future internet

MDPI

Article

Combining Facial Expressions and Electroencephalography to Enhance Emotion Recognition

Yongrui Huang, Jianhao Yang, Siyu Liu and Jiahui Pan *

School of Software, South China Normal University, Guangzhou 510641, China; huangyongrui@m.scnu.edu.cn (Y.H.); 20152005029@m.scnu.edu.cn (J.Y.); 20162005055@m.scnu.edu.cn (S.L.)

* Correspondence: panjiahui@m.scnu.edu.cn

Received: 28 February 2019; Accepted: 16 April 2019; Published: 2 May 2019

Abstract: Emotion recognition plays an essential role in human–computer interaction. Previous studies have investigated the use of facial expression and electroencephalogram (EEG) signals from single modal for emotion recognition separately, but few have paid attention to a fusion between them. In this paper, we adopted a multimodal emotion recognition framework by combining facial expression and EEG, based on a valence-arousal emotional model. For facial expression detection, we followed a transfer learning approach for multi-task convolutional neural network (CNN) architectures to detect the state of valence and arousal. For EEG detection, two learning targets (valence and arousal) were detected by different support vector machine (SVM) classifiers, separately. Finally, two decision-level fusion methods based on the enumerate weight rule or an adaptive boosting technique were used to combine facial expression and EEG. In the experiment, the subjects were instructed to watch clips designed to elicit an emotional response and then reported their emotional state. We used two emotion datasets—a Database for Emotion Analysis using Physiological Signals (DEAP) and MAHNOB-human computer interface (MAHNOB-HCI)—to evaluate our method. In addition, we also performed an online experiment to make our method more robust. We experimentally demonstrated that our method produces state-of-the-art results in terms of binary valence/arousal classification, based on DEAP and MAHNOB-HCI data sets. Besides this, for the online experiment, we achieved 69.75% accuracy for the valence space and 70.00% accuracy for the arousal space after fusion, each of which has surpassed the highest performing single modality (69.28% for the valence space and 64.00% for the arousal space). The results suggest that the combination of facial expressions and EEG information for emotion recognition compensates for their defects as single information sources. The novelty of this work is as follows. To begin with, we combined facial expression and EEG to improve the performance of emotion recognition. Furthermore, we used transfer learning techniques to tackle the problem of lacking data and achieve higher accuracy for facial expression. Finally, in addition to implementing the widely used fusion method based on enumerating different weights between two models, we also explored a novel fusion method, applying boosting technique.

Keywords: emotion recognition; EEG; facial expressions; decision-level fusion; transfer learning

1. Introduction

In recent years, emotion recognition has gained greater significance in many areas, representing a crucial factor in human–machine interaction systems. These applications are manifested across a variety of levels and modalities [1]. Different modalities, which are both non-physiological and physiological, are used to detect emotion in these applications. The focus of this paper is on a multimodal fusion between facial expressions and an electroencephalogram (EEG) for emotion recognition.

Psychologists have proposed various emotion representation models, including discrete and dimensional models [2]. Discrete emotions, such as happiness and disgust, are straightforward to understand because they are based on language. However, these models can fall short in expressing certain emotions in different languages and for different people. In contrast, emotions can be presented in multidimensional spaces derived from studies, which identify the axes that exhibit the largest variance of all possible emotions. The two-dimensional model of arousal and valence is the most widely used dimensional emotion model [3]. The arousal–valence coordinates system model maps discrete emotion labels in a two-dimensional space. Valence ranges from unpleasant to pleasant, and arousal ranges from calm to excited, and can describe the intensity of emotion [4]. Given that most of the variance in emotions comes from two dimensions, our research implemented emotion recognition using these two dimensions.

The early works of emotion recognition were more performed by voluntary signals, such as facial expression, speech, and gesture. In one study, Abhinav Dhall et al. used videos of faces to classify facial expressions into seven categories (anger, disgust, fear, happiness, neutral, sadness, and surprise) and achieved 53.62% accuracy based on a test set [5]. Pavitra Patel et al. applied the Boosted-GMM (Gaussian mixture model) algorithm to speech-based emotion recognition and classified emotion into five categories (angry, happy, sad, normal, and surprise) [6]. Although non-physiological signals are easier to obtain, they are less reliable. For example, we can fake our facial expression to cheat the machine. Recently, more researches were done based on involuntary signals. Wei-Long Zheng el al. investigated stable patterns of EEG over time for emotion recognition. They systematically evaluated the performance of various popular feature extraction, feature selection, feature smoothing, and pattern classification methods with a Database for Emotion Analysis using Physiological Signals (DEAP) and a newly developed dataset called SJTU Emotion EEG Dataset (SEED) for this study [7]. They found that a discriminative graph regularized extreme learning machine with differential entropy features can achieve the best average accuracies of 69.67% and 91.07% on the DEAP and SEED datasets, respectively. Yong-Zhang el al. presented their attempt to investigate feature extraction of EEG-based emotional data by focusing on an empirical mode decomposition (EMD) and autoregressive (AR) model, and constructed an EEG-based emotion recognition method to classify these emotional states, reporting an average accuracy between 75.8% to 86.28% in the DEAP dataset [8].

The methods based on physiological signals seem to be more effective and reliable, but physiological signals often mix with noisy signals. A common instance is that the movement of facial muscles often causes the fluctuation of EEG.

In recent years, with the development of multisource heterogeneous information fusion processing, it has become possible to fuse features from multicategory reference emotion states. Jinyan Xie el al. proposed a new emotion recognition framework based on multi-channel physiological signals, including electrocardiogram (ECG), electromyogram (EMG), and serial clock line (SCL), using the dataset of Bio Vid Emo DB, and evaluated a series of feature selection methods and fusion methods. Finally, they achieved 94.81% accuracy in their dataset [9]. In a study using the MAHNOB-human computer interface (MAHNOB-HCI) dataset, Sander Koelstra et al. performed binary classification based on the valence-arousal-dominance emotion model using a fusion of EEG and facial expression and found that the accuracies of valence, arousal, and control were 68.5%, 73%, and 68.5%, respectively [10]. Mohammad et al. proposed a method for continuously detecting valence from EEG signals and facial expressions in response to videos and studied the correlations of features from EEG and facial expressions with continuous valence in the MAHNOB-HCI dataset [11]. In addition, our previous study proposed two multimodal fusion methods between the brain and peripheral signals for emotion recognition and reached 81.25% and 82.75% accuracy for four categories of emotion states (happiness, neutral, sadness, and fear). Both of these accuracies were higher than the accuracies of facial expression (74.38%) and EEG detection (66.88%) [12]. In our previous study, we applied principal component analysis to analyze facial expression data and extract high-level features, and we used fast Fourier transform to extract various power spectral density (PSD) features from raw EEG signals. Due to the limited amount of data, we used a simple model, namely, a two-layer neutral network for facial expression and a support vector machine (SVM) for EEG rather than a deep learning model, and two simple fusion rules were applied to combine EEG and facial expressions. Simply, the method can help prevent overfitting in cases of limited data availability.

