IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2020, accepted September 7, 2020, date of publication September 28, 2020,
date of current version December 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027026

Automatic Emotion Recognition Using Temporal Multimodal Deep Learning

BAHAREH NAKISA[1], MOHAMMAD NAIM RASTGOO[2,3], ANDRY RAKOTONIRAINY[3], (Member, IEEE), FREDERIC MAIRE[2], (Member, IEEE), AND VINOD CHANDRAN[2]

[1]School of Information Technology, Deakin University, Geelong, VIC 3217, Australia
[2]School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia
[3]Centre for Accident Research and Road Safety-Queensland, Queensland University of Technology, Brisbane, QLD 4000, Australia

Corresponding author: Bahareh Nakisa (bahar.nakisa@deakin.edu.au)

The work of Frederic Maire was supported by the Queensland University of Technology (QUT) through the Centre for Robotics.

ABSTRACT Emotion recognition using miniaturised wearable physiological sensors has emerged as a revolutionary technology in various applications. However, detecting emotions using the fusion of multiple physiological signals remains a complex and challenging task. When fusing physiological signals, it is essential to consider the ability of different fusion approaches to capture the emotional information contained within and across modalities. Moreover, since physiological signals consist of time-series data, it becomes imperative to consider their temporal structures in the fusion process. In this study, we propose a temporal multimodal fusion approach with a deep learning model to capture the non-linear emotional correlation within and across electroencephalography (EEG) and blood volume pulse (BVP) signals and to improve the performance of emotion classification. The performance of the proposed model is evaluated using two different fusion approaches – early fusion and late fusion. Specifically, we use a convolutional neural network (ConvNet) long short-term memory (LSTM) model to fuse the EEG and BVP signals to jointly learn and explore the highly correlated representation of emotions across modalities, after learning each modality with a single deep network. The performance of the temporal multimodal deep learning model is validated on our dataset collected from smart wearable sensors and is also compared with results of recent studies. The experimental results show that the temporal multimodal deep learning models, based on early and late fusion approaches, successfully classified human emotions into one of four quadrants of dimensional emotions with an accuracy of 71.61% and 70.17%, respectively.

INDEX TERMS Emotion recognition, electroencephalography, blood volume pulse, convolutional neural network, long short-term memory, temporal multimodal fusion.

I. INTRODUCTION

Automated emotion recognition using lightweight body sensors and advanced machine learning technologies has been used in different application domains such as computer games [1], e-health [2], [3] and road safety [4]. Lightweight wireless sensors in headbands and smart watches can be used by individuals as they carry on their daily life activities. These sensors can record physiological signals like blood volume pulse (BVP), electroencephalograms (EEG), skin temperature and skin conductance in a minimally invasive manner. Among the various physiological signals available, EEG and BVP have been found to be useful in inferring emotional states. A strong correlation has been observed between such physiological signals and basic emotions like sadness, anger, surprise etc. [5], [6].

In recent research, multimodal data are utilised to improve the performance of emotion classification [7]–[9]. Data from multiple sources are correlated and can provide complementary emotion-related information. To capture such information, it is important to capture the correlation between modalities with a compact set of latent variables. However, learning the latent emotion information in heterogeneous physiological data like EEG and BVP signals is a challenging

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

problem. This is because EEG and BVP signals are comprised of heterogeneous time-series data and there are some emotion structures within and across modalities *over time*.

Several approaches have been presented to address the multimodal fusion problem. Early fusion refers to feature concatenation in early integration. As this sort of approach addresses the classification problem based on the extracted features from each modality separately, it is not able to learn patterns that exist across multiple data modalities. Therefore, the early fusion approach is not able to capture the non-linear correlation across modalities. This is due to the fact that the correlation between features within each modality is stronger than the cross correlation [10].

Therefore, to build a robust emotion recognition system using multimodal physiological signals, it is essential to propose a multimodal fusion model that can capture and learn the inherent changes within each modality as well as across modalities. We believe that a good fusion model based on multimodal data should be able to simultaneously learn a joint representation of multimodal data, including temporal structure within each modality.

Recently, deep learning techniques and architecture are becoming well-known in capturing non-linear correlation across multimodal data such as audio-visual [11] and physiological signals [12], [13] and have obtained state-of-the-art performance [10], [14]. The proposed multimodal fusion methods are able to jointly learn the highly correlated representation across modalities. Physiological signals are inherently temporal in nature, which means that the current pattern in the signal is influenced by the previous ones. However, the multimodal networks like deep Autoencoder, and the Boltzmann Machine do not model the temporal multimodal fusion.

To address these challenges, we employ temporal deep learning models with the aim to improve the performance of emotion classification based on the fusion of EEG and BVP signals from lightweight sensors.

Fig. 1 shows a simple illustration of the temporal multimodal fusion model. The raw EEG and BVP signals are segmented into consecutive windows. In each window (time slice), the EEG signals and BVP signals are jointly learned using the proposed networks. The learned joint representations across modalities in different windows are directly connected from start to end, which makes the current window learn using the previous window.

To build an automatic emotion recognition model based on the conventional models, first features are extracted from physiological signals. The features are concatenated and then the generated multimodal feature set is passed into a classifier to determine the emotional states. However, our system is trained in an end-to-end fashion. Using end-to-end learning, the constructed features using ConvNet are trained jointly with the classification step as a single network. Moreover, in an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction.

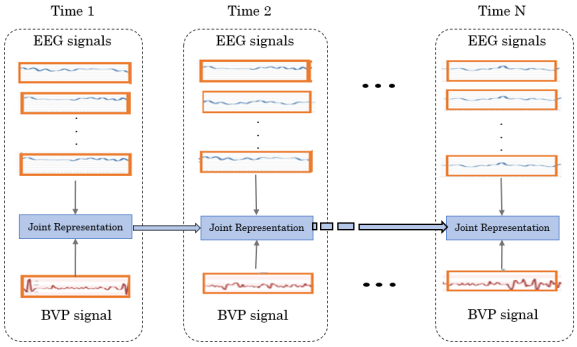

FIGURE 1. The proposed model demonstrates temporal multimodal fusion. The EEG channels and BVP signal are segmented into windows. The sequence of windows from each channel is fed into a deep learning network and the output forms the joint representation across modalities. The generated joint representation based on the current window depends on the previous windows.

