IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2020, accepted October 27, 2020, date of publication November 9, 2020, date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036877

Multimodal Attention Network for Continuous-Time Emotion Recognition Using Video and EEG Signals

DONG YOON CHOI, (Graduate Student Member, IEEE), DEOK-HWAN KIM, (Member, IEEE), AND BYUNG CHEOL SONG, (Senior Member, IEEE)
Department of Electronic Engineering, Inha University, Incheon 22212, South Korea

Corresponding author: Byung Cheol Song (bcsong@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government [Ministry of Science and ICT (MSIT)], Artificial Intelligence Convergence Research Center (Inha University), under Grant 2020-0-01389, and in part by the Industrial Technology Innovation Program through the Ministry of Trade, Industry, and Energy (MI, South Korea), Development of Human-Friendly Human–Robot Interaction Technologies Using Human Internal Emotional States, under Grant 10073154.

ABSTRACT Emotion recognition is a very important technique for ultimate interactions between human beings and artificial intelligence systems. For effective emotion recognition in a continuous-time domain, this article presents a multimodal fusion network which integrates video modality and electroencephalogram (EEG) modality networks. To calculate the attention weights of facial video features and the corresponding EEG features in fusion, a multimodal attention network, that is utilizing bilinear pooling based on low-rank decomposition, is proposed. Finally, continuous domain valence values are computed by using two modality network outputs and attention weights. Experimental results show that the proposed fusion network provides an improved performance of about 6.9% over the video modality network for the MAHNOB human computer interface (MAHNOB-HCI) dataset. Also, we achieved the performance improvement even for our proprietary dataset.

INDEX TERMS Emotion recognition, video, EEG, multimodality, multimodal fusion, attention.

I. INTRODUCTION

Recognition of human emotions is a key technology for ultimate human–robot interaction (HRI). In addition, emotion recognition has received much attention in the field of artificial intelligence. Conventional emotion recognition algorithms distinguished emotion categories by detecting changes in facial expressions [1], [2]. Recently, various emotion recognition mechanisms based on convolutional neural network (CNN) which are trained in an end-to-end manner have been developed and showed reliable performance [3], [4].

On the other hand, there were many attempts to recognize human emotions from tone information of voice signals [5]. However, since the voice information is temporally sparse, those voice tone-based emotion recognition schemes have a fundamental limitation in extracting consecutive emotions. Recently, several emotion recognition algorithms using EEG, which is an electrical bio-signal generated in the human brain have been reported [6]–[8]. For example, a frequency-domain feature such as power spectral density (PSD) is extracted, and a typical machine learning algorithm is applied to recognize emotions [6]. A few EEG-based algorithms [7], [8] employed the inherent asymmetry characteristics between EEG channels as salient features for deep learning-based emotion classification. However, the conventional techniques have a structure that recognizes only a single emotion per tens of seconds of video clip. So it is hard to say that they can ultimately perceive emotional changes in the continuous-time domain.

Busso *et al.* proposed an emotion recognition mechanism based on multimodal signals where two or more signals among video, voice, and bio-signals are employed for emotion recognition [10]. It was reported that multimodal approach was superior to conventional unimodal approaches. On the other hand, a world-wide emotion recognition challenge called Emotion Recognition in the Wild (EmotiW) [11] ranks competing algorithms [12], [13] through performance evaluation for Acted Facial Expressions in the Wild (AFEW) dataset that is composed of wild audio-visual data excerpted from sitcoms and movies. The AFEW dataset

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

consists of seven discrete-domain emotion classes. Such a discrete-domain dataset does not represent complex emotions due to the limitation of the number of emotional classes, and it cannot express the intensity of emotions. Therefore, studies using continuous emotional labels such as valence and arousal have become active recently. Psychologically, valence is an index that can differentiate between positive and negative emotions, and arousal is an index that can discriminate between high and low emotions. Note that valence and arousal can be mapped to discrete-domain emotion classes if necessary and have an advantage of expressing the emotion intensity.

Meanwhile, Soleymani *et al.* [6] proposed a valence regression algorithm, which applied long short-term memory (LSTM) to multimodal data consisting of EEG signals and facial landmarks. Soleymani *et al.* adopted relatively simple modality fusion, such as data concatenation and simple averages. So, their method has a critical disadvantage that it cannot effectively utilize mutual information between different modalities.

In order to maximize complementarity between video and EEG modalities, this article proposes a multimodal attention network that adaptively uses two modality signals according to the input state. The proposed multimodal attention network analyzes intermediate features obtained from a video modality network and an EEG modality network, and determines the attention weight of each modality. As a result, the multimodal attention network contributes to improve the overall emotion recognition accuracy by selecting a more reliable one between video and EEG. Experiments on the MAHNOB-HCI dataset [14] shows that the emotion recognition scheme based on the proposed multimodal attention network outperforms a state-of-the-art multimodal emotion recognition method [6]. In addition, we evaluated the proposed method for the ASIA dataset which consists of video and EEG signals acquired from 32 Asian subjects. As a result, the proposed method works well even for the ASIA dataset which was produced according to the same protocol as MAHNOB-HCI by our research team.

The contributions of this article are as follows.

- We propose an algorithm that recognizes emotional changes in continuous-time domain by using video and EEG signals simultaneously, and experimentally verify that the combination of different modality signals can be synergistic in improving emotion recognition performance.
- We propose a modality attention network based on low-rank decomposition of multi-layer structure to obtain attention weights for video and EEG modalities.
- A proprietary dataset, i.e., ASIA dataset is produced by concurrently acquiring Asian facial expressions and their EEG signals, and the performance of the proposed method is evaluated, on the ASIA dataset.

This article is organized as follows. In Section II, some related works are explained. Next, we describe in detail the MAHNOB-HCI dataset and the ASIA dataset, and we then introduce the proposed algorithm. Finally, we conclude this article with experimental results and a discussion.

II. RELATED WORKS

This section briefly introduces previous studies on emotion recognition and multimodal deep learning.

A. VIDEO-BASED EMOTION RECOGNITION

For a long time, human emotions have been regarded as being the same as facial expressions. So, facial video–based methods have been intensively developed for emotion recognition. Tong *et al.* [1] detected activated areas in the face, and defined action units (AU) according to their positions and shapes, and then classified facial expressions according to AU types or AU combinations. With rapid development of algorithms to extract landmarks, i.e., locations of facial key points, many facial expression recognition (FER) algorithms using landmark information have been proposed [2]. Recently, CNN-based FER algorithms have been developed because they can be trained in an end-to-end manner without extracting any features such as AUs and landmarks [3], [4].

On the other hand, CK+ [15] and MMI [16] datasets were commonly used for training the CNN-based FER techniques. However, most of the facial expressions in the datasets were artificially created or acted. In other words, CK+ and MMI are distant from natural facial expressions in reality. So, a few datasets collecting natural facial expressions have been attracting attention recently. The representative dataset is the AFEW dataset [11] which consists of video clips collected from movies and sitcoms and are classified under seven labels: 'anger', 'disgust', 'fear', 'happiness', 'sadness', 'surprise', and 'neutral'. For example, three-dimensional CNNs [13] and CNN-LSTM–based networks [46] were proposed for FER on the AFEW dataset. However, since those algorithms basically follow a way of classifying several discrete emotions, they cannot represent complex emotions or emotion intensities.

Recently, studies on FER in the continuous domain have been started. We introduce two well-known datasets for continuous domain FER. One is the AFEW valence arousal (AFEW-VA) dataset [47] that reconfigures discrete labels of the AFEW dataset, and the other is the Affect-in-the-Wild (Aff-wild) dataset, which annotates human emotional responses for YouTube contents [46]. Continuous domain emotions such as valences and arousals are regressed for AFEW-VA and Aff-wild datasets. To do this, CNN-LSTM–based networks are widely used [46], [48]. On the other hand, the AFEW-VA dataset composed of video clips from movies and sitcoms is called content-centered dataset, and the Aff-wild dataset is user-centered because it consists of videos that monitor the actual emotional responses.

B. EEG-BASED EMOTION RECOGNITION

People may have negative or positive emotions without revealing facial expressions. In order to accurately recognize authentic emotion, it is preferable to use bio-signals as well as facial videos. EEG is a representative bio-signal used for emotion recognition. By analyzing locations and waveforms

of the activated EEG signals, the emotional state can be estimated.

