IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 April 2024, accepted 28 June 2024, date of publication 3 July 2024, date of current version 12 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422403

RESEARCH ARTICLE

Multi-Task Learning by Leveraging Non-Contact Heart Rate for Robust Facial Emotion Recognition

YERIM JI[1] AND SUH-YEON DONG[2], (Member, IEEE)
[1]Department of Information Technology Engineering, Sookmyung Women's University, Seoul 04310, South Korea
[2]Division of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul 04310, South Korea

Corresponding author: Suh-Yeon Dong (sydong@sookmyung.ac.kr)

This work was supported in part by the National Research Foundation of Korea's Brain Korea 21 FOUR Program at Sookmyung Women's University; and in part by the Ministry of Science and ICT (MSIT), South Korea, through the ICT Challenge and Advanced Network of HRD (ICAN) Program supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2024-RS-2022-00156299.

ABSTRACT Building a robust facial expression recognition (FER) system remains a challenging problem due to the emotional ambiguity of facial expressions. Recent approaches employ both facial expressions and physiological signals to design multi-modal emotion recognition systems. However, these approaches require physical contact with the skin as they need to use sensor modalities. To meet the demands for a non-contact emotion recognition system, we use a convolutional recurrent neural network (CRNN) to extract facial features and utilize these features for estimating the heart rate (HR) from face image sequences. In particular, unlike the conventional feature fusion method, we propose a multi-task learning (MTL) framework to simultaneously predict the emotion and HR from face image sequences using a single model. Experiments on the DEAP and MAHNOB-HCI datasets demonstrate that the proposed multi-task framework improves FER accuracy by up to 6.85% and achieves superior performance against the state-of-the-art methods.

INDEX TERMS Facial expression recognition, multi-task learning, heart rate, deep neural network.

I. INTRODUCTION

With the rapid development of human-computer interaction (HCI), affective computing has attracted considerable attention. Facial expressions are a rich source of emotional information and they can be easier acquired compared to the other modalities. Hence, facial emotion recognition (FER) is the most widely used and intuitive approach for identifying one's emotional states. However, FER is still a challenging task because it tends not to perform well when there is no meaningful change in facial expression [1]. Another problem is that facial expressions can be intentionally manipulated or hidden, unlike the actual emotional state [2], [3].

In recent years, many studies have been conducted to develop reliable FER systems combining facial expressions and physiological signals [4], [5], [6]. Physiological signals are generated spontaneously by the human body, so they are more objective in capturing true emotional states. However, physiological signals are measured by specific devices or sensors in contact with the skin, which can cause discomfort. This complexity and inconvenience of measurement strongly limit its scope of application.

To address the above problems, this study leverages physiological signals extracted from facial videos using remote photoplethysmography (rPPG). The rPPG is a non-contact heart rate (HR) measurement technique that detects facial color changes caused by cardiac activity [7]. HR is an important physiological indicator that reflects the physical and emotional states of humans. Therefore, it is possible to simultaneously predict the physiological state and the emotional state from facial images.

Recently, some researchers attempted to improve the FER accuracy by combining facial features and rPPG signal features [2], [3], [8]. These recent studies have shown that the non-contact HR or HRV features obtained from rPPG can improve FER accuracy. However, since these methods

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

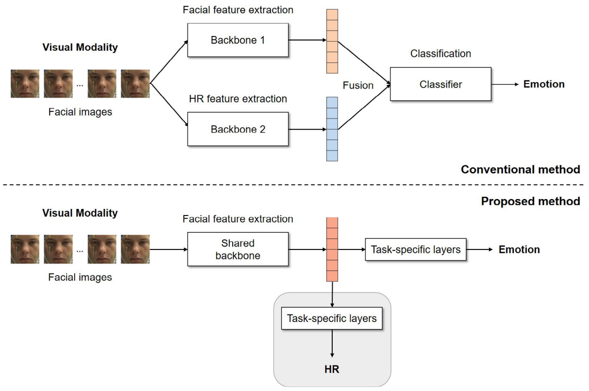

FIGURE 1. **Comparison of the conventional method and the proposed method.**

are based on multimodal approaches, an auxiliary backbone network is required for HR extraction, which makes complex the overall system and increases the computational cost. On the other hand, some studies have developed models to improve FER performance through a multi-task learning (MTL) method that jointly learns tasks related to emotion recognition [9], [10], [11]. MTL shares parameters in a shared backbone model and implements task-specific layers. In other words, the shared backbone model learns multiple tasks simultaneously, increasing the generalization ability of the model and improving the performance of the primary task.

