논문 22-47-01-11 The Journal of Korean Institute of Communications and Information Sciences '22-01 Vol.47 No.01
https://doi.org/10.7840/kics.2022.47.1.79

Deep Learning Based Data Fusion Methods for Multimodal Emotion Recognition

Judith Nkechinyere Njoku◆, Angela C. Caliwag*, Wansu Lim°, Sangho Kim**, Han-Jeong Hwang***, Jin-Woo Jeong****

ABSTRACT

Multimodal emotion recognition is a robust and reliable method as it utilizes multimodal data for more comprehensive representation of emotions. Data fusion is a key step in multimodal emotion recognition, because the accuracy of the recognition model mostly depends on how the different modalities are combined. The goal of this paper is to compare the performances of deep learning (DL) based models for the task of data fusion and multimodal emotion recognition. The contributions of this paper are two folds: 1) We introduce three DL models for multimodal fusion and classification: early fusion, hybrid fusion, and multi-task learning. 2) We systematically compare the performance of these models on three multimodal datasets. Our experimental results demonstrate that multi-task learning achieves the best results across all modalities; 75.41%, 68.33%, and 78.75% for classification of three emotional states using the combinations of audio-visual, EEG-audio, and EEG-visual data, respectively.

Key Words : Data-fusion, deep learning, emotion recognition, multimodal, EEG

Ⅰ. Introduction

Emotion recognition plays a significant role in diverse domains, e.g entertainment, health, and learning[1-3]. It enables the responses of software applications to be adapted to the emotional states of the end-users[4]. Unfortunately, many methods of emotion recognition focus on a single modality (speech, facial expression, posture, electroence-phalograph (EEG), etc.)[5]. This greatly limits the accuracy of the emotion recognition task[3].

Due to the complex nature of emotions, single modality data is not capable of comprehensively describing emotions. Thus, the use of multimodal data has become a promising approach. In principle, the use of multimodal data substantially improves the accuracy and reliability of an emotion recognition model[6,7]. This is because the use of multiple modalities allows for more complementary information to be captured for emotion recognition. Moreover, multimodal models can be employed even if one of the modalities is missing[7].

Multimodal data fusion is one of the main steps taken to realize multimodal emotion recognition. Various researchers have utilized different types of data fusion strategies such as feature level fusion

※ This work was supported by the National Research Foundation of Korea (2020R1A4A101777511)
◆ First Author : Kumoh National Institute of Technology, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, 학생회원
° Corresponding Author : Kumoh National Institute of Technology, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, wansu.lim@kumoh.ac.kr, 정회원
* Kumoh National Institute of Technology, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, 학생회원
** Kumoh National Institute of Technology, School of Industrial Engineering
*** Korea University, Department of Electronics and Information Engineering
**** Seoul National University of Science and Technology, Department of Data Science, 정회원
논문번호 : 202109-242-C-RN, Received September 15, 2021; Revised October 26, 2021; Accepted November 4, 2021

and decision level fusion. In [9], feature level fusion and decision level fusion were applied to merge different physiological data types. In [10], the authors employed MAX fusion, SUM fusion, and fuzzy integral fusion for multimodal emotion recognition. In [11], a hierarchical classifier that combines feature level and decision level for emotion recognition in the wild was proposed. However, the effective fusion of multiple modalities poses a huge challenge, due to the heterogeneous nature of the data employed for emotion recognition.

With the rapid advancement of deep learning (DL), its potential for the fusion of multimodal data has been being explored[12,13]. Many deep learning models have been applied to multimodal data fusion for emotion recognition, however, the characteristics of these different DL-based fusion approaches when applied to different data types, have not yet been fully studied. To establish an efficient multimodal emotion recognition system, it is important to select the best data fusion strategy.

In this paper, we compare the recognition performance of Early Fusion, Hybrid Fusion, and Multi-Task Learning (MTL) strategies for multimodal emotion recognition. Early fusion concatenates feature vectors from various modalities into a single long vector, which is fed into the DL model. Hybrid fusion combines early and decision (late) fusion strategies to yield the outputs of fused tasks. MTL explores the commonalities and differences across different tasks to yield he outputs of fused tasks. The main contributions of this paper can be summarized as follows:

1) We introduce three DL models, which all follow the early, hybrid, and MTL fusion strategies, and apply them for multimodal emotion recognition.