The Database for Emotion Analysis using Physiological signals (DEAP) dataset [17] and the MAHNOB-HCI dataset [18], which sensed EEG signals with facial videos when humans feel emotions, were released in 2012. Since then, many EEG-based emotion recognition algorithms have been developed [6]–[8], [52], [53]. Soleymani *et al.* proposed a real-time method to regress the valence values in the MAHNOB-HCI dataset [6]. In [6], PSD features were extracted from EEG signals, and the emotion states were estimated by applying a continuous-time conditional random field (CRF) and LSTM to PSD features. However, Soleymani *et al.*'s method had lower accuracy than conventional video-based methods. Kim and Jo explored that the channels on which EEG signals are activated differ depending on the emotional state [7]. They extracted significant connectivity features between EEG channels, and then applied convolutional LSTM (ConvLSTM) to the connectivity features so as to regress emotional states. However, since Kim and Jo's technique could estimate a single emotion per minute, it is not suitable for continuous-time emotion recognition. Song *et al.* presented an emotion recognition algorithm that applied a dynamical graph convolutional neural network to EEG features [8]. Chen *et al.* combined temporal and frequency features extracted from EEG signals, and applied CNN to the combined features [52]. Zhong *et al.* proposed a regularized graph neural network that analyzes topological relations between EEG channels [53]. However, the previous methods do not aim at emotion recognition in the continuous-time domain.

In addition to the DEAP and MAHNOB-HCI datasets, several multimodal datasets associated with bio-signals were produced. For example, Subramanian *et al.* built a multimodal dataset to recognize personality as well as emotion [56]. This dataset consists of video signals as well as bio-signals such as EEG, ECG, and GSR. Miranda-Correa *et al.* acquired multimodal data on individual as well as group [57]. Each data in this dataset has emotion, personality, mood and social context information. However, since the emotion labels of the above-mentioned datasets were annotated by subject self-assessment, the datasets are not suitable for emotion recognition in the continuous-time domain, i.e., the purpose of this study.

Note that Soleymani *et al.* proposed a multimodal emotion recognition algorithm using video and EEG signals simultaneously [6]. To our knowledge, this is the only method for recognizing emotions through the fusion of video modality and EEG modality in the continuous-time domain.

However, according to [6], the performance of the feature-level fusion and decision-level fusion networks is worse than that of the video modality network only.

C. MULTIMODAL DEEP LEARNING

With rapid development of deep learning technology, the so-called multimodal deep learning method (which simultaneously processes multimodal data) is actively being studied. For example, Ngiam *et al.* proposed a bimodal deep auto-encoder [55]. The bimodal deep auto-encoder was trained to recover each modality information from fused features. Ren *et al.* proposed an LSTM network that can learn multimodal data, i.e., a multimodal LSTM [54]. However, the above-mentioned networks have a limitation in fusion performance because they increase the number of modality networks in proportion to the number of modalities. In other words, they have a naïve fusion structure. As shown in Fig. 1, recent multimodal deep learning networks are based on a fully-connected (FC) layer structure, adopting an approach of concatenating input modalities or intermediate features for delivery to the next layer [18].

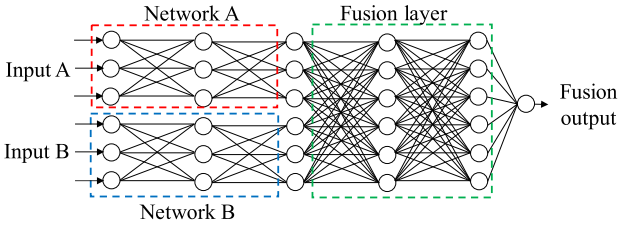

FIGURE 1. A general fully connected layer–based multimodal deep learning structure.

FC layer–based multimodal deep learning has an advantage of simplicity, but it does not consider the characteristic where features of different modalities may have different distributions and intensities. So, the representation capacity of the fused feature may not be sufficient [40]. In order to overcome this problem, feature fusion using bilinear pooling [51] was used, which is popular in the field of visual question answering (VQA) [37], [40]–[42]. In general, bilinear pooling is expressed as seen in Eq. (1):

$$h_i = z_A^T \boldsymbol{W}_i z_B = \boldsymbol{W}_i \cdot (z_A \otimes z_B) \tag{1}$$

where h_i indicates the i-th value of a fused feature vector $\boldsymbol{h} \in \mathbb{R}^N$, and $z_A \in \mathbb{R}^L$ and $z_B \in \mathbb{R}^M$ are the feature vectors from two different modality networks, A and B, as shown in Fig. 1. $\boldsymbol{W}_i \in \mathbb{R}^{L \times M}$ is the projection matrix that is the trainable parameter. In Eq. (1), since all possible pairwise interactions between two feature vectors z_A and z_B are obtained through the outer product, a richer representation than the concatenated feature can be obtained by bilinear pooling [42]. Recently, bilinear pooling has been used as a technique to recognize emotions via fusion of video, audio, and text features [39].

On the other hand, since N projection matrices are required to obtain a fused feature vector $\boldsymbol{h}$ of length N, the weight parameter $\boldsymbol{W} \in \mathbb{R}^{N \times L \times M}$ is a 3D tensor. So, bilinear pooling requires a greater number of weight parameters than the FC layer–based approach, which can be a memory burden in the calculation process. Bilinear pooling may also cause overfitting in the learning process. To solve these problems, various studies were presented. For instance, Yu *et al.* proposed a multimodal factorized bilinear pooling that performs element-wise multiplication after projecting feature vectors

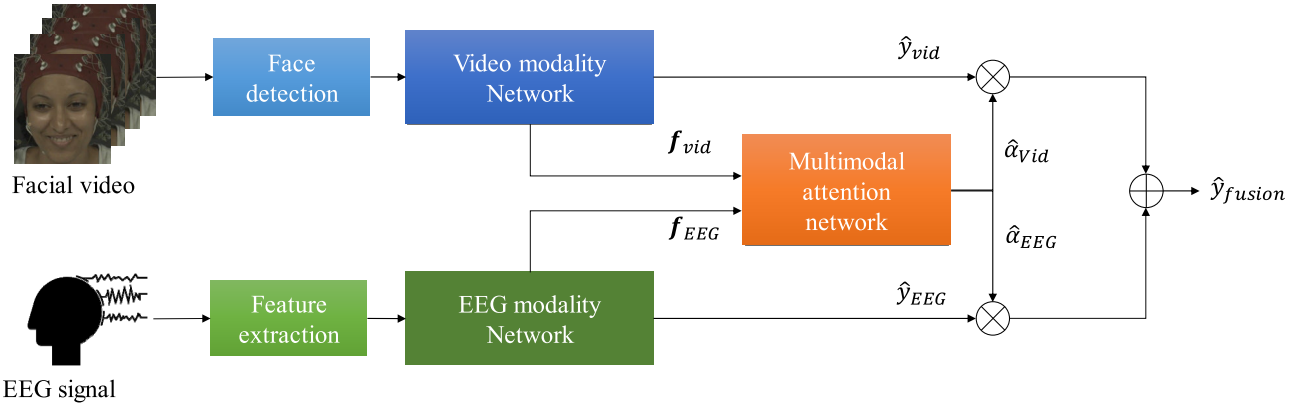

FIGURE 2. A block diagram of the proposed method.

of each modality in advance [40]. Fukui *et al.* achieved the compacter operation by applying fast Fourier transform (FFT) to the outer product process of the features of each modality [41]. Kim *et al.* [42] and Liu *et al.* [39] adopted low-rank decomposition. Ben-Younes *et al.* proposed a method to simplify the bilinear pooling model using block-term decomposition [37].

The bilinear pooling techniques mentioned so far have focused on the fusion of different modality features. There were several studies from the perspective of the fusion network architecture. Vielzeuf *et al.* proposed CentralNet, which converges different modality features step by step by using several levels of interim features available in each modality network [35]. There was also an attempt to use reinforcement learning–based AutoML to find the optimal fusion network architecture [38]. AutoML is effective in finding the optimal combination of hyper-parameters from each network layer and the layer from which the features of each modality are extracted. However, the AutoML approach has a disadvantage in that the learning time is too huge.

On the other hand, there are various ways to use fused information in multimodal deep learning. The first example is to extract the target output directly from the fused feature [39]. Also, each modality weight can be extracted from the fused feature and a weighted sum or hard thresholding is applied to the multimodal output according to modality weight. The attention-based method [36], [50] and the gated fusion method [20]–[22] are the representative hard-thresholding approaches. As an example of an application to emotion recognition using video and audio modalities, Chen and Jin concatenated each modality input and interim features, and then adjusted the output weights of the video and audio modalities by calculating the attention weight using the FC layer [36].