This is the first emotion recognition work that fuses EEG and BVP signals from lightweight sensors using an end-to-end temporal multimodal fusion model. The temporal multimodal fusion models based on convolutional neural networks (ConvNet LSTM networks) can fuse EEG and BVP signals temporally to capture the temporal structures of emotions within and across the modalities. Two types of temporal multimodal fusion methods, early and late fusion, are investigated in this study and compared with other recent methods.

In this study, emotional states based on dimensions of arousal and valence are categorised into four quadrants (Fig. 2): HA-P which is High Arousal-Positive emotions; LA-P which is Low Arousal-Positive emotions; HA-N which is High Arousal-Negative emotions; and LA-N which is Low Arousal-Negative emotions.

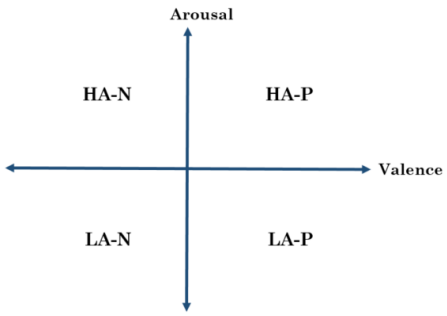

FIGURE 2. Emotions categorised into four quadrants.

In summary, the contributions of the proposed framework are as follows:

- We compare two temporal multimodal deep learning models based on early and late fusion approaches using a ConvNet LSTM model with end-to-end learning. The goal of these two emotion classification models is to improve the performance by obtaining temporal, emotion-related information from EEG and BVP signals.

- We evaluate the performance of two temporal multimodal fusion models using different window sizes and a sliding window strategy. We compare the proposed models with non-temporal multimodal deep learning models based on a trial-wise strategy. In trial-wise training, the entire duration of a raw physiological signal per video clip, called a trial, is used as input and the corresponding trial emotion label is used as the target for training.
- We demonstrate that temporal multimodal fusion models can outperform, in regard to accuracy, the handcrafted features extraction method to classify emotions into four quadrants from a dataset collected from lightweight physiological sensors, namely Empatica (E4) wrist bands and Emotiv headsets.

The paper is organised as follows: Section 2 provides a theoretical background and a review of the related works. The proposed methods are presented in Section 3. In Section 4, we evaluate the performance of our systems for emotion recognition based on our dataset collected using wearable sensors.

II. BACKGROUND AND RELATED WORKS

Human emotion recognition based on physiological signals is becoming more popular as a research topic [15]. Two major components in our body are responsible for any changes related to inner emotion: the Autonomic Nervous System (ANS) and the Central Nervous System (CNS). Inner emotional states can affect the body's physiological signals such as EEG and BVP signals which originate from these two components [5], [6]. As EEG signals come directly from the CNS, these signals can strongly capture emotional states. It has been shown that emotion classification has often improved when EEG signals are combined with different modalities [15]–[17]. One of the best indicators of different emotions is the BVP signal [5], [18], [19]. The BVP signal indicates the blood flow rate controlled by heart pumping activity and is regulated by the ANS. External stimuli and emotional states modulate the activity of the ANS. The BVP signal is measured using a photoplethysmography (PPG) sensor. Although the accuracy of BVP is lower than that of electrocardiograms (ECGs), due to its simplicity BVP is widely used in biosensors developed for applications like office workers' mental workload prediction [20].

A. EMOTION CLASSIFICATION FRAMEWORK

To build an automatic emotion recognition system, three main steps should be considered: pre-processing, feature extraction and emotion classification. In the pre-processing step, the raw physiological signals are prepared for data modelling. In this step, the noise and artefacts are removed to form purer signals. In the next step, a set of features of the denoised signals are extracted. Then, a classifier is applied to classify different emotions.

One of the most challenging steps in the pipeline of automatic emotion classification is feature extraction. There are two main approaches for feature extraction: handcrafted feature extraction methods and deep learning techniques. To date, most of the reported approaches to recognise different emotions rely on extracting handcrafted features. This process is accomplished either by taking advantage of human expert knowledge or using a conventional feature extraction algorithm.

Several useful EEG features from the time and frequency domains are proposed and used to recognise different emotional states. In our recent study [21], we reviewed a comprehensive set of extractable features from EEG signals, and found the best salient subset of features and channels using different evolutionary algorithms. There are some studies that focus on extracting features from the time and frequency domains from the BVP signal [22]–[24]. Some of the time-domain features such as standard deviation, mean, and variance from peak have been used in recognising different emotions. It has been shown that the power spectrum density from three sub-frequencies: VLF (0–0.04 Hz), LF (0.05–0.15 Hz) and HF (0.16–0.4 Hz), and the ratio of LF/HF can accurately distinguish different emotions.

It should be noted that the performance of the emotion recognition model significantly depends on the quality of the extracted features. As a result, it is always desirable to extract the most relevant and critical features. However, extracting salient features needs expert knowledge which is time-consuming. Moreover, extracting features from different physiological signals is challenging as the extracted features are not always robust to variations like noise and signal resolution.

Recently, deep learning (DL) methods have increasingly emerged to solve challenging problems. DL methods have strong capabilities in constructing reliable features in different domains like speech recognition [25] and time-series data analysis [26], [27]. It has been shown that DL techniques are more reliable for effective modelling compared to the popular feature extraction-based methods [11], [28], [29]. One of the DL methods, which has been successfully used for automatic feature extraction, is a convolutional neural network (ConvNet). As a ConvNet has a strong capability for learning features, it is suitable for multidimensional signal processing applications. By using its convolution component, a ConvNet can learn local patterns in data. It firstly extracts local, low-level features from the raw input, and then increasingly extracts more global and high level features in deeper layers. Some studies have applied a ConvNet with a different number of layers to physiological signals to classify different emotions [28], [30]–[32]. However, in these studies, they extracted some features from raw physiological signals, and then applied ConvNet techniques to extract higher-level features and classify different emotions.

Another advanced technique that has achieved high accuracy is LSTM-based emotion recognition. This technique has been applied to EEG signals to recognise emotions in three dimensions: arousal, valence and liking. Another study applied a stacked autoencoder to extract better EEG

features [33]. It should be noted that some features such as power spectral density were extracted from raw 32-channel EEG signals. PCA techniques were applied to reduce the high dimensionality and improve the performance of the model. Although different deep learning techniques have been applied and have achieved good performance, these techniques have been applied to full-scale EEG signals with 32 channels.

