# Deep Physiological Affect Network for the Recognition of Human Emotions

Byung Hyung Kim and Sungho Jo, *Member, IEEE*

**Abstract**—Here we present a robust physiological model for the recognition of human emotions, called Deep Physiological Affect Network. This model is based on a convolutional long short-term memory (ConvLSTM) network and a new temporal margin-based loss function. Formulating the emotion recognition problem as a spectral-temporal sequence classification problem of bipolar EEG signals underlying brain lateralization and photoplethysmogram signals, the proposed model improves the performance of emotion recognition. Specifically, the new loss function allows the model to be more confident as it observes more of specific feelings while training ConvLSTM models. The function is designed to result in penalties for the violation of such confidence. Our experiments on a public dataset show that our deep physiological learning technology significantly increases the recognition rate of state-of-the-art techniques by 15.96% increase in accuracy. An extensive analysis of the relationship between participants' emotion ratings and physiological changes in brain lateralization function during the experiment is also presented.

**Index Terms**—Emotion Recognition, Affective Computing, Physiological Signals, EEG, PPG, Convolutional, LSTM, Emotional Lateralization, Inter-hemispheric Asymmetry, Valence, Arousal.

## 1 Introduction

Past research on the recognition of human affect has made use of a variety of techniques utilizing physiological sensors. Recently, miniaturized physiological sensors and advanced mobile computing technologies have enabled the continuous monitoring of physiological signals using so-called "everyday technology" [1], [2], [3]. These sensors provide us with electroencephalography (EEG), heart rate variability, pulse oximetry, and galvanic skin response data, which have been used as reflections of emotional changes. These data help us to better understand the etiology of mental health pathologies, such as stress. However, building reliable automated systems for understanding affective dynamics is a challenging problem, as the mechanisms by which emotions are elicited and the characteristics of the related physiological signals are complex.

Emotions are multicomponent phenomena, may be expressed in different manners, and can even be withheld over time. The complexity of the neural mechanisms of emotional processing has led to difficulties in measuring and accurately understanding emotions. Changes in physiological signals are affected by human emotions, although these signals are often subject to noises from various artifacts, low signal-to-noise ratio (SNR) of sensors, and inter- and intra-subject variability in physiological activation.

In response to these challenges, here we present a robust physiological model for the recognition of human emotions, called the Deep Physiological Affect Network. This model is based on convolutional long short-term memory (LSTM) networks [4] and a new temporal margin-based loss function. This system helps to bridge the gap between the low-level physiological sensor representations and the high-level context-sensitive interpretation of emotion. Formulating the problem of the recognition of emotional changes as a spectral-temporal sequence classification problem, wherein the input is a physiological signal sequence and the targets are discrete numbers of emotional states, we focus on the time-frequency analysis of bipolar EEG signals underlying brain lateralization and photoplethysmogram (PPG) signals. Brain lateralization refers to the idea that the two halves of the brain (left and right cerebral cortex) have differences in function. More specifically, emotional lateralization is the asymmetrical representation of emotion perception and expression processing in the cerebral hemispheres. The major advantage of the differential lateralization in EEG signals is that the minimum configuration requires only two electrodes. This simplicity enables the development of everyday technology such as a lightweight EEG device that can be worn easily and allows users to act freely in everyday situations. Created with its potentiality for applications in everyday technology in mind, our system learns the differential physiological activations in inter-hemispheric EEG signals and a PPG signal and quantifies them for recognizing emotions.

Furthermore, we present a new temporal margin-based classification loss function to better recognize and localize emotions temporally. Typical LSTM models have shown their superiority in memorizing useful patterns of previous observations and providing longer-range context for the current prediction. However, using only classification loss in training such models typically fails to properly penalize incorrect predictions. This is because LSTMs only implicitly consider the context that is passed along over time in the form of the previous hidden state and memory as well. This implicitness in training LSTM models is especially critical for learning long-term sequential data, such as physiological signals that contain complex emotional elicitation mechanisms. We have added an explicit temporal constraint into

• *B. Kim and S. Jo are with the School of Computing, KAIST, Republic of Korea. S. Jo is the corresponding author.*
*E-mail: {bhyung, shjo}@cs.kaist.ac.kr*

*Manuscript received April 19, 2005; revised August 26, 2015.*

our LSTM training so that the trained model better captures the explicit progression of emotions globally from the onset of the emotion until the current time.

In summary, the contributions of the proposed system, as an alternative to the existing systems are as follows:

- Robust model for capturing and tracing emotional changes: We present the Deep Physiological Affect Network (DPAN), which is based on convolutional long short-term memory (ConvLSTM) modeling of multi-modal physiological features. The goal of this model is to identify emotional states according to a two-dimensional emotion model whose axes are valence and arousal [5].

- Temporal margin-based classification loss function: We propose a new classification loss function to better learn models that can discriminate emotional states. We show that our model has significant improvements over a ConvLSTM model trained only using classification loss in emotion recognition tasks.

- Analysis of the effect of emotional lateralization on emotion recognition: We present the correlations between emotional lateralization and emotional valence and arousal obtained from the classified results of our system for potential applications in everyday technology, providing better understanding of the threshold of the differentiator which has suffered from inter- and intra-subject variability.

The rest of this paper is organized as follows: In Section 2, we provide theoretical background and previous studies in emotion recognition related to our proposed system. Section 3 presents our DPAN system and consists of the following subsections: 1) formulation of the emotion recognition problem, 2) physiological feature extraction, 3) ConvLSTM using our proposed temporal margin-based loss function. In Section 4, we evaluate the performance of our system using the public dataset, Database for Emotion Analysis using Physiological Signals (DEAP) [6]. The potential of our model is reflected in improved recognition accuracy for several physiological phenomena. In Section 5, we explore how the brain is lateralized, and how this is correlated with emotional changes. We then examine the physiological phenomena using theoretical studies on emotional lateralization. Furthermore, we investigate the effect of the convolutional structure in DPAN on emotion recognition using kernels of different sizes. We conclude this article with perspectives on future work.

## 2 Background and Related Work

Multiple theories have been proposed to understand emotion due to its multifaceted nature. Russell [7] and Panksepp [8] described the multifaceted nature of emotion as an "umbrella" concept when referring to the roles of psychological constructs. The concept includes the various processes that produce the different components of emotion, their associations, and the categorization of these elements as a specific emotion. the proposed DPAN is a deep learning model based on a particular theory of emotion, emotional lateralization. Our proposed model takes multi-modal physiological signals as input data from EEG and PPG signals given in the DEAP. Upon these, the overall aim of this section is to provide 1) theoretical background and 2) its related work of emotional lateralization, previous studies in emotion recognition using 3) multimodal physiological signals and 4) deep learning methods from a methodological perspective. At last, we cover the DEAP database with a summary of the baseline classifier and its corresponding accuracy classification.

### 2.1 Emotional Lateralization

Emotional lateralization is the asymmetrical representation of emotional processing between the left and right hemispheres. Previous research has shown the asymmetrical activation of these two distinct cerebral hemispheres. The oldest theory regarding emotional lateralization claims that the left hemisphere is associated with cognitive processes, while the right hemisphere is involved in the processing of emotion. This theory has been supported by several studies based on experiments on facial expression [9], [10]. However, many alternative studies have reported different patterns of brain asymmetry beyond the dominant role of the right hemisphere in understanding human emotions, in particular those concerning positive and negative affect.

The valence hypothesis posits that there is a center for positive feelings in the left hemisphere and a center for negative feelings in the right hemisphere. Davidson and colleagues have tested this hypothesis and have shown the asymmetrical activation in the frontal brain regions [11]. Another alternative to the above hypothesis is the motivational approach-withdrawal hypothesis [12], [13]. According to this hypothesis, emotions are intimately associated with the behavior and motivational direction of the individual in their environment, and are categorized using evolutionary concepts; happiness, surprise, and anger are categorized as approach emotions due to their tendency to induce movement toward environmental stimuli, whereas sadness, fear, and disgust are associated with withdrawal behaviors because of their tendency to lead to avoidance of environmental sources of aversive stimulation.

The positive/negative and the approach/withdrawal hypotheses have many similar aspects, but they strongly disagree on the classification of the emotion of anger. In the positive/negative model, anger is considered as a negative emotion along with sadness, fear, and disgust. However, anger is classified as an approach emotion in the approach/withdrawal model. It is assigned to the same category as happiness and surprise because it leads the individual to fight and is a source of stimulation. Despite this disagreement, the hypotheses are complementary and have been supported by many studies in the past few decades [12].

### 2.2 Inter-hemispheric Asymmetry-based Features

The finding of inter-hemispheric asymmetry related to emotion described in the above section has led to implementing related EEG features, such as differential and rational

asymmetry in symmetric EEG electrodes. Lin *et al.* [14] have proposed an EEG-based framework to recognize four emotional states during music listening. They have also investigated the most relevant independent features of emotional processing across different subjects and tested the efficacies of multiple classifiers. They claim that a spectral power asymmetry-based feature is superior to other features in characterizing brain dynamics in response to four emotional states (joy, anger, sadness, and pleasure). Clerico *et al.* [15] have presented a method for the automated recognition of affective states during four different classification tasks. In this model, the mutual information between spectral-temporal physiological patterns in inter-hemispheric electrodes is quantified. Although these feature-based approaches have been widely used in the field of affective computing and have been developed using advanced signal processing, most studies have difficulties when attempting to develop subject-specific differentiators for different emotions and therefore rely on a different and usually small dataset [16]. To solve this problem, our DPAN is built on deep learning technology, which has been beneficial to capturing inter- and intra-class variability.