Inspired by the fact that both FER and non-contact HR measurement can utilize facial features extracted from facial images, we propose an MTL framework that simultaneously performs emotion recognition and HR prediction. 1 illustrates the comparison of conventional methods and our proposed method. Most of the previous methods use facial images as input to each backbone model, whereas the proposed method uses facial images as input to a shared backbone model to generate features shared by two task-specific layers. The shared backbone model generates efficient representations related to emotion recognition and HR prediction tasks using only a single visual modality as input.

The main contributions can be summarized as follows: (i) We propose an end-to-end multi-task learning framework with a task-shared convolutional recurrent neural network (CRNN) model and two task-specific branches for FER and HR estimation tasks. Physiological signals (HR), as well as emotion labels, are predicted from facial images to provide natural and non-contact emotion recognition; (ii) Extensive experiments conducted on benchmark DEAP and MAHNOB-HCI datasets demonstrate that the proposed MTL model achieves a better recognition accuracy without substantially increasing the computational cost.

II. RELATED WORK

A. FER USING RPPG SIGNALS

There are some existing studies utilizing rPPG signals to improve FER performance. For example, Du et al. used a bidirectional long and short-term memory (Bi-LSTM) and a convolutional neural network (CNN) to extract HR and facial features from face images [2]. A deep neural network (DNN) called SOM-BP was used for feature fusion. Yu et al. employed two 3D-CNNs to extract facial features from the facial image sequence and HR features from the forehead image sequence, respectively [3]. Ouzar et al. proposed the Xception network with squeeze and excitation (SE) to extract facial expression features and used a multi-task sequential shift convolutional attention network (MTTS-CAN [12]) to extract the rPPG signal. They extracted the heart-rate variability (HRV) features from the rPPG signal and combined them with facial features for emotion classification [8]. Therefore, we assume that the non-contact HR or HRV features obtained from rPPG will help the FER task.

B. MULTI-TASK LEARNING FOR FER

Unlike conventional single-task learning (STL), multi-task learning is a method of learning multiple tasks simultaneously using one model. Most of the emotion recognition studies using multi-task learning proposed a method of predicting values by simultaneously considering axes of emotion dimensions such as valence, arousal, and dominance [13], [14], [15]. However, it is difficult for the model to learn additional information because the information required for each prediction task is almost similar [16]. In previous FER studies, facial action unit detection was mainly used as an auxiliary task to improve emotion recognition performance [9], [10], [17]. On the other hand, Sang et al. showed that smile detection and gender classification tasks share mutual features with emotion classification [11]. Through these studies, it was found that joint learning with other tasks related to emotions improves accuracy compared to learning only emotion recognition tasks. However, since the aforementioned auxiliary tasks utilize the information that can be obtained from facial expressions, they are not helpful in situations where facial expressions are intentionally changed or hidden.

Therefore, in this study, we propose a multi-task learning model that can efficiently perform emotion recognition and HR prediction tasks simultaneously. To the best of our knowledge, this is the first study to demonstrate the efficiency of the non-contact HR prediction task in the FER study. This is to utilize physiological information together to build a reliable FER and is different from previous studies in that it does not utilize similar information related to facial expressions as an auxiliary task.

III. METHOD

In this section, we present an end-to-end multi-task deep learning framework for simultaneous learning of emotion recognition and HR estimation.

The overall architecture of the proposed MTL framework is shown in 2. The proposed framework consists of a CRNN backbone followed by two task-specific branches. Both tasks utilize facial features extracted by the CRNN.

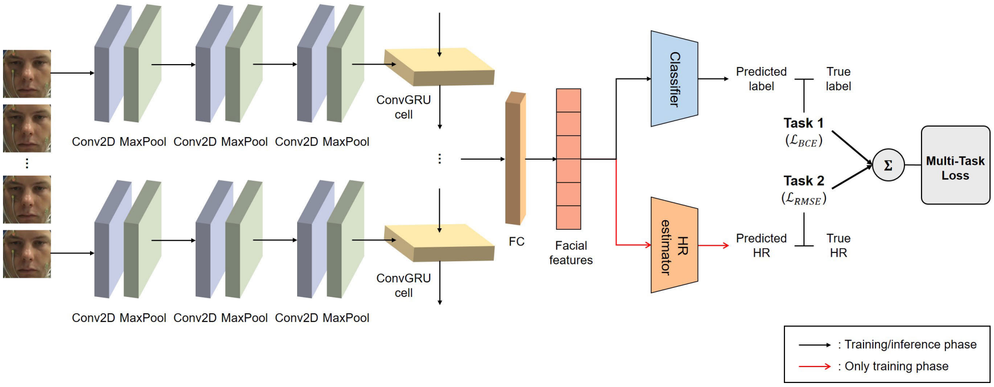

FIGURE 2. The overall architecture of the multi-task learning model.