B. MULTIMODAL DATA FUSION FOR EMOTION CLASSIFICATION

Using multimodal input has improved the accuracy of emotion recognition compared to using inputs of a single modality. This is because the multimodal data provide additional information which results in higher accuracy of the overall result or decision. Generally, there are two types of fusion for different modalities: early fusion and late fusion. In early fusion, different features are first extracted from each modality, then all features are concatenated to construct a joint feature vector. The joint feature vector is then used to build an effective classifier. Using the early fusion approach, we are able to identify the correlated features that improve recognition accuracy. However, with this approach there is little control over the contribution of each feature set from each modality on the final result. Moreover, the joint feature space from different modalities can result in high dimensionality and a more difficult classifier design. Therefore, large training sets are typically required for model training.

Furthermore, the derived features from the various modalities are different in many aspects like sampling rate. Therefore, multimodal learning at this level is sometimes difficult. Some studies have shown early fusion can improve the performance of emotion recognition based on different modalities [7], [34], [35].

In contrast, in a late fusion approach, the feature set of each modality is examined and classified independently, then the results from each modality are fused into a decision vector. The benefit of using late fusion compared to early fusion is that it is easier to combine asynchronous data. Another advantage of this fusion model is that every modality utilises its best classifier which is suitable for the task. This may help to increase the performance of the model. Late fusion is most commonly used with a the combination of gestures and speech [36]. However, it is almost certainly incorrect to use late fusion in real-time approaches. Using late fusion method, each modality is treated independently, and then combine their results at the end. In a real-time environment, people produce audio, video and tactile interactive signals containing both complementary and redundant information. Both early and late fusion approaches have a weakness in capturing the non-linear correlation across modalities [10].

Deep learning has become the most effective method used in fusing different modalities in different domains, particularly in audio-visual speech recognition [10], [37] and affective computing [11], [37]–[39]. For example, two Deep Belief Networks (DBNs) with the multimodal Restricted Boltzmann machine are combined [37], where each DBN is used to train one modality. The proposed multimodal DBN has surpassed the multi-stream HMM models. The Multimodal Restricted Boltzmann machine has shown to be effective in fusing both audio and visual modalities to generate a joint representation. However, in the proposed deep learning networks, the temporal information was not considered, which apparently deviates from the natural properties of time-series data.

From the literature review on emotion classification using physiological signals, most of the proposed solutions found are not based on an end-to-end method. It means that some features are first extracted from the physiological signals and then deep learning techniques are applied. Moreover, the advanced fusion techniques based on deep learning approaches are not well explored in this domain.

In this study, we propose multimodal deep learning models based on early and late fusion using an end-to-end learning approach. We investigate the performance of the models on data comprised of the fusion of EEG signals with only 5 channels and BVP signals captured via lightweight sensors. This investigation can help build an application to be used in real-time situations.

III. MODELS

This section presents the proposed temporal multimodal (EEG and BVP signals) fusion with deep learning models to capture the temporal emotion structures within and across the modalities. The proposed deep learning models are based on end-to-end ConvNet LSTM networks and two different fusion approaches: early and late fusion. Using an early fusion model, the raw EEG and BVP data are fed into a ConvNet network to extract features, and then all the generated features are concatenated to form a joint feature vector. The created joint feature vector is fed into the LSTM network followed by a dense and soft-max layer for emotion classification.

Using a late fusion model, the raw EEG and BVP data are fed into a ConvNet followed by the LSTM network and dense layer. The generated features from each network are combined and then fed into a dense and softmax layer for emotion classification.

These models were evaluated on the dataset collected using wearable physiological sensors (Empatica E4 and Emotiv Insight). The Empatica E4 and Emotiv capture BVP and EEG signals, respectively.

The dataset is described in Section *A* and the data preparation for the temporal multimodal deep learning is described in Section *B*. We then present two new frameworks for temporal multimodal deep learning based on early and late fusion (Sections *C* and *D*). Finally, the ConvNet architecture designed for this study is presented in Section *E*.

A. DESCRIPTION OF DATASET

We investigated the performance of the temporal multimodal model on a dataset collected from 20 subjects, aged between 20 and 38. Each participant watched nine video clips used

in the MAHNOB dataset [40] to induce different emotions. These video clips contain movie scenes selected from popular movies such as Gangs of New York, Earworm, and The Pianist and the videos are annotated by psychology experts. The videos were selected based on the highest number of tags in different emotion classes. For example, the video clips with the highest number of happiness tags were selected to induce happiness. While the participants watched the video clips their brain activity (EEG) and heart activity (BVP) was captured using lightweight sensors. We used Emotiv insight and Empatica E4 to capture EEG and BVP signals, respectively (see Fig. 3). The Emotiv insight contains only 5 channels (AF3, AF4, T7, T8, and Pz) with 2 reference channels. The channels are located based on the international 10-20 system (see Fig. 4).

(a) (b)

FIGURE 3. (a) The Emotiv Insight headset (link), (b) the Empatica wristband (link).

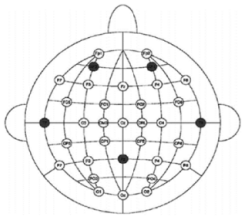

FIGURE 4. The location of the five channels of the Emotiv sensor are indicated by black dots. The nose is placed at the centre front, with an ear on each side.

To acquire the raw BVP and EEG signals, Empatica Connect and TestBench software was used. After watching each video clip, each participant was asked to express their emotional state (anger, happiness, disgust, surprise, neutral, anxiety, amusement, sadness and fear) using a keyboard.

In this study, the presented emotional states were mapped into four quadrants of dimensional emotions. In the first step, the participants were asked to close their eyes and relax for about one minute while their baseline EEG and BVP signals were recorded with the least amount of ocular noise. One minute silence was allowed between each video clip to help to prevent mixing up the current emotion with the previous emotion. Fig. 5 shows the experimental protocol.

The collected signals were analysed manually to ensure data quality. Noisy and low quality EEG signals from the lightweight Emotiv sensors were ignored. The generated noise in the EEG signals may have been the result of shifting electrodes or a loose contact. After removing noisy data,

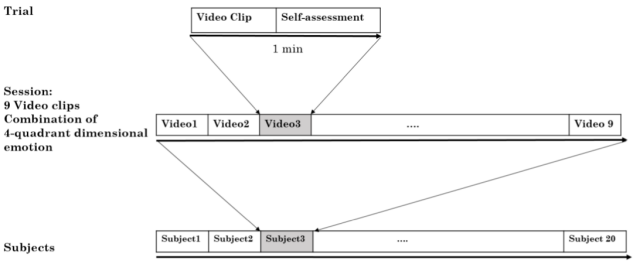

FIGURE 5. The experimental protocol for emotion elicitations with 20 participants. Each participant watched nine video clips and was asked to report their emotions after each clip (via self-assessment).