### 2.3 Physiology and Multi-modality for Recognizing Human Affect

Several theories of emotion indicate that physiological activity is key to understanding emotions. As a result, studies on human affect using physiological signals have been widely carried out and have advanced significantly in many ways over the past few decades [17], [18].

To understand human emotions in this study, we focus on identifying patterns in the physiological activity that corresponds to the expression of different emotions using machine learning techniques. Most affect recognition methods involve changes in the central nervous system (CNS) [19], [20] and the autonomic nervous system (ANS) elicited by specific emotional states. The two systems are considered to be major components in affective computing studies. The use of CNS-based methods is justified by the fact that the cerebral cortex contains several areas used to regulate human emotions. In particular, physiological signals obtained from EEG and PPG have been widely used in emotion recognition, as each has its own merits.

EEG measures the electrical activity of the brain. It refers to the recording of the brain's spontaneous electrical activity with multiple electrodes placed on the scalp. Despite its low spatial resolution on the scalp, its very high temporal resolution is valuable to clinical applications. For instance, epilepsy and other sleep disorders can be identified by detecting temporal abnormalities in EEG readings [21], [22]. Moreover, the non-invasiveness and mobility of EEG have extended its usage to the field of brain-computer interfaces (BCIs), external devices that communicate with the users brain [23]. It has been pursued extensively by many studies associated with its control strategies such as motor imagery [24] and visual evoked potential [25].

Most EEG-related studies have relied on feature-based classifiers. Upon electrode selection based on neuroscientific assumptions, features are extracted and selected to classify discrete emotions. For instance, Liu *et al.* [26] have described a real-time EEG-based emotion recognition system based on their proposed standardized database of movie clips. Similarly, Wang *et al.* [20] investigated the characteristics of EEG features for emotion classification and techniques to track the trajectory of emotion changes. They extracted features to assess the association between EEGs and emotional states. Their work indicates that the right occipital lobe and the parietal lobe are mainly associated with emotions related to the alpha band, the parietal and temporal lobes are associated with emotions related to the beta band, and the left frontal and right temporal lobes are associated with emotions related to the gamma band. In these approaches, spectral power in specific frequency bands associated with emotional states has been used for emotion recognition. Unlike Wang *et al.*'s work, Petrantonakis and Leontios [19] developed adaptive methods for EEG signal segmentation in the time-frequency domain and the assessment of associations between these segments and emotion-related information. They exploited the frontal EEG asymmetry and the multidimensional directed information approach to explain causality between the right and left hemispheres. These results have shown that emotional lateralization in the frontal and temporal lobes can be a good differentiator of emotional states.

EEG-based emotion recognition systems have often had improved results when different modalities have been used [6], [27], [28]. Among the many peripheral physiological signals, PPG, which measures blood volume, is widely used to compute heart rate (HR). It uses optical-based technology to detect volumetric changes in blood in peripheral circulation. Although its accuracy is considered lower than that of electrocardiograms (ECGs), due to its simplicity as shown in Fig. 4, it has been used to develop wearable biosensors in clinical applications such as detecting mental stress in daily life [29]. HR, as well as heart rate variability (HRV), has been shown to be useful for emotion assessment [30], [31], [32]. Over the past two decades, some reports have shown that HRV analysis can provide a distinct assessment of autonomic function in both the time and frequency domains. However, these assessments require high time and frequency resolutions. Due to these requirements, HRV has only been suitable for analyzing long-term data. Several researchers have focused on overcoming this limitation. Valenza *et al.* [33] have recently developed a personal probabilistic framework to characterize emotional states by analyzing heartbeat dynamics exclusively to assess real-time emotional responses accurately.

In these studies, distinct or peaked changes of physiological signals in the time or frequency domains at a single instantaneous time have been considered as candidates. However, this approach is limited and cannot be used to fully describe emotion elicitation mechanisms due to their complex nature and multidimensional phenomena. To overcome this problem, we formulate emotion recognition in Section 3 as a spectral-temporal physiological sequence learning problem.

### 2.4 Deep Learning Approaches for Emotion Recognition

Recently, deep learning (DL) methods have increasingly emerged in the fields of computer vision, robotics, and neu-

rosciences. In emotion recognition, DL technologies have been studied to develop models of affect more reliable and accurate than the popular feature extraction-based affective modeling.

Martínez *et al.* [34] presented the use of DL methodologies for modeling human affect from multiple physiological signals. For training models of affect, they used a multilayer convolutional neural network (CNN) with denoising auto-encoders. They hypothesized that the automation of feature extraction via DL would yield physiological affect detectors of higher predictive power, which, in turn, will deliver affective models of higher accuracy. They evaluated the DL method on a game data corpus, which contained players' physiological signals and subjective self-report of affect, and showed that DL outperforms manual ad-hoc feature extraction as it yields significantly more accurate affective models.

DL has also proven to be beneficial to learning non-stationary data streams for complex tasks that require an understanding of temporal changes in the data. The non-stationary nature of brain activity in the context of emotion as revealed by EEG has been investigated in a recent study by Zheng *et al.* [35]. The authors investigated meaningful frequency bands and channels for emotion recognition using deep belief networks with differential entropy features extracted from multichannel EEG data. Meng *et al.* [36] presented a time-delay neural network (TDNN) to model the temporal relationships between consecutive affect predictions in their two-stage automatic system for predicting affective values continuously from facial expression videos. They aimed to separate the emotional state dynamics from an individual emotional state prediction step using TDNN, which makes the temporal information unbiased and unaffected by the high variability between features of consecutive frames.

In line with other works, our DPAN is a DL model to recognize various human emotions. Unlike others, we address the importance of understanding the characteristics of emotions, which have not yet been fully studied for building DL-based emotion recognition models.

### 2.5 DEAP: Database for Emotion Analysis using Physiological Signals

DEAP is a multimodal dataset for analyzing various emotions from physiological signals. The DEAP dataset was produced by recording 32-channel EEGs at a sampling rate of 512Hz using active AgCl electrodes placed according to the international 10-20 system and 13 other peripheral physiological signals (e.g., plethysmographs) from 32 participants while they watched 40 one-minute-long excerpts of music videos (for some participants, a frontal face video was also recorded). The dataset contained continuous valence, arousal, liking, and dominance ratings on scales from 1 to 9 and discrete familiarity ratings on scales from 1 to 5 rated directly after each trial. Self-Assessment Manikins [37] were used to visualize the ratings. For example, thumbs up and thumbs down icons were used for liking.

The authors of the dataset also presented the methodology and results of single-trial classification using three different modalities, EEG signals, peripheral physiological signals, and Multimedia Content Analysis (MCA), for automated affective tagging of videos in their dataset. For classification, they used a naive Bayes classifier as a baseline classifier. From the different modalities, physiological features including MCA were extracted and were used to classify low and high states of arousal, valence, and liking. The low and high states were determined by the threshold, which was placed on the middle of the 9-point rating scales. Using the baseline naive classifier, they achieved an average accuracy of 67.7% over participants for each modality and rating scale and a best accuracy of 65.2% from the multimodal fusion, concluding that there are still some obstacles to making highly accurate single-trial classifications, such as signal noise, individual physiological differences, and the limited quality of self-assessment. Their baseline classifier has limited capability to solve these problems since the naive Bayes model's independence assumptions using the maximum-likelihood method can lead to the overlooking of the maximization of posterior probabilities between different emotions. Details are further described in Section 4.3.

Recent works have strived to improve the accuracy of classifying EEG-based emotional states using the dataset. The fusion technology of different modalities has been further studied by Verma and Tiwary [38]. They investigated 3D emotion representation models and developed a multimodal fusion framework for recognizing and predicting various emotions from the measured physiological signals using a wavelet-based multiresolution approach. Yoon and Chung [39] proposed a probabilistic Bayes-based classifier that uses 32-channel EEG data with 61 additional virtual channels such as C3-C4 and C4-P4, which were generated by the transversal bipolar montage and the longitudinal bipolar montage, respectively, to achieve average accuracies of 70.9% and 70.1% for classifying two levels (high and low) in the valence and arousal ratings, respectively. To overcome the limited quality of self-assessment, Jirayucharoensak *et al.* [40] presented a deep learning network (DLN) using a stacked autoencoder to discover unknown feature correlations from 32-channel EEG input signals, which showed better performance compared to that of naive Bayes classifiers. Zheng *et al.* [41] investigated stable EEG patterns, which are considered as neural activities that share commonality across individuals and sessions under different emotional states, and evaluated how well models differentiated EEG signals among the various emotions. Results showed an average accuracy of 69.67% for classifying four states (high valence/high arousal, high valence/low arousal, low valence/high arousal, and low valence/low arousal) on the DEAP using differential entropy features.

## 3 Deep Physiological Affect Network

The proposed DPAN describes affect elicitation mechanisms used to detect emotional changes reflected by physiological signals. The inputs to this model are a sequence of bipolar EEG signals and a PPG signal. The model then learns the representations of the signals according to the known emotional valence-arousal model.