III. PROPOSED METHOD

The proposed method works as shown in Fig. 2. Video modality and EEG modality have independent networks. The output features of the two networks are fused through the attention network, which calculates the attention weight of each modality. The weighted average of two modality outputs becomes the final emotion information.

TABLE 1. The comparison of ideal fusion with each modality for the MAHNOB-HCI dataset in terms of RMSE. Here, ideal fusion indicates that RMSE is calculated by manually selecting a better output among video and EEG modalities.

Modality	RMSE
Video	0.0393±0.0044
EEG	0.0485±0.0047
Ideal fusion	0.02963±0.0022

Table 1 shows the 10-fold validation result of each modality network on the MAHNOB-HCI dataset. The video and EEG networks, which were trained independently, provided RMSEs (root mean of squared errors) of 0.0393 and 0.0485, respectively. Details of the two modality networks will be described in Section IV.B and IV.C. On the other hand, the ideal fusion where an RMSE is calculated by manually selecting a better output among video modality and EEG modality had an excellent RMSE of 0.0296. Based on this result, if we can analyze the characteristics of two modalities and calculate their weights, we can achieve a synergy of video modality and EEG modality for emotion recognition.

In order to effectively fuse two modality features like ideal fusion, we employ bilinear pooling based on low-rank decomposition. The bilinear pooling is realized via a multimodal attention network with a multi-layer structure, which will be depicted in Section IV.D. Experimental results demonstrate that the proposed multimodal attention network accomplishes high performance with low computational complexity and less memory.

A. OVERALL FRAMEWORK

The proposed method regresses valences in continuous-time domain. Considering causality, a short sequence from the current time stamp to the previous 2 seconds becomes an input unit for learning the proposed method. As shown in Fig. 2, the proposed method briefly consists of a video modality network, an EEG modality network, and a multimodal attention network. The multimodal attention network, which is the key module of the proposed method, takes the intermediate features of two modality networks, and determines the weights of two modalities by comparing the features.

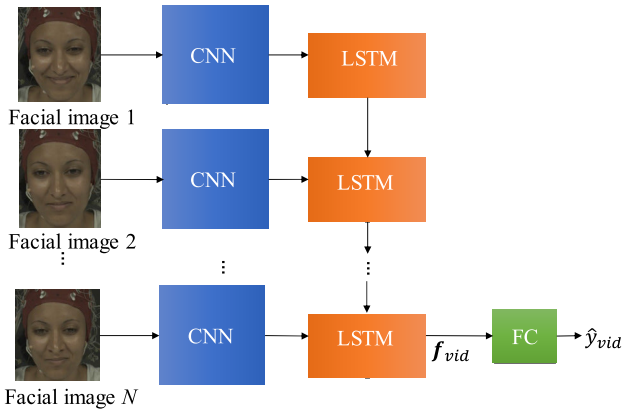

FIGURE 3. A block diagram of the video modality network.

TABLE 2. EEG features according to domain.

Domain	EEG Features
Time (5)	Mean, max, min, 1st difference, normalized 1st difference
Frequency (12)	Mean, max, integral of PSD in 4 bands (slow alpha, alpha, beta, gamma)
Time–frequency (20)	Mean, max, absolute, log, abs(log) of DWT in 4 bands (slow alpha, alpha, beta, gamma)

Since two modality networks are trained independently and fusion is done through the multimodal attention network at the output stage, the proposed method is easier to learn than the conventional method [14], which conducts end-to-end learning with pre-fused multimodal input. Also, because the proposed method determines the weight of each modality based on the feature of the modality, it can get higher accuracy than the previous work [6] using a simple average or fixed-weight sum.

B. VIDEO MODALITY NETWORK

The video modality network operates as in Fig. 3. First, the facial images are cropped from an input video sequence. Here, a face detection algorithm called the Single Shot Scale-invariant Face Detector (S3FD) [24] is used. Next, CNN extracts the features from each facial image, and the CNN features pass through LSTM [25] to generate a video feature $\boldsymbol{f}_{vid}$ whose length is 2048. Finally, a valence value $\hat{y}_{vid}$ is obtained by FC.

1) DEEP CONVOLUTIONAL ENCODER

Prior to the LSTM, each facial image is converted into a one-dimensional (1D) feature vector by CNN-based deep convolutional encoder. The deep convolutional encoder is based on a famous DenseNet [26]. The 1D feature vector passing through the FC2 layer of DenseNet becomes the output of this encoder. Here, the size of each input image is 224×224 and the 1D feature vector is 4096 in length.

On the other hand, the deep convolutional encoder is not trained with the MAHNOB dataset or the ASIA dataset. Since the datasets were acquired from a small group of subjects (20 to 30 persons), they are too small to train any deep neural networks. Besides, there is a possibility of overfitting due to data similarity during the training process. The datasets were actually made by sampling several similar frames from a video sequence. Note that the main purpose of the deep convolutional encoder is to convert an image data into an efficient 1D feature vector. Therefore, the deep convolutional encoder is pre-trained with a well-known ImageNet [27], which is a huge–scale image classification dataset with 1000 classes, and the pre-trained network is fine-tuned with the FER2013 dataset [28], which consists of 32,298 images with 7 emotional classes. During this fine-tuning process using the FER2013 dataset, the parameters of all layers of the deep convolutional encoder are updated.

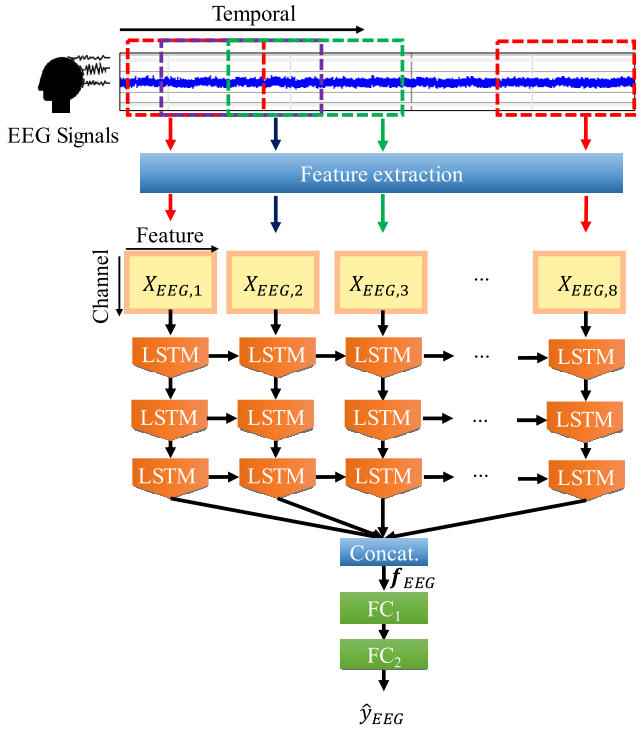

FIGURE 4. The EEG feature extraction process and the LSTM model of EEG modality network.

C. EEG MODALITY NETWORK

1) EEG FEATURE EXTRACTION ENCODER

As shown in Fig. 2, EEG features must be extracted before the EEG modality network. The upper part of Fig. 4 illustrates how EEG features are extracted from EEG signals in a sliding-window manner. The length of the sliding window is set to two seconds. Features of three domains (time, frequency, and time–frequency) are utilized as in Table 3. Totally 37 EEG features are extracted per channel, and they are converted into 1D vectors. The dimension of 1D vector in the MAHNOB-HCI dataset, is 1184 (= 32channels × 37features). The dimension of the 1D vector in the ASIA dataset is 444 (= 12 × 37).

Frequency domain features were mainly used because of the superior spatial resolution of EEG signals. Since the outputs in different frequency bands are good for identifying

TABLE 3. EEG features of the asia dataset.

Extracted features
PSD in alpha (8-12 Hz), beta (13-30 Hz), gamma (30-45 Hz) bands for all EEG channels (12 channels): F7, F8, Fz, T3, C3, C4, T4, T5, Pz, T6, O1 and O2

different emotional states, the frequency domain features such as PSD were dominantly used in previous studies. This article also adopts PSD which is divided into four bands: slow alpha (8-10Hz), alpha (8-12.9Hz), beta (13-29.9Hz), and gamma (30-50Hz). We additionally employ the features of time and time–frequency domains for better performance as in Table 3. The time domain features are advantageous for detecting emotional changes over time [16], [17]. On the other hand, the time–frequency features based on the discrete wavelet transform (DWT) has been widely used in the speech processing field [19] and recently in the field of emotion recognition [45]. Mean, max, abs are calculated in each frequency band of DWT, and log, Abs (Log) of DWT are used in this article.

