SIP (2021), vol. 10, e12, page 1 of 12
doi:10.1017/ATSIP.2021.11

ORIGINAL PAPER

3D skeletal movement-enhanced emotion recognition networks

JIAQI SHI,[1,2] CHAORAN LIU,[3] CARLOS TOSHINORI ISHI[2,3] AND HIROSHI ISHIGURO[1,3]

Automatic emotion recognition has become an important trend in the fields of human–computer natural interaction and artificial intelligence. Although gesture is one of the most important components of nonverbal communication, which has a considerable impact on emotion recognition, it is rarely considered in the study of emotion recognition. An important reason is the lack of large open-source emotional databases containing skeletal movement data. In this paper, we extract three-dimensional skeleton information from videos and apply the method to IEMOCAP database to add a new modality. We propose an attention-based convolutional neural network which takes the extracted data as input to predict the speakers' emotional state. We also propose a graph attention-based fusion method that combines our model with the models using other modalities, to provide complementary information in the emotion classification task and effectively fuse multimodal cues. The combined model utilizes audio signals, text information, and skeletal data. The performance of the model significantly outperforms the bimodal model and other fusion strategies, proving the effectiveness of the method.

Keywords: Deep learning, emotion recognition, gesture, skeleton

Received 28 December 2020; Revised 14 July 2021

I. INTRODUCTION

With the rapid development of artificial intelligence technology and the widespread popularity of smart devices, the study of human–computer natural interaction has been widely concerned. Human–computer natural interaction aims to provide effective and natural interaction between humans and computers so that the machine can understand the user's intention and generate natural feedback based on the user's needs and behavior. As an important subject of human–computer interaction, emotion recognition attracts increasing attention due to its vital role and wide application in intelligent interaction systems, mental health care, and so on [1, 2]. For example, an interactive system with the ability of emotion detection can decipher the emotional thinking by analyzing the user's emotional state and generate appropriate behaviors, which is conducive to providing users with more efficient and comfortable services [3, 4].

Humans recognize emotions through a variety of different modalities during natural interaction, e.g. facial expressions, voice tone, and body movement [5]. Through the acquisition of information from different modalities, humans can obtain multiple related but different aspects of emotional information, to judge the emotional state more accurately. In the research of automatic emotion recognition, it is also a common practice to improve the performance of the system by fusing multimodal information and leveraging the strengths of each modality [6–9]. Although these studies have used information from multiple modalities to achieve better prediction performance, what is the best mechanism to effectively integrate multimodal cues is still an unclear and promising research topic.

Gesture is one of the most important forms of nonverbal communication, which plays an extremely important role in the recognition of emotions [10]. Exploring the relationship between gesture and emotion through affective computing is a very meaningful and challenging subject. Most of the existing body-tracking methods are based on video data, which makes it extremely challenging and usually amounted to single-frame analysis [11, 12]. On the contrary, skeletal movement data are the most natural and intuitive depiction of body movements, which can represent the interrelationships of body parts and joint movements [13, 14]. However, the existing research in the field of multimodal emotion recognition mainly focuses on analyzing features of text information, speech signals, and facial expressions, and the role of gesture in emotion detection is rarely considered. One of the important reasons gesture modality is seldom considered is the lack of large open-source emotional databases containing three-dimensional (3D) skeletal movement data.

[1]Graduate School of Engineering Science, Osaka University, Osaka, Japan
[2]Guardian Robot Project, RIKEN, Kyoto, Japan
[3]Advanced Telecommunications Research Institute International, Kyoto, Japan

Corresponding author:
J. Shi
Email: shi.jiaqi@irl.sys.es.osaka-u.ac.jp

There are some multimodal emotional databases containing 3D skeleton data, such as emoFBVP database [15] and Multimodal Database of Emotional Speech, Video and Gestures [16]. Although they contain multiple modalities including skeleton data representing body movements, they all have some disadvantages, i.e. they have a relatively small size, and there is no dialog and interaction between people. IEMOCAP [17] is a database with over 10 000 samples, which contains improvised behaviors and relatively natural conversations in hypothetical interaction scenarios. However, only MOCAP (motion capture) data recorded by the sensors on the head and hands of the participants, rather than the skeleton data of the joints, are included in this database, which ignores the movements of the spine, arms, and shoulders that play a very important role in emotional expression and prediction.