2) We apply these three DL models to three different modalities, including EEG data, speech data, and facial expression for multimodal emotion recognition.

3) We systematically compare the recognition performance of these DL models on the RAVDESS audio-visual dataset[14] and an EEG dataset from [5].

The remainder of this paper is organized as follows. The related works are presented in Section II. Section III introduces the methods employed in detail. Section IV describes the experimental settings. Section V presents the experimental results, while Section VI concludes the paper, and gives insight into the direction for future research.

Ⅱ. Related Works

Multimodal fusion has been applied in diverse fields such as event detection, video classification, image segmentation, etc, due to its promising potentials for performance improvement. At the fusion level, there are three traditional approaches for data fusion: 1)Feature level fusion, 2)decision level fusion, and 3)hybrid fusion. Recently DL has been applied to explore these traditional approaches amongst others, for effective data fusion.

2.1 Traditional based data fusion

Feature-level fusion is a commonly adopted strategy. It is quite straightforward when used in the fusion of different modalities. This strategy is also known as early fusion because the fusion occurs before classification. The features extracted from the various modalities are combined into a high dimensional feature and forwarded to the classification model[10]. In [10], the authors' employed MAX fusion, SUM fusion, and fuzzy integral fusion for multimodal emotion recognition. They analyzed confusion matrices to discover the complementary attributes of EEG and eye movement features. This strategy is effective because it can capture and utilize the correlations between multiple modalities at an early stage. Furthermore, the fused data hold more information than a single modality. As a result, the performance of the multimodal fusion strategy was much better than single modality emotion recognition[11]. Despite these advantages, there are some drawbacks to this approach. Firstly, features with high dimensions might lead to a computationally complex model training process. Secondly, it is challenging to effectively represent the time synchronizations between the features of multiple modalities[9]. Lastly, when applied to small datasets, this strategy could suffer from the curse of

dimensionality, resuling in significant performance drop.

The decision-level fusion strategy employs different kinds of classifiers for each data type. An ensemble model is then used to assemble all these classifiers. Specifically, the outcomes of each classifier are merged into a single decision. The bedrock of this strategy is founded on rules. This strategy is also known as late fusion because the fusion is performed after the classification task. In [15], the classifiers of three different psychological signals were fused for efficient emotion recognition. In [16], a decision-level weight fusion strategy was employed in the fusion of several physiological signals. The classification result of each classifier was fused linearly based on the weight matrix. One advantage of this strategy is that it is easy to compare the decisions from different classifiers, enabling each modality to utilize the most suitable classifier. However, one drawback to this strategy is that it is difficult to design an efficient rule. A rule which is too simple might fail to discover the relationships between the different modalities.

Hybrid fusion strategy combines the feature level and decision level fusion strategies. In [11], feature and decision level fusion was used to build a hierarchical classifier for the task of emotion recognition. In [17], the authors employed the hybrid fusion strategy for the fusion of facial expressions, EEG, and galvanic skin response. They demonstrated that the proposed model can determine the correct emotional state when natural deceptive facial expressions are adopted. In [18], a hybrid classifier that combined support vector machine (SVM) and fuzzy cognitive map were employed for classifying different emotions that have compressed sensing representations.

2.2 Deep Learning based data fusion

Different multimodal fusion approaches have been developed using DL[6,19-21]. Some of these DL approaches adopt the strategies employed by traditional approaches. In [6], the authors employed a hybrid model which extracted audio and visual features using CNNs and 3D-CNN respectively. These features were then fused using a deep belief network (DBN). In [19], the early fusion approach was adopted for multimodal emotion recognition using a convolutional neural network (CNN) and long short term memory (LSTM) model. The hybrid network was jointly trained to learn audio-visual feature representations that are discriminative. In [20], DL-based transfer learning was employed to fuse several bio-sensing and video data. The research showed that the proposed method helped to overcome the inconsistencies between the datasets used.

One DL approach which is now being explored for emotion recognition is MTL. MTL is based on the premise that related tasks are often inter-independent and yield better results when a joint framework is applied[21]. MTL can learn the differences and commonalities across different tasks. In [22], MTL was applied to a categorical emotion recognition task and a valence-based emotion recognition task. The loss functions from these two tasks were combined to form the objective function of the MTL framework. In [23], a deep MTL framework was proposed for the dual task of sentiment and emotion analysis. In this paper, we apply DL models based on early fusion, hybrid fusion, and MTL for multimodal emotion recognition from EEG, video, and audio data.