2) MULTI-LAYER LSTM-BASED REGRESSOR

The extracted 1D vectors are input to the subsequent EEG modality network. The structure of the EEG modality network is shown in the lower part of Fig. 4. It basically consists of three LSTM layers and two FC layers. First, the EEG features corresponding to each time window is divided into 0.25sec intervals and input into the LSTM layers. Since the length of the sliding window is two seconds, a total of eight sub-features are input to the LSTM layers. The structure of the multi-layer LSTM is designed to enable deep inference, and the number of layers is determined experimentally. The hidden layer size of the LSTM is set to 425. A dropout layer to prevent overfitting is inserted between LSTM layers. The final output of the LSTM layers is concatenated to generate $\boldsymbol{f}_{EEG}$. The dimension of $\boldsymbol{f}_{EEG}$ is 3400 and the output of FC1 becomes a 1D vector with dimension 8. FC2 layer outputs a valence value, $\hat{y}_{EEG}$. The range of the valence value is $[-0.5, 0.5]$.

D. MULTIMODAL ATTENTION NETWORK

Vielzeuf *et al.* [35] and Pérez-Rúa *et al.* [38] proposed to fuse multi-modal features through multiple layers and refine the fused features. Inspired by [35], [38], we propose a new multi-modal attention network that extends the existing bilinear pooling-based fusion process into a multi-layer structure.

Figure 5 describes the proposed multimodal attention network. The network calculates the attention weight of each modality jointly using the heterogeneous features produced by the video modality network and the EEG modality network.

First, an initial fused feature $\boldsymbol{f}_{fusion1}$ is created by fusing the video feature $\boldsymbol{f}_{vid}$ and the EEG feature $\boldsymbol{f}_{EEG}$. At the same time, each modality feature is compressed via the FC layer (FC1), and then $\boldsymbol{f}_{vid1}$ and $\boldsymbol{f}_{EEG1}$ are derived. Second, three interim features $\boldsymbol{f}_{vid1}$, $\boldsymbol{f}_{EEG1}$, $\boldsymbol{f}_{fusion1}$ are fused to

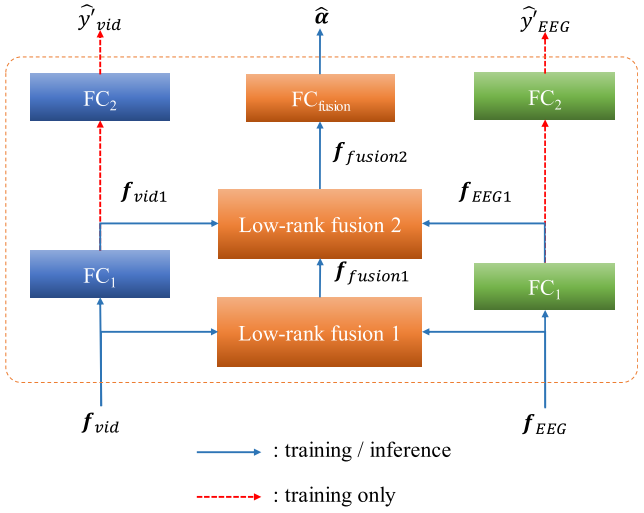

FIGURE 5. The EEG feature extraction process and the LSTM model of EEG modality network.

produce $\boldsymbol{f}_{fusion2}$. Finally, modality attention weight $\hat{\alpha}$ is generated as $\boldsymbol{f}_{fusion2}$ passes through an FC layer (FC$_{\text{fusion}}$). Both of the above-mentioned fusions are realized by bilinear pooling based on low-rank decomposition. The first fusion is based on the bilinear model, and the second fusion is based on the trilinear model. The details are in the following subsection.

If attention weight $\hat{\alpha} = [\hat{\alpha}_{vid}, \hat{\alpha}_{EEG}]$ is available, valence information $\hat{y}_{fusion}$ is finally produced in the form of a weighted sum according to Eq. (2):

$$\hat{y}_{fusion} = \bar{\alpha}_{Vid}\hat{y}_{vid} + \bar{\alpha}_{EEG}\hat{y}_{EEG} \tag{2}$$

where $\hat{y}_{vid}$ and $\hat{y}_{EEG}$ are valences output from video and EEG modality networks, respectively; and $\bar{\alpha}$ normalizes $\hat{\alpha}$ for weighted sum operation.

1) LOW-RANK DECOMPOSITION–BASED BILINEAR POOLING

Conventional bilinear pooling–based multimodal fusion requires massive weight parameter sizes and memory, and often causes overfitting. To solve this problem, we propose a new multimodal attention network using low-rank decomposition [39], [42]. First, based on Eq. (1), the outer product of $\boldsymbol{f}_{vid}$ and $\boldsymbol{f}_{EEG}$ is computed. This can be expressed by Eq. (3):

$$\boldsymbol{F} = \boldsymbol{f}_{vid} \otimes \boldsymbol{f}_{EEG} \tag{3}$$

On the other hand, if low-rank decomposition is applied to weight $\boldsymbol{W}$ configured in the tensor form, $\boldsymbol{W}$ can be approximated as in Eq. (4):

$$\boldsymbol{W} \simeq \sum_{i=1}^{r} \boldsymbol{w}_{vid}^{(i)} \otimes \boldsymbol{w}_{EEG}^{(i)} \tag{4}$$

where r is the rank of the tensor. Substituting Eq. (3) and Eq. (4) into Eq. (1), $\boldsymbol{f}_{fusion1}$ is represented by

$$\begin{aligned} \boldsymbol{f}_{fusion1} &= \left(\sum_{i=1}^{r} \boldsymbol{w}_{vid}^{(i)} \otimes \boldsymbol{w}_{EEG}^{(i)}\right) \cdot \left(\boldsymbol{f}_{vid} \otimes \boldsymbol{f}_{EEG}\right) \\ &= \left(\sum_{i=1}^{r} \boldsymbol{w}_{vid}^{(i)} \cdot \boldsymbol{f}_{img}\right) \circ \left(\sum_{i=1}^{r} \boldsymbol{w}_{EEG}^{(i)} \cdot \boldsymbol{f}_{EEG}\right) \end{aligned} \tag{5}$$

where ∘ is an element-wise multiplication operator. In Eq. (5), each decomposed weight is multiplied by the corresponding modality feature vector, and fusion is done as element-wise multiplication is applied to both modalities. As a result, bilinear pooling based on low-rank decomposition depends on multiplication of 1D vectors and 2D matrices. So, compared to the conventional bilinear model, which is calculated by multiplication of 2D matrix and 3D tensor, the amount of computation is greatly reduced by the proposed method, and the weight parameter size is also significantly decreased.

A trilinear model fusion of $\boldsymbol{f}_{fusion1}$, $\boldsymbol{f}_{vid1}$, and $\boldsymbol{f}_{EEG1}$ is also approximated by low-rank decomposition as follows:

$$\boldsymbol{f}_{fusion2} = \left(\sum_{i=1}^{r} \boldsymbol{w}_{vid1}^{(i)} \cdot \boldsymbol{f}_{vid1}\right) \circ \left(\sum_{i=1}^{r} \boldsymbol{w}_{EEG1}^{(i)} \cdot \boldsymbol{f}_{EEG1}\right) \circ \left(\sum_{i=1}^{r} \boldsymbol{w}_{fusion1}^{(i)} \cdot \boldsymbol{f}_{fusion1}\right) \quad (6)$$

Rank r is empirically set to 8 through an ablation study. $\boldsymbol{f}_{vid}$ and $\boldsymbol{f}_{EEG}$ are projected through $\boldsymbol{w}_{vid}^{(i)}$ and $\boldsymbol{w}_{EEG}^{(i)}$ into 512-length feature vectors in a low-rank fusion module. The lengths of $\boldsymbol{f}_{fusion1}$ and $\boldsymbol{f}_{fusion2}$ are set to 256.