All studies have shown that the performance of emotion recognition systems can be improved by employing multimodal information fusion. However, for facial expression, the abovementioned works all used manually assessed features (e.g., action units [10] or generic linear features [11]). Convolutional neural networks (CNNs) are able to extract effective features automatically during learning and achieve better performance for images. However, due to the lack of data, deep learning models are difficult to apply in emotion experiments. In addition, certain limitations still existed in our previous study. On the one hand, the data used in our experiment were limited, and our model was not tested using different types of datasets, which made it impossible to explore some advanced models like deep learning model. On the other hand, our fusion methods were simple but less reliable. We used the average output of two models and the expert knowledge-based product rules for fusion.

In this study, we adopted a multimodal emotion recognition framework by combining facial expression and EEG. For facial expression, we applied a pre-trained, multi-task CNN model to automatically extract facial features and detect the value of valence and arousal using a single modal framework. For EEG, we first used the wavelet transform to capture the time-domain characteristics of the EEG when extracting PSD features, followed by using two different SVM models to recognize the value of valence and arousal. Finally, we obtained the parameters of decision-level fusion based on training data on both fusion methods. As a follow-up study, we solved some existed limitation in our previous research. Unlike our previous research, we used advanced deep learning models (e.g., CNN) to detect facial expression and used more datasets (i.e., DEAP and MAHNOB-HCI) to test our models [13,14]. Additionally, an online experiment was also conducted. On the other hand, the fusion methods we used in this study were based on training data instead of expert knowledge, which makes them more robust to data. Moreover, the contribution of this work is as follows. To begin with, we combined facial expression and EEG to improve the performance of emotion recognition. Furthermore, we used transfer learning techniques to solve the problem of lacking data and achieve higher accuracy for facial expression. Finally, in addition to implementing the widely used fusion method based on enumerating different weights between two models, we also explored a novel fusion method applying boosting technique.

2. Materials and Methods

Figure 1 gives an overview of the workflow in our work. In the beginning, the video clips were used to elicit the emotional response of the subjects. During this process, the signals (face image and EEG) were recorded. At the end of the video, the subjects were asked to report their emotional status of valence and arousal. The status of valence and arousal both ranged from 1 to 9. In this study, we performed binary classification on the statuses of valence and arousal, which are the threshold into high (rating 6–9) and low (rating 1–5) classes. For facial expression, in the training stage, we first pre-trained a CNN model in a large dataset, followed by fine tuning the model in the target dataset. In the test stage, we found the faces from the images obtained by camera, and then used the extracted faces to predict the results of the facial expression using CNN. For EEG, we applied wavelet transform to obtain PSD features from raw EEG data. We further selected the extracted features using recursive feature elimination (RFE). The selected features were used to classify by SVM to get the results of the EEG. Finally, we fused the results of facial expression and EEG to get the fusion results. All the results were compared to the ground truth labels reported by subjects and the performance (e.g., accuracy) of models was obtained.

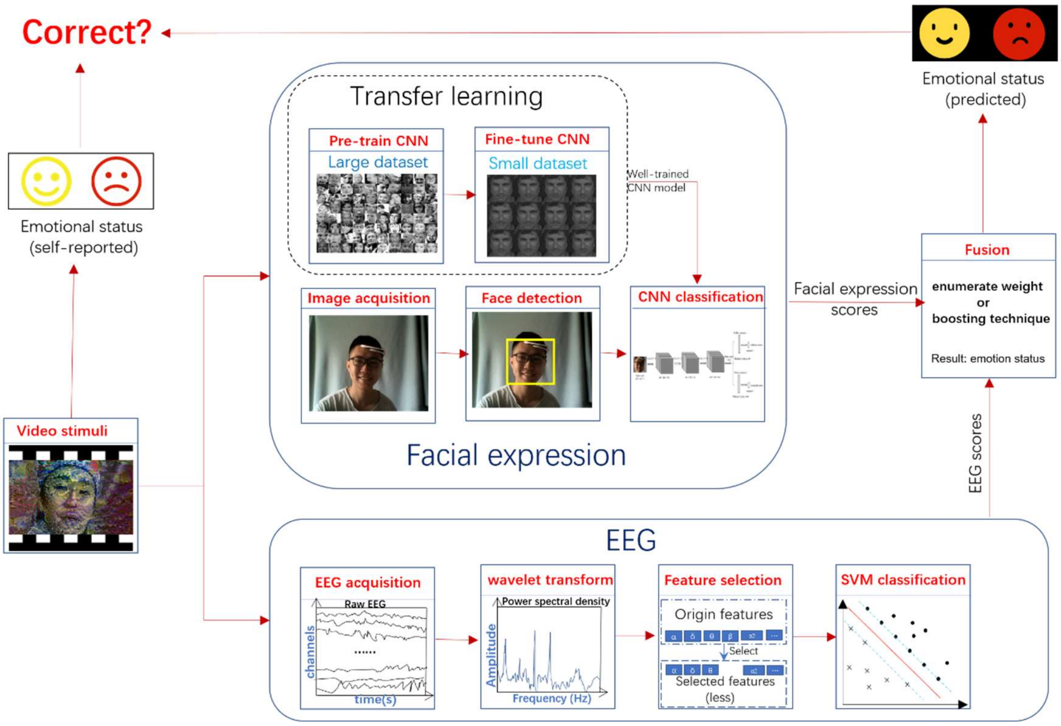

Figure 1. An overview of the workflow in this work.

2.1. Dataset Description and Data Acquisition

In this paper, we ran offline experiments in MAHNOB-HCI and DEAP dataset. The MAHNOB-HCI dataset contained EEG, video, audio, gaze, and peripheral physiological recordings of 30 participants [13]. In this dataset, each participant watched 20 clips extracted from Hollywood movies and video websites, such as youtube.com and blip.tv. The stimulus videos ranged in duration from 35 to 117 s. After watching each stimulus, the participants used self-assessment manikins (SAMs) to rate their perceived arousal and valence on a discrete scale of 1 to 9 [15]. We then divided these selections into a high class (ratings 6–9) and a low class (ratings 1–5). For 6 of the 30 participants, various problems (such as a technical failure) occurred during the experiment [13]. Here, only the 24 participants for whom all 20 trials are available were used.

The DEAP data set includes the EEG and peripheral physiological signals of 32 participants when watching 40 one-minute music videos [14]. This data set also contains the participants' ratings of each video in terms of the levels of arousal and valence. We divided these values into a high class (ratings 6–9) and a low class (ratings 1–5). Note that we used only 13 participants for whom both the face data and EEG data were available for all 40 trials [14].

Additionally, we performed on online experiment. In the online experiment, we collected data from 20 participants (50% female, 50% male, aged 17–22). Their facial images were captured by a Logitech camera (25 FPS, 800 × 600 image size) while the EEG signals were recorded using EMOTIV INSIGHT, a five channel mobile EEG device (San Francisco, CA, USA). Five channels were placed on the scalp at positions AF3, AF4, T7, T8, and Pz to record the EEG signals. The sensors were placed according to the standard 10–20 system. The impedances of all electrodes were maintained below 5 kΩ.