EEG signals from 17 out of the 20 participants, nine females and eight males, were used in this study. The expected benefit of these sensors is due to their lightweight, and wireless nature, making them possibly the most suitable for free-living studies in natural settings.

B. DATA PREPARATION FOR THE PROPOSED MODELS

There were nine trials for each participant as each participant watched nine video clips. Each trial was labelled with a different emotion class. Six channels of signals were recorded for each trial: five EEG channels and one BVP channel. To prepare the data for temporal multimodal learning, a sliding window strategy was used on each channel per each trial. We applied a sliding window and created a set of successive fixed-size windows with a fixed degree of overlap. Let us denote the 6-channel inputs as sequences of length T, namely

$$
\begin{aligned}
EEG\_ch_1 &= (ch_1^1, \ldots, ch_1^{t-1}, ch_1^t, \ldots, ch_1^T),\\
EEG\_ch_2 &= (ch_2^2, \ldots, ch_2^{t-1}, ch_2^t, \ldots, ch_2^T),\\
&\ldots\\
EEG\_ch_5 &= (ch_5^1, \ldots, ch_5^{t-1}, ch_5^t, \ldots, ch_2^T),\\
BVP &= (BVP^1, \ldots, BVP^{t-1}, BVP^t, \ldots, BVP^T)
\end{aligned}
$$

where $\mathrm{ch}_1^t, \ldots, \mathrm{ch}_5^t$ and BVP^t denote the window of $\mathrm{EEG\_ch}_1, \ldots, \mathrm{EEG}_{\mathrm{ch}_5}$ and BVP at time slice t.

All of the generated windows are considered to be the new training data examples with the same labels as their original trials. We then segmented each channel into consecutive windows with different window sizes (2 sec, 3 sec, 5 sec and 10 sec) and 50% overlap. Pre-processing techniques such as band-pass filtering (6^{th} order Butterworth filtering), Notch filtering and ICA were applied to the EEG signals. To remove noise and artefacts from the BVP signals, a 3 Hz low-pass Butterworth filter was applied. In addition, we normalised our data with a zero mean and unit variance.

C. TEMPORAL MULTIMODAL DEEP LEARNING BASED ON EARLY FUSION

This section presents the proposed temporal multimodal deep learning model with an early fusion approach. In the proposed model, the temporal physiological signals were fused into joint representation sequences at an early stage (after ConvNet). Fig. 6 depicts the architecture of the proposed model.

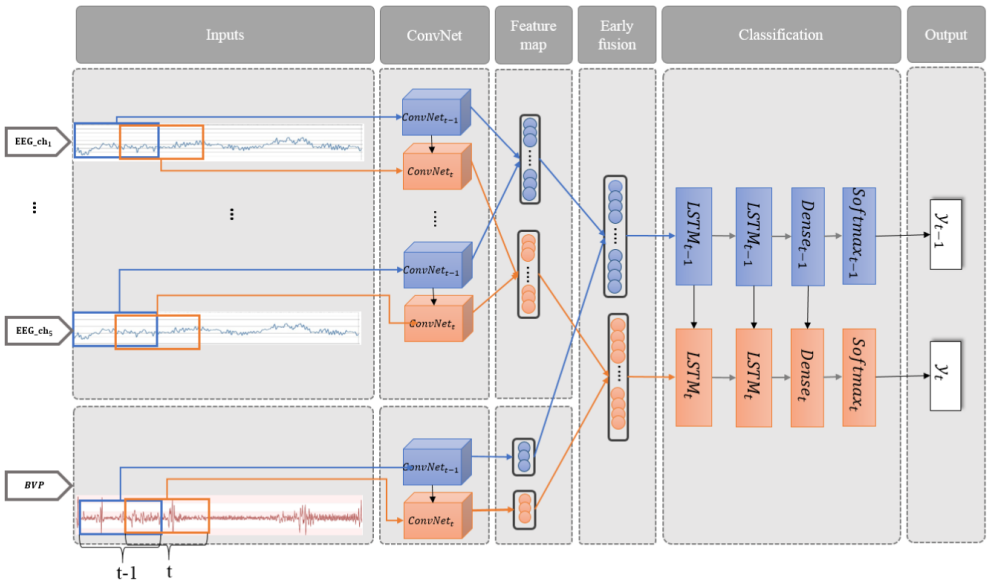

FIGURE 6. The overall end-to-end pipeline of the temporal multimodal deep learning model based on early fusion using ConvNet LSTM. The inputs to this system are the EEG (five EEG channels) and BVP (one channel) signals. The inputs are segmented into successive fixed-size windows with some degree of overlap (50% overlap). The output of this model is one of four dimensional emotions (HA-P, HA-N, LA-P and LA-N). The created window at time *t* from each of the six channel is fed into a two-block ConvNet to extract feature maps. The output feature maps at time *t* from each channel are concatenated to build the joint representation. The created joint representation at time slice *t* is then fed into the two layers of LSTM followed by a dense layer and a soft-max layer for emotion classification.

This model consists of four layers: input layer, ConvNet, feature map, early fusion and classifier.

Input. The temporal multimodal deep learning model strongly depends on its inputs. To apply the temporal multimodal learning, the sliding window strategy was applied to each of the EEG and BVP channels. The window at time t from each channel is considered as an input to be fed into the ConvNet for training.

ConvNet. For each channel, the input, the window at time t from each channel, was fed into the 2-block ConvNet feature extractor. The hierarchical features through convolution, activation, normalisation and max-pooling layers were then learned. Since in this study physiological signals are used, we applied a 1D convolution layer.

There are more details about the ConvNet architecture in Section 3.5. Based on this architecture, the window at time t from each EEG and BVP signal was individually fed into the ConvNet architecture. The output of the ConvNet from each channel at time t was the corresponding feature map.

Feature map. If $ConvNet^{t}_{ch1}$ denotes the ConvNet for EEG_ch$_1$ and if FM^{t}_{ch1} denotes the corresponding feature maps at window t, then:

$$FM^{t}_{ch1} = ConvNet^{t}_{ch1}\left(\mathrm{ch}^{\mathrm{t}}_{1}\right)$$

To achieve temporal ConvNet learning for each channel, both the current input and its history are considered. To obtain the feature maps representation, the recent per-modality history ($ConvNet^{t-1}$) at window t is appended to the current window.