In this study, we add a new modality of skeletal data for IEMOCAP database, and propose a skeletal movement enhanced network to verify the effectiveness of this modality for emotion prediction. This paper mainly makes the following contributions:

(1) We extract 3D skeletal movement data from raw video based on pose estimation, and the method can be used to expand existing databases by adding a new modality. The extracted data are a representation of body movements, in the form of 3D joints positions sequence.
(2) We propose an attention-based convolutional network for obtaining informative representations of the skeleton data to identify emotional classes.
(3) We propose a multimodal network that fuses the 3D skeletal movement data extracted by the proposed model, with the audio and text features extracted by existing methods. The performance of the model outperforms the prior model significantly, which proves the effectiveness of the extracted modality.

The rest of the paper is organized as follows: Section II introduces some related studies; Section III describes the method of extracting skeleton data and our uni-modal and multi-modal models; the experiment and results are described in Section IV; we perform an ablation study and make a further discussion about the gesture-based emotion recognition model in Section V; and finally, we conclude this paper with a brief summary and mention some future research.

II. RELATED STUDIES

A) Relationship between emotion and gesture

Many studies have shown that people can analyze emotional information from nonverbal expressions, such as facial expressions, and use the information to infer others' emotional states fairly accurately [18, 19]. Similarly, as an important part of nonverbal expression, gesture also has a significant relationship with emotion. Not only static body posture can promote emotion perception [20, 21], but also the dynamic characteristics of body movement, e.g. amount, speed, force, fluency and size, can help to accurately identify emotions [22].

B) Emotion recognition using body motion information

Some body movement analyzing based emotion recognition methods have been proposed in recent years. These methods can be categorized into hand-crafted features-based methods and deep-learning methods using an end-to-end manner. The first type of methods design some hand-crafted features to capture the properties of body movement, for example, kinematic-related features, spatial extent-related features, and leaning-related features [23, 24]. Inspired by the great performance of end-to-end deep learning in many tasks, some researchers also use end-to-end deep-learning-based methods to analyze the emotional features of joint motion [13, 25]. However, these studies are limited by the relatively small amount of data and the lack of interaction and dialog between people, so it is difficult to study the real emotions expressed by body movement in the scene of natural interaction. Our method effectively alleviates the lack of data in the field of gesture emotion recognition by extending the existing large emotional interaction database, which has a positive effect on boosting the research in this field.

C) Multimodal fusion for emotion recognition

Many studies have fused various types of information from multiple modalities to improve the performance of multimodal emotion recognition. Fusion strategies in these studies can be divided into three typical categories, namely feature-level (early) fusion, decision-level (late) fusion, and model-level fusion [26]. In feature-level fusion, the concatenation of features from different modalities is used to construct a joint feature vector and is fed into a single classifier. Decision-level fusion makes the prediction of each modality separately and then combines the predictions of different modalities to obtain the multimodal prediction. Model-level fusion is a compromise between the two extremes and uses the concatenation of high-level feature representations from various modalities. For neural networks, model-level fusion could be a concatenation of different hidden layers from multiple modalities [27]. Feature-level fusion does not perform well if the input features from different modalities differ in the temporal characteristics, and the high-dimensional feature set may easily suffer from the problem of data sparseness. Therefore, most research focuses on the decision-level fusion and the model-level fusion [28, 29]. The network proposed in [30] encoded the information from audio and text and directly concatenated the features of the two modalities to predict the emotion class. In [31], the proposed tensor fusion network was used to model both intra-modality and inter-modality dynamics directly. In [32] and [33], it is shown that

cross-modal attention can be used to learn interactive information between audio and text modalities to improve the emotion recognition performance. Siriwardhana *et al.* [34] introduced an attention-based fusion mechanism that can combine multimodal self-supervised learning features for emotion recognition. However, relatively few studies have investigated the fusion of information from body movements and more effective fusion methods still need further research. In our study, we fuse the information from body movement with audio and text modalities and employ a graph attention [35] based fusion method that considers the high-level features from each modality as a node and assigns each node in the graph different weights according to the features of neighboring nodes, so that the model can not only find the inter-modal and intra-modal relationships, but also effectively utilize the advantages of each modality.

III. METHODOLOGY

This section describes the method of extracting gesture modality from the video and the structure of the proposed models. We extract skeleton data from the original video files and perform data cleaning aimed at removing noise. The preprocessed data are fed into our spatial attention-based convolutional network to extract features related to emotional expression. The features are concatenated with representations of text and audio to form a multimodal feature vector used for emotion prediction.