2.2. Spontaneous Facial Expression Detection

For predicting the emotion status based on facial expression, the following steps were taken. We first resampled the video to 4 Hz and used OpenCV's Viola & Jones face detector (frontal and profile) to find the face's position in each image frame [16]. The extracted faces were resized into a width and height of 48 pixels and fed directly to the multi-task CNN model to conduct a forward propagation. The outputs of the forward propagation were the status of valence and arousal.

The training process and architecture of the multi-task CNN were as follows. The training process of the multi-task CNN was conducted in two steps. In the first step, supervised pre-training was performed. We discriminatively pre-trained the CNN using a large auxiliary data set (namely, fer2013) with image-level annotations [17]. In the second step, domain-specific fine tuning was performed. We continually conducted stochastic gradient descent (SGD) training for the CNN parameters using only faces extracted from the video with a relatively low learning rate (0.001).

The illustration in Figure 2 shows the architecture of this network. Given the input 48 × 48 grayscale image patch, the first layer was a convolutional layer with 32 kernels of size 3 × 3 × 1. The second layer was a convolutional layer with 32 kernels of size 3 × 3 × 32. The third convolutional layer had 64 kernels of size 3 × 3 × 32, followed by 2 × 2 max pooling. The fourth layer was fully connected with 64 neurons. In all convolutional and fully connected layers, the rectified linear unit (ReLU) non-linearity approach was applied [18]. The network was subsequently split into two branches.

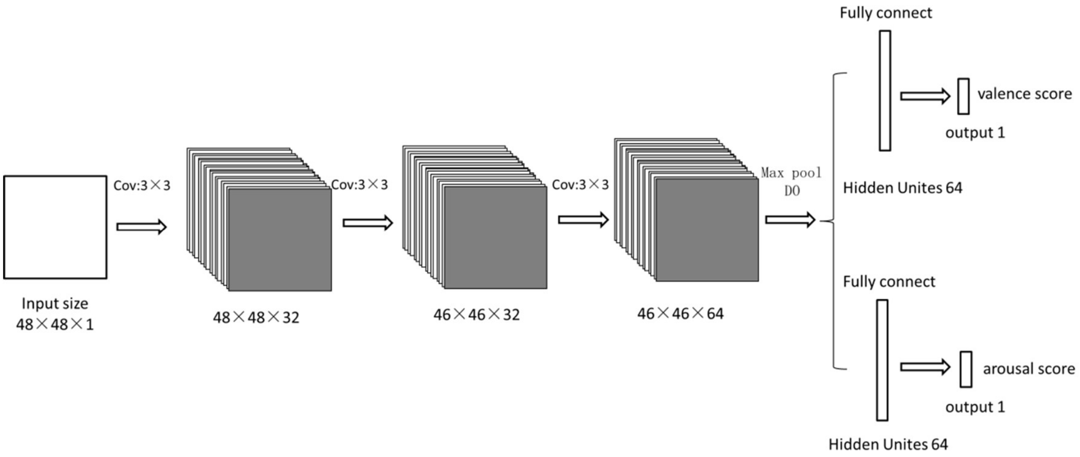

Figure 2. The CNN network for multitask classification; "DO" indicates that the layers drop out [15]. Note that we apply padding in each layer except the second layer. Therefore, excluding the second layer, the output of each layer is of the same size as the previous layer.

The first branch learns the valence decision and contains two additional fully connected layers of sizes 64 and 1. The output is then input into a sigmoid function, and the cross-entropy loss L_1 is minimized:

$$L_1 = -\sum_{i=1}^{m} (1 - y_{1i} \log \hat{y_{1i}} + \hat{y_{1i}}) \log(1 - \hat{y_{1i}}) \tag{1}$$

where y_{1i} represents the true label of valence of the kth sample, $\hat{y_{1i}}$ represents the sigmoid output of the ith sample, and m represents the size of the training sample.

The second branch learned the arousal decision and contained two additional fully connected layers of sizes 64 and 1. The output was fed to a sigmoid function, and we again minimized the cross-entropy loss L_2:

$$L_2 = -\sum_{i=1}^{m} (1 - y_{2i} \log \hat{y_{2i}} + \hat{y_{2i}}) \log(1 - \hat{y_{2i}}), \tag{2}$$

where y_{2i} represents the true label of arousal in the ith sample, $\hat{y_{2i}}$ represents the sigmoid output of the ith sample, and m represents the size of the training sample.

When multi-task learning was performed, we minimized the following linear combination of losses:

$$L = \sum_{p=1}^{2} \alpha_p L_i, \tag{3}$$

where α_p are linear weights currently set to 1. Note that if we set the second weights to 0, we return to the conventional approach of single-task learning.

After training, we could obtain the predicted result from the output of the network S_{face}, as shown in (4), where r_{face} represents the result of valence or arousal based on facial expression:

$$r_{face} = \begin{cases} high & S_{face} \geq 0.5 \\ low & S_{face} < 0.5 \end{cases} \tag{4}$$

2.3. EEG Detection

The EEG-based detection included three progressive stages: PSD features extraction, feature selection, and classification.

2.3.1. PSD Features Extraction

The raw electrode data were filtered using a wavelet transform to obtain the PSD features. The wavelet transform is appropriate for multiresolution analysis, in which the signal can be examined at different frequencies and time scales. We used Daubechies' wavelet transform coefficients for feature extraction because EEG signals contain information at various frequency bands. The wavelet, which is a mathematical transformation function, divided the data into diverse frequency components. With the resolution matched to the scale, these components were analyzed separately. A wavelet transform is the representation of a function by mother wavelets. The one-dimensional sustained wavelet transform denoted by $w_f(s, \tau)$ of a one-dimensional function $f(t)$ is defined in (5) [19]:

$$w_f(s, \tau) = \int_{-\infty}^{\infty} f(t)\varphi_{s\tau}(t)dt, \tag{5}$$

where φ is the mother wavelet function, s is the scale parameter, and t is the translation parameter. To reconstruct the original signal from the transformed signal, the inverse sustained wavelet transform is defined in (6):

$$f(t) = \frac{1}{C_\varphi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_f(s, \tau)\varphi_{s\tau}(t)\frac{dtds}{s^2}, \tag{6}$$

such that $C_\varphi = \int_{-\infty}^{\infty}\left(\left|\varphi(u)\right|^2 / |u|du\right)$, where $\varphi(u)$ is the Fourier transform of $\varphi(t)$ [20].

For the experiment run on offline dataset MAHNOB-HCI and DEAP, we chose only 14 electrodes (Fp1, T7, CP1, Oz, Fp2, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, and PO4) and three symmetric pairs (i.e., T7-T8, Fp1–Fp2, and CP1-CP2) for the EEG feature extraction. The logarithms of the PSD from the theta (4 Hz < f < 8 Hz), slow alpha (8 Hz < f < 10 Hz), alpha (8 Hz < f < 12 Hz), beta (12 Hz < f < 30 Hz), and gamma (30 Hz < f < 45 Hz) bands were extracted from all 14 electrodes and three symmetric pairs to serve as features. The total number of EEG features in the offline dataset using 14 electrodes and three corresponding asymmetric features were 14 × 5 + 3 × 5 = 85.