Ⅲ. Emotion Recognition Using Data Fusion

The entire emotion recognition system is illustrated in Fig. 1. As shown in the figure, first, data processing algorithms are applied to all channels. Then, features are extracted from the preprocessed data. These extracted features are then forwarded to an Feedforward Neural Network (FNN) fusion model, which is also trained to classify the fused data. The model is evaluated by testing on a held-out sample data for generalization performance. The data fusion technique used in this study is the intermediate fusion. Intermediate fusion can be employed at different stages of model training. Intermediate fusion transforms the input data into higher level features by passing them

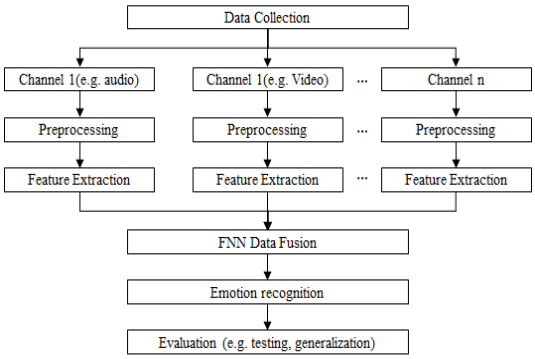


Fig. 1. Flowchart of end-to-end multimodal emotion recognition framework

through multiple layers. Each of these layers operates linear and nonlinear functions that transform the scale, skew and swing of the input data. Thus, giving a new representation to the original input data. These data fusion models are discussed in the following subsections.

3.1 Early-Fusion Model

In the early fusion model employed, a fully connected layer receives the feature vectors of two different data types (EEG, video, audio). As illustrated in Fig. 2, these feature vectors are concatenated in the first layer of the CNN. The concatenated feature is passed through two more layers that are activated with a linear activation function. The last layer that is activated by a Rectified Linear Unit (ReLu) activation function outputs the final prediction of the fused modalities. The output vector is a 3-dimensional vector representing the three classes of emotion; neutral, negative, and positive. The main idea in this model was to fuse the features before performing the classification task.

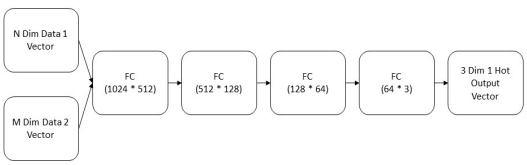


Fig. 2. Early fusion model

3.2 Hybrid Fusion Model

In the hybrid fusion model employed, each data type is received by a fully connected (FC) layer per modality to build and explore the correlations which exist between similar features. Another FC layer merges the outputs from the previous layers by concatenation. The function of this layer is to correlate the essence of the different modalities. The next two FC layers all activated with the ReLu activation function, outputs modality-specific vectors. Each of these 3-dimensional vectors represents the prediction of each data type. A decision rule is applied using an audio/video priority heuristic method. In this method, a weighted majority rule is applied. Each of the final FC layers are assigned specific weights. The predicted labels for each layer is multiplied by the weights and the average is computed. Based on these weight averages, the final emotion label is assigned.

The main idea in this model was to combine the concepts of early fusion and decision fusion for emotion recognition tasks. The hybrid fusion model is illustrated in Fig. 3.

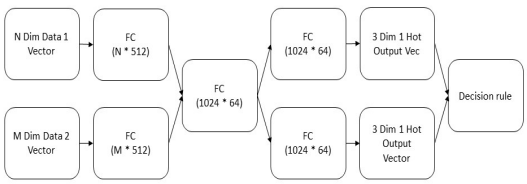


Fig. 3. Hybrid fusion model

3.3 Multi-Task Learning Model

In the MTL model, each data type is received by an FC layer, which explores the correlation between similar features. These features are forwarded through two more layers corresponding to two tasks, e.g. audio and video emotion recognition tasks. these layers are activated with the linear activation function. The outputs of these layers are forwarded to a single FC layer, which concatenates the outputs and serves as the prediction layer. The output of the final layer is a 3-dimensional vector that represents the different types of emotions. In this model, hard parameter sharing is applied by sharing the network layers between the two tasks. This helps to reduce overfitting. The multi-task learning model is

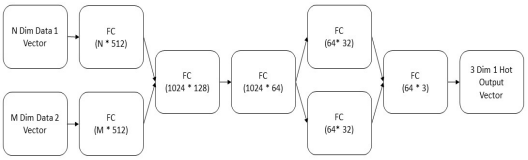


Fig. 4. Multi-Task learning model

illustrated in Fig. 4.