Fig. 1 illustrates the model that we have used for the recognition of emotions. This model contains two major components: 1) physiological feature extraction, which is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing



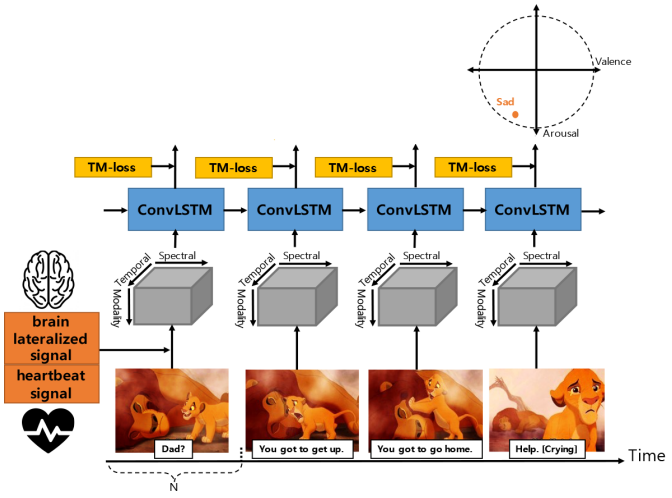



Fig. 1. An overview of DPAN. After every time interval $N$, the proposed DPAN first extracts two physiological features (brain lateralized and heartbeat features) and constructs a spectral-temporal tensor. These features are then fed into ConvLSTM to compute affective scores of emotions via our proposed loss model, temporal margin-based loss (TM-loss). The output at the final sequence is selected to represent an emotion over a 2-dimensional valence-arousal model for the entire sequence.

based on the formulation of emotion recognition problems focusing on the time-frequency analysis of bipolar EEG signals underlying brain lateralization and a PPG signal; and 2) ConvLSTM and our proposed temporal margin-based classification loss function that computes affective scores based on the features of the current frame and the hidden states and memory of ConvLSTM from the previous time step. We use a ConvLSTM described in [4] that applies dropout on non-recurrent connections.

### 3.1 Formulation of the Emotion Recognition Problem

To describe the complex affect mechanisms, DPAN focuses on the time-frequency analysis of bipolar EEG signals underlying brain lateralization and a PPG signal. At each time frame, the network takes the two-channeled EEG signals, and the PPG signal as inputs and outputs the one-dimensional vector, which represents emotional states scaled from 1 to 9. To detect physiological changes in emotion, frequencies appearing as peaks or distinct from others in the PSD occurring at a single instantaneous time have been considered as candidates. However, this approach can not handle inter- and intra-subject variability problems due to complex and multidimensional phenomena of emotion elicitation mechanisms. We believe that the estimation considering local neighbors of frequencies in temporal sequences would outperform methods that estimate frequencies at any single time. We have thus formulated emotion recognition as a spectral-temporal sequence learning problem.

Suppose that we obtain physiological signals from EEG and PPG sensors at time $N$ over a spectral-temporal region represented by an $M \times N$ with $P$ different modalities. The observation at the given time can then be represented by a tensor $\mathcal{X} \in \mathbb{R}^{M \times N \times P}$, where $\mathbb{R}$ denotes the domain of the observed physiological features (see Fig. 2) extracted by

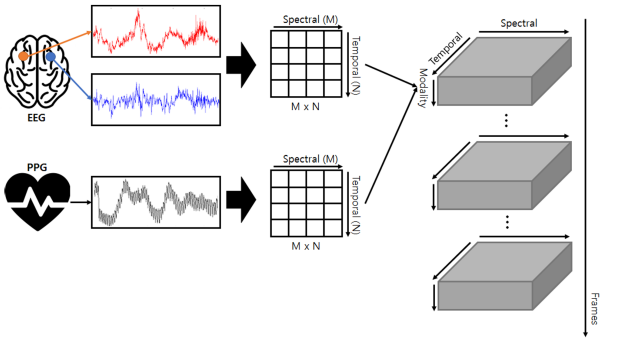



Fig. 2. The physiological feature extraction process and the formulation of the emotion recognition problem. At time $N$, brain-lateralized and heartbeat features represented by $M \times N$ grids are extracted from the spectrograms of bipolar EEG signals and a PPG signal, respectively. The tensor $\mathcal{X} \in \mathbb{R}^{M \times N \times P}$ is then constructed using the spectral-temporal features from the two modalities ($P = 2$).

the following section. The learning problem is the identification of the correct class based on the sequence of tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t$

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} P(y|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t) \tag{1}$$

, where $\mathcal{Y}$ is the set of valence-arousal classes.

### 3.2 Physiological Feature Extraction

We extract physiological features from the two modalities ($P = 2$) of EEG and PPG sensors. The extracted features are represented by $\mathcal{X}$ in (1) over the $M \times N \times P$ spectral-temporal domain, where $M$ represents frequency and $N$ represents time. From the two-channeled EEG signals, $E_t$, at each time frame $t$, we extract brain asymmetry features, $B_t \in \mathbb{R}^{M \times N}$, which underlie the spectral and causal asymmetry in the left-right channel pairs. They have provided differential and causal interaction in the brain [16]. Our system fuses them into $B_t$ to describe causal directionality and magnitude of emotional lateralization in a feature space.

$$B_t = \xi_{rl} \circ \frac{(\zeta_l - \zeta_r)}{(\zeta_l + \zeta_r)} \tag{2}$$

, where '$\circ$' denotes the Hadamard product and the matrix $\xi_{rl}$ is the causal asymmetry between the $r$ and $l$ EEG bipolar channels. It is used to measure the directed interactions from the channel $r$ to the channel $l$ which means the channel $r$ affects the channel $l$. It takes values between 0 and 1 where high values reflect a directionally linear influence from $r$ to $l$. Therefore, the asymmetry provides information on the directionality of causal interaction between two channels.

To measure this causality from $r$ to $l$, we use the Partial Directed Coherence (PDC) measure, which is based on the concept of Granger causality [42] as follows.

$$\xi_{rl}(m,n) = \frac{|A_{rl(m,n)}|}{\sqrt{a_k^H(m,n)a_k(m,n)}} \tag{3}$$

, where $m = 1, \cdots, M, n = 1, \cdots, N$, $A_{rl}$ is the $rl$th element of $A(m,n)$, $a_k^H$ denotes the Hermitian transpose of

the vector $a_k$, which is the $k$th column of the matrix $A(m,n)$ defined as follows:

$$A(m,n) = I - \sum_{d=1}^{p} A_d(n) z^{-d}|_{z=e^j 2\pi f} \quad (4)$$

, where $I$ is the identity matrix and the frequency $m$ varies within the range of 0 to the Nyquist rate. The matrices $A_d$ are given by

$$A_d = \begin{bmatrix} a_{11}^d & \cdots & a_{1M}^d \\ \vdots & \ddots & \vdots \\ a_{M1}^d & \cdots & a_{MM}^d \end{bmatrix} \quad (5)$$

, which are calculated using a causal multivariate autoregressive (MVAR) model, The MVAR model is the expression of Granger causality-based measures such as Granger Causality Index (GCI) and Directed Transfer Function (DTF) as well as PDC. These measures are defined in the framework of a MVAR model. Using the PDC measure is suitable for our study since it is defined in the frequency domain (not for GCI) and directional, which means that $\xi_{rl} \neq \xi_{lr}$. The $a_{rl}^d$ reflects the linear relationship between channels $r$ and $l$ at the delay $d$. This allows us to consider the direction of information flow between EEG channels as well as direct and indirect influences. A detailed explanation of MVAR models can be found in [42]. $\frac{(\zeta_l - \zeta_r)}{(\zeta_l + \zeta_r)}$ represents the spectral asymmetry between the $l$ and $r$ EEG channels. The asymmetry describes the degree of hemispheric lateralization. $\zeta_l$ and $\zeta_r$ are the logarithms of the spectral powers of the specific bands of the left and right hemispheres, respectively. An increase in the asymmetry feature leads to an increase in the left hemisphere activation more than the right. Therefore, the brain asymmetry feature in (2) describes directionality and magnitude of emotional lateralization between two hemispheres.

We extract the heart rate features $H_t$ over the $M \times N$ spectral-temporal domain, where frequencies with peaks in the PSD of the PPG signal are regarded as candidates of the true heart rate, from the PPG signal $P_t$ at each time frame $t$. These data form a candidate set over time.

### 3.3 Convolutional LSTM (ConvLSTM)

We apply ConvLSTM to recognize emotional states formulated in (1). ConvLSTM is an extension of the fully connected LSTM (FC-LSTM) [43], which has convolutional structures in both the input-to-state and state-to-state transitions [4]. ConvLSTM denotes inputs, hidden states, outputs, and other gates as three-dimensional (3D) tensors whose last two dimensions are spatial dimensions (rows and columns). ConvLSTM uses a convolution operator in the state-to-state and input-to-state transitions for determining the future state of a certain cell in the grid based on the inputs and the past states of its neighbors. This operator in the convolutional structure enables to capture local dependencies in spatio-temporal data, which is equivalent to our formulation in (1). Handling spatio-temporal data has been a major drawback of FC-LSTM. The full connections of FC-LSTM in input-to-state and state-to-state transitions contains too much redundancy to encode spatial information.