For the online experiment, all five electrodes (AF3, AF4, T7, T8, and Pz) provided by the device were used. The logarithms of the PSD from the theta (4 Hz < f < 8 Hz), slow alpha (8 Hz < f < 10 Hz), alpha (8 Hz < f < 12 Hz), beta (12 Hz < f < 30 Hz), and gamma (30 Hz < f < 45 Hz) bands were extracted from all five electrodes. The total number of EEG features in the online experiment were 5 × 5 = 25.

2.3.2. Features Selection

SVM-RFE (recursive feature elimination) was used to select features from the set of features for a better classification of each subject. This was done by iteratively calculating the feature weights using a linear SVM classifier (the penalty parameter C of the error term was equal to 1.0) and subsequently removing 10% of the features with the lowest weights. This process continued until the target number of features remained. The number of selected features was optimized using 10-fold inner cross-validation for the training set [21]. Finally, we selected half of the features from the features extracted from the last stages (i.e., 42 for the offline dataset MAHNOB-HCI and DEAP, and 12 for the online experiment).

2.3.3. Classification

After the final features were chosen, these features were input to train an SVM classifier by applying the radial basis function (RBF) kernel (hype-parameter: C = 1.0, gamma = 1/the number of features). After training, the features were classified into high/low classes in the valence/arousal space. Two learning tasks (valence and arousal) were used to implement the above method. Each task yielded results based on the SVM output S_{EEG} according to (7), where r_{EEG} represents the result of valence or arousal based on EEG,

$$r_{EEG} = \begin{cases} high & S_{EEG} \geq 0.5 \\ low & S_{EEG} < 0.5 \end{cases} \tag{7}$$

2.4. Classification Fusion

After the two classifiers for facial expressions and EEG data were obtained, various modality fusion strategies were used to combine the outputs of these classifiers at the decision level. We employed two fusion methods for both EEG and facial expression detection as follows.

For the first fusion method, we applied the enumerate weight fusion approach for decision-level fusion. It was widely used in a lot of the research about multimodal fusion [10,11]. Specifically, the output result of this method is given in (8):

$$\begin{aligned} S_{enum} &= kS_{face} + (1-k)S_{EEG} \\ r_{enum} &= \begin{cases} high & S_{enum} \geq 0.5 \\ low & S_{enum} < 0.5 \end{cases}, \end{aligned} \tag{8}$$

where r_{enum} represents the predicted result (high or low), S_{face} and S_{EEG} represent the predicted output scores for the facial expression and EEG, respectively, and k (ranging from 0 to 1) represents the importance degree of the facial expression. The key objective of this method is to find a proper k that can lead to satisfactory performance. To achieve this objective, k is varied between 0 and 1 in steps of 0.01, and the value that provides the highest accuracy for the training samples is selected. We applied this method separately for the two learning tasks (valence and arousal) and obtained two different k values, one for valence space and the other for arousal space. Note that we trained models with different parameters for different subject because we treated it as a subject-dependence problem. This means the value of k is different for each subject. There is an optimal value for this parameter, but it is different from subject to subject.

For the second fusion method, we applied adaptive boosting (adaboost) techniques to decision-level fusion [22,23]. Our goal for this approach was to train w_j (j = 1, 2, . . . , n) for every sub-classifier and obtain the final output, as given in (9):

$$\begin{aligned} S_{boost} &= 1/(1+\exp(-\sum_{j=1}^{n} w_j s_j)) \\ r_{boost} &= \begin{cases} high & S_{boost} \geq 0.5 \\ low & S_{boost} < 0.5 \end{cases}, \end{aligned} \tag{9}$$

where r_{boost} represents the prediction results (high or low) of the adaboost fusion classifier. $s_j \in \{-1, 1\}$ $(j = 1, 2, \ldots, n)$ represents the corresponding output result of the j sub-classifier. In this case, S_1 is the facial expression classifier and S_2 is the EEG classifier.

To obtain w_j (j = 1, 2...n) from a training set of size m, $s(x_i)_j \in \{-1, 1\}$ designates the output of the jth classifier for the ith sample, and y_i denotes the true label of the ith sample.

We first initialized the weights α_i for each data point using (10),

$$\alpha_i = 1/m, \tag{10}$$

where α_i represents the weight of the ith data point.

Each sub-classifier was used to predict the training set and calculate the error rate ε_j using (11),

$$\varepsilon_j = \sum_{i=1}^{M} t_i \alpha_i, \tag{11}$$

where t_i is calculated using (12),

$$t_i = \begin{cases} 0 & s(x_i)_j = y_i \\ 1 & s(x_i)_j \neq y_i \end{cases}. \tag{12}$$

The weight of the jth sub-classifier was assigned as shown in (13):

$$w_j = \ln((1 - \varepsilon_j)/\varepsilon_j)/2. \tag{13}$$

Subsequently, we updated α_i as shown in (14):

$$\alpha_i = \begin{cases} \alpha_i \exp(-w_j)/\sum_{i=1}^{m} \alpha_i \; s(x_i)_j = y_i \\ \alpha_i \exp(w_j)/\sum_{i=1}^{m} \alpha_i \; s(x_i)_j \neq y_i \end{cases}. \tag{14}$$

We continued to calculate the weight of the subsequent sub-classifier after updating α_i.

The facial expression classifier and the EEG classifier were both used as the sub-classifiers of the boosting classifier. We applied this method separately for the two learning tasks (valence and arousal).

We applied two fusion methods, both of which can give the final results of valence and arousal (i.e., two fusion methods gave their results independently). For the first fusion method, we used Equation (8) to calculate the final results after the output of facial expression and EEG and the value k were obtained. For the second method, we took both classifiers (facial expression and EEG) as the sub-classifier of adaboost algorithm to train the weight of the adaboost algorithm. The final results were calculated using Equation (9).

2.5. Experiment

2.5.1. Experiment Based on Public Datasets

For each subject in the MAHNOB-HCI dataset, leave-one-trial-out cross-validation was performed for binary classification. In this process, tests were performed on one trial, and the other 19 trials were used for training. The video data were used to fine tune the facial expression classifier (CNN), pre-trained based on the fer2013 data set, and the EEG data were used to train the EEG classifier (SVM). For each subject, we used the number of trials that were predicted correctly compared to the total number of trials as a metric to measure model performance.

For each subject in the DEAP dataset, we randomly selected 20 trials as a training set, and the rest of the 20 trials were used as a test set. A leave-one-trial-out cross-validation was performed for the training set to select the best hyperparameter for the model and then the trained models were tested in test set. For each subject, we used the accuracy of the test set as a metric to assess model performance.

2.5.2. Online Experiment

Twenty subjects (50% female, 50% male), whose ages ranged from 17 to 22 (mean = 20.15, std = 1.19), volunteered to participate in the experiment. We first introduced the meanings of "valence" and "arousal" to the subjects. The subjects were instructed to watch video clips and reported their emotion status of valence and arousal at the end of each video. During the experiments, the subjects were seated in a comfortable chair and instructed to avoid blinking or moving their bodies. We also conducted device testing and corrected the camera position to ensure that the faces of the subjects appeared in the center of the screen.