The main aim of this model is to exploit the inter-dependence between features to increase the confidence of the emotion recognition task in prediction.

Ⅳ. Experimental Settings

4.1 Datasets

Two different datasets were employed in this study. The audio and video data were obtained from the RAVDESS dataset. The RAVDESS dataset was developed by Livingstone and Russo[14]. The dataset contains audio and video files of 24 professional actors, all vocalizing specific statements in a neutral North American accent. The emotions recorded include calm, happy, sad, angry, fearful, surprise, disgust, and neutral. The EEG data was introduced by [5]. The dataset is made up of EEG brainwave data, which has been processed using statistical extraction. The data was collected from two people using a Muse EEG headband. The emotions recorded include neutral, positive, and negative.

4.2 Feature Extraction

The raw audio files were extracted from the videos in the RAVDESS dataset, by using Librosa's load function. Due to the different video lengths, each video was sampled at a maximum threshold length of 3 seconds. The audio signals were sampled at 48000Hz, each signal representing an array of length 14400. MFCC features were extracted for each speech sample. These features were then flattened and normalized to a 1D feature array of 3887 dimensions.

Two pre-trained CNNs were used to extract visual features from the video samples. A 3D ResNexT-101 model which has been pre-trained on the Kinetics dataset was used for extracting visual features. The video frames were sampled at a rate of 1 frame per second. With each video having 4 frames, the final feature matrix was of 4x2048 dimensions. This was flattened to return a 1D array of 8192 dimensions.

The EEG features were already extracted and pre-processed by the developers[5], so no further pre-processing was done.

The labels of the audio and visual datasets were modified to match the three emotional classes in the EEG data. Thus all the emotions with calm, happy, and surprised labels were relabelled as positive. The sad, angry, fearful, and disgust labels were relabelled negative, while the neutral emotion remained neutral.

4.3 Model Training

All models were trained using custom cross-entropy loss functions. The early fusion and MTL models were trained using the cross entropy loss function:

$$CE = -\sum_{i}^{C} t_i \log(f(s_i)) \tag{1}$$

where C is the number of emotion labels, t_i is the ground truth, and s_i is the model score for each class i in the fused data. The function $f(s_i)$ refer to the activations applied to the output before the loss computation.

The hybrid fusion model was trained using the cross-entropy loss function:

$$CE = -\sum_{i}^{C} t_i \log(f(s_i) + f(v_i)) \tag{2}$$

where C is the number of emotion labels, t_i is the ground truth, and s_i and v_i are the model scores for each class i in each data type. The functions $f(s_i)$ and $f(v_i)$ refer to the activations applied to the outputs of each data type before the loss computation.

4.4 Training setup

All mod All models were implemented in Pytorch with GPU 1xTesla K80, having 2496 CUDA cores and 12GB GDDR5 VRAM in Google Colaboratory. All datasets were first saved in Google drive and imported into Google Colaboratory. The entire dataset consists of 1440 samples for each data type. Among these, 1200 samples were used for training, while 240 samples were used for the evaluation process. Table 1 describes the statistics of emotion labels in the training and test datasets.

All models were trained from scratch with randomly initialized weights in 50 epochs, using the Adaptive moment optimizer. The batch size for each iteration was set to 120 and the learning rate was initialized at $10 * 10^{-4}$. A metric was employed to reduce the learning rate once the accuracy stops increasing. All models were trained to optimize their specific cross-entropy loss functions. To determine the best hyperparameters for training the models, a brute force method was used. In this technique, several models with different hyperparameters are tested until the model with optimal hyperparameters are obtained All models were implemented using the same GPU, to have an objective result. The results of all experiments are detailed in the following section.