The task-specific layers are implemented according to emotion recognition and HR estimation tasks, respectively. Therefore, a hard parameter-sharing strategy was applied to the facial features extraction layers while maintaining task-specific layers. The HR estimation task plays a supporting role during the training phase, but is not used at inference phase. The ablation study will show that this combination of emotion recognition and HR estimation tasks improves performance on the emotion recognition task, even though the HR estimation task is not used at the inference phase. In the following subsections, we describe the details about each part of the proposed MTL framework architecture.

A. FACIAL FEATURE EXTRACTION

We denote the i-th sample of facial video frames as $\boldsymbol{F_i} = \{f_i|1, 2, \ldots, n_i; f_i \in \boldsymbol{R}^{3\times h\times w}\}$, where n_i is the length of the video, and 3 represents the number of RGB channels. Each sample $\boldsymbol{F_i}$ has the size of $n \times 3 \times h \times w$, where n denotes the sequence length. The sample $\boldsymbol{F_i}$ is sent as input to 3-layer CNN (pre-trained on AFEW-VA dataset [18]) and then passed through ConvGRU [19] thus extracting the facial features from both spatial and temporal dimensions. The filter number of three convolutional layers is set as 32, 64, and 128, respectively, where all the kernels have the same size of 3×3. Each layer is followed by a ReLU activation function and a 3×3 max pooling layer. For each timestep t, the activation h_t^l of the ConvGRU is defined as:

$$z_t^l = \sigma(W_z^l * x_t^l + U_z^l * h_{t-1}^l) \tag{1}$$

$$r_t^l = \sigma(W_r^l * x_t^l + U_r^l * h_{t-1}^l) \tag{2}$$

$$\tilde{h}_t^l = tanh(W^l * x_t^l + U * (r_t^l \odot h_{t-1}^l)) \tag{3}$$

$$h_t^l = (1 - z_t^l)h_{t-1}^l + z_t^l\tilde{h}_t^l \tag{4}$$

where $*$ is the 2D convolution operation, $\odot$ is element-wise multiplication, and $W^l, W_z^l, W_r^l, U, U_z^l, U_r^l$ are all 2D convolution kernels. z_t^l is the update gate, which dictates the degree (0 to 1) to which the unit updates. r_t^l is the reset gate, which determines the degree of ignoring the state information of the previous time. $\tilde{h}_t^l$ is the candidate activation, which can be regarded as new information at the current time. A followed fully connected (FC) layer accepts the final hidden state and outputs the facial features.

B. EMOTION RECOGNITION TASK

Given a facial feature map, the emotion recognition task predicts the emotion label corresponding to the input image sequence. Each sequence was associated with rating values grouped according to the valence and arousal scales, which are the emotional states quantified using Russell's circumplex model [20]. We divide the rating values of 1-9 into two classes, low and high, using the value = 5 as the threshold. Low and high values are then assigned as 0 and 1, respectively.

In the classification layers, two FC layers with 20 neurons were used and a ReLU activation function and a dropout layer were used between the FC layers. The sigmoid function was applied for the last layer to output a probability value. The loss function for the emotion recognition task is the binary cross-entropy loss, which is calculated as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \ln \hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i)) \tag{5}$$

where N denotes the batch size, y denotes the ground-truth label and $\hat{y}$ denotes the predicted probability of the i-th sample.

C. HEART RATE ESTIMATION AUXILIARY TASK

Shared facial features are also fed into the HR estimation branch. A fully connected layer was applied to estimate HR from the extracted facial feature map. For the HR estimation task, the root mean square error (RMSE) loss is used to measure the performance of the model. The ground-truth HR was calculated from the PPG/ECG signals. The RMSE loss is calculated as follows:

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{HR}_i - HR_i)^2} \tag{6}$$

where $\hat{HR}$ is the output of the HR estimation task, HR is the ground-truth HR of each sample.