With $\mathcal{X}, \mathcal{C}$, and $\mathcal{H}$ representing the inputs, cell outputs, and hidden states, respectively, the key equations of ConvLSTM are shown as follows:

$$\begin{aligned} i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f) \\ \mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o) \\ \mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t) \end{aligned} \quad (6)$$

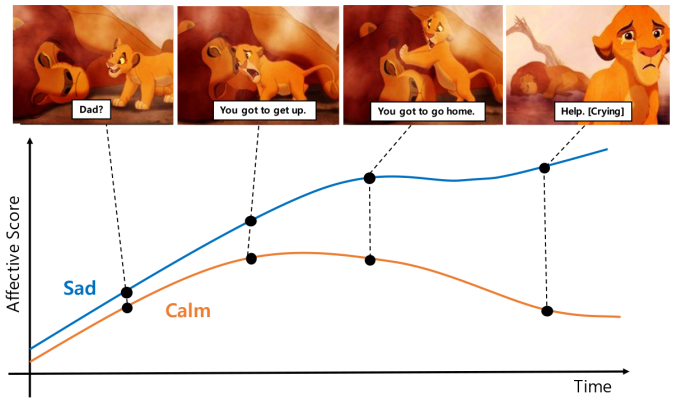
, where $i_t$, $f_t$, and $o_t$ are gates of ConvLSTM represented by 3D tensors. '$*$' denotes the convolution operator and '$\circ$' denotes the Hadamard product. Through activating the input, output, and forget gate, cells can store and retrieve information over long periods of time. This gives access to long-range context information and solves the vanishing gradient problem. Note that the traditional FC-LSTM, represented by [4], can be viewed as a special case of ConvLSTM on a single cell if we represent the hidden states and cell outputs of FC-LSTM using 3D tensors with the last two dimensions being 1. ConvLSTM has outperformed FC-LSTM in capturing spatiotemporal – or spectro-temporal in our case – correlations better.

To identify emotional states, ConvLSTM with a linear layer compute affective scores based on the physiological features of the current time frame $t$ and the hidden states and memory of ConvLSTM from the previous stage. In our work, we use a softmax layer for the final linear layer, so the affective score is the softmax output of the model.

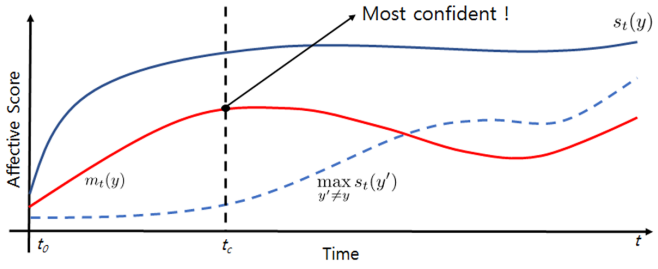
### 3.4 Temporal Margin-based Classification Loss

While ConvLSTM is efficient, it still fails to penalize incorrect predictions properly when using only classification loss in training. The model penalizes the same error no matter how much emotional context the model has already processed. For example, according to the Külber-Ross model [44], the feeling of grief entails denial, anger, bargaining, depression, and acceptance in order. Since the emotion contains bargaining and acceptance, a feeling of satisfaction could be the most likely incorrect output yielded by the model. LSTMs, given the sequence of the grief emotion, will output the same penalty regardless of how much of the sequence it has already processed. For instance, if the LSTM has processed the grief emotion up to depression, the incorrect label 'feeling satisfaction' would receive the same penalty as if the model had processed up to anger. However, outputting the incorrect emotion after seeing emotions up to depression should be penalized more than outputting the incorrect emotion after seeing emotions up to anger. Applying correct penalizations is required because LSTMs only implicitly consider the context that is passed over time in the formulation of the previous hidden state and memory. Without correct penalization, LSTM models struggle to learn from long-term sequential data, such as physiological signals which contain complex emotional elicitation mechanisms. These mechanisms are not considered as an antecedent stage to emotion but thought of as a constitutive stage of emotion for a relatively long period. Therefore, learning the progression patterns of the emotions in training is very important to develop reliable affect models.

To solve this critical problem, we have modified the existing classification loss function and formulated a new




(a) An example of our intuition of the proposed loss function




(b) Discriminative margin over time

Fig. 3. An example of our rationale for the proposed loss formulation and discriminative margin of an emotion over time. (a) As DPAN sees more of the emotion *sad*, it should become more confident of the presence of the correct emotion state (blue line) and the absence of the incorrect states (orange line). (b) The discriminative margin $m_t(y)$ (red line) of an emotion $y$ started at $t_0$. The margin $m_t(y)$ is computed as the difference between the ground truth affective score $s_t(y)$ (blue line) and the maximum scores $\max_{y' \neq y} s_t(y')$ (dashed blue line) of all incorrect emotion states between $t_0$ and $t$. The model becomes more and more confident in classifying emotion states until the time $t_c$. However, after the time $t_c$, $\mathcal{L}_t$ are non-zero due to the violation of the monotonicity of the margin.

loss function based on the temporal margin between the correct and incorrect emotional states. As shown in Fig. 3, our reasoning for using the formulation is as follows:

- *When more of a particular emotion is observed, the model should be more confident of the emotional elicitation as the recognition process progresses.*

Fig. 3a shows an example sequence of sad scenes in the movie "The Lion King (1994)". While desperately trying to rescue his son Simba in the midst of a stampede, Mufasa is thrown and killed by his brother Scar. This sequence of the movie contains stages of complex emotions such as sadness and calm. As the sequence progresses, sadness wells up and it reaches a peak at the scene in which Simba recognizes his father Mufasa's death while the calm feeling is fading. Our function constrains the affective score of the correct emotional state to discriminate its margin, which does not monotonically decrease with all others while the emotion progresses. We thus present a temporal margin-based classification loss that discriminates between the correct and incorrect emotion classes.

$$\mathcal{L}_t = -\log s_t(y) + \lambda \max(0, \max_{t' \in [t_0, t-1]} m_{t'}(y) - m_t(y)) \quad (7)$$

, where $-\log s_t(y)$ is the conventional cross-entropy loss function commonly to train deep-learning models. $y$ is the ground truth of emotion rating and $s_t(y)$ is the classified affective score of the ground truth label $y$ for the time $t$. $m_t(y)$ is the discriminative margin of the emotion label $y$ at time $t$.

$$m_t(y) = s_t(y) - \max\{s_t(y') | \forall y' \in \mathcal{Y}, y' \neq y\} \quad (8)$$

$\lambda \in \mathbb{Z}^+$ is a relative term to control the effects of the discriminative margin. Eq. (7) describes a model that becomes more confident in discriminating between the correct state and the incorrect states. The model is encouraged to maintain monotonicity in the affective score as the emotion training progresses. As shown in Fig. 3b, after the time $t_c$, the loss becomes non-zero due to the violation of the monotonicity of the margin. Note that the margin $m_t(y)$ of the emotion $y$ spanning $[t_0, t]$ is computed as the difference between the affective score $s_t(y)$ for the ground truth $y$ and the maximum classification scores $\max_{y' \neq y} s(y')$ for all incorrect ratings at each time point in $[t_0, t]$.

During training, we compute the gradient of the loss with respect to $s_t(y)$ with back propagated through time to compute the gradients with respect to the model parameters. For simplicity, we do not compute and back propagate the gradients of the loss with respect to $s_t(y')$.

## 4 Experiments

For the quantitative evaluation, we used the public dataset, DEAP [6], which has been widely used to analyze human affective states [38], [41], [45]. From comparing with two existing models, our experimental results show that our model is effective in recognizing human emotions. We asked the following questions regarding our model:

- Q.1. Does our model consistently outperform FC-LSTM networks and the state-of-the-art method?
- Q.2. What effect would the temporal margin-based classification loss have?
- Q.3. How do the convolutional kernels capture the spectral-temporal physiological patterns?

### 4.1 DEAP dataset

For our study, we used the eight symmetrical pairs of electrodes on the right (F8, FC2, FC6, C4, T8, CP2, CP6, and PO4) and left hemisphere (F7, FC1, FC5, C3, T7, CP1, CP5, and PO3) from the DEAP dataset. From the 32 electrodes, we selected the eight symmetrical pairs of electrode channels for which there have been reports of significant correlations with emotion in [6]. In addition to the EEG signals, we also used plethysmographs, which measure blood volume in the participant's left thumb, in order to compute the HR-related physiological features. Fig. 4 illustrates the EEG electrode placement and a plethysmograph used to acquire the physiological signals used in our study. The continuous ratings of valence and arousal are converted into discrete ratings rounding them toward negative infinity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing



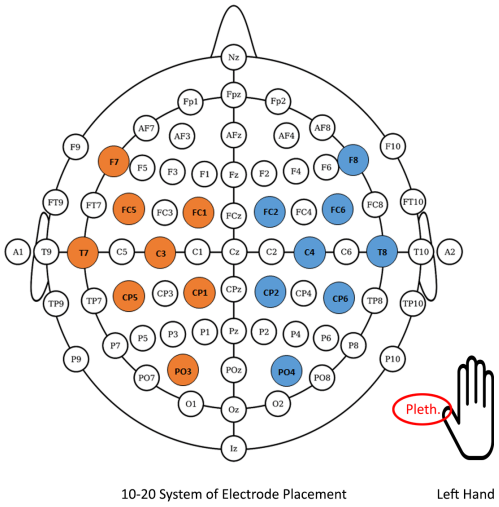



Fig. 4. Placement of EEG electrodes and a plethysmograph. EEG signals from the 8 pairs of electrodes were used to extract physiological features in (1).

