects in attention-deprived experiment have different scan-paths. In particular, the scan-path of n_1 whose CPT-IP accuracy is 85%, is very similar to the average eye movement trajectory of the 16 subjects with attention.

The two parameters of DBSCAN, *eps* and *min_samples*, are set to 0.53 and 2, respectively. Finally, the subjects with similar scan-path are clustered into a cluster, while other subjects are scattered around them, among which 3 subjects, i.e. p_{16}, n_1 and n_5 do not match the experimental category to which they belong.

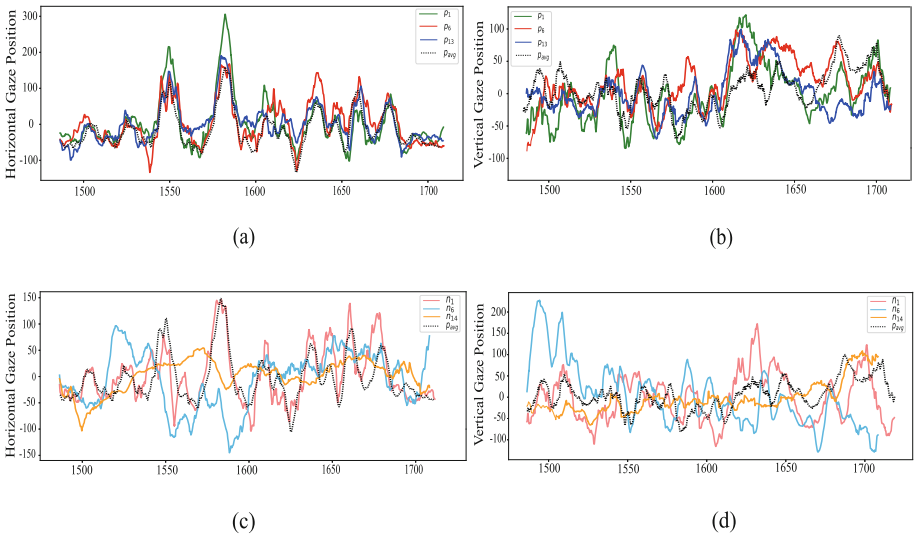

Fig. 4. Similarity comparison of scan-paths between different subject. The black dot line in each figure stands for the average scan-paths of all subjects with attention. (a), (b) The subjects with attention share a similar scan-path in horizontal gaze position and vertical gaze position. (c), (d) The subjects in the attention-deprived experiments have different eye movement trajectories.

4.2 Results of Feedback

According to the clustering results, we filtered out the EEG data of the subjects corresponding to the outliers. In order to verify the validity of the feedback, we test on different classifiers. The hyper-parameters and their corresponding range of these models are shown in Table 1.

Table 1. The hyper-parameters and their corresponding range of different models.

Model	Linear SVM	LSTM	CNN
c	$2^{-10} \sim 2^{10}$	-	-
Learning rate	-	$10^{-6} \sim 10^{-3}$	$10^{-5} \sim 10^{-3}$
Hidden layer	-	2	1 conv
Hidden size	-	$128 \sim 512$	$32 \sim 64$
Time step	-	$5 \sim 30$	-
Epoch	-	500	300

Table 2. Classification accuracies (%) of different models with and without EEG data filtering. The size of training data set is 60390 before EEG data filtering, and becomes 34221 after EEG data filtering.

Model	Linear SVM	LSTM	CNN
Accuracy without filtering	72.39	75.81	81.1
Accuracy with filtering	**76.30**	**80.39**	**82.3**

The recognition performance of these four models are shown in Table 2. As shown above, although the training data is reduced by nearly 50%, the accuracy in testing set is improved after feedback when the testing set keeps unchanged. We believe the reason for the increase in accuracy is that the EEG data removed by our attention evaluation algorithm contain less emotional patterns.

To explore the relationship between attention and emotion recognition accuracy, we train a linear SVM classifier for each subject's EEG data. Firstly, we use the EEG data for the first nine clips as the training set and the EEG data for the remaining six clips as the test data, then we sort the classification accuracy of each subject from high to low. There is a correlation between EEG accuracy and the average cosine distance, as shown in Fig. 5, when the accuracy rate of emotion recognition decreases, the divergence of scan-path between subjects tends to increase.