2) TRAINING PROCESS OF THE ATTENTION NETWORK

The training process of the multimodal attention network is as follows. Assume that the model parameters of each modality network are fixed after a specific learning process. Since the outputs of the video and EEG networks are available, the error of each modality with GT can be computed. If the error of video modality is smaller than that of EEG modality, the attention label $\boldsymbol{\alpha}$ is set to [1,0]. Otherwise, $\boldsymbol{\alpha}$ is set to [0,1]. Then, obtain $\hat{\alpha}$ by inputting the same $\boldsymbol{f}_{vid}$ and $\boldsymbol{f}_{EEG}$ to the attention network. Finally, the error between $\hat{\alpha}$ and $\boldsymbol{\alpha}$ is computed, which is defined as a term of training loss.

On the other hand, as shown in Fig. 5, auxiliary layer FC2 is applied to each modality to output continuous emotion values, i.e., $\hat{y'}_{vid}$ and $\hat{y}'_{EEG}$. So we add the errors to the final training loss. As a result, the training loss for fusion L_{fusion} is defined by

$$L_{fusion} = L_{MSE}(\boldsymbol{\alpha}, \hat{\alpha}) + L_{MSE}(y, \hat{y'}_{vid}) + L_{MSE}(y, \hat{y'}_{EEG}) \quad (7)$$

Here, L_{MSE} indicates the loss function in terms of MSE. L_{fusion} also plays a role in refining two inputs of the second fusion layer, i.e., $\boldsymbol{f}_{vid1}$ and $\boldsymbol{f}_{EEG1}$, which induces the additional effect of improving $\boldsymbol{f}_{fusion2}$. Note that only the operation corresponding to the blue path in Fig. 5 is performed in the inference stage, and the operation on the red path, that is, FC2, is not performed.

IV. DATASETS

Two datasets were used in this study. One is the MAHNOB-HCI dataset, which is valence-tagged as a public dataset, and the other is a proprietary ASIA dataset that our research team created in house to recognize Asians' emotions.

A. MAHNOB-HCI DATASET

Twenty famous commercial movies were chosen to derive emotions from subjects. Each video clip was about 34 to 117 seconds. In the preliminary study, participants helped in selecting videos by reporting their emotions through subject self-assessment.

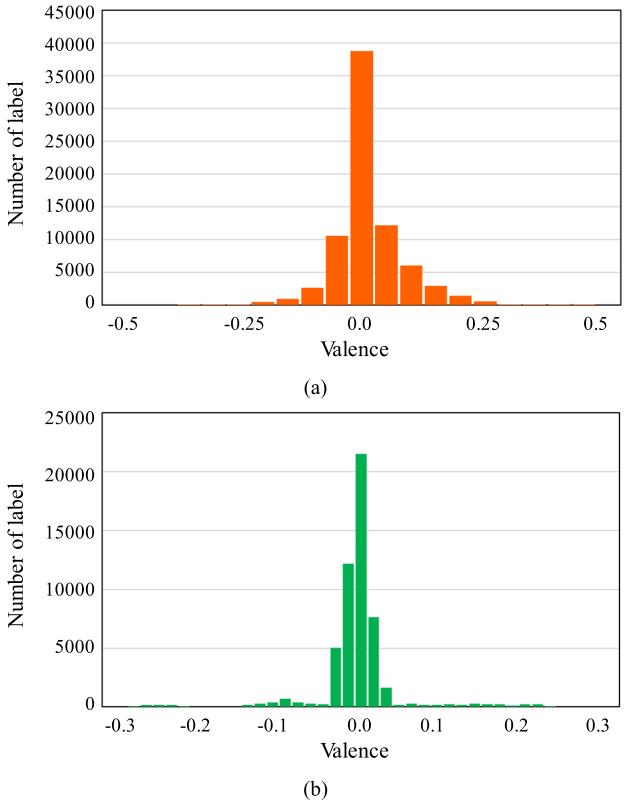

FIGURE 6. Histogram of valence label. (a) MAHNOB-HCI dataset, (b) ASIA dataset.

The total of 28 healthy subjects comprised 12 men and 16 women. EEG signals were acquired using 32 activated electrodes located according to the guidelines of the Biosemi Active II system [33] and the 10-20 International System [34]. While sensing EEG signals, the frontal faces were simultaneously filmed at 720 × 580.60Hz. For a detailed description of database production, refer to [14]. As a result, a total of 239 video–EEG sequences were produced, and all the data included corresponding label information. On the other hand, five educated commentators were employed for consecutive annotations of participants' facial expressions, and they determined the valences of the facial expressions using FEELTRACE [23] and joysticks. Figure 6(a) shows the label distribution of the MAHNOB-HCI dataset. Since 'neutral' emotion is dominant, the distribution is like Gaussian. As far as we know, MAHNOB-HCI is the only dataset that contains bio-signals with emotion labels in the continuous-time domain.

B. ASIA DATASET

We produced a proprietary dataset for Asians according to the same protocol as MAHNOB-HCI. To determine the stimuli videos for Asians, we collected many video candidates from YouTube, and then conducted pilot experiments several times. Finally, eight stimuli videos were selected through questionnaire surveys. They include four videos to cause

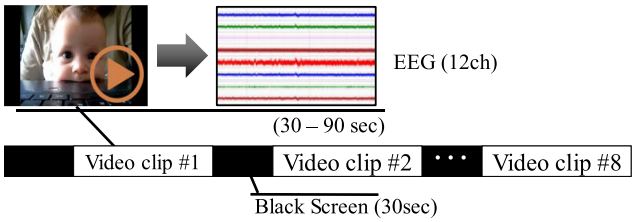

FIGURE 7. Experimental protocol of a trial with the ASIA dataset.

negative emotions such as 'sad' and 'angry', and three videos to cause positive emotions such as 'happy', and one video to cause 'neutral' emotion. The length of each video clip was about 30 to 90 seconds. Each video clip taken was relatively short to reduce the time to adapt to the stimulated emotions. They were short in length but designed to be stimulating enough to evoke the subjects' emotions. Because the video clips that induce each emotion were made with strong stimuli, a black screen was placed between the video clips to prevent emotions from continuing into the next clip (see Fig. 7).

The 32 healthy subjects comprised 19 males and 13 females. EEG signals were obtained with 12 active electrodes placed in accordance with the 10-20 International System standard using Biopack's M150 instrument. Together with the EEG signals, the full-frontal faces of the subjects were shot at 1080p@30Hz. On the other hand, since video running time is somewhat short, EOG was used to remove eye blinking artifacts. As a result, a total of 236 segments were prepared, and each segment included the corresponding valence label information. Table 3 summarizes the EEG features from the ASIA dataset. As with the MAHNOB dataset, we employed five trained annotators for continuous annotation of participants' facial expressions, and the annotators determined the valences of facial expressions using a joystick. The average of five labels obtained from annotators becomes the ground truth (GT) of valence. In order to verify the reliability of GT, we measured the RMSE between GT and the label of each annotator. As a result, the average RMSE of the five labels was sufficiently small, 0.0268. This indicates that five labels obtained from annotators were consistent. Finally, GTs of the ASIA dataset were judged to be reliable enough. On the other hand, Fig. 6(b) is a valence histogram of the ASIA dataset. Compared to Fig. 6(a), the valence labels of the ASIA dataset are more concentrated on 'neutral' emotion than those of the MAHNOB-HCI. This phenomenon indirectly shows that Asians have less emotional responses through facial expressions than Westerners. The ASIA dataset built by our research team will be released later.

V. EXPERIMENTS

We evaluated the performance of the proposed method on the ASIA dataset as well as the MAHNOB-HCI dataset. The proposed method was trained as follows. Note that the video modality network, the EEG modality network, and the multimodal attention network were trained independently. The video modality network and the EEG modality network were trained in advance. Next, the multimodal attention network was trained with parameters of the pre-learned video and EEG networks fixed. For each network learning, Adam optimizer was used. The learning rate was set at 0.001 and was decreased to 1/4 scale every 10 epochs. Also, L_2 regularization was performed, and weight decay was set to 5e-4. The Max epoch was set to 100 cycles. The proposed method was implemented using Pytorch [31], and the computing environments for learning were a Xeon E5-2560 and a GTX1080Ti.