D. MULTI-TASK LOSS

The proposed MTL aims to optimize two tasks, emotion recognition and HR estimation. In principle, after joint learning, the shared features will have the ability to classify emotion and measure HR at the same time, that is, the generalization ability of facial features becomes stronger, which makes the two tasks mutually reinforcing. The final multi-task loss function is defined as:

$$\mathcal{L}_{ALL} = \alpha\mathcal{L}_{BCE} + \beta\mathcal{L}_{RMSE} \tag{7}$$

We aggregate the weighted sum of the losses from two tasks to compute the overall loss. In equation (7), the $\mathcal{L}_{BCE}$ indicates the binary cross entropy (BCE) for emotion recognition task and the $\mathcal{L}_{RMSE}$ indicates RMSE loss for the HR estimation task. The α and β are two weighting factors that balance two tasks. In this study, we set α and β to be 1 and 0.01, respectively.

IV. EXPERIMENTS

A. DATASETS

We conduct extensive experiments on multimodal datasets (DEAP [21] and MAHNOB-HCI [22]) to validate the performance of our proposed framework. These two datasets are the only publicly available datasets that contain facial videos, synchronized physiological signals, and emotion labels. Face videos were used as input of the model. Emotion labels were used for the primary task. Photoplethysmography (PPG) and electrocardiograms (ECG) were used to calculate HR to be compared with predicted HR in HR estimation auxiliary task.

B. IMPLEMENTATION DETAILS

For all the experiments, we first crop the face based on facial landmarks and resize facial images to 64 × 64. We implement the MTL framework using PyTorch. We used ConvGRU-based CRNN as the backbone for extracting shared facial features. The model for each task is trained during 15 epochs with the Adam optimizer with a learning rate of 0.001. The batch size is 64. The experiment was performed on a computer with an Intel(R) Core(TM) i7-10700K CPU @ 3.80 GHz 3.79 GHz and NVIDIA GeForce RTX 3080 graphics processing unit (GPU). We report the classification accuracy on the FER task.

V. RESULTS

A. PERFORMANCE COMPARISON

We evaluated the performance of our proposed method and the conventional methods on the same benchmark datasets. table 1 shows the experimental results for the DEAP dataset. Our proposed method achieved promising results and performed better than the reference methods. The classification accuracy of our proposed framework gave 24.4% and 24.1% better results in valence and arousal classification compared to the reference models using facial modality such as [23], [24], and [25]. We also compared the studies using the PPG modality in that we performed an rPPG-based HR estimation task. The proposed method showed significantly higher results in both valence and arousal classification than those using the PPG signals [26], [27], [28].

TABLE 1. Results (in %) on the DEAP dataset.

Method	Modality	Valence	Arousal
MTL-CNN	Face	72.31	71.15
VGG-16	Face	72.28	74.47
3D-CNN	Face	71.00	71.11
R-ELM	PPG	64.06	63.28
mRMR+RF	PPG	70.23	68.59
CNN	PPG	75.8	76.3
Ours	Face	**96.26**	**96.22**

TABLE 2. Results (in %) on the MAHNOB-HCI dataset.

Method	Modality	Valence	Arousal
MTL-CNN	Face	73.33	69.79
VGG-16	Face	85.13	81.57
NCA	ECG	64.1	66.1
WMD-DTW	HR + EDA	93.69	94.00
Ours	Face	**96.97**	**96.75**

Table 2 shows the experimental results for the MAHNOB dataset. The proposed methods yielded significantly better results in most cases. However, Albraikan et al. showed high performance by combining EDA signals with HR [29]. However, our proposed framework showed results up to 23.1% higher in valence and up to 27% higher in arousal using only facial modality.

B. SINGLE-TASK LEARNING VERSUS MULTI-TASK LEARNING

To verify that the improvement of recognition accuracy really benefits from the multi-task learning framework. We compare the multi-task learning model and the single-task learning model. The comparison results are given in table 3. This proves that multi-task learning plays a big role in improving FER performance. Our proposed method respectively achieves the improvements of 4.1% and 3.2% in terms of average recognition accuracy on valence and arousal classification. Moreover, many studies on using deep learning to recognize emotions have considered accuracy but ignored efficiency. However, our proposed MTL framework has no significant difference in the number of parameters compared to a single-task learning (STL) model.

C. EFFECT OF MULTI-TASK LEARNING

In this section, we perform experiments to verify the efficacy of the proposed method. The following experiments adopt CRNN as the shared backbone and are trained for two tasks on the DEAP dataset.