### 4.2 Experimental Setup

Since we focus on brain lateralization, data from the eight selected electrodes and a plethysmograph recorded during 1,280 videos, along with the 64 combinations of physiological signals per video, lead to the generation of 81,920 physiological data points. We split the total data set into fifths (16,384 physiological data points) for testing. We used one-fifth (13,107 physiological data points) of the remaining data for our validation and four-fifths (52,429 physiological data points) of the data as a training set. We note that the training and testing data were subject-independent, which means they were chosen entirely randomly. The validation data was chosen randomly while keeping the distribution of ratings balanced. The highlighted one-minute EEG and plethysmography signals were split into 6 frames of 10 seconds each. They were down-sampled to 256 Hz and their power spectral features were extracted.

For EEG signals, same as in [6], high-pass filtered with a 2 Hz cutoff frequency using EEGlab toolbox and the same blind source separation technique for removing eye artifacts were applied. For plethysmograph signal, constrained independent component analysis (cICA) algorithm [46] was applied to refine the signal removing motion artifacts. The cICA algorithm is an extension of ICA and has been applicable in cases where prior knowledge about the underlying sources is available [47]. The logarithms of the spectral powers of bands ranging from 4 to 65 Hz were extracted from the selected electrodes and the participant's thumb. Using every two spectral-temporal datasets per frame, we generated two sets of $50 \times 50$ spectral-temporal features from $B_t$ and $H_t$ as inputs $\mathcal{X}_t$ with the corresponding ground truths of valence and arousal as the two targets.

We evaluated the performance of our model and compared it with 1) FC-LSTM and 2) Koelstra *et al.*'s method [6]. Our DPAN model uses a 1-layer network with 256 hidden states and the input-to-state and state-to-state kernel sizes of $5\times$ 5. To train our model, we used learning batches of 32 sequences. Back-propagation through time was performed for ten timesteps. The momentum and weight decay were set to 0.7 and 0.0005, respectively. The learning rate starts at 0.01 and is divided by 10 after every 20,000 iterations. We also performed early-stopping on the validation set. The above configuration was chosen as the best configuration, which yielded the minimum loss in the training set.

We also tried other configurations, such as $3 \times 3$, $7 \times 7$, and $9 \times 9$, to investigate their effects on capturing spectral-temporal correlations between emotions and physiological signals in Section 5. For the FC-LSTM, we used three 1,700-node LSTM layers with the softmax layer as the output. Other configurations, such as those of momentum, weight decay, and learning rate, were the same as those used in our model. For the Koelstra *et al.*'s method, we used the same classifier as those in [6], That classifier was the naive Bayes classifier with the fusion of single-modality. Two modalities are processed independently by the naive Bayes classifier, and each modality is set to contribute equally to the final decision.

### 4.3 Experimental Results

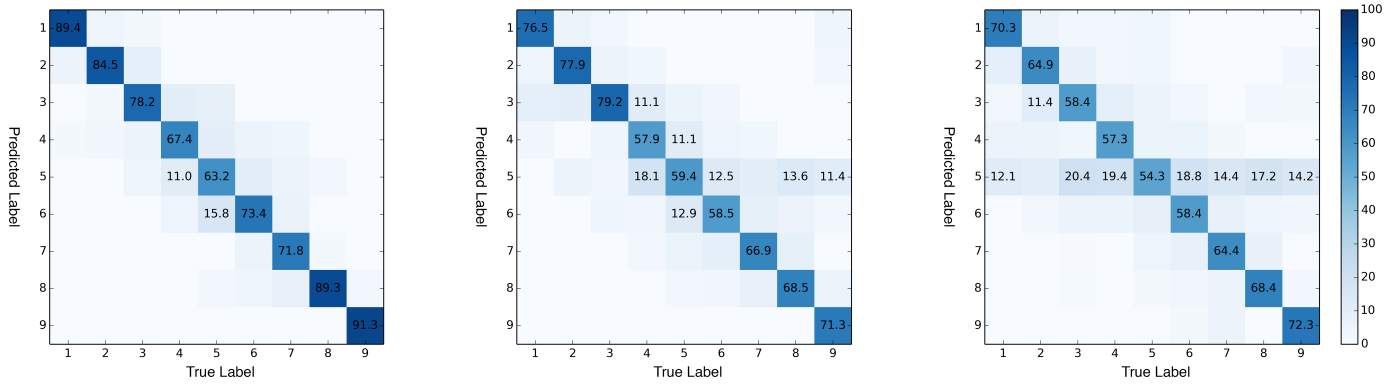
Fig. 5 and 6 show the confusion matrices of the valence and arousal ratings resulting from our DPAN model, the FC-LSTM, and the Koelstra *et al.*'s method. Our proposed system achieved overall accuracies of 78.72% and 79.03% for recognizing valence and arousal emotions, respectively. These values are much higher than those obtained using the other two methods; the FC-LSTM (68.45% and 66.56% for the valence and arousal ratings, respectively) and the Koelstra *et al.*'s method (63.23% and 62.59% for the valence and arousal ratings, respectively).

Our experiments show that the proposed system performs consistently better than the others, answering uestion Q.1. This superiority is mainly due to two reasons. First, our model, which is based on ConvLSTM, is able to learn complex spectral-temporal patterns of emotion elicitation mechanisms with the help of the nonlinear and convolutional structure of the network. The input-to-state and state-to-state kernels of the convolutional structure can capture localized spectral-temporal patterns and keep local consistencies reducing inter- and intra-subject variability in the physiological measures.

In contrast, the naive Bayes model in Koelstra *et al.* has difficulty in understanding the complexity of the signals and in training. Estimations of parameters underlying independence assumptions using the maximum-likelihood method can lead to overlooking of the maximization of posterior probabilities between emotion classes. This limitation is significant when the valence rating is 5, and the arousal rating is 1. The classifier shows poor predictive performance results for specific instances, such as when the valence rating 5 and the arousal rating is 1. This leads to errors in identification, as the classifier learns representations of two ratings excessively. It thus loses the ability to exploit the interactions between physiological features. The fully-connected structure of FC-LSTM has too many redundant connections and makes it very unlikely for the optimization to capture important local consistencies in spectral-temporal patterns. Another reason for the superiority of our DPAN

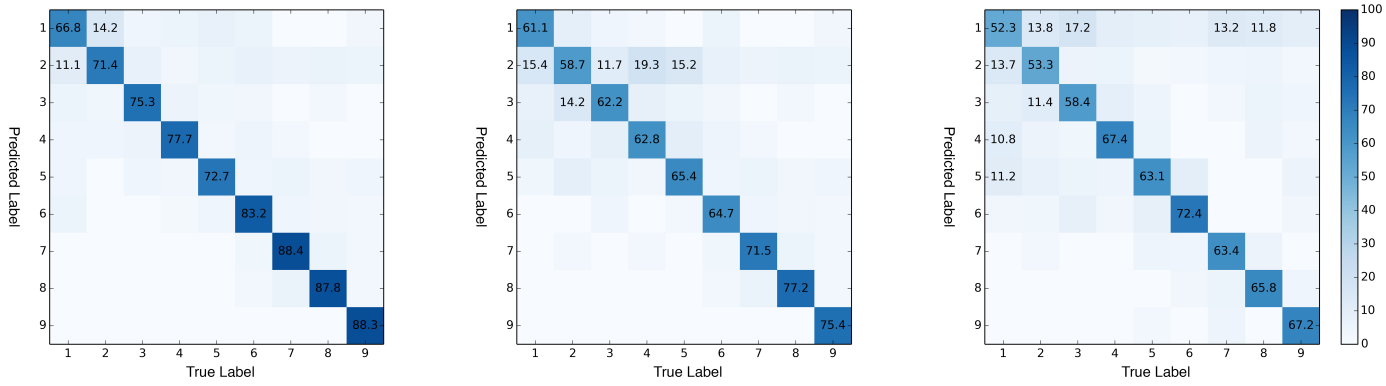
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing






(a) DPAN (b) FC-LSTM (c) Koelstra *et al.*

Fig. 5. The confusion matrices of valence ratings resulting from the (a) DPAN, (b) FC-LSTM, and (c) Koelstra *et al.*'s method to the DEAP. Note that the averaged accuracies of the three different models are 78.72%, 68.45%, and 63.23%, respectively. For better visualization, numbers are displayed only if their percentag is higher than 10 percent.




(a) DPAN (b) FC-LSTM (c) Koelstra *et al.*

Fig. 6. The confusion matrices of arousal ratings resulting from the (a) DPAN, (b) FC-LSTM, and (c) Koelstra *et al.*'s method to the DEAP. Note that the averaged accuracies of the three different models are 79.03%, 66.56%, and 62.59%, respectively. For better visualization, numbers are displayed only if their percentage is higher than 10 percent.

model is that it can discriminate among physiological patterns by imposing penalties for incorrect classification.

Regarding question Q.2, our proposed temporal margin-based classification loss globally increases physiological distinctness during training. The distinctness, however, can hardly be achieved by naive Bayes models and LSTMs that only use classification loss. Furthermore, this problem becomes more severe when the two systems use classifications closer to the valence ratings between 4 and 6 and the arousal ratings between 1 and 3. The emotion elicitation specifically worked well for the high arousal/high valence (HAHV) and high arousal/low valence (HALV) conditions, as emotional stimuli for the conditions induce strong physiological changes [6]. The other two systems have difficulties capturing small physiological changes elicited by the neutral conditions and learn their representations.