The prepared feature maps at time t from each EEG channel are concatenated to form an EEG joint representative feature map at time t.

Early fusion. In this layer, at each time step (t) the EEG joint representation feature map and BVP feature map are concatenated to build a single feature map vector.

Classification. In this layer, two layers of LSTM network followed by a dense and softmax layer were used to model the overall temporal dynamics of the multimodal feature representation at time *t*. It should be noted that an LSTM network can help in learning the temporal emotion structures, because the LSTM network consists of hidden states or memory which helps in storing the previous information (hidden layers) and learning the temporal emotion structures.

Therefore, the output of the LSTM at time t depends on the preceding hidden states (*t-1*) as well as the current state, which can capture the temporal aspect of the previous joint representations as well. It should be emphasised that the proposed model is able to learn the temporal pattern from each channel separately as well as the temporal patterns across modalities using the joint representations.

D. TEMPORAL MULTIMODAL DEEP LEARNING BASED ON LATE FUSION

In this section, a temporal multimodal deep learning model based on late fusion is presented. In the proposed architecture, the EEG channels and BVP signal are temporally fused based on a late fusion approach. The architecture of this model consists of input, ConvNet, feature map, LSTM networks, late fusion and output layers (see Fig. 7). The input ConvNet layers in this architecture are the same as for the temporal multimodal learning model based on early fusion.

Input. First the windows from each 6 channels (5 EEG channels and 1 BVP signal) at time t are fed into the ConvNet.

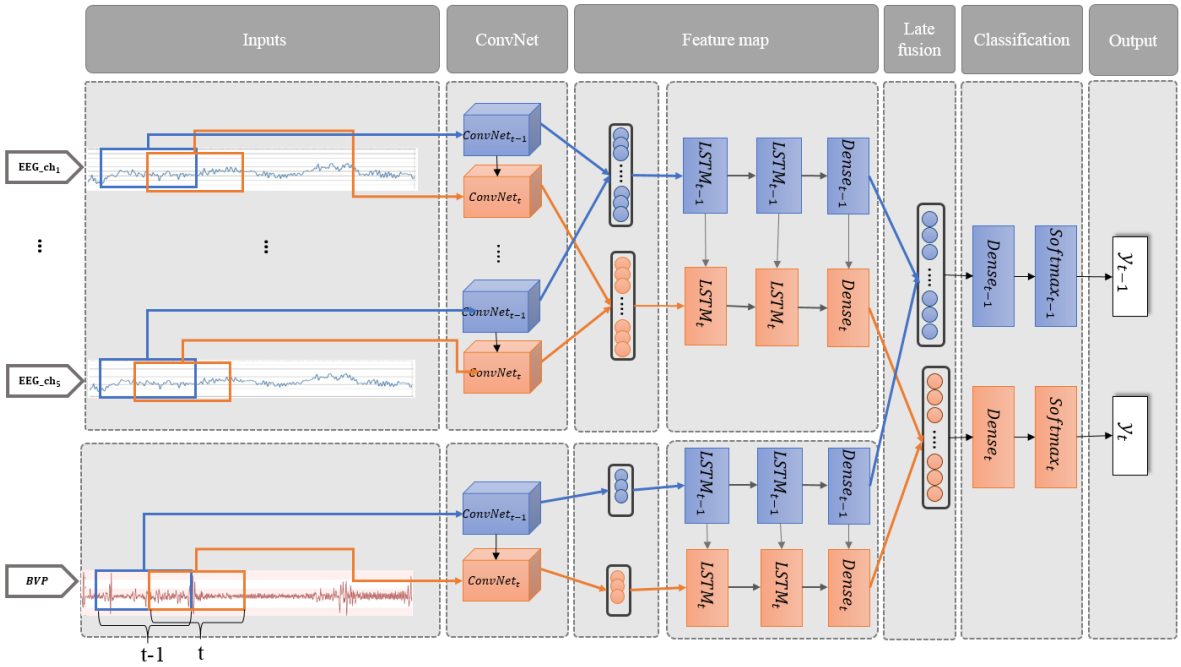

FIGURE 7. The overall end-to-end pipeline of the temporal deep learning model based on late fusion using ConvNet LSTM networks. The inputs to this system are the EEG (5 EEG channels) and BVP (one channel) signals. These signals are divided into successive fixed-size windows with a degree of overlap (50% overlap). The output of this model is one of four dimensional emotions (HA-P, HA-N, LA-P and LA-N). The generated window at time *t* from each 6-channel are passed into individual two-block ConvNet to extract feature maps. The output of feature maps from EEG channels over the window *t* are combined to build the joint representation. The created joint respresetation at time slice *t* from each modality (EEG and BVP) are fed into a two layers of LSTM and a dense layer. The output of dense layer at time *t* from two modalities are combined to create a joint representative and then is fed into a dense layer followed by a Softmax layer for emotion classification.

ConvNet. For each channel, the input, the sliced window from each channel at time t, is fed into the 2-block feature extractor. The architecture of the ConvNet in this architecture is the same as early fusion.

Feature map. The feature map of each channel is generated by a ConvNet. The feature map of each EEG channel is concatenated to form a joint representative feature map for EEG modality. The feature maps from each modality are fed into two layers of LSTM networks followed by a dense layer.

In the *late fusion layer*, the higher level feature maps generated from the two-layer LSTM and a dense layer from each modality at time t are combined to build a joint representative layer.

Classification. To classify different emotions, the joint representative layer at time t is fed into a dense layer and a softmax layer.

E. ConvNet ARCHITECTURE FOR THE RAW PHYSIOLOGICAL SIGNALS

Using a ConvNet, the local non-linear features are firstly learned, then the higher-level features are generated from the lower-level features. The ConvNet consists of convolutional layers, which can produce lower-level features using a set of learnable filters, and multiple layers of processing, which can represent the higher-level features. In addition, many ConvNet networks use a pooling layer to control overfitting. A pooling layer reduces the number of parameters in the network and the spatial size of representation which can help to avoid overfitting.

TABLE 1. Two-blocks of ConvNets architecture.