A) Skeletal data extraction

Considering that skeleton data can be directly used in emotion prediction instead of processing image sequence, we adopt a human pose estimation-based method to extract human skeletal movement data from raw videos. The skeleton is essentially a coordinate representation of the joint positions of the human body, which can be used to describe body movements. The data require some preprocessing operations in order to be fed into the emotion classification model. The extracted data can be used not only for emotion classification but also for the study of emotional motion generation and action interaction.

A.1 Human pose estimation

Human pose estimation is used to reconstruct human joints and limbs based on images, obtaining the coordinate representation of each joint, and creating gestures by forming connections between joints. We detect the two-dimensional (2D) joint position from the image sequence of video, and then project the joint position in 3D coordinates from the 2D pose data.

In this study, AlphaPose is used as a 2D pose detector. AlphaPose [36–38] is an open-source pose estimation system with extremely high accuracy. The AlphaPose detector pretrained on the COCO dataset [39] is applied to detect the 2D keypoints of the same person across the frames of the video. For 3D pose estimation, we used the pretrained temporal convolution model proposed in [40], which is proved to be effective at predicting 3D poses in videos. The model takes 2D joint sequences as input, applies dilated temporal convolution to obtain long-term information, and generates 3D pose estimation results. In this way, we obtained the position data of the joints in the 3D coordinate system from the original videos.

A.2 Preprocessing

Due to video quality and error of detection, there is high-frequency noise in the detection results of 2D key points, which leads to fluctuations in the estimated 3D joint position data. In order to filter out the noise and get clean data, a low-pass filter is applied to the 3D position result across time for each joint. Here, we use the filter order of 8 and normalized cut-off frequency of 0.1 as the parameters of the low-pass filter. The low-pass filter significantly reduced the influence of noise during the detection process.

Besides, the lower body of the actors in the IEMOCAP dataset is invisible in the video. Therefore, the predicted pose of the lower body is not reliable and only the data of 10 joints of the upper body is used in this study. As a result of the different lengths of the video clips, the skeleton data in each sample are a variable-length sequence, which cannot be directly used as the input of convolutional network. To unify the length of the sequence, zero padding is applied to the data.

B) Spatial multi-head attention-based convolutional network

The structure of our spatial multi-head attention-based convolutional network (SMACN) is shown in Fig. 1. It takes the time sequence of skeletal movement as input, extracts emotion-related features through the convolutional layers and the attention layer, and predicts the emotion class. The convolutional layers are trained to detect emotional features from sequence data. The attention mechanism reduces the feature dimension by evaluating the effectiveness of each feature vector and weighting it. The final feature vector is used to predict the emotion class.

In recent years, convolutional neural networks (CNNs) have achieved excellent performance in many tasks related to digital image processing, e.g. target detection [41–43] and human pose estimation [40, 44]. CNNs are capable of compressing images with large amounts of data to a relatively small dimension, without damaging most of the effective features. Considering that to some extent, the skeleton data can also be regarded as a special kind of image data, CNNs are employed to extract the high-level features of skeletal movement data in the spatial and temporal domains.

The sequence of skeleton data is fed into the model as the input of size $T \times V$, where T represents the number of time steps in the sequence and V denotes the number of joint positions in the skeleton data. In our model, we use a 2D convolutional layer and four convolutional units to extract features from the data. Each convolutional unit contains a 2D convolutional layer and a 2D maxpooling layer. The output size of the convolutional layers is $T' \times V' \times C$, where T'