## 5 Discussion

Since the proposed loss is designed for our system to be confident in elicitation of the emotion as the recognition progresses when more of a specific emotion is observed, it is necessary to analyze the effects of the temporal margin-based classification loss over the evolving time scale of the training in the model.

### 5.1 Improvement of the Temporal Margin-based Classification Loss

Fig. 7 shows the average accuracies of the valence and arousal ratings for which recognition performance is improved following the use of the proposed loss during training. This demonstrates that our proposed loss is beneficial for training a better ConvLSTM model for emotion recognition. Significant improvements of approximately 6% (6.3% and 6.2% for the valence and arousal ratings, respectively) are achieved consistently when compared with the ConvLSTM model trained only using classification loss. Furthermore, the proposed loss is effective for the valence and arousal ratings between 4 and 6. As shown in Fig. 5 and Fig. 6, such ratings are difficult for recognition systems to classify because of the physiological similarities between the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing



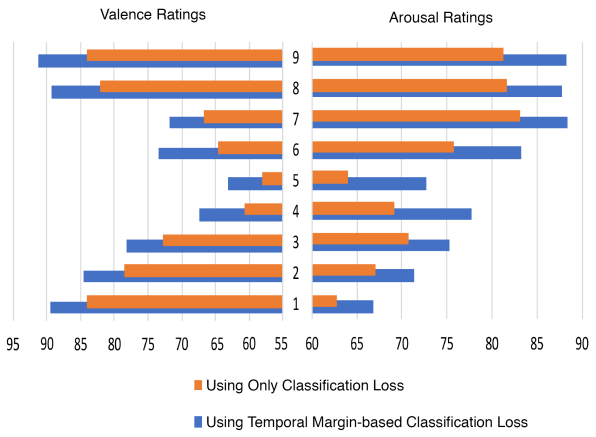


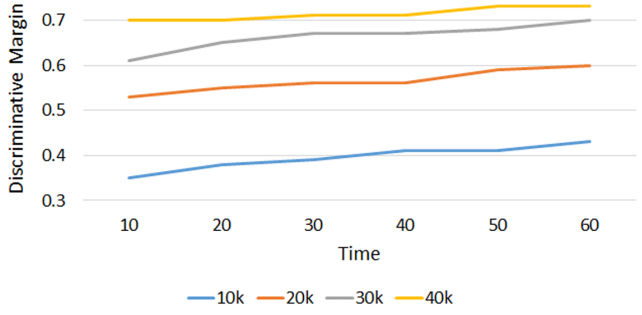

Fig. 7. Comparison of true positive accuracies of the valence (left) and arousal (right) ratings resulting from the use of our proposed temporal margin-based classification loss and those resulting from the use of classification loss only. The average improvements in the accuracies of the valence and arousal ratings following the use of our proposed loss are averagely 6.3% and 6.2%, respectively.

ratings. Our proposed temporal margin-based loss improves the recognition performance of these ratings more than those of the other ratings. This indicates that the benefits of the proposed loss are applicable to various types of emotions in recognition.

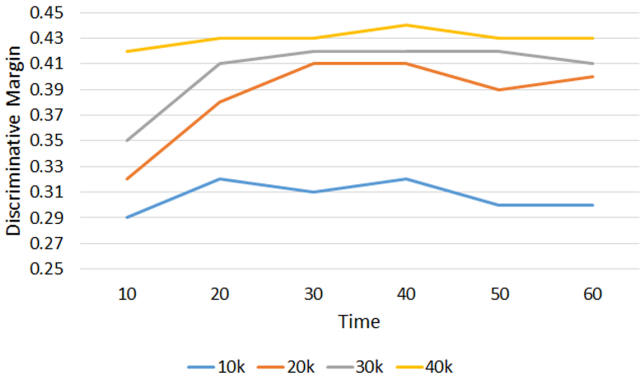
### 5.2 Effects of the Temporal Margin-based Classification Loss

We also analyzed the changes in the discriminative margins between correct ratings and incorrect ratings. We computed the discriminative margins at every frame in each test sequence in our tests using our proposed DPAN model and ConvLSTMs trained after 10,000, 20,000, 30,000, and 40,000 iterations. The same testing data is used to calculate the margins by taking snapshots of the two models trained after every 10,000*th* iteration a total of 4 times. Therefore, this produces a curve of the discriminative margin as a function of time for each test sequence. We note that the discriminative margins are averaged over the entire test set.

Fig. 8 displays the discriminative margins obtained using the DPAN trained with (a) the proposed loss and (b) the classification loss only after 10,000, 20,000, 30,000, and 40,000 iterations. The margins of (a) tend not to decrease. This monotonicity becomes more apparent as we train over more iterations. The absolute values of the discriminative margins also increase as we train over more iterations. The margin scores obtained using the proposed loss for recognition is significantly higher than those obtained using the model without the proposed loss. However, the margins of (b) tend to be flat after approximately 20 seconds of the recognition progress. These results indicate that the temporal margin-based classification loss has beneficial impacts on discriminating margins between recognition scores conforming to our rationale that "*When more of a particular emotion is observed, the model should be more confident of the elicitation of the emotion as the recognition progresses.*" This temporal distinction may be useful in real applications of emotion




(a) Temporal margin-based classification loss




(b) Classification loss only

Fig. 8. Averaged discriminative margins as functions of time over all test sequences. At every frame in each test sequence, the discriminative margins are computed using our DPAN trained by after 10,000, 20,000, 30,000, and 40,000 iterations with (a) our proposed temporal margin-based classification loss, and (b) a general classification loss.

recognition, as we are unable to recognize when an emotion is elicited in real life easily.

### 5.3 Effects of Emotional Lateralization and DPAN

Our results indicate that DPAN can learn physiological changes when an emotion is evoked during an emotion elicitation process. In this section, we investigate the physiological phenomena that are observed when emotions are classified using DPAN. Heart-related physiological features have served as essential elements reflecting the function of the ANS. However, in our study, we focus more on the brain lateralization feature $B_t$ in (2), as it has relatively large inter- and intra-subject variability. There are several explanations for this finding, which have been supported by some related studies. In this section, we explore how the brain is lateralized, and how this lateralization is correlated with emotional changes. We will also examine physiological phenomena by discussing theoretical studies on emotional lateralization.

To investigate correlations between subjective ratings and emotional lateralization derived in (2), we computed Spearman correlation coefficients between emotional lateralization in the four frequency bands and the subjective valence and arousal ratings. We also computed the $p$-values for the positive and negative correlation tests. This was performed for each participant separately, and assuming independence [6], the 32 resulting $p$-values per correlation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing



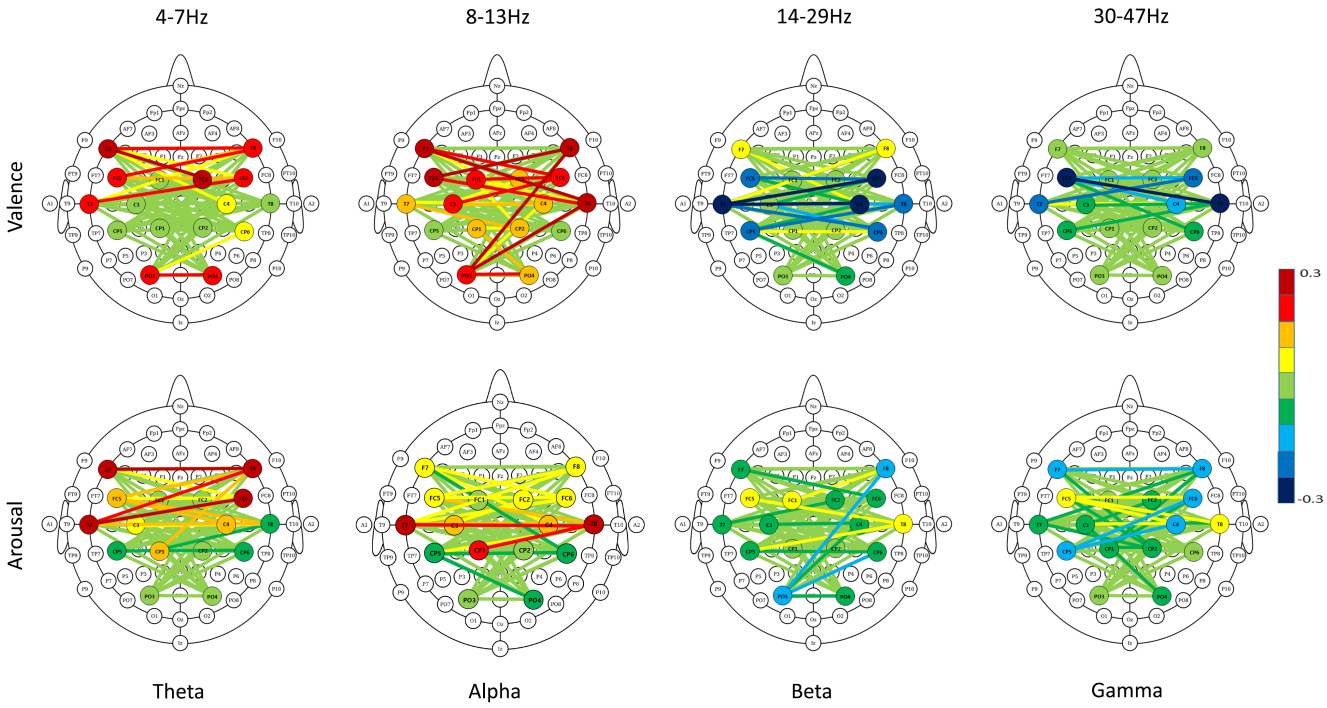



Fig. 9. Correlates of emotional lateralization and their ratings in the four frequency bands ($p < 0.05$)