In the MTL, we jointly optimize emotion recognition task loss $\mathcal{L}_{BCE}$, HR prediction task loss $\mathcal{L}_{RMSE}$, and overall loss $\mathcal{L}_{ALL}$. In 3, it can be seen that the emotion recognition task loss $\mathcal{L}_{BCE}$, converges faster in the MTL framework,

TABLE 3. Results of ablation study experiments on the DEAP and MAHNOB-HCI datasets.

Method	# Params	DEAP		MAHNOB-HCI	
		Valence	Arousal	Valence	Arousal
STL (w/o $\mathcal{L}_{RMSE}$)	1,052,473	90.4 ± 0.5	91.8 ± 0.6	96.4 ± 1.1	96.1 ± 0.7
MTL (w/ $\mathcal{L}_{RMSE}$)	1,052,490	**96.3 ± 0.5**	**96.2 ± 1.0**	**97.0 ± 0.5**	**96.7 ± 0.5**

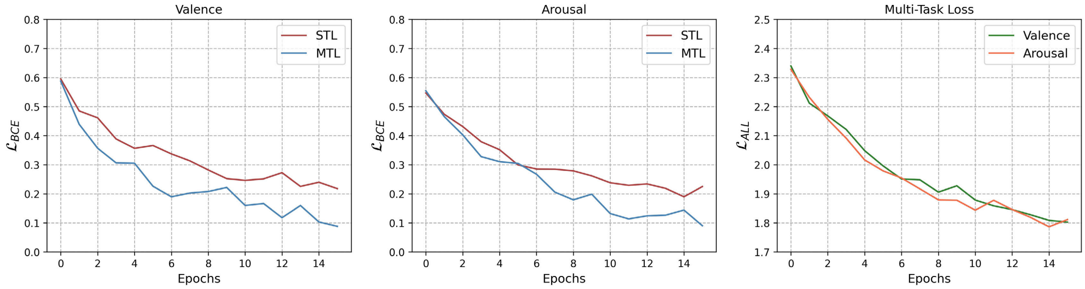

FIGURE 3. Comparison of validation loss of MTL and STL models.

TABLE 4. Performance of models with different coefficients.

	acc	loss
$\beta = 0$	93.67%	0.1607
$\beta = 0.001$	96.41%	0.0945
$\beta = 0.01$	**96.55%**	**0.0919**
$\beta = 0.1$	95.51%	0.1235
$\beta = 1$	94.49%	0.1514

even though the HR prediction task loss $\mathcal{L}_{RMSE}$ is considered only during model training. As a result, we confirmed that the auxiliary use of the non-contact HR estimation task in the model learning phase improves the FER performance by allowing the FER task, which is the main task, to converge more quickly. It can be seen that the $\mathcal{L}_{RMSE}$ performed as regularization in the multi-task loss function and showed better results in the validation dataset compared to the single model.

D. THE IMPACT OF THE HR LOSS

Table 4 shows accuracy and BCE loss for four weight coefficients of β. $\beta = 0$ corresponds to STL where the model only performs the FER task. Since the FER result is a probability value between 0 and 1, and the result of HR estimation is the heart rate in bpm. So, if the model is trained with the weight equally fixed at 1:1, the RMSE loss value is larger than the BCE loss value. Therefore, the MTL model achieved the best performance when β was 0.01, where the two tasks are balanced.

VI. CONCLUSION AND DISCUSSION

In this study, we introduced a novel MTL framework for emotion recognition from facial videos, concurrently addressing FER and HR estimation within a shared convolutional recurrent neural network. Our findings demonstrate significant advancements over conventional methods, particularly in datasets such as DEAP and MAHNOB-HCI. The MTL approach not only outperformed single-task models but also surpassed a multi-modal system in the case of the MAHNOB-HCI dataset, highlighting its efficacy in integrating diverse sources of information for robust emotion recognition.

However, our study also reveals important limitations and challenges. One critical observation is the sensitivity of the model's performance to the relative weights assigned to different tasks in MTL. We found that adjusting the multi-task loss weight significantly impacts individual task performance. Specifically, equalizing the weights between FER and HR estimation (1:1 ratio) led to a notable decrease in HR RMSE but also resulted in reduced accuracy in emotion recognition. This trade-off underscores the delicate balance required in MTL to optimize performance across disparate tasks with varying initial loss values.