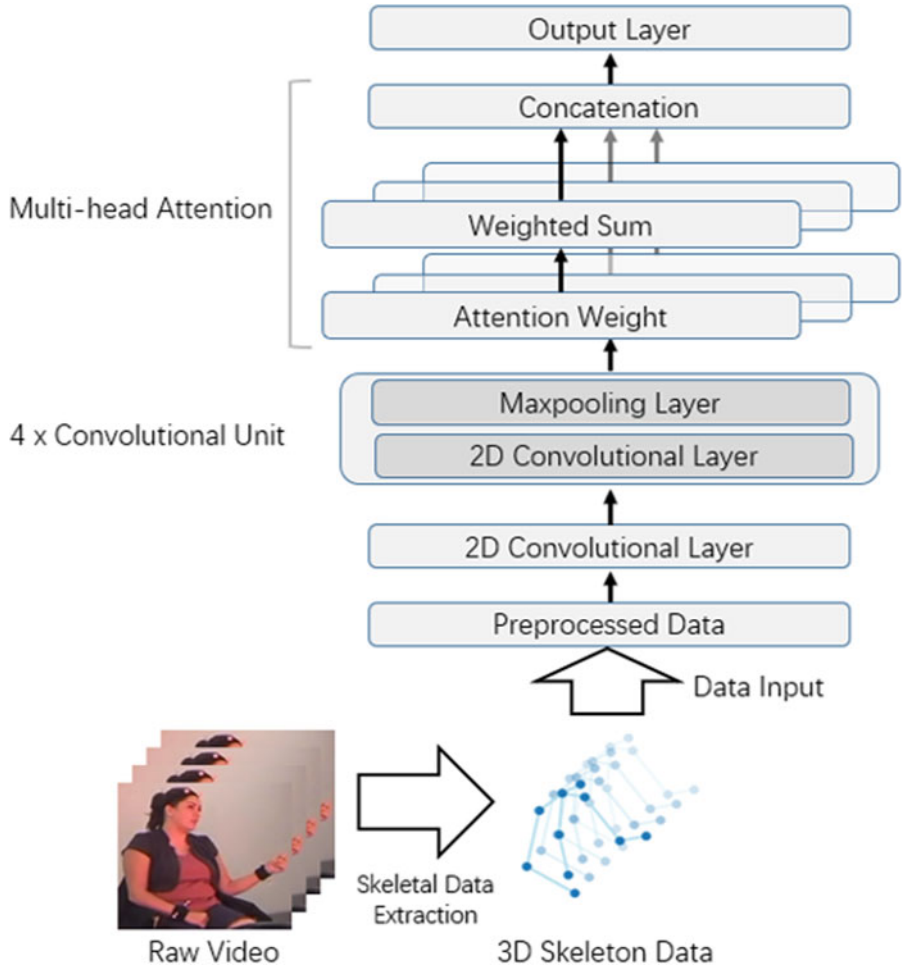

Fig. 1. Architecture of the proposed SMACN.

and V' indicate the feature size of the temporal domain and the spatial domain, respectively, and C is the channel size, i.e. the number of feature maps.

In the spectrogram representation-based speech emotion recognition task of [45] and [46], the attention pooling method can reduce the number of network parameters and make the model perceiving which parts of the sequence are more emotion-relevant. Similarly, not all the temporal–spatial regions of skeletal motion data contribute equally to emotional states. Therefore, we use a multi-head spatial attention layer on the output of our convolution unit to enable the network to find more effective parts. The multi-head attention mechanism not only allows the model to find multiple features in different aspects, but also has a low-computational cost.

The input size of the attention layer is $T' \times V' \times C$. We represent the vector composed of the elements at the same position in the feature map of each channel as $a_i \in \mathbb{R}^C$, whose amount is $L = T' \times V'$:

$$\mathbf{A} = \{a_1, ..., a_L\}. \tag{1}$$

Then we apply a linear transformation to a_i, and use the nonlinear activation function tanh to calculate the new representation of a_i:

$$y_i = \tanh(\mathbf{W}a_i + b), \tag{2}$$

where $\mathrm{W} \in \mathbb{R}^{F \times C}$ represents the weight of the linear transformation, and the bias is $b \in \mathbb{R}^F$. Then the learnable matrix $\mathbf{U} \in \mathbb{R}^{H \times F}$ is multiplied by this vector to calculate the importance weight vector $E_i \in \mathbb{R}^H$:

$$E_i = \mathbf{U}y_i, \tag{3}$$

$$E_i = \left[e_i^1, ..., e_i^H\right]^T. \tag{4}$$

After this, the softmax function is applied for each head to normalize the attention weights:

$$\alpha_i^{head} = \frac{\exp e_i^{head}}{\sum_{k=1}^{L} \exp e_k^{head}}, \tag{5}$$

All the weights on each attention head form a 2D spatial attention map $M^{head} \in \mathbb{R}^{T' \times V' \times 1}$. We concatenate the weighted sums of the input feature vectors with attention

weight on each head as the emotional vector representation $v_s \in \mathbb{R}^R, R = H \times L$:

$$v_s = \sum_{i=1}^{L} \alpha_i^1 a_i \oplus \cdots \oplus \sum_{i=1}^{L} \alpha_i^H a_i, \tag{6}$$

where $\oplus$ represents the concatenation operation. Finally, the emotion vector is passed into the fully connected output layer to obtain the prediction result.