TABLE 1
Pairs of electrodes for which emotional lateralization is significantly correlated with the subjective ratings ($p < 0.01$).

| Emotion | Theta | | | Alpha | | | Beta | | | Gamma | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Elec. pair | $R^+$ | $R^-$ | Elec. pair | $R^+$ | $R^-$ | Elec. pair | $R^+$ | $R^-$ | Elec. pair | $R^+$ | $R^-$ |
| **Valence** | (F7, F8) | 0.53 | -0.04 | (F7, T8) | 0.69 | -0.07 | (T7, FC6) | 0.1 | -0.59 | (FC5, T8) | 0.47 | -0.11 |
| | (F7, FC2) | 0.67 | -0.11 | (FC5, F8) | 0.61 | -0.11 | (T7, C4) | 0.11 | -0.53 | | | |
| | (FC5, F8) | 0.55 | -0.02 | (FC1, FC6) | 0.39 | -0.02 | (FC5, FC6) | 0.09 | -0.48 | | | |
| | (T7, FC6) | 0.49 | -0.17 | (C3, FC6) | 0.43 | -0.08 | | | | | | |
| | (PO3, PO4) | 0.48 | -0.05 | (PO3, F8) | 0.56 | -0.1 | | | | | | |
| **Arousal** | (F7, F8) | 0.62 | -0.03 | (C3, C4) | 0.41 | -0.08 | (F8, PO3) | 0.01 | -0.29 | (F7, F8) | 0.11 | -0.29 |
| | (T7, FC6) | 0.58 | -0.16 | (T7, C4) | 0.34 | -0.03 | | | | | | |
| | | | | (T7, T8) | 0.66 | -0.11 | | | | | | |
| | | | | (CP1, T8) | 0.44 | -0.05 | | | | | | |

direction, frequency band, and electrode were then combined into one $p$-value using Fisher's method.

Fig. 9 shows the average correlation coefficients, with significantly ($p < 0.05$) correlating pairs of two electrodes highlighted. The significant correlations ($p < 0.01$) are also reported in TABLE 1. We found significant correlations with valence in all of the four frequency bands. In the theta and alpha frequency bands over the frontal and occipital regions, an increase in valence led to an increase in the lateralization power. The positive correlation is consistent with the findings in [6]. The authors of the above study reported that an increase in valence in the theta and alpha bands leads to an increase in the frequency power over the left, rather than the right, temporal and occipital regions. Our observation might be consistent with the so-called valence hypothesis of hemispheric asymmetry, which claims that there is a center for positive feelings in the left hemisphere and a center for negative feelings in the right hemisphere [48]. In contrast, in the beta and gamma bands, we observed negative correlations between temporal lobes. This indicates that increased beta and gamma power over the right temporal region, when compared with the left temporal region is associated with positive emotion. This observation is in line with those of similar studies [49], [50], although it is inconsistent with the previous valence hypothesis. This shows that not only has the valence hypothesis been highly debated, but also that several alternatives have been suggested in reports on the neuro-physiological correlates of affective states. Although valence-based distinction has been key in understanding the bidimensional theories of emotion, the understanding of brain mechanisms underlying the valence hypothesis has always had ambivalent attributes. For instance, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2790939, IEEE Transactions on Affective Computing



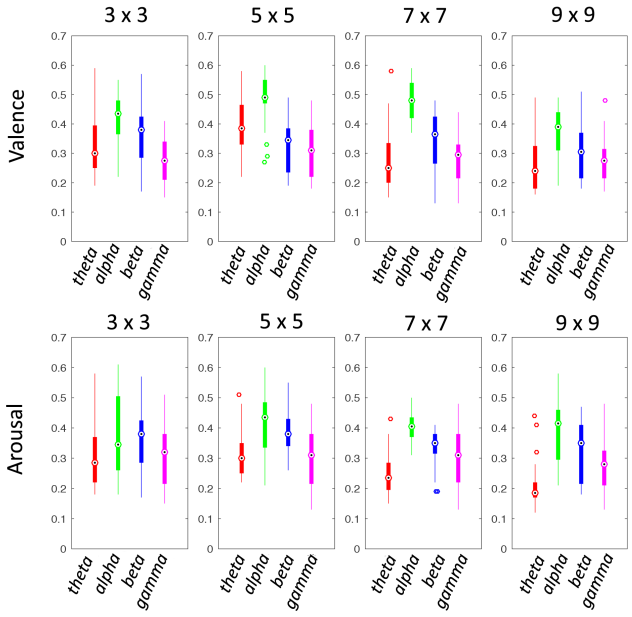



Fig. 10. Effect of the convolutional structure of DPAN. The average correlations between emotional lateralization in four frequency bands and the valence and arousal ratings, along with different convolutional kernel sizes.

appraised valence of an event and the valence of the feeling are not always congruent. Individuals can feel the emotion "interest" but the emotion can also be elicited by stimuli appraised as negative.

While emotional lateralization is correlated with valence in the four frequency bands, it has limited relationships with arousal. There are significant positive correlations in the theta band over the temporal and frontal regions and in the alpha frequency bands over the temporal regions. However, there are no significant relationships between brain lateralization and arousal ratings in the beta and gamma bands. Besides, there are no significant relationships in the alpha band over frontal areas. The more restricted correlations for arousal, when compared with valence can be explained as follows. Several studies have shown that arousal states are more associated with patterns of ANS activity, which regulate body functions, such as heart rate, respiratory rate, and pupillary responses, than with those of the CNS, which comprises the brain and spinal cord. Several researchers have reported that heart-related features, such as heart rate and heart rate variability are good indicators of arousal.

Our findings may also be justified by the fact that when some negative but approach-related emotions, such as "anger" which would be lateralized to the left hemisphere, are induced, they lead to increases in alpha band activity in the left anterior and the left temporal regions in the beta and gamma bands. In other words, this observation may be a reflection of the inter-correlations between valence and arousal, as reported in [6].

### 5.4 Effect of the convolutional structure of DPAN

We have shown the efficacy of choosing the best kernel size during training for DPAN. Our results indicate that using this strategy leads to the minimization of the loss. For answering question Q.3, it is thus necessary to investigate the relationships between different sizes of convolutional kernels and elicited emotions, along with valence and arousal, in the four frequency bands. Fig. 10 shows the average correlations between emotional lateralization and the valence and arousal ratings, along with different convolutional kernel sizes.

Interestingly, for valence and arousal, theta is the most sensitive frequency band over the different kernel sizes ($\sigma_{valence} = 0.056$ and $\sigma_{arousal} = 0.045$). This is due to the narrowness of the theta band, which usually consists of three frequencies (4-7 Hz). The limited number of frequencies in the theta band result in increased sensitivity in the correlation analysis. In contrast, the gamma frequency band is the least sensitive to changing kernel sizes($\sigma_{valence} = 0.009$ and $\sigma_{arousal} = 0.011$). This finding is also explained by the frequency size of the band. The gamma band had the largest frequency range (19 frequencies in our study). We believe that, in addition to the above physical limitations, the inconsistencies between the two frequency bands may be due to electrooculogram and electromyogram activities, which are dominant in low and high frequencies, although these frequency bands are also correlated with valence and arousal, as seen in Fig. 9 and TABLE 1.

The alpha band had the highest correlations with both valence and arousal and had decreased variance in the correlations. When using the $5 \times 5$ kernel for valence and arousal, the highest correlation is achieved using the smallest variance of the 32 resulting $p$-values. This may imply that the physiological signals in the alpha band captured by the DPAN using the $5 \times 5$ convolutional kernel have a central role in recognizing affective states and minimizing inter-subject variability.

## 6 Conclusion

Here, we presented a robust physiological model for the recognition of human emotions, called DPAN. This model is based on ConvLSTM and a new temporal margin-based loss function. Our proposed system helps to bridge the gap between the low-level physiological sensor representations and the high-level context-sensitive interpretations of human emotions. It extracts physiological spectral-temporal features from bipolar EEG signals underlying brain lateralization and a PPG signal. The model then recognizes emotions in such a way that it becomes increasingly confident as it observes more of a specific feeling. Our experimental results obtained using a public dataset showed that our deep physiological learning technology enables recognition rates that significantly outperform state-of-the-art techniques. In fact, we observed an average 15.96% increase in accuracy. An extensive analysis of the relationships between participants' emotion ratings and physiological signal frequencies during the experiment is also presented. We showed that our model captures spectral-temporal correlations better while recognizing emotions.