ConvNet
Convolutional Layer: Filter=20,kernel size=(10,1), stride=2
Exponential Linear Units(ELU): Alpha=0.1
Batch Normalization+ Dropout (0.15)
Max-Pooling: Pool-size=(2,1), stride=2
Convolutional Layer: Filter=20,kernel size=(10,1), stride=2
Exponential Linear Units(ELU): Alpha=0.1
Batch Normalization+ Dropout (0.15)
Max-pooling: Pool-size=(2, 1), stride=2

In this study, we proposed a ConvNet network consisting of a two-block convolutional max-pooling layer (see Table 1.). A convolutional layer, an Exponential Linear Unit (ELU), a batch normalisation layer and a max-pooling layer forms each block of the ConvNet. In the convolution layer, the current input/window at time t or the outputs of the previous layer with the set of filters (K) are convolved to be learned. This layer is able to capture the temporal information using trainable filters. The output of each filter is computed according to

$$y = frame^t * K + b$$

where b is the bias term, and $*$ is the convolution operator. The activation function used in the proposed method is Exponential Linear Units (ELU) that maps the output of the

previous layer by the following function:

$$ELU\ (x) = alpha * (\exp\ (x) - 1)\ x < 0, ELU\ (x) = xx \geq 0.$$

There is also a batch normalisation layer that normalises the output of the previous feature maps. We used a max pooling layer, to reduce the number of parameters. In fact, it downsamples the input by taking maximum feature maps over the defined window (local neighbourhood). Table 1 presents the two-blocks of ConvNet architecture.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed models, we used the dataset comprised of physiological signals (EEG and BVP signals) collected using wearable sensors to analyse human affective states. We evaluated the temporal multimodal deep learning models (early and late fusion approaches) and compared them with multimodal learning models based on a trial-wise strategy and handcrafted feature extraction methods. Our experimental results show that the proposed temporal multimodal learning models are effective in building an automatic human emotion recognition system using EEG and BVP signals in an end-to-end manner.

To evaluate the performance of the two proposed models, first the efficacy of the sliding window strategy was investigated. The performance of the proposed models based on early and late fusion approaches with different window sizes (Section 4.1) was evaluated and compared with each other. Moreover, these two models were also evaluated based on a non-temporal strategy, where the input layer is based on a trial-wise strategy instead of a sliding window strategy.

In Section 4.2, the confusion matrices of the temporal multimodal deep learning models most successful in classifying human emotions into four quadrant dimensional emotions are presented. Lastly, in Section 4.3, the best average performance of both of the ConvNet LSTM models are compared with models that use the conventional handcrafted feature extraction method.

A. EXPERIMENTAL SETUP

An extensive experiment was conducted to determine if the proposed temporal multimodal deep learning models based on early and late fusion can be used as an effective fusion method for automatic emotion classification using BVP and EEG signals. We focused on four classes of dimensional emotions (LA-P, HA-P, LA-N and HA-N), used a subject-independent approach and applied the leave-one-subject-out cross validation (LOSO) method 17 times. This means that the video clips for one subject were used for testing and the video clips for the remaining subjects were used for training. The proposed models were learned using the training dataset and then evaluated using the test dataset. This process was repeated 17 times until all the participants data were used as the test dataset.

To prepare the input data, the EEG and BVP signals were divided into successive fixed size windows with a fixed 50% overlap. Different window sizes were selected to evaluate the performance of the proposed temporal models. Before the segmentation, some noise reduction techniques such as Butterworth, Notch filtering and ICA were applied. The proposed multimodal learning models (early and late fusion) based on temporal and non-temporal approaches use two-layer LSTM networks followed by a dense layer with 100 and 20 hidden states, respectively. To train our models, learning batches of 10 sequences were used. Early stopping was also set for the validation process. This is the configuration which resulted in minimum loss and highest accuracy. The training was performed for 300 iterations.

B. COMPARISON OF TEMPORAL AND NON-TEMPORAL MODELS BASED ON EARLY AND LATE FUSION

In this section, the effectiveness and performance of the two temporal multimodal deep learning models (early and late fusion) are evaluated based on different window sizes and compared with the non-temporal multimodal learning models. The aim of this comparison is to present the efficacy of the sliding window strategy on emotion classification.

To evaluate the efficacy of the sliding window strategy on emotion classification using multimodal learning models, we investigated the performance of the two temporal multimodal deep learning models using different window sizes: 2 sec, 3 sec, 5 sec and 10 sec. The small window sizes were selected to make it possible for real-time applications.

The architecture of the non-temporal multimodal learning models is the same that of the temporal models, only the input layer in the latter models is different. The input layer in the non-temporal models is based on a trial-wise strategy. In the trial-wise training, the input is the whole duration of the raw physiological signals per video clip, called a trial, whereas, the target, the corresponding trial label is used for training the ConvNet. The EEG channels and BVP signal for different video clips were used for training. In our data collection, we were given 9 trials per 17 subjects. Therefore, the total number of training and testing samples was $9 * 17{=}153$.

As the lengths of the video clips varied, data from each video clip was transformmed into the same length. This approach helped in preparing the input data for ConvNet for the trial-wise strategy. To transform video vlips data with different length into the same length, zero-padding approach and the maximum length are considered.

Fig. 8 presents the distribution accuracy of the two multimodal learning models based on non-temporal (trial-wise) and temporal data with different window sizes, 2 sec, 3 sec, 5 sec and 10 sec, with 300 iterations.

As shown in Fig. 8, as the window size increases, the performance of emotion classification based on both temporal multimodal deep learning models improves, with the best performance achieved with a window size of 10 sec. It also shows that both temporal models using ConvNet can capture and better learn spontaneous patterns with a longer window size.

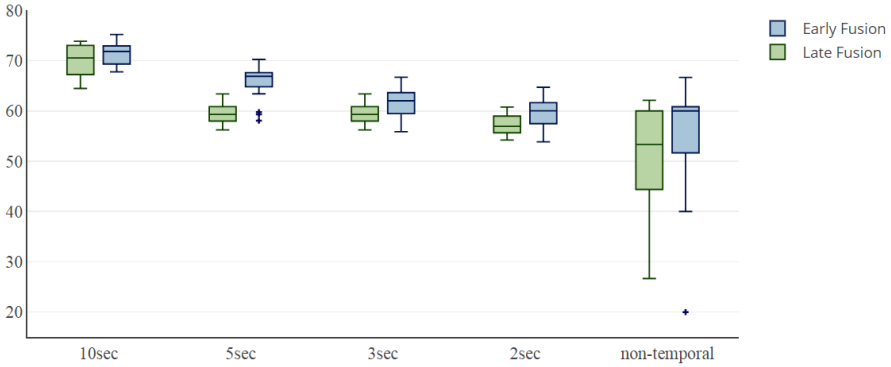

FIGURE 8. The figure shows the accuracy distribution of the temporal multimodal deep learning model with different window sizes and the non-temporal multimodal deep learning model based on early and late fusion.