A. MAHNOB-HCI DATASET

We compared the proposed method to Soleymani *et al.*'s [6] for the MAHNOB-HCI dataset. For a fair comparison with [6], we adopted the same environment and dataset. To do this, we received continuous annotation information of MAHNOB-HCI from the authors of [6]. Each sequence in the dataset was annotated with a valence value of 4 Hz. A total of 239 video sequences and EEG segments were used for the experiment. The accuracy of the proposed method was measured using 10-fold cross validation as in [6]. First, 10% of the entire dataset was randomly chosen as a test dataset, and 60% and 40% of the remaining dataset were allocated to a training dataset and a validation dataset, respectively. Here, sampling was performed in video sequence units. Next, such a random selection was iterated ten times. Finally, a total of ten trainings and tests were conducted, and their average becomes the final result. On the other hand, each video sequence was composed of 24 frames by sub-sampling the 60Hz video at 12Hz. For the EEG signals, two seconds was used as unit data with time synchronization at 256Hz and 32-channel data.

In order to evaluate the performance of the proposed multimodal attention network, we additionally implemented and compared decision-level fusion (DLF), as was done in [6]. The experiment for feature-level fusion was not performed because the sampling rates of the video data and the EEG feature were different. In DLF, the output values of two modalities were weighted-averaged. Here, the video and EEG modality weights were set to 0.6 and 0.4, respectively. We experimentally verified that this ratio provides the best performance among various ratios.

Table 4 shows the accuracy comparison in terms of RMSE and Pearson correlation coefficient (PCC). DLF of the proposed method has an RMSE as small as about 0.007, compared to DLF in [6]. The proposed multimodal attention network further reduces RMSE by 0.0007. Also, from the PCC perspective, we can observe a similar trend. The main reason that the proposed method is superior to [6] is in improving the performance of each modality network. Table 4 demonstrates that video modality as well as EEG modality in the proposed method provide higher performance than those of [6]. Another reason is that the proposed bilinear-pooling attention network based on low-rank decomposition greatly improved the fusion effect. In the case of [6], the performance of the video modality network only is rather superior to the fusion performance, i.e., DLF performance. This indicates that Soleymani *et al.*'s method

TABLE 4. Performance comparison according to modality and fusion style with the MAHNOB dataset. Here, the numerical result of 10-fold validation is represented in terms of mean and standard deviation. The results of [6] are from the paper. Here ideal fusion indicates that RMSE and PCC are calculated by manually selecting a better output among video and EEG modalities.

	Soleymani et al. [6]		Proposed method	
	PCC	RMSE	PCC	RMSE
Video modality only	0.48±0.37	0.0430±0.0260	0.52±0.07	0.0393±0.0044
EEG modality only	0.24±0.37	0.0530±0.0290	0.29±0.08	0.0485±0.0047
Feature level fusion (FLF)	0.40±0.33	0.0470±0.0250	-	-
Decision level fusion (DLF)	0.45±0.35	0.0440±0.0260	0.52±0.07	0.0373±0.0034
Multimodal attention network	-	-	**0.53**±0.07	**0.0366**±0.0032
Ideal fusion	-	-	0.70±0.05	0.0296±0.0022

failed to derive performance-wise synergy through fusion of video and EEG signals. Finally, the proposed method reduced 0.0027 RMSE compared to video modality by using the EEG modality together, which can be seen as a performance improvement of 6.9%.

Next, we evaluated performance by plotting the valence estimation results for single modality and those for multimodal fusion through the proposed multimodal attention network. Seeing interval [21], [26] and interval [36], [51] in Fig. 8(a), fusion is similar to the EEG modality and GT, because the attention weight is strongly given to the EEG modality. Also, in interval [15], [35] of Fig. 8(b), the regression result for the negative valence improves as the attention weight of the EEG modality increases. On the other hand, in interval [30], [43] and interval [53], [80] of Fig. 8(c), as the attention weight is evenly distributed between the video and EEG modalities, the fusion result becomes similar to GT. As a result, Fig. 8(c) is an example that the non-binary attention weights provide better performance.

For further analysis, Fig. 9 shows the RMSE performance of video and EEG modality networks according to valence values. In the section where the EEG modality has a large valence, e.g., −0.25 or less, or 0.3 or more, the EEG modality has lower RMSE than the video modality. So, in the large valence sections of Figs. 8(a) and (b), we can observe that the fusion result was closer to GT than the video modality and showed similar performance to the EEG modality. This proves experimentally that the proposed multimodal attention network determines more important modality in favor of overall performance.

Figure 10 is plotting estimated valences and attention weights over time. In interval [27], [44], as the attention of video modality increases, video modality approaches GT. On the contrary, in interval [46], [67], EEG modality is closer to GT than video modality, because a larger attention weight is given to EEG modality. On the other hand, we can see that the weights of video modality are extremely strong in the vicinity of 54, 55, and 59 seconds. In terms of valence, EEG modality is estimated to be closer to GT than video modality. Nevertheless, attention weight of video modality is much larger. Looking at the subject's facial expressions at these moments, they are almost neutral. As a result, this

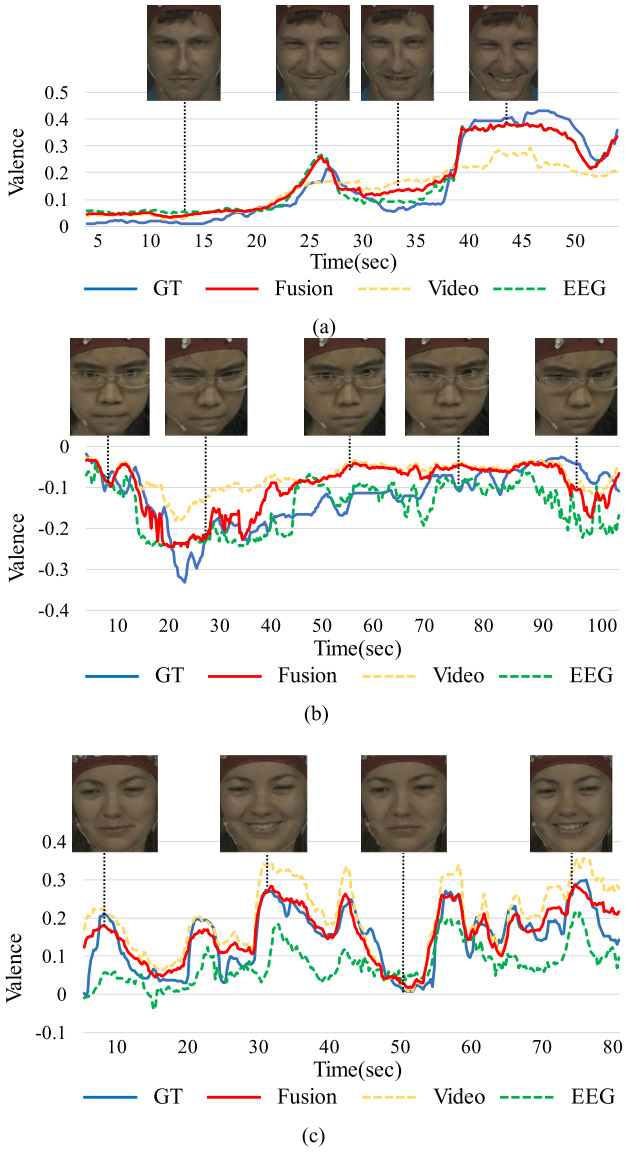

FIGURE 8. Valance result plotting of the MAHNOB-HCI dataset: (a) sequence index: 01_08, (b) sequence index: 07_34, and (c) sequence index: 24_22.

phenomenon is expected to give greater attention weight to video modality in the case of 'neutral', because the portion of neutral emotions in MAHNOB-HCI is very large. In other

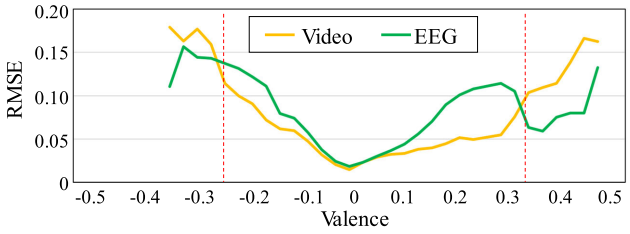

FIGURE 9. RMSE vs. valence of video and EEG modality. For this experiment, MAHNOB-HCI dataset was used.

words, this phenomenon is caused due to an imbalance of training data of a specific emotion.

B. ASIA DATASET

This section evaluated the performance of the proposed method for the ASIA dataset. A total of 236 sequences in the dataset were randomly sampled in a sequence basis, and then training data and test data were separated at a ratio of 7:3. Since a video data of 30Hz was sub-sampled to 15Hz, a two-second video sequence consisting of 30 frames was used in this experiment. Also, EEG signals of two seconds were used as unit data. Figure 11(a) is the result of positive valences. Even though two modalities show large errors, the multimodal attention network makes the fusion valence similar to GT.