In a preliminary study, 40 video clips were manually selected from commercially produced movies as stimuli. They were separated into two parts for the calibration run and evaluation run, respectively. Each part contained 20 videos. The movie clips range in duration from 70.52 to 195.12 s (mean = 143.04, std = 33.50).

In order to conduct an evaluation run, we first needed data to train the model. Therefore, we first conducted a calibration run to collect data before performing an evaluation run. In the calibration run, the collected data consisted of 20 trials for each subject. At the beginning of each trial, there was a 10 s countdown in the center of the screen to capture each subject's attention and to serve as an indicator of the next clip. After the countdown was complete, movie clips that included different emotional states were presented on the full screen. During this time, we collected four human face images in each second using a camera and 10 groups of EEG signals each second using an Emotive mobile device. Each movie clip was presented for 2–3 min. At the end of each trial, a SAM appeared in the center of the screen to collect the subject's label of valence or arousal [15]. Subjects were instructed to fill in the entire table and to click the "submit" button to proceed to the next trial. There was also a 10 s countdown in the center of the screen between two consecutive trials for emotional recovery. The collected data (EEG signal, face image and corresponding valence, and arousal label) were used to train the model described above.

The evaluation run was composed of 20 trials for each subject. The procedure of each trial was similar to that of data collection. Note that different movie clips were used for stimuli because reusing these stimuli would have reduced the impact of the movie clips by increasing the knowledge that subjects had of them. At the end of each trial, four different detectors (face expression detector, EEG detector, first fusion detector, and second fusion detector) were used to predict the valence and arousal based on the face image and EEG signal. By comparing the predicted result and the ground truth label, the accuracy was subsequently calculated.

During the 10 s countdown, the sensors stopped recording signals. The data between two 10 s countdowns were used as a sample for the fusion level. For facial expression, the face data was used to fine-tune the CNN model. For EEG, the EEG data was used to train the SVM model. The two sub-models were trained independently and then they were fused together in the decision level. This guaranteed the face and EEG data had been co-registered for a trial.

3. Results

3.1. Result Analysis

Figure 3 provides a graphical overview of the accuracy of various detection methods among various data sets for each subject. From top to bottom, the accuracy of the MAHNOB-HCI data set, the DEAP data set, and the online experiment in both valence and arousal space are presented in order.

The emotion recognition results of the MAHNOB-HCI data set are presented in Table 1, the results of the DEAP data set are presented in Table 2, and the results of the online experiment are presented in Table 3. Although the overall accuracy of facial expression was high, poor performance was still observed for some subjects (such as subjects 1, 3, 5, 11, and 12 for the valence space in the DEAP data set), indicating that high volatility is associated with facial expressions during emotion experiments. The performance of the modality can be improved by combining the findings with those of another modality (e.g., EEG). Except for one case (arousal space in the online experiment, where the accuracy of the first fusion method was lower than that of EEG (signal modality)), the accuracy of the fusion method was higher than that of the signal modality approach.

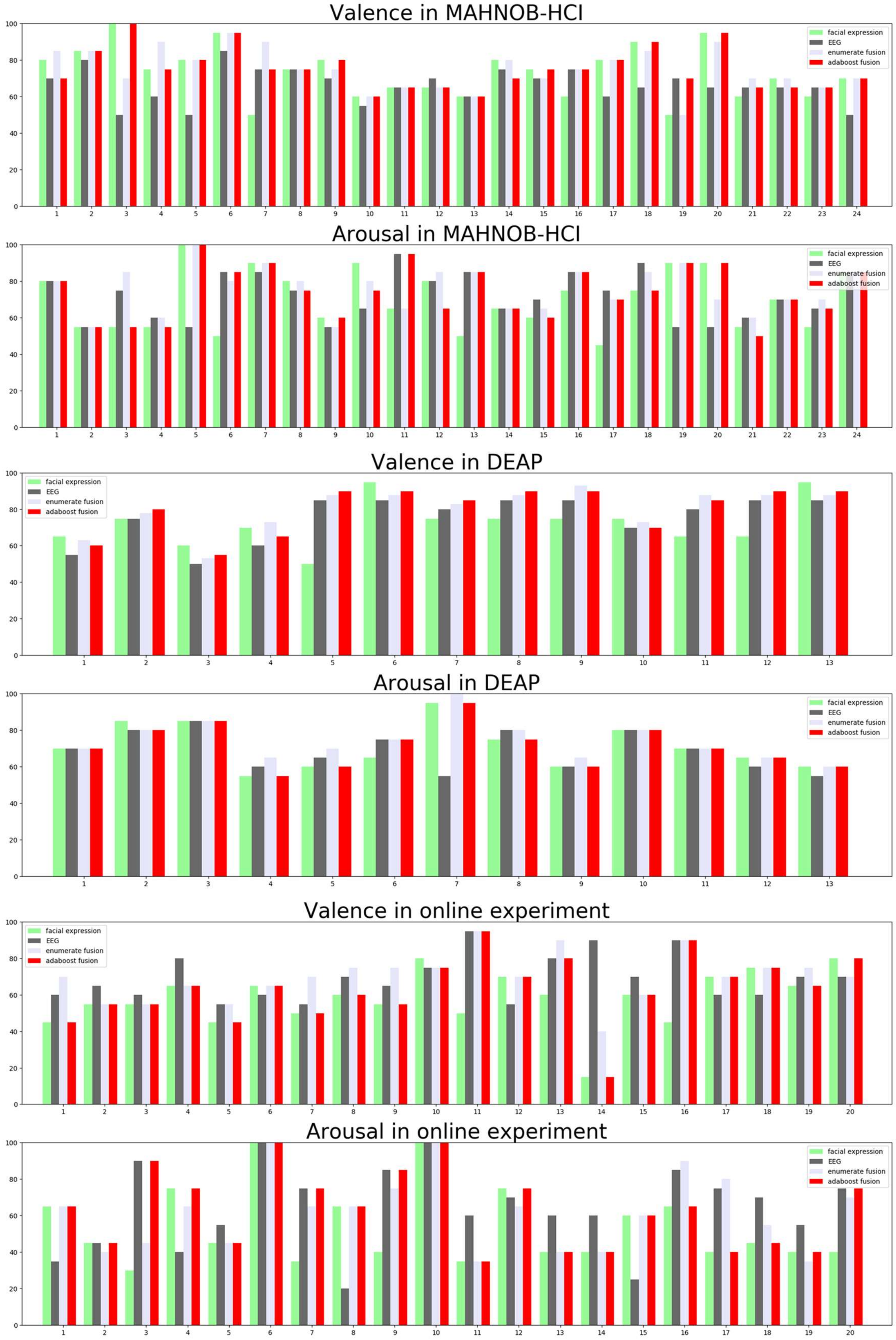

Figure 3. Accuracy of various subjects among various data sets. The X-axis of each sub-figure represents the subject I.D., and the Y-axis represents the accuracy.