Table 1. Statistics of dataset

Statistics	Training	Testing
Neutral	240	48
Negative	640	128
Positive	320	64

V. Experimental Results

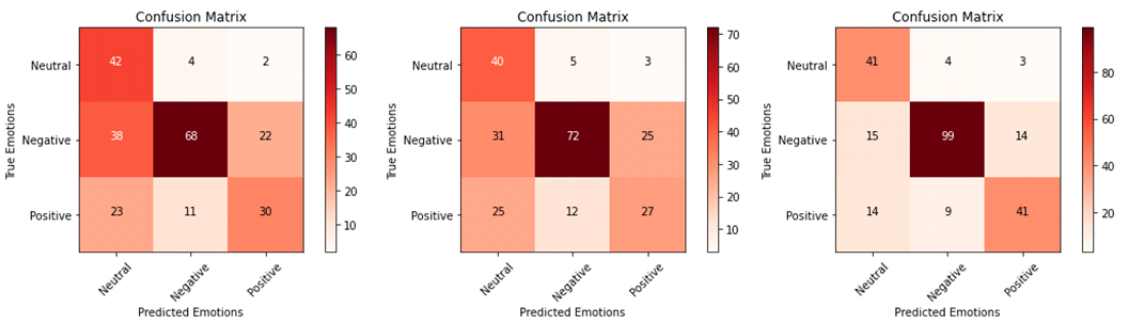
In the first experiment, for audio-visual emotion recognition, the features of the three modalities were combined in pairs and applied to the three different DL models. The three resulting pairs were audio-visual, audio-EEG, and EEG-visual modalities. Their performances are compared in terms of accuracy as represented in Table 2. From the results, the early fusion model achieved its best performance on the audio-visual data, while the hybrid fusion and MTL models achieved their best performances on the EEG-visual data. The audio-EEG data had the poorest performance accuracy across all models.

From the results, while the early fusion model achieved the least performance accuracies for all modalities, the MTL model has the highest accuracies.

To further analyze the performance, we compare the confusion matrices of all the experiments. This

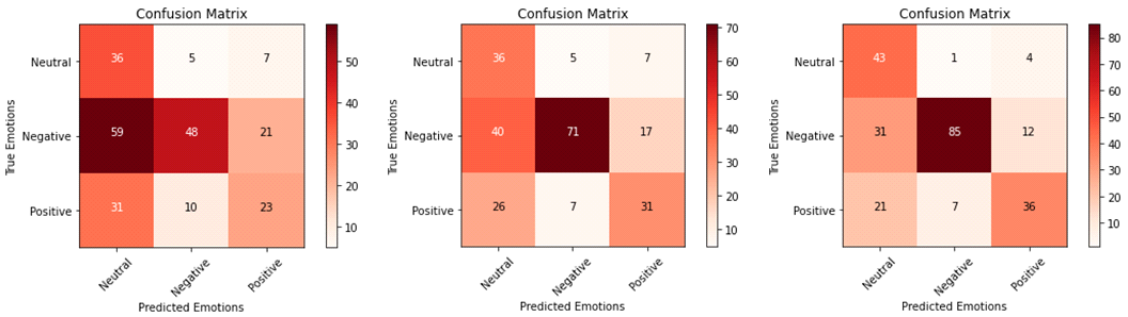
Table 2. Performance comparison of all models on all modalities

Data/ Model	Early Fusion Acc (%)	Hybrid Fusion Acc (%)	MTL Acc (%)
Audio-visual	58.33	57.91	75.41
Audio-EEG	44.58	57.5	68.33
EEG-visual	47.5	63.75	78.75