Neuro-scientific findings of emotional lateralization as a differentiator of valence levels have motivated recent studies to implement related features for realizing the lateralization. However, as we describe in Section 2, the emotional lateralization is not always congruent with the valence of the feeling. Furthermore, accurate determination of the threshold of the differentiator has suffered from inter- and intra-subject variability. This issue has hindered our ability to develop reliable affect models for emotion classification. Through our discoveries in Section 5, we were able to shed light on the fact that the effects of physiological changes captured by our system on emotion are partially consistent with other theoretical studies described in Section 2. From our experimental results in Section 4, we showed that learning physiological spectral-temporal patterns and progression patterns of emotion in training could improve performance in emotion recognition concerning emotional lateralization.

Furthermore, our foundation may extend the outlook on the lateralization mechanism from a theoretical to a methodological perspective. For instance, not only vertical but also diagonal symmetry between the electrodes may be a valid indicator for the detection of emotional changes. As shown in Fig. 9 and TABLE 1, a diagonal pair of electrodes (F8 and PO3) have relative activation due to the emotional stimulus of a pleasurable sound. One of the other potential factors that extend improvement in emotion recognition from a theoretical to a methodological perspective is the learning of temporal physiological patterns in emotion elicitation progression. Mechanisms involved in emotion elicitation and its effects on the emotional response are described as two-step processes whereby the presence of a stimulus elicits a particular emotion and produces an emotional response. Elicited emotions are typically considered brief episodes with quick onsets and short durations [18]. The short duration of feeling has been a challenging issue in the field of affective neuroscience. There has been some research on changes in brain activity and functional connectivity induced by instantaneous emotions, although these studies have not probed emotion after the instant episodes [51].

We showed our temporal margin-based classification loss beneficially impacts our ability to discriminate margins between recognition scores conforming to our rationale about the monotonicity described in Section 3. We extend the use of the monotonic increment of the margin scores in training to realizing the theoretical perspective in emotion elicitation and its duration. Since our proposed system learns emotion elicitation progression, it enables us not only to recognize emotional states but also to detect the start point of an emotion after observing only a fraction of the emotion. For many real applications, it is desirable to detect the emotion as early as possible for better interactions with humans. Such interactions can be used in a health care system to manage stress-related illnesses before the development of long-term mental sickness. This early detection ability in emotion recognition is based on the detection of the emotion segment after observing only a fraction of the emotion. It is important since physiological signals have a relatively long period in the affect mechanism. Our next work will study early detection and investigate the efficacy of our proposed system for recognizing emotions in real life and show the progression patterns of the recognition in training.

## ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00432, Development of non-invasive integrated BCI SW platform to control home appliances and external devices by user's thought via AR/VR interface) and (No.2017-0-01778, Development of Explainable Human-level Deep Machine Learning Inference Framework).

## REFERENCES

[1] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica e3a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *Proceedings of the 2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*. IEEE, 2014, pp. 39–42.

[2] R. W. Picard, S. Fedor, and Y. Ayzenberg, "Multiple arousal theory and daily-life electrodermal activity asymmetry," *Emotion Review*, 2015.

[3] O. Yürüten, J. Zhang, and P. H. Pu, "Predictors of life satisfaction based on daily activities from mobile sensor data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2014, pp. 497–500.

[4] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*. MIT Press, 2015, pp. 802–810.

[5] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 03, pp. 715–734, 2005.

[6] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[7] D. Sander and K. Scherer, *Oxford companion to emotion and the affective sciences*. Oxford University Press, 2009.

[8] J. Panksepp, "Affective consciousness: Core emotional feelings in animals and humans," *Consciousness and Cognition*, vol. 14, no. 1, pp. 30–80, 2005.

[9] H. A. Sackeim, R. C. Gur, and M. C. Saucy, "Emotions are expressed more intensely on the left side of the face," *Science*, vol. 202, no. 4366, pp. 434–436, 1978.

[10] R. Adolphs, H. Damasio, D. Tranel, and A. R. Damasio, "Cortical systems for the recognition of emotion in facial expressions," *Journal of neuroscience*, vol. 16, no. 23, pp. 7678–7687, 1996.

[11] R. J. Davidson and K. Hugdahl, *Brain asymmetry*. Mit Press, 1996.

[12] H. A. Demaree, D. E. Everhart, E. A. Youngstrom, and D. W. Harrison, "Brain lateralization of emotional processing: historical roots and a future incorporating dominance," *Behavioral and Cognitive Neuroscience Reviews*, vol. 4, no. 1, pp. 3–20, 2005.

[13] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen, "Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I." *Journal of Personality and Social Psychology*, vol. 58, no. 2, p. 330, 1990.

[14] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[15] A. Clerico, R. Gupta, and T. H. Falk, "Mutual information between inter-hemispheric eeg spectro-temporal patterns: A new feature for automated affect recognition," in *Proceedings of the 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2015, pp. 914–917.

[16] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[17] R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Inc, 1996.
[18] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005.
[19] P. C. Petrantonakis and L. J. Hadjileontiadis, "Adaptive emotional information retrieval from eeg signals in the time-frequency domain," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2604–2616, 2012.
[20] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.
[21] S. Smith, "Eeg in the diagnosis, classification, and management of patients with epilepsy," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 2, pp. ii2–ii7, 2005.
[22] N. Lovato and M. Gradisar, "A meta-analysis and model of the relationship between sleep and depression in adolescents: recommendations for future research and clinical practice," *Sleep Medicine Reviews*, vol. 18, no. 6, pp. 521–529, 2014.
[23] J. Wolpaw and E. W. Wolpaw, *Brain-computer interfaces: principles and practice*. OUP USA, 2012.
[24] Y. Chae, J. Jeong, and S. Jo, "Toward brain-actuated humanoid robots: asynchronous direct control using an eeg-based bci," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1131–1144, 2012.
[25] B. H. Kim, M. Kim, and S. Jo, "Quadcopter flight control using a low-cost hybrid interface with eeg-based classification and eye tracking," *Computers in biology and medicine*, vol. 51, pp. 82–92, 2014.
[26] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from eeg signals," *IEEE Transactions on Affective Computing*, 2017.
[27] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.
[28] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, 2016.
[29] Z. Zhang, Z. Pi, and B. Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.
[30] H. Chigira, A. Maeda, and M. Kobayashi, "Area-based photoplethysmographic sensing method for the surfaces of handheld devices," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2011, pp. 499–508.
[31] Y. Lyu, X. Luo, J. Zhou, C. Yu, C. Miao, T. Wang, Y. Shi, and K.-i. Kameyama, "Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, 2015, pp. 857–866.
[32] D. Sun, P. Paredes, and J. Canny, "Moustress: detecting stress from mouse motion," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2014, pp. 61–70.
[33] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics," *Scientific Reports*, vol. 4, 2014.
[34] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, 2013.
[35] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.
[36] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.
[37] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[38] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, 2014.
[39] H. J. Yoon and S. Y. Chung, "Eeg-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm," *Computers in biology and medicine*, vol. 43, no. 12, pp. 2230–2237, 2013.
[40] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *The Scientific World Journal*, vol. 2014, 2014.
[41] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *arXiv preprint arXiv:1601.02197*, 2016.
[42] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.
[43] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.
[44] E. Kübler-Ross and D. Kessler, *On grief and grieving: Finding the meaning of grief through the five stages of loss*. Simon and Schuster, 2014.
[45] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using eeg signals and facial expressions," in *Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.
[46] W. Lu and J. C. Rajapakse, "Approach and applications of constrained ica," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 203–212, 2005.
[47] J. A. Urigüen and B. Garcia-Zapirain, "Eeg artifact removalstate-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, no. 3, p. 031001, 2015.
[48] J. Armony and P. Vuilleumier, *The Cambridge handbook of human affective neuroscience*. Cambridge university press, 2013.
[49] H. W. Cole and W. J. Ray, "Eeg correlates of emotional tasks related to attentional demands," *International Journal of Psychophysiology*, vol. 3, no. 1, pp. 33–41, 1985.
[50] J. A. Onton and S. Makeig, "High-frequency broadband modulation of electroencephalographic spectra," *Frontiers in Human Neuroscience*, vol. 3, p. 61, 2009.
[51] H. Eryilmaz, D. Van De Ville, S. Schwartz, and P. Vuilleumier, "Impact of transient emotions on functional connectivity during subsequent resting state: a wavelet correlation approach," *NeuroImage*, vol. 54, no. 3, pp. 2481–2491, 2011.

**Byung Hyung Kim** received the B.S. degree in computer science from Inha University, Incheon, Korea, in 2008, and the M.S. degree in computer science from Boston University, Boston, MA, USA, in 2010. He is currently working toward the Ph.D. degree at KAIST, Daejeon, Korea. His research interests include affective computing, brain-computer interface, computer vision, assistive and rehabilitative technology, and cerebral asymmetry and the effects of emotion on brain structure.

**Sungho Jo** (M'09) received the B.S. degree in school of mechanical & aerospace engineering from the Seoul National University, Seoul, Korea, in 1999, the S.M. in mechanical engineering, and Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2001 and 2006 respectively. While pursuing the Ph.D., he was associated with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Laboratory for Information Decision and Systems (LIDS), and Harvard-MIT HST NeuroEngineering Collaborative. Before joining the faculty at KAIST, he worked as a postdoctoral researcher at MIT media laboratory. Since December in 2007, he has been with the department of computer science at KAIST, where he is currently Associate Professor. His research interests include intelligent robots, neural interfacing computing, and wearable computing.