Moreover, the diversity across datasets such as DEAP and MAHNOB-HCI presents another layer of complexity. These datasets capture emotional responses under different conditions—controlled laboratory settings versus naturalistic environments—highlighting the need for models that generalize well across diverse emotional states and contexts. Our framework demonstrated robust performance across these datasets, yet future research should continue exploring methods to enhance model adaptability and generalization capabilities across broader emotional spectrums.

The insights gained from this study suggest several avenues for future research. Firstly, exploring dynamic approaches to adjust task weights during training could mitigate the challenges associated with static weight assignment in MTL. Recent advancements in dynamic multi-task learning offer promising directions to improve the flexibility and performance stability of such frameworks. Secondly, further investigation into enhancing dataset diversity representation within training protocols could yield more resilient models capable of handling varied emotional expressions and environmental conditions.

In conclusion, while our MTL framework represents a significant advancement in non-contact emotion recognition technology, it also highlights ongoing challenges and opportunities in the field. Addressing issues such as task weight balance and dataset diversity comprehensively is crucial. By doing so, our approach can pave the way for more effective and adaptive HCI systems. These advancements are expected to be particularly significant for applications ranging from healthcare to virtual reality, where accurate emotion recognition is important for enhancing user experience and interaction.

REFERENCES

[1] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.

[2] G. Du, S. Long, and H. Yuan, "Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments," *IEEE Access*, vol. 8, pp. 11896–11906, 2020.

[3] W. Yu, S. Ding, Z. Yue, and S. Yang, "Emotion recognition from facial expressions and contactless heart rate using knowledge graph," in *Proc. IEEE Int. Conf. Knowl. Graph (ICKG)*, Aug. 2020, pp. 64–69.

[4] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1359–1367.

[5] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 2, pp. 52–58, 2021.

[6] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103029.

[7] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.

[8] Y. Ouzar, F. Bousefsaf, D. Djeldjli, and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2459–2468.

[9] D. Kollias, "ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2327–2335.

[10] P. T. Dac Thinh, H. M. Hung, H.-J. Yang, S.-H. Kim, and G.-S. Lee, "Emotion recognition with sequential multi-task learning technique," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3586–3589.

[11] D. V. Sang, L. T. B. Cuong, and V. Van Thieu, "Multi-task learning for smile detection, emotion recognition and gender classification," in *Proc. 8th Int. Symp. Inf. Commun. Technol.*, Dec. 2017, pp. 340–347.

[12] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. NeurIPS*, 2020, pp. 19400–19411.

[13] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, Aug. 2017, pp. 1103–1107.

[14] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 217–225, Jul. 2019.

[15] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6705–6709.

[16] Y. Li, H. Yang, J. Li, D. Chen, and M. Du, "EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-CAM," *Neurocomputing*, vol. 415, pp. 225–233, Nov. 2020.

[17] G. Pons and D. Masip, "Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4764–4771, Jun. 2022.

[18] 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.

[19] N. Ballas, L. Yao, C. J. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–11.

[20] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.

[21] 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. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[22] 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.

[23] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, May 2019.

[24] T.-P. Jung and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2022.

[25] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression," in *Proc. 4th Int. Conf. Mach. Learn. Soft Comput.*, Jan. 2020, pp. 81–85.

[26] Z. Guendil, Z. Lachiri, and C. Maaoui, "Computational framework for emotional VAD prediction using regularized extreme learning machine," *Int. J. Multimedia Inf. Retr.*, vol. 6, no. 3, pp. 251–261, Sep. 2017.

[27] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *J. Med. Biol. Eng.*, vol. 40, no. 2, pp. 149–157, Apr. 2020.

[28] M. Lee, Y. K. Lee, M.-T. Lim, and T.-K. Kang, "Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features," *Appl. Sci.*, vol. 10, no. 10, p. 3501, May 2020.

[29] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8402–8412, Oct. 2019.

YERIM JI received the B.S. and M.S. degrees from the Department of IT Engineering, Sookmyung Women's University, Seoul, South Korea, in 2021 and 2023, respectively.

She is currently an AI Research Engineer with the Department of AI Research, EXOSYSTEMS, Seongnam, Gyeonggi-do, South Korea. Her research interests include biomedical engineering and artificial intelligence.

SUH-YEON DONG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2010, 2011, and 2016, respectively.

She is currently an Associate Professor with the Division of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul, South Korea. Her research interests include machine-learning-based biosignal processing and cognitive neuroscience.