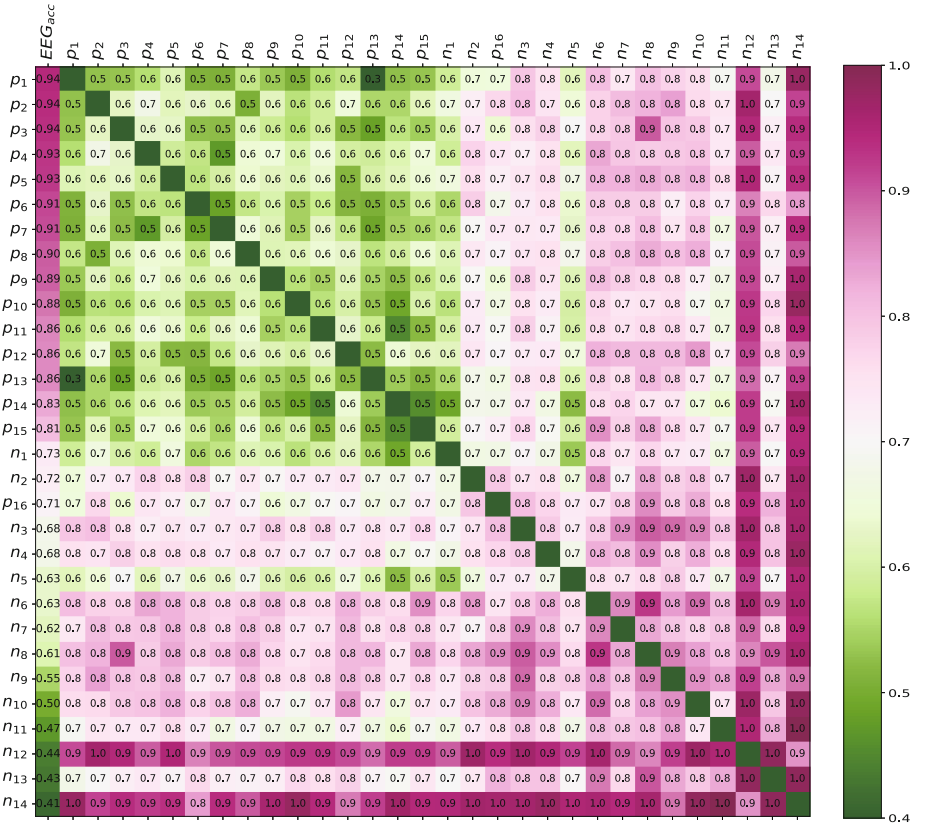

Fig. 5. The matrix shows the relationship between scan-path similarity and the emotion recognition accuracy. The first column is the EEG accuracy. The rest columns form a 30*30 matrix, in which each element is the average cosine distance between two scan-path.

5 Conclusion

In this paper, we have proposed a modified framework to evaluate the quality of EEG data for emotion recognition by using spacial-temporal scan-path analysis. The performance of emotion recognition using the selected EEG data with high engagement is better than that of original training dataset without filtering. The experimental results have demonstrated the effectiveness of our proposed framework and have indicated that the scan-path is related to the quality of data as well as the attendance level of participants.

Acknowledgments. This work was supported in part by the grants from the National Key Research and Development Program of China (Grant No. 2017YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266), and the Fundamental Research Funds for the Central Universities.

References

1. Borg, I., Groenen, P.: Modern multidimensional scaling: theory and applications. J. Educ. Meas. **40**(3), 277–280 (2003)
2. Cornblatt, B.A., Risch, N.J., Faris, G., Friedman, D., Erlenmeyer-Kimling, L.: The continuous performance test, identical pairs version (CPT-IP): new findings about sustained attention in normal families. Psychiatr. Res. **26**(2), 223–238 (1988)
3. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 6th International IEEE/EMBS Conference on Neural Engineering, pp. 81–84. IEEE (2013)
4. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. A Wiley-Interscience Publication, New York (1973)
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
6. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, pp. 49–56 (2008)
7. Koelstra, S., et al.: Patras: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
8. Lu, Y., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: IJCAI, vol. 15, pp. 1170–1176 (2015)
9. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
10. Philipp, A.M., Kalinich, C., Koch, I., Schubotz, R.I.: Mixing costs and switch costs when switching stimulus dimensions in serial predictions. Psychol. Res. **72**(4), 405–414 (2008)
11. Qian, G., Sural, S., Gu, Y., Pramanik, S.: Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In: Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 1232–1237. ACM (2004)
12. Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cognit. Emot. **24**(7), 1153–1172 (2010)
13. Shi, Z.F., Zhou, C., Zheng, W.L., Lu, B.L.: Attention evaluation with eye tracking glasses for EEG-based emotion recognition. In: 8th International IEEE/EMBS Conference on Neural Engineering, pp. 86–89. IEEE (2017)
14. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
15. Yan, X., Zheng, W.L., Liu, W., Lu, B.L.: Investigating gender differences of brain areas in emotion recognition using LSTM neural network. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) ICONIP 2017. LNCS, vol. 10637, pp. 820–829. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70093-9_87
16. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)