TABLE 2. The average performance of the temporal multimodal deep learning models with different window sizes, and the non-temporal models.

Fusion Models	Temporal model								Non-temporal model	
	10-sec		5-sec		3-sec		2-sec			
	Accuracy	Valid Loss	Accuracy	Valid Loss	Accuracy	Valid Loss	Accuracy	Valid Loss	Accuracy	Valid Loss
Early fusion	**71.61± 2.71**	**0.62± 0.08**	65.5± 3.3	0.74± 0.03	61± 2.7	0.81± 2.4	56 ± 3.4	0.93± 0.07	55.07± 4.3	0.96± 0.13
Late fusion	70.17± 3.7	0.63± 0.10	64.4± 3.7	0.74± 0.09	59.4± 1.9	0.87± 2.4	55.9± 3.4	0.94± 0.05	52.28± 4.6	0.98± 0.15

From the figure it can be seen that the performance of all the temporal models with different window sizes is higher than that of the non-temporal models. It shows that the sliding window strategy is essential for building an accurate emotion classification model. Moreover, the overall accuracy of the early fusion temporal multimodal deep models is slightly better compared to the late fusion models. This means that the accuracy of the classification of emotions into four quadrants of dimensional emotions is increased using an early level fusion approach, as this approach can capture the correlated emotional information across modalities.

The average performance of the temporal and non-temporal models, including average accuracy ± standard deviation, average loss value ± standard deviation is presented in Table 2. As shown in Table 2, the overall accuracy of the multimodal learning models based on early fusion is higher than those based on late fusion for both the temporal and non-temporal approaches. It also shows that the performance of both temporal models based on early and late fusion with a window size longer than 3 sec is significantly improved. This not only confirms the efficacy of the sliding window strategy in improving emotion classification using multimodal learning models, but also shows there is no generic window size that can be used to achieve high performance for this emotion classification problem.

Since the goal of this study is to apply the proposed model to automatic emotion classification with the potential for real-time applications, it is necessary to choose the smallest window size which gives the highest accuracy.

A window size of 10 sec or 5 sec could be the acceptable for classifying emotions into four quadrants of dimensional emotions. However, the performance of the model using a 10-sec window size is more accurate than using a 5-sec window size.

C. EVALUATION OF TEMPORAL MULTIMODAL DEEP LEARNING MODELS USING EARLY AND LATE FUSION ON EMOTION CLASSIFICATION

In this section, the performance of emotion recognition using the two proposed models with a 10-sec window size and 300 iterations is evaluated and compared. Fig. 9 and 10 show the confusion matrices of the four quadrants of dimensional emotions for the temporal multimodal deep learning models based on early and late fusion, respectively.

FIGURE 9. Temporal deep learning model based on early fusion.

As shown in Fig. 9, recognising high-arousal negative (HA-N) emotions is more difficult than recognising the other three quadrant emotions. It also shows that HA-N emotions

TABLE 3. The comparison of the best average performance of our proposed model with other state-of-the-art methods.

Models	Model	Class No.	Accuracy
Li et al. [12]	CRNN	2	Valence: 72.06% Arousal: 74.12%
Xing et al. [13]	LSTM	2	Valence: 81.10% Arousal: 74.38%
Alhagry et al. [41]	LSTM RNN	2	Valence: 72.06% Arousal: 74.12%
Nakisa et al. [21]	Handcrafted feature extraction model	4	65.04∓ 3.19%
Our proposed model	**ConvNet LSTM (early fusion)**	**4**	**71.61± 2.71%**

FIGURE 10. Temporal deep learning model based on late fusion.

are often misclassified as HA-P emotions. The LA-P quadrant emotions are often recognised as either HA-N or HA-P. It is also shown that the model is able to recognise LA-N emotions more accurately compared to the other four quadrants of dimensional emotions.

Fig. 10 shows that the late fusion model is able to classify HA-N and LA-P emotions better than HA-P emotions and the HA-P quadrant is often misclassified as the HA-N quadrant. It also shows that the performance of this architecture in recognising LA-N is as good as the temporal multimodal deep learning model based on early fusion.

Overall, the performance of the temporal model based on early fusion in classifying emotions into four quadrant dimension emotions is better than the late fusion model, particularly in classifying HA-N.

D. COMPARISON OF TEMPORAL MULTIMODAL DEEP LEARNING MODELS WITH THE LATEST APPROACHES

Finally, the most highly tuned configuration of our system was compared with some of the latest methods. The experimental results for emotion classification are presented in Table 3.

Based on the comparison with other methods, although these other methods have achieved high accuracy using only EEG signals with 32 channels, they used medical sensors (medical cap) which have a higher resolution in terms of data collection. Moreover, the results from the other models are based on two-class labels (arousal and valence) of emotions. Whereas, our model was evaluated using data captured by lightweight mobile sensors with only 5 EEG channels. Moreover, our proposed model was able to achieve similar accuracy on four-class emotion classification (HA-N, HA-P, LA-P and LA-N).

Li *et al.* [12] developed a ConvNet and recurrent neural network (RNN) model based on only EEG signals.

However, in this model the wavelet energy of the EEG signals was used as input for the ConvNet RNN model. Moreover, they evaluated the performance of the model on two-class emotion classification (arousal and valence).

In contrast, our model is based on raw EEG and BVP signals consumed in an end-to-end temporal manner. This means we fed the raw signals into the ConvNet LSTM model without any feature extraction. Moreover, the performance of the proposed model was evaluated on the classification of emotions into four dimensional emotions.

Xing *et al.* [13] extracted some features from the frequency domain such as Power Spectrum density (PSD) to feed into their Stack Autoencoder LSTM model. The model was evaluated on arousal and valence, with 81.10 and 74.38 per cent accuracy achieved, respectively. Although, we achieved a lower performance (71.61 per cent), the output of our model was four-class classification of emotions.

In our previous study [21], we analysed the performance of conventional handcrafted feature extraction models using EEG and BVP signals. Table 3 shows that the temporal multimodal deep learning models using both early and late fusion improved the accuracy of emotion classification by about 4% compared to the handcrafted feature extraction methods. This improvement confirms that temporal multimodal deep learning methods are able to better capture the latent emotion structure within and across EEG and BVP signals.