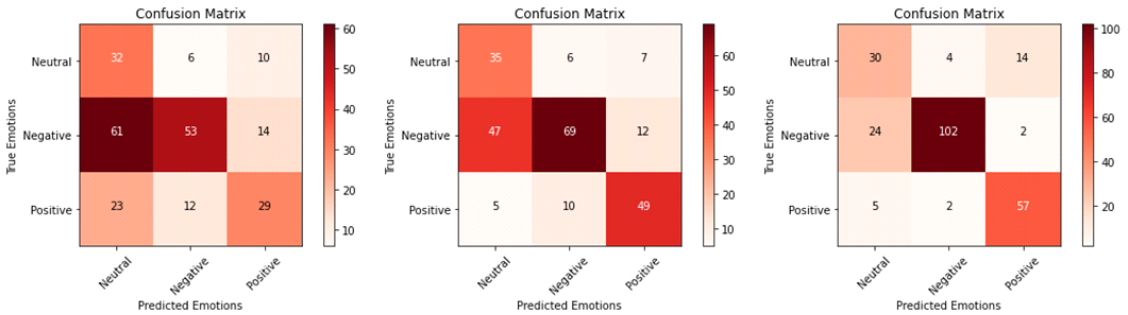
(a) Early (b) Hybrid (c) MTL

Fig. 5. Confusion matrices of Audio-Visual emotion recognition



(a) Early (b) Hybrid (c) MTL

Fig. 6. Confusion matrices of EEG-Audio emotion recognition



(a) Early (b) Hybrid (c) MTL

Fig. 7. Confusion matrices of EEG-Visual emotion recognition

provides more insight into the misclassification performance of the different modalities and models.

In Fig. 5, the neutral emotion was the least misclassified emotion across all models, with an accuracy of 87.5%, 83.33%, and 85.41% on the Early fusion, Hybrid fusion, and MTL models respectively. The positive emotion achieved the highest misclassification error with an accuracy of 46.86%, 42.18%, and 64.06% on all three models.

Fig. 6 illustrates the performance of the EEG-Audio modalities on all models. The neutral emotion was also the least misclassified with an accuracy of 75%, 75%, and 89.6% on the Early fusion, Hybrid fusion, and MTL models respectively. The positive emotion is also the most misclassified.

In Fig. 7, the performance of all models on the EEG-Visual modalities was illustrated. In this experiment, the positive emotion achieved the best accuracy of 89.06% on the MTL model. The poorest performance was by the Early fusion model on the negative emotion.

A summary of model accuracies across all modalities is shown in Fig. 8. As shown in the figure, the MTL model achieved the best recognition accuracy of 75.41%, 68.33%, and 78.75% on the Audio-visual, EEG-audio, and EEG-visual modalities respectively.

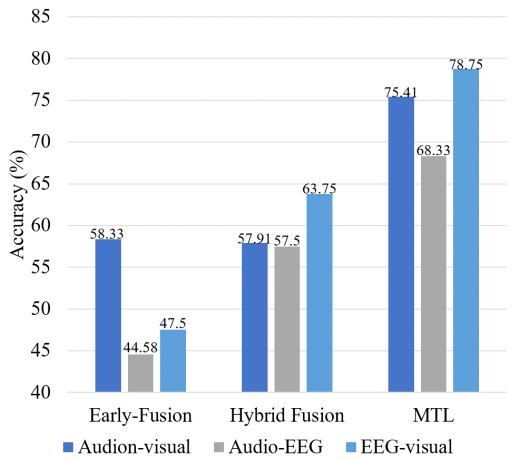


Fig. 8. Model accuracies for audio-visual, audio-EEG and EEG-visual modalities

Ⅵ. Conclusion and Future Works

In this paper, we have implemented three DL-based approaches for data fusion; early fusion, hybrid fusion and multitask learning. Our main goal was to compare their performances of emotion recognition on multimodal data. To achieve this, we extracted features from three modalities (video, audio, EEG) and fed them to each model with pair-wise modalities, i.e., audio-video, EEG-audio, and EEG-video. Experimental results suggested that the neutral emotion is the easiest to recognize, while the multitask learning approach achieved the highest performance for all modalities.

In future studies, we would like to explore more modalities for emotion recognition using multitask learning in order to improve the performance of emotion recognition.

References

[1] J. Kim, "Multimodal parametric fusion for emotion recognition," *J. Advanced Smart Convergence*, vol. 9, no. 1, pp. 193-201, Jan. 2020.

[2] E. Seo, et al., "Reading the mind in the eyes test: Relationship with neurocognition and facial emotion recognition in non-clinical youths," *Psychiatry Investig.*, vol. 17, no. 8, pp. 835-839, Aug. 2020.

[3] E. Han, et al., "Adaptive feature generation for speech emotion recognition," *IEIE Trans. Smart Process. Comput.*, vol. 9, no. 3, pp. 185-192, Mar. 2020.