Table 1. Emotion recognition accuracy for the MAHNOB-HCI data set.

Target	Facial expression	EEG	First fusion	Second fusion
Valence	73.33 ± 13.59	66.25 ± 9.04	75.00 ± 11.08	75.21 ± 10.94
Arousal	69.79 ± 15.64	71.88 ± 12.48	75.63 ± 11.92	74.17 ± 14.04

Table 2. Emotion recognition accuracy for the DEAP data set.

Target	Facial expression	EEG	First fusion	Second fusion
Valence	72.31 ± 12.02	75.38 ± 12.16	80.30 ± 11.37	80.00 ± 12.40
Arousal	71.15 ± 11.62	68.85 ± 10.02	74.23 ± 10.34	71.54 ± 11.16

Table 3. Emotion recognition accuracy for the online experiment.

Target	Facial expression	EEG	First fusion	Second fusion
Valence	55.75 ± 19.25	69.25 ± 11.96	69.75 ± 12.79	68.00 ± 17.39
Arousal	54.00 ± 20.28	64.00 ± 22.39	61.75 ± 19.76	70.00 ± 20.51

Moreover, we first performed a normality test of accuracy for each of the four detectors, and the data were considered normal when the result of the normality test was below 0.05. This test was followed by a paired t-test (normal data) or Nemenyi procedure (not normal). The statistical test results are presented in Tables 4–9. The output of the paired t-test or Nemenyi procedure was considered p. The results were considered significant when the p values were below 0.05. The statistical analysis based on the t-test and Nemenyi procedure indicated that in some cases (but not all), significant differences existed between the fusion and signal modalities. In the MAHNOB-HCI data set, there was a significant difference between the first fusion method and the EEG in the MAHNOB-HCI data set in the valence space (p = 0.005) and the second fusion method and the EEG in the MAHNOB-HCI data set in the valence space (p = 0.004). For the DEAP data set, no significant difference was observed. In the online experiment, there was a significant difference between the first fusion method and the facial expressions in the valence space (p = 0.012). Additionally, for the second fusion method versus the facial expression, there was a difference in the valence space (p = 0.026) and the arousal space (p = 0.007). Although reduced accuracy was found in the online experiment, some significant improvement was achieved through information fusion.

Table 4. The results of statistical tests in the MAHHNOB-HCI dataset in valence space.

	Face	EEG
First fusion method	0.65	<0.05
Second fusion method	0.60	<0.05

Table 5. The results of statistical tests in MAHHNOB-HCI dataset in arousal space.

	Face	EEG
First fusion method	0.16	0.30
Second fusion method	0.32	0.56

Table 6. The results of statistical tests in the DEAP dataset in valence space.

	Face	EEG
First fusion method	0.11	0.31
Second fusion method	0.13	0.37

Table 7. The results of statistical tests in the DEAP dataset in arousal space.

	Face	**EEG**
First fusion method	0.50	0.20
Second fusion method	0.93	0.54

Table 8. The results of statistical tests in the online experiment in valence space.

	Face	**EEG**
First fusion method	0.01	0.90
Second fusion method	0.03	0.76

Table 9. The results of statistical tests in the online experiment in arousal space.

	Face	**EEG**
First fusion method	0.24	0.74
Second fusion method	<0.05	0.32

In terms of the computational cost, the proposed CNN model contains 1,019,554 parameters, and it took 0.0647 s to conduct one forward analysis on a single sample. For the training, it took about 6 h and 30 min in the pre-training step. In the fine tuning step, it took 4–6 min for each subject. All the experiments were conducted with a GeForce GTX 950.

3.2. Comparison with Other Literature

In the literature, based on the MAHNOB-HCI data set, the authors of [10] developed a method of mapping face action units to high-level emotional states for facial feature extraction and used SVM-RFE to select EEG features before classifying them using a Gaussian naive Bayes classifier. Following fusion, they achieved 73.0% valence detection and 68.5% arousal detection rates. The authors of [24] proposed a novel emotion recognition method using a hierarchical Bayesian network to accommodate the generality and specificity of emotions simultaneously and achieved 56.9% and 58.2% accuracy for valence and arousal, respectively. In the literature, based on the DEAP data set, the authors of [25] addressed the single-trial binary classification of emotion dimensions (arousal, valence, dominance, and liking) using EEG signals and achieved 76.9% and 68.4% accuracy for valence and arousal, respectively. The authors of [24] also tested their model on the DEAP data set and achieved 58.0% and 63.0% accuracy for valence and arousal, respectively. Our performance for both valence and arousal space surpassed that of previous results

4. Discussion

In this study, we explored two methods for the fusion of facial expressions and EEGs for emotion recognition. Two data sets containing facial videos and EEG data were used to evaluate these methods. Additionally, an online experiment was conducted to validate the robustness of the model. In a binary classification of valence and arousal, significant results were obtained for both single modalities. Moreover, two fusion methods outperformed the single modalities. Compared with studies in the recent literature [5–7], which used single modalities to detect emotion, our major novelty is combining EEG with facial expression. In addition, in emotion experiments, deep learning methods are often difficult to apply due to the limits of the samples. Our other highlight was solving this problem by pre-training our deep learning model based on a large data set before fine tuning with the target data set. Thus, we achieved high performance in detecting emotion compared to [10], especially for the face-based method, in which [10] achieved 64.5% accuracy while we achieved 73.33% accuracy in valence space. In addition to implementing the widely used fusion method based on enumerating different weights between two models [10,11], we also explored a novel fusion method applying a boosting technique, which can reduce computational cost with the growth of the fusion modal (such as EMG).

Our performance for both valence and arousal surpassed that of previous studies who used EEG-based, face-based emotion recognition in terms of binary valence/arousal classification based on two public datasets, partly because we performed end-to-end deep leaning mapping of the face from image pixels directly to emotion states instead of trying to manually extract features from the images. CNN is widely used in image-based recognition tasks and often achieves better performance than traditional methods [26]. On the other hand, as we argued in the introduction section, non-physiological signals (videos of faces) can be easily faked. In this respect, the drawbacks of facial expression detection can be compensated for by the EEG detection to a very large extent. Thus, the facial expression detection and EEG detection were irreplaceable and complementary to each other, and the multimodal fusion should achieve higher accuracies using both detections than using one of the two detections.

In terms of the computational cost, the bottleneck of our model is the CNN, because other models (SVM and fusion strategies) are simply compared to the CNN. However, in the fusion technique, we repeatedly used the output of the CNN. We did not conduct repeated forward analysis in the CNN model. Instead, we saved the output of the CNN model as intermediate variables to reduce the computational cost. According to our results, the time cost of the CNN is acceptable.