Similarly, Fig. 11(b) shows the result for negative valences. Video modality shows almost no emotional change. In case of EEG modality, negative output increases with time. So the fusion result shows a similar trend to EEG modality. RMSE and PCC for the entire dataset are shown in Table 5. Although the improvement was smaller than MAHNOB-HCI, the proposed multimodal attention network provided an average improvement of about 2.5% over the video modality network for the ASIA dataset. Also, PCC increased by up to 0.0418, compared to video modality.

Figure 12 shows the weight distribution according to valance values. In case of MAHNOB-HCI in Fig. 12(a), we can see that weights of video modality are large in most

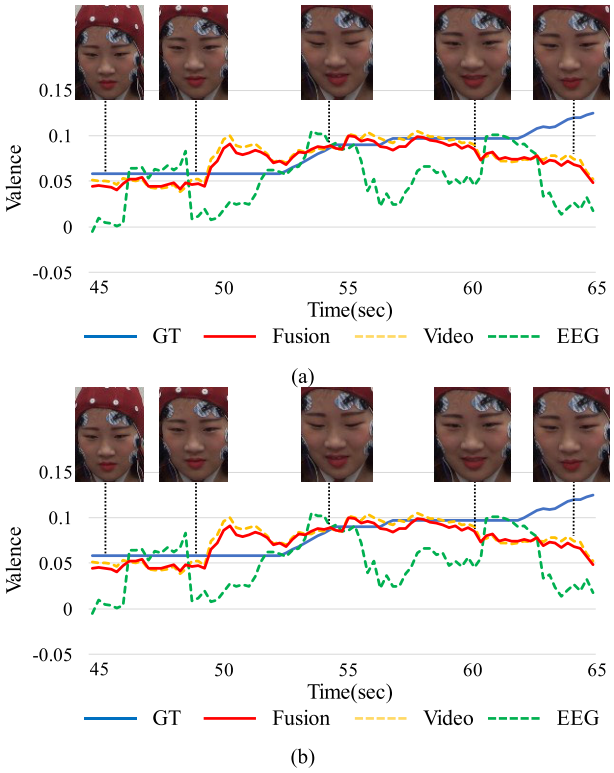

FIGURE 11. Valance results for the ASIA dataset: (a) an example of a positive reaction, and (b) an example of a negative reaction.

valence bands. However, weights of EEG modality tend to increase in the section where the valence is less than −0.25 or more than 0.3, that is, the section with a large valence. This phenomenon is similar to the performance trend of single modality according to valences as in Fig. 9. Thus, it was proved that the proposed multimodal attention network induces synergy between heterogeneous video and EEG signals by assigning a larger attention weight to the modality favorable to performance. We can observe a similar trend for the ASIA dataset of Fig. 12(b).

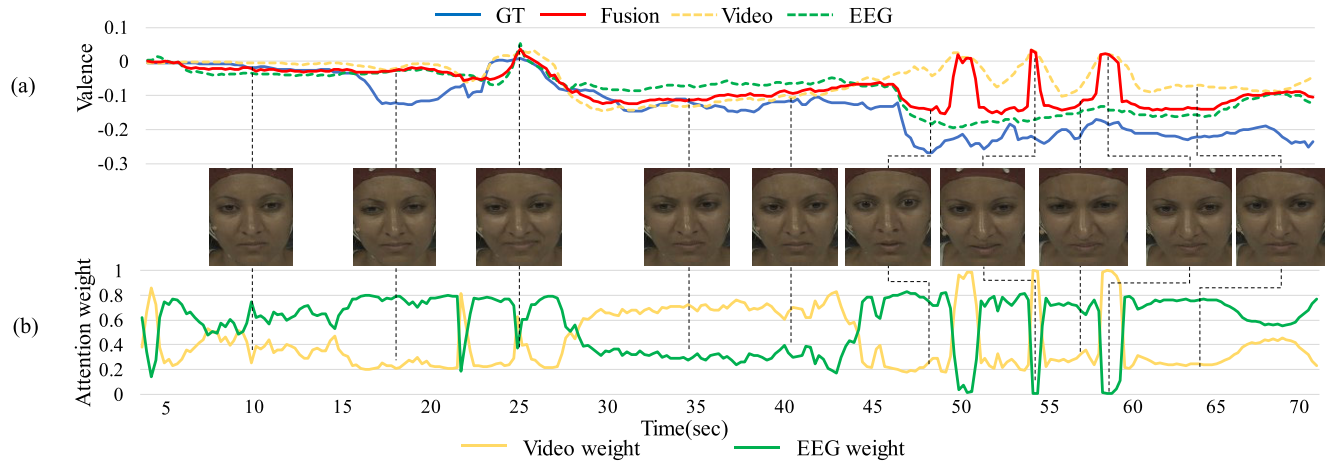

FIGURE 10. (a) Valence output for sample 23_20 of the MAHNOB dataset, and (b) fusion weights of the two modalities.

TABLE 5. Performance comparison according to modality and fusion style for the ASIA dataset.

	Video Only	EEG Only	Proposed
RMSE	0.0568	0.0770	0.0554
PCC	0.1114	0.1072	0.1532

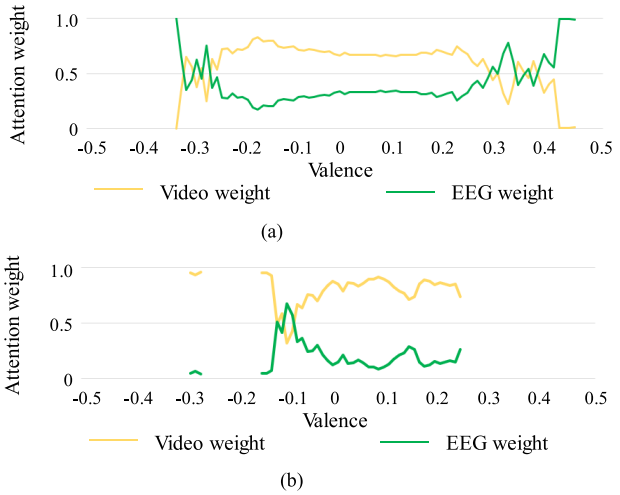

FIGURE 12. Attention weight distribution by modality according to valence value: (a) the MAHNOB-HCI dataset, and (b) the ASIA dataset. In the figure, the valence section without weight occurs because there is no such data and valence label in the dataset.

C. ABLATION STUDY

This section verifies the accuracy of valence regression according to the structure and output type of the multimodal attention network. Performance was evaluated by a well-known 10-fold validation on the MAHNOB-HCI dataset. Note that with the model parameters of the video and EEG networks fixed, the only fusion performance of the multimodal attention network was verified. On the other hand, in order to investigate the performance change according to the output type in the multimodal attention network, we implemented a technique that directly extracts the emotion output from the fused feature, which is named *valence output*. We also compared it with the proposed method.

Firstly, three different fusion methods were evaluated: feature concatenation & FC layer, bilinear pooling [51], and low-rank decomposition. Among them, the low-rank decomposition approach is the proposed one. Table 6 shows the RMSE result. The low-rank fusion with the two-layer structure shows the best performance (0.0366).

Bilinear pooling has the lowest efficiency, with an RMSE of 0.0374, despite having the largest model size. The low-rank fusion of three layers shows no actual performance difference with the low-rank fusion of the two layers. However, the model size of the former is 33% larger than the latter. This indicates that increasing the number of layers in the proposed low-rank fusion is not necessarily advantageous. As a result, the low-rank fusion of the two layers is optimal in terms of performance and model size.

On the other hand, Table 6 shows that RMSE of *valence output* is lower than that of attention weight, irrespective of the fusion method. In concat. & FC, bilinear pooling, and single-layer low-rank fusion, the difference between attention weight and *valence output* is not so significant. However, in the case of low-rank fusion of two and three layers, the performance difference between the two output types becomes noticeable. This indicates that for the proposed low-rank fusion, the more layers, the more the performance is affected by the output type of the attention network.

TABLE 6. Emotion recognition accuracy according to attention network structure and output type.