This study confirms that the fusion of signals captured via lightweight mobile body sensors as input to a deep learning model can accurately classify emotions into four quadrants of dimensional emotions, thus confirming the feasibility of using these sensors for non-critical applications.

V. CONCLUSION AND FUTURE WORK

In this study, we proposed two new frameworks using temporal multimodal learning models based on early and late fusion in the context of emotion recognition. The proposed

temporal multimodal deep learning models are based on ConvNet LSTM networks using an end-to-end method. We evaluated the performance of the proposed model on our dataset collected using wireless wearable sensors (Emotiv and Empatica wristband). Our dataset was comprised of the physiological signals of 17 participants recorded while they watched nine video clips. A sliding window strategy was utilized to apply temporal multimodal learning models to these physiological signals. Hence, the raw physiological signals were divided into successive fixed-size windows with a 50% overlap. The performance of the proposed models with different window sizes was investigated and compared with the non-temporal multimodal learning models using a trial-wise strategy. In the trial-wise training, the entire duration of the raw physiological signal per video clip, called a trial, was used as inputs to the model and the corresponding trial label was used as the target for training. It was shown that the performance of the temporal multimodal deep learning models using early and late fusion was higher than that of the multimodal learning models based on a non-temporal strategy, with recorded accuracies of 71.61± 2.71 and 70.17± 3.7 versus 55.07± 4.3 and 52.28± 4.6, respectively. Moreover, the average accuracies of temporal multimodal deep learning models based on early fusion with different window sizes were higher than those for the late fusion model. The results showed that the temporal multimodal models based on early fusion with longer window sizes perform better than those with shorter window sizes. In this study, the best results for the multimodal learning model based on EEG and BVP signals were achieved with a 10-sec window, resulting in an accuracy of 71.61 ± 2.71. The proposed models outperformed models based on the handcrafted feature extraction method, because these models can better capture the latent emotion structure within and across EEG and BVP signals.

Despite the promising results, there is still a need to investigate other deep learning techniques and evaluate their performance. Moreover, this study was based on a limited number of participants and it would be worthwhile to expand the study and investigate the performance and the methods with a larger sample of participants.

REFERENCES

[1] R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 4, pp. 329–347, Apr. 2007.

[2] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE Trans. Robot.*, vol. 24, no. 4, pp. 883–896, Aug. 2008.

[3] A. Luneski, P. D. Bamidis, and M. Hitoglou-Antoniadou, "Affective computing and medical informatics: State of the art in emotion-aware medical applications.," *Stud. Health Technol. Informat.*, vol. 136, p. 517, Oct. 2008.

[4] B. T. Nugraha, R. Sarno, D. A. Asfani, T. Igasaki, and M. N. Munawar, "Classification of driver fatigue state based on EEG using emotiv EPOC," *J. Theor. Appl. Inf. Technol.*, vol. 86, no. 3, p. 151, 2016.

[5] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Affective Dialogue System*. Springer, 2004, pp. 36–48.

[6] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, May 2004.

[7] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5040–5043.

[8] S. Haq and P. J. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, 2011, pp. 398–423.

[9] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[10] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[11] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[12] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 352–359, doi: 10.1109/BIBM.2016.7822545.

[13] *Frontiers|SAE+LSTM: A New Framework for Emotion Recognition From Multi-Channel EEG | Frontiers in Neurorobotics*. Accessed: Jul. 23, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2019.00037/

[14] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2141–2149.

[15] S. Koelstra, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *Proc. Int. Conf. Brain Informat.*, 2010, pp. 89–100.

[16] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Multimedia Content Representation, Classification And Security*. Cham, Switzerland: Springer, 2006, pp. 530–537.

[17] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *Proc. IEEE Int. Conf. Ind. Technol.*, 2004, pp. 1138–1143, doi: 10.1109/ICIT.2004.1490720.

[18] A. M. Khan and M. Lawo, "Recognizing emotion from blood, volume pulse, and skin conductance sensor using machine learning algorithms," in *Proc. Conf. Med. Biol. Eng. Comput.*, 2016, pp. 1297–1303.

[19] K. Takahashi, "Remarks on SVM-based emotion recognition from multi-modal bio-potential signals," in *Proc. 13th IEEE Int. Workshop*, 2004, pp. 95–100. Accessed: Jun. 15, 2015. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1374736

[20] F. Zhang, "The effects of higher temperature setpoints during summer on office workers' cognitive load and thermal comfort," *Building Environ.*, vol. 123, pp. 176–188, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132317302834

[21] B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors," *Expert Syst. Appl.*, vol. 93, pp. 143–155, Mar. 2018.

[22] T. Hui and R. Sherratt, "Coverage of emotion recognition for common wearable biosensors," *Biosensors*, vol. 8, no. 2, p. 30, Mar. 2018.

[23] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Anal. Appl.*, vol. 9, no. 1, pp. 58–69, May 2006.

[24] S. Akselrod, D. Gordon, F. Ubel, D. Shannon, A. Berger, and R. Cohen, "Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control," *Science*, vol. 213, no. 4504, pp. 220–222, Jul. 1981.

[25] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[26] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2014, pp. 298–310.

[27] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.
[28] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013, doi: 10.1109/MCI.2013.2247823.
[29] S. E. Kahou, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interface*, vol. 10, no. 2, pp. 99–111, 2016.
[30] F. Ringeval, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, May 2015.
[31] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, 2015, pp. 49–56.
[32] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," 2019, *arXiv:1905.07039*. [Online]. Available: http://arxiv.org/abs/1905.07039
[33] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 627892. [Online]. Available: http://www.hindawi.com/journals/tswj/2014/627892/abs/
[34] G. Caridakis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *Proc. Int. Conf. Artif. Intell. Appl. Innov.*, 2007, pp. 375–388.
[35] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. Late fusion," in *Proc. Int. Conf. Syst., Man Cybern.*, vol. 4, Feb. 2005, pp. 3437–3443.
[36] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration—A statistical view," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 334–341, Dec. 1999.
[37] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7596–7599.
[38] S.-E. Moon, S. Jang, and J.-S. Lee, "Convolutional neural network approach for eeg-based emotion recognition using brain connectivity and its spatial information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2556–2560.
[39] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," in *Proc. AAAI*, 2017, pp. 4746–4752.
[40] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.
[41] S. Alhagry and A. Aly, "Emotion recognition based on EEG using LSTM recurrent neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 335–358, 2017, doi: 10.14569/IJACSA.2017.081046.