For decision-level fusion, the method we applied for facial expression and EEG was also important. The number of samples extracted from a single object is limited. Therefore, it is challenging to train a complex model, such as a CNN, using only a small amount of training data without overfitting. Our approach to addressing this problem was based on previous studies [27–30] that have consistently shown that supervised fine tuning with a relatively small data set based on a network pre-trained with a large image data set of generic objects can lead to a significant improvement in performance. In addition, we did not obtain the results of valence and arousal separately (as in most previous studies). Instead, the results for valence and arousal were obtained in the same network. Such a multi-task learning scheme can further improve the accuracy of the classifier [31,32]. To further examine what the model learned, we also visualized the spatial pattern that maximally excited different neurons. Figure 4 displays the activations of the network during the forward pass. Noted that we selected the top nine activated neurons for each layer. For the first and second layers, the low-level features (e.g., edge, light) were detected. For the final layer, the high-level features (e.g., eyes, glasses, mouth) were detected. Our results are partially consistent with the findings of previous work [33,34]. For instance, Khorrami el al. found that when the CNN model was used to detect facial expression, the output of the final layer of the model resembles facial action units (FAUs) [33]. For EEG, we applied a wavelet transform instead of using the Fourier transform, which was used in most previous studies. Relative to the Fourier transform, the wavelet transform can analyze signals of various frequencies and time scales. To further illustrate the influence of valence and arousal in different brain regions, we projected the PSD features onto the scalp to obtain the brain patterns of the five selected frequency bands in Figure 5. Specifically, we calculated the correlation between EEG spectral power in different bands and 14 electrodes with the continuous valence and arousal labels for the subjects who achieved the highest accuracy in corresponding emotion space from the MAHNOB-HCI and DEAP dataset. The topographs in Figure 5 show that lower frequency (theta, slow alpha, alpha) components from electrodes positioned on the frontal have a higher correlation with valence measurements. In terms of the arousal space, higher frequency (beta, gamma) components from electrodes positioned on all positions have a higher correlation. Our results are partially consistent with the findings of previous work [35]. Pan et al. found that the neural patterns had significantly higher responses in low frequencies at prefrontal sites and significantly higher theta and alpha responses at parietal sites for sad emotions and valence is a measure of both positive and negative emotion [35]. The reason why two sub-classifiers were fit for boosting is two-fold. On one hand, they are two different models (CNN and SVM) and how they process data is quite different. On the other hand, the features they use (facial expression and EEG) are different. The diversity of the sub-classifiers helped improve the performance of the fusion model. For fusion, the first method (in which we enumerate the different weights for all models) has been widely used in previous studies. However, this method suffers from high computational costs as the type of the modals increases (e.g., taking eye movement as another modal). We explored the possibility of fusion using a different method (adaboost fusion), which is not limited by the computational cost. This can be validated because the time complexity of the first fusion method is O(m100n), while the time complexity of the second fusion method is O(nm), where n is the number of the modal and m is the number of the sample.

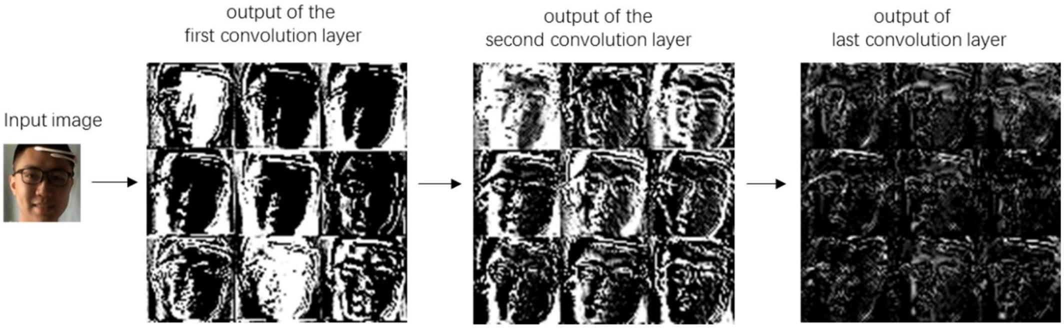
output of the first convolution layer
output of the second convolution layer
output of last convolution layer
Input image

Figure 4. Visualization of spatial patterns that activate the nine selected outputs in the three convolutional layers of our pre-trained network.

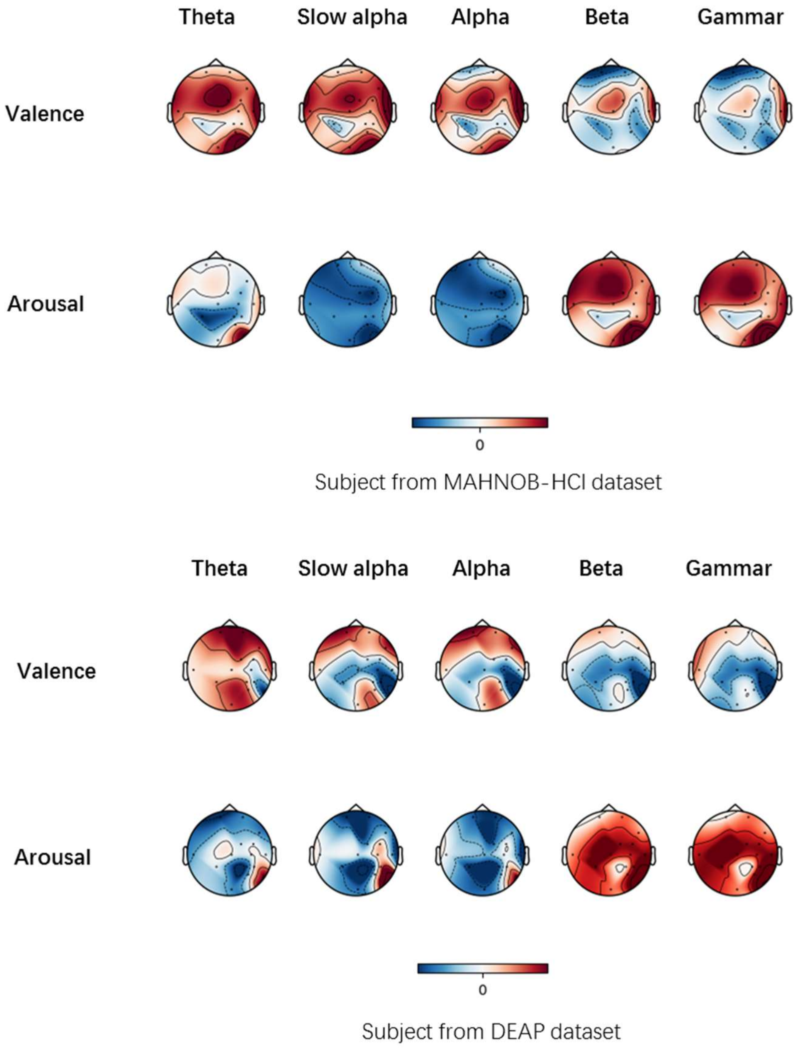

Figure 5. The correlation maps between PSD and continuous valence for theta, slow alpha, alpha, beta, and gamma bands. The correlation values are averaged over all sequences.