Fusion methods	Output of multimodal attention network	Model size	RMSE
Concat. & FC	Att. weight	24Mb	0.0371±0.0035
	Valence output		0.0374±0.0039
Bilinear pooling	Att. weight	273Mb	0.0374±0.0036
	Valence output		0.0381±0.0045
Low-rank fusion (1 layer)	Att. weight	13Mb	0.0372±0.0035
	Valence output		0.0378±0.0038
Low-rank fusion (2 layers)	Att. weight	51Mb	**0.0366±0.0032**
	Valence output		0.0494±0.0055
Low-rank fusion (3 layers)	Att. weight	68Mb	**0.0368±0.0034**
	Valence output		0.0591±0.0067

TABLE 7. Emotion recognition accuracy according to rank value in 2-layer low-rank fusion method.

Fusion method	Rank value	Model size	RMSE
Low-rank fusion (2 layers)	1	33Mb	0.0373±0.0035
	2	35Mb	0.0370±0.0034
	4	40Mb	0.0371±0.0035
	8	51Mb	**0.0366±0.0032**
	16	71Mb	0.0371+0.0035

Next, we performed another ablation study to determine the optimal rank value of low-rank decomposition. Table 7 shows the model sizes and RMSEs for several rank values. For this experiment, MAHNOB-HCI was employed and the multimodal attention network is fixed to 'low-rank fusion (2 layer)' in Table 6. As the rank value increases, the model size increases but the RMSE performance improves. However, when the rank value reaches 16, the RMSE performance starts to decrease again. As a result, the best tradeoff between model size and RMSE performance was when the rank value is 8, so we decided the rank value to be 8.

VI. DISCUSSION

In the MAHNOB-HCI and ASIA datasets used in this article, the continuous-time valence labels were determined by the annotators observing facial expressions of subjects. Since the same valence label was used simultaneously as GTs of video modality and EEG modality, the reliability of EEG modality could be inevitably reduced. To overcome this, we need a way to annotate emotion labels in continuous-time domain without relying on facial expressions, which can be our future work. On the other hand, the proposed multimodal fusion network predicted the best modality favorable to the overall

performance and then assigned a larger attention weight to the predicted modality, which led to synergy between video and EEG modalities in emotion recognition. Therefore, the performance of the proposed fusion method is superior to the single modality even when the annotation reliability of EEG modality is not guaranteed.

In the two datasets, recognition of negative emotions with little change in facial expressions tended to be inferior to recognition of positive emotions. In order to compensate for this, emotion recognition technology using additional bio-signals, such as the electromyogram (EMG) and galvanic skin response (GSR), has recently been studied [8], [32]. Thus, it is necessary to study the fusion technology between the video signal and multiple bio-signals.

Despite selecting stimuli videos that cause strong emotions for the ASIA dataset, labels are concentrated on 'neutral' section. So, it was difficult to generalize the characteristics of modality according to valence size. In order to more effectively recognize Asian emotions, the improved annotation skill is required and a new dataset including various emotions must be built.

VII. CONCLUSION

We proposed a multimodal attention network that effectively integrates video modality and EEG modality networks for multimodal emotion recognition. We experimentally demonstrated that the proposed method improves emotional recognition performance over single-modality networks for the MAHNOB-HCI dataset and the ASIA dataset. This means that the proposed multimodal attention network generates synergy in terms of emotion recognition by effectively merging the video and EEG signals at the feature level. If additional modalities (such as voice data and more informative bio-signals) are adopted in the future, emotion recognition performance will be further improved.

On the other hand, absolute labeling is limited due to subjective annotation, and annotation depending on visual information causes uncertainty in evaluations. If quantitative measurement and annotation methods of accurate emotions proceed, research on emotion recognition in the future can be advanced.

REFERENCES

[1] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[2] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, Jun. 2013.

[3] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.

[4] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2983–2991.

[5] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, "Emotion recognition from audio signals using support vector machine," in *Proc. IEEE Recent Adv. Intell. Comput. Syst.*, Dec. 2015, pp. 139–144.

[6] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan. 2016.

[7] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 230–243, Apr./Jun. 2020.

[8] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[9] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul. 2014.

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interface*, Oct. 2004, pp. 205–211.

[11] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: EmotiW 5.0," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 524–528.

[12] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 553–560.

[13] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 529–535.

[14] 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.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[16] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. 3rd Intern. Workshop EMOTION (Satellite LREC), Corpora Res. Emotion Affect*, 2010, p. 65.

[17] S. Koelstra, C. Muhl, 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 Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[18] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1868–1877.

[19] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 352–364, Feb. 2018.

[20] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3029–3037.

[21] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[22] X. Zhang, H. Dong, Z. Hu, W. S. Lai, F. Wang, and M. H. Yang, "Gated fusion network for joint image deblurring and super-resolution," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2018, pp. 1–13.

[23] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*. Sep. 2000, pp. 1–6.

[24] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 192–201.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, Dec. 2015, pp. 59–63.

[29] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, Mar. 2017.
[30] 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 *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng.*, Apr. 2015, pp. 914–917.
[31] N. Ketkar, "Introduction to Pytorch," in *Deep Learning With Python*. Berkeley, CA, USA: Apress, 2017, pp. 195–208.
[32] A. Mavratzakis, C. Herbert, and P. Walla, "Emotional facial expressions evoke faster orienting responses, but weaker emotional responses at neural and behavioural levels compared to scenes: A simultaneous EEG and facial EMG study," *NeuroImage*, vol. 124, pp. 931–946, Jan. 2016.
[33] *Biosemi Active II System*. Accessed: Oct. 18, 2019. [Online]. Available: https://www.biosemi.com/products.htm
[34] U. Herwing, P. Satrapi, and C. Honfeldt-Lecuona, "Using the international 10-20 EEG system for positioning of transcranial magnetic simulation," *Brain Topogr.*, vol. 16, no. 2, pp. 95–99, Dec. 2003.
[35] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 575–589.
[36] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 571–575.
[37] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," 2019, *arXiv:1902.00038*. [Online]. Available: http://arxiv.org/abs/1902.00038
[38] J.-M. Pérez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6966–6975.
[39] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*. [Online]. Available: http://arxiv.org/abs/1806.00064
[40] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1821–1830.
[41] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*. [Online]. Available: http://arxiv.org/abs/1606.01847
[42] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*. [Online]. Available: http://arxiv.org/abs/1610.04325
[43] S.-H. Oh, Y.-R. Lee, and H.-N. Kim, "A novel EEG feature extraction method using Hjorth parameter," *Int. J. Electron. Electr. Eng.*, vol. 2, no. 2, pp. 106–110, Jun. 2014.
[44] M. Cukic, D. Pokrajac, M. Stokic, S. Simic, V. Radivojevic, and M. Ljubisavljevic, "EEG machine learning with Higuchi fractal dimension and sample entropy as features for successful detection of depression," 2018, *arXiv:1803.05985*. [Online]. Available: http://arxiv.org/abs/1803.05985
[45] E. Derya Übeyli, "Analysis of EEG signals by combining eigenvector methods and multiclass support vector machines," *Comput. Biol. Med.*, vol. 38, no. 1, pp. 14–22, Jan. 2008.
[46] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: AFF-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 907–929, Jun. 2019.
[47] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017.
[48] J. Li, Y. Chen, S. Xiao, J. Zhao, S. Roy, J. Feng, S. Yan, and T. Sim, "Estimation of affective level in the wild with multiple memory networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1–8.
[49] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, Oct. 2010, pp. 1459–1462.
[50] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4193–4202.
[51] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000.
[52] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019.
[53] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, early access, May 11, 2020, doi: 10.1109/TAFFC.2020.2994159.
[54] J. Ren, Y. Hu, Y. W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn-a multimodal LSTM for speaker identification," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.
[55] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.
[56] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.
[57] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, early access, Nov. 30, 2018, doi: 10.1109/TAFFC.2018.2884461.

DONG YOON CHOI (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Inha University, Incheon, South Korea, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electronic engineering. His research interests include image processing, computer vision, and multimodal deep-learning.

DEOK-HWAN KIM (Member, IEEE) received the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology. He is currently a Professor with Inha University, South Korea. His research interests include embedded systems, storage systems, image processing, bio-signal processing, and multimedia systems.

BYUNG CHEOL SONG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2008, he was a Senior Engineer with the Digital Media Research and Development Center, Samsung Electronics Company Ltd., Suwon, South Korea. In March 2008, he joined the Department of Electronic Engineering, Inha University, Incheon, South Korea, where he is currently a Professor. His research interests include image processing and computer vision.