[4] P. A. Abhang, et al., *Introduction to EEG-and speech-based emotion recognition*, Academic Press, 2016.

[5] J. J. Bird, et al., "Mental emotional sentiment classification with an eeg-based brain-machine interface," in *Proc. Digital Image and Sign. Process.*, Oxford, UK, Apr. 2019.

[6] H. Ranganathan, et al., "Multimodal emotion recognition using deep learning architectures," in *Proc. Applicat. Comput. Vision*, USA, Mar. 2016.

[7] S. Zhang, et al., "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Cir. Sys. Video Techn.*, vol. 28, no. 10, pp. 3030-3043, Oct. 2018.

[8] P. Tzirakis, et al., "End-to-End multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Sign. Process*, vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[9] P. Bota, et al., "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?," *Sensors*, vol. 20, no. 17, pp. 4723, Aug. 2020.

[10] Y. F. Lu, et al., "Combining eye movements and EEG to enhance emotion recognition," in *Proc. Artificial Intell.*, Buenos Aires, Argentina, Jul. 2015.

[11] B. Sun, et al., "Combining feature level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 125-137, Nov. 2016.

[12] M. J. Lucia, et al., "Vibrotactile captioning of musical effects in audio-visual media as an alternative for deaf and hard of hearing People: an EEG study," *IEEE Access*, vol. 8, pp. 190873-190881, Oct. 2020.

[13] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder," *IEEE Access*, vol. 8, pp. 164130-164143, Sep. 2020.

[14] S. R. Livingstone, et al., "The Ryerson audio-visual database of emotional speech and song : A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, pp. 1-35, May 2018.

[15] J. Xie, et al., "WT feature based emotion recognition from multi-channel physiological signals with decision fusion," in *Proc. Affective Comput. and Intell. Interaction*, China, May 2018.

[16] W. Wei, et al., "Emotion recognition based on weighted fusion strategy of multichannel physiological signals," *Comput. Intell.*

Neurosci., vol. 2018, pp. 1-9, 2018.

[17] Y. Cimtay, et al., "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865-168878, Sep. 2020.

[18] K. Guo, et al., "A hybrid fuzzy cognitive map support vector machine approach for EEG based emotion classification using compressed sensing," *Int. J. Fuzzy Syst.*, vol. 21, pp. 263-273, Nov. 2018.

[19] B. Nakisa, et al., "Automatic emotion recognition using temporal multimodal deep learning," *IEEE Access*, vol. 8, pp. 225463-225474, Sep. 2020.

[20] S. Siddharth, et al., "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affective Comput.*, vol. 1, pp 1-12, May 2019.

[21] Y. Zhang, et al., "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.,* (Early Access) 2021.

[22] R. Xia, et al., "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 3-14, Mar. 2017.

[23] M. Akhtar, et al., "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. North Am. Chapter of the Assoc. for Computation Linguistics: Human Lang. Technol.*, USA, Jun. 2019.

주 디 스 (Judith Nkechinyere Njoku)
2019년 9월~현재 : 금오공대, 항공기계융합전공
<관심분야> 임베디드 시스템, 지능형 제어
[ORCID:0000-0002-2294-9204]

안 젤 라 (Angela C. Caliwag)
2019년 10월~현재 : 금오공대, 항공기계융합전공
<관심분야> 임베디드 시스템, 지능형 제어
[ORCID:0000-0002-0279-935X]

임 완 수 (Wansu Lim)
2014년 9월~현재 : 금오공대 전자공학부 부교수
<관심분야> 임베디드 시스템, 지능형 제어
[ORCID:0000-0003-2533-3496]

김 상 호 (Sangho Kim)
1996년 3월~현재 : 금오공대, 산업공학부 교수
<관심분야> 인간공학, 감성인식, 인터랙션디자인
[ORCID:0000-0003-0599-289X]

황 한 정 (Han-Jeong Hwang)
2020년 3월~현재 : 고려대 전자및정보공학과 부교수
<관심분야> 인공지능기반 신경신호분석
[ORCID:0000-0002-1183-1219]

정 진 우 (Jin-Woo Jeong)
2021년 3월~현재 : 서울과기대 산업공학과/데이터사이언스학과 부교수
<관심분야> 멀티모달 인터랙션, 딥러닝 응용
[ORCID:0000-0001-9313-6860]