However, although performance was improved through information fusion, we did not find strong statistical evidence indicating that multimodal fusion offers a significant improvement over the single-modal approach (e.g., when conducting an independent two-sample t-test of the accuracy distribution, the p value is not always less than 0.05). In most cases, (e.g., [8]), high volatility is associated with facial expressions during emotion experiments because subjects are able to trick the machine if they know how to mimic certain facial expressions. In this respect, the gap between the error associated with facial expressions and the Bayesian error of true emotion detection can be filled by adding information sources (e.g., EEG) [36]. Unfortunately, in our experiment, the experimental subjects were asked to behave normally rather than mimic certain facial expressions, which may be the main reason that we could not find strong statistical evidence indicating significant improvement after fusion. In addition, for only the DEAP data set, no significant difference was found; we believe this result was observed because we only used 14 subjects, which was a limited sample size. For the other two data sets, we found significant differences between single and multiple modes.

5. Conclusions

Certain limitations of this study should be considered in the future. Currently, due to the limited time participants can use the equipment before fatigue sets in and before the effectiveness of the electrode gel is degraded, it is often challenging to obtain a large data set for EEG. Thus, the number of samples was insufficient to provide a definitive answer regarding the benefits of fusion in this study. In the future, we will attempt to either collect more EEG data or generate EEG data using a generative model for semi-supervision learning. Furthermore, identifying the additional benefits of information fusion for multimodal emotion recognition is an interesting topic that we intend to investigate. Finally, applying emotion recognition technology to special populations, such as infants or people who suffer from disorders of consciousness, is also an interesting area we can work on.

Author Contributions: Conceptualization, J.P. and Y.H.; methodology, J.P., Y.H. and J.Y.; software, Y.H. and J.Y.; validation, Y.H. and J.Y.; formal analysis, J.P., Y.H. and J.Y.; investigation, J.P., Y.H.; resources, J.P.; data curation, J.P., Y.H. and J.Y.; writing—original draft preparation, J.P., Y.H. and S.L.; writing—review and editing, J.P., Y.H. and S.L.; visualization, J.P., Y.H. and S.L.; supervision, J.P., Y.H. and S.L.; project administration, J.P.; funding acquisition, J.P.

Funding: This research was funded by The National Natural Science Foundation of China under Grants 61876067 and 61503143, the Guangdong Natural Science Foundation under Grant 2014A030310244, the Pearl River S&T Nova Program of Guangzhou under Grant 201710010038 and the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) under Grant 2017pdjhb0134.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gratch, J.; Marsella, S. Evaluating a computational model of emotion. *Auton. Agents Multi-Agent Syst.* **2005**, *11*, 23–43. [CrossRef]
2. Scherer, K.R. What are emotions? And how can they be measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [CrossRef]
3. Gunes, H.; Schuller, B.; Pantic, M.; Cowie, R. Emotion representation, analysis and synthesis in continuous space: A survey. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops IEEE, Santa Barbara, CA, USA, 21–25 March 2011; pp. 827–834.
4. Russell, J.A.; Mehrabian, A. Evidence for a three-factor theory of emotions. *J. Res. Personal.* **1977**, *11*, 273–294. [CrossRef]
5. Dhall, A.; Goecke, R.; Ghosh, S.; Joshi, J.; Hoey, J.; Gedeon, T. From individual to group-level emotion recognition: EmotiW 5.0. In Proceedings of the ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 524–528.
6. Patel, P.; Chaudhari, A.; Kale, R.; Pund, M. Emotion recognition from speech with gaussian mixture models & via boosted gmm. *Int. J. Res. Sci. Eng.* **2017**, *3*, 47–53.

7. Zheng, W.-L.; Zhu, J.-Y.; Lu, B.-L. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* **2017**, *1*. [CrossRef]
8. Zhang, Y.; Zhang, S.; Ji, X. EEG-based classification of emotions using empirical mode decomposition and autoregressive model. *Multimed. Tools Appl.* **2018**, *77*, 26697–26710. [CrossRef]
9. Xie, J.; Xu, X.; Shu, L. WT Feature Based Emotion Recognition from Multi-channel Physiological Signals with Decision Fusion. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–6.
10. Koelstra, S.; Patras, I. Fusion of facial expressions and EEG for implicit affective tagging. *Image Vis. Comput.* **2013**, *31*, 164–174. [CrossRef]
11. Soleymani, M.; Asghariesfeden, S.; Pantic, M.; Fu, Y. Continuous emotion detection using EEG signals and facial expressions. In Proceedings of the IEEE International Conference on Multimedia and Expo, Chengdu, China, 14–18 July 2014; pp. 1–6.
12. Huang, Y.; Yang, J.; Liao, P.; Pan, J. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 2107451. [CrossRef] [PubMed]
13. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]
14. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
15. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
16. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
17. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems Conference (NIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
19. Bhatnagar, G.; Wu, Q.M.J.; Raman, B. A new fractional random wavelet transform for fingerprint security. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2012**, *42*, 262–275. [CrossRef]
20. Verma, G.K.; Tiwary, U.S. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *Neuroimage* **2014**, *102*, 162–172. [CrossRef] [PubMed]
21. Duan, K.B.; Rajapakse, J.C.; Wang, H.; Azuaje, F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. NanoBiosci.* **2005**, *4*, 228–234. [CrossRef]
22. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 148–156.
23. Ponti, M.P., Jr. Combining classifiers: From the creation of ensembles to the decision fusion. In Proceedings of the 2011 24th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Alagoas, Brazil, 28–30 August 2011; pp. 1–10.
24. Gao, Z.; Wang, S. Emotion recognition from EEG signals using hierarchical bayesian network with privileged information. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; ACM: New York, NY, USA, 2015; pp. 579–582.
25. Rozgić, V.; Vitaladevuni, S.N.; Prasad, R. Robust EEG emotion classification using segment level decision fusion. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 1286–1290.
26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
27. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.

28. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In Proceedings of the International Conference on Machine Learning, Beijing, China, 16–21 June 2014; pp. 647–655.
29. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
30. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
31. Zhang, C.; Zhang, Z. Improving multiview face detection with multi-task deep convolutional neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Steamboat Springs, CO, USA, 24–26 March 2014; pp. 1036–1041.
32. Ranjan, R.; Patel, V.M.; Chellappa, R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 121–135. [CrossRef] [PubMed]
33. Khorrami, P.; Paine, T.; Huang, T. Do deep neural networks learn facial action units when doing expression recognition? In Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPR), Santiago, Chile, 7–13 December 2015; pp. 19–27.
34. Yosinski, J.; Clune, J.; Fuchs, T.; Lipson, H. Understanding neural networks through deep visualization. In Proceedings of the International Conference on Machine Learning (ICML) Workshop on Deep Learning, Lille, France, 6–11 July 2015.
35. Pan, J.; Xie, Q.; Huang, H.; He, Y.; Sun, Y.; Yu, R.; Li, Y. Emotion-Related Consciousness Detection in Patients with Disorders of Consciousness through an EEG-Based BCI System. *Front. Hum. Neurosci.* **2018**, *12*, 198. [CrossRef] [PubMed]
36. Wellendorff, J.; Lundgaard, K.T.; Møgelhøj, A.; Petzold, V.; Landis, D.D.; Nørskov, J.K.; Bligaard, T.; Jacobsen, K.W. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **2012**, *85*, 23. [CrossRef]