In our experiment, the 2D convolutional layer has 128 channels for output. The four convolutional layers in the convolutional units have 256 channels for output. Each convolutional layer is followed by a ReLU activation function. We apply dropout with a rate of 0.5 to avoid overfitting in the training process.

C) Skeletal movement-enhanced emotion recognition network

In order to confirm the utility of emotion-related representations extracted from skeleton data, we construct Skeleton Movement-enhanced Emotion Recognition Network (SMERN) for integrating multi-modal information, including text, audio, and skeleton information (see Fig. 2). For text and audio data, Multimodal Dual Recurrent Encoder (MDRE) [30] is used as the basic model. The MDRE model is composed of Audio Recurrent Encoder (ARE) and Text Recurrent Encoder (TRE). It takes MFCC features, prosodic features, and textual transcripts as input at the same time, considering the relevance of sequential audio features, statistical audio features, and text information. ARE takes MFCC features as input. The concatenation of the final hidden state of the audio encoder and the prosodic features is passed into the fully connected layer to form the vector representation A. For text modality, the sequence of word embedding vectors, that is formed by the transcription script being passed into the embedding layer, is fed to the text encoder. The final hidden state after a fully connected layer is the vector representation T of the text. The concatenation of vectors A and T contains both audio and text information and is used for emotion prediction.

In our study, we propose a two-phase hierarchical network to consider the features of audio, text, and gesture at the same time. The uni-modal features are fed to ARE, TRE, and the proposed CNN, respectively, to obtain the uni-modal feature representations (200, 200, and 256 dimensions, respectively) in the first phase. In the second phase, the feature representations are passed through three fully connected layers respectively to reduce the dimensions by half, and then concatenated to pass to the output layer for final emotion prediction.

D) Skeletal movement-enhanced emotion recognition network with graph attention

To make better use of the multimodal features and fuse cross-modal information more effectively, we employ a multi-head graph attention on the extracted features (Fig. 3). We first apply a linear transformation to adjust the dimensions of the uni-modal feature vector representations from ARE, TRE, and the proposed CNN. The graph attention module takes a set of feature vectors $\mathbf{v} = \{v_a, v_t, v_s\}, v_i \in \mathbb{R}^{F_{in}}$ as input, where F_{in} is the number of input features, regards each vector as a node and computes the interactive information between modalities. In the graph attention module, the attention coefficient α_{ij} is formulated as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[\mathbf{W}v_i || \mathbf{W}v_j]))}{\sum_{v_k \in \mathbf{v}} \exp(\text{LeakyReLU}(a^T[\mathbf{W}v_i || \mathbf{W}v_k]))}, \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{F_{out} \times F_{in}}$ is a trainable weight matrix for linear transformation and $a \in \mathbb{R}^{2F_{out}}$ is a learnable weight vector. The weights of edges are calculated based on the features of each pair of connected nodes and are normalized for each node.

Then we calculate the weighted sum of all nodes and average the output tensor across attention heads:

$$v_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{v_k \in \mathbf{v}} \alpha_{ij}^k \mathbf{W}^k v_j + b\right), \tag{8}$$

where σ represents ELU activation function, K is the number of attention heads, and $b \in \mathbb{R}^{F_{out}}$ is the bias.

In our proposed Skeletal Movement enhanced Emotion Recognition Network with Graph Attention (SMERN-GA), the input of the graph attention module is three 128-dimensional vectors (audio, text, and gesture feature vectors) and the output is three 64-dimensional feature vectors. The number of the attention heads is set to 4. The three output vectors are then passed into the output layer to predict emotions.

IV. EXPERIMENTS AND RESULTS

A) Dataset description

We use the interactive emotional dyadic motion capture database (IEMOCAP), which contains more than 10 h of audio and video data from 10 actors. For simulating natural binary interaction between people, the dialogs between a male and a female in scripted or hypothetical scenarios are recorded in the database. The emotion label set includes 10 classes, i.e. neutral emotion, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and other. The category of each sample is evaluated by 3–4 annotators. We adopt four emotional labels of them, i.e. happy, sad, angry, and neutral, and merge the excitement subset into the happiness subset to keep it consistent with previous research. In some videos of the database, the image of one of two actors is missing, hence these samples are removed. The final dataset includes a total of 5492 utterances (1606 happy, 1081 sad, 1102 angry, and 1703 neutral).

B) Feature extraction

To make a fair comparison with the previous model, our feature extraction follows the study of [30]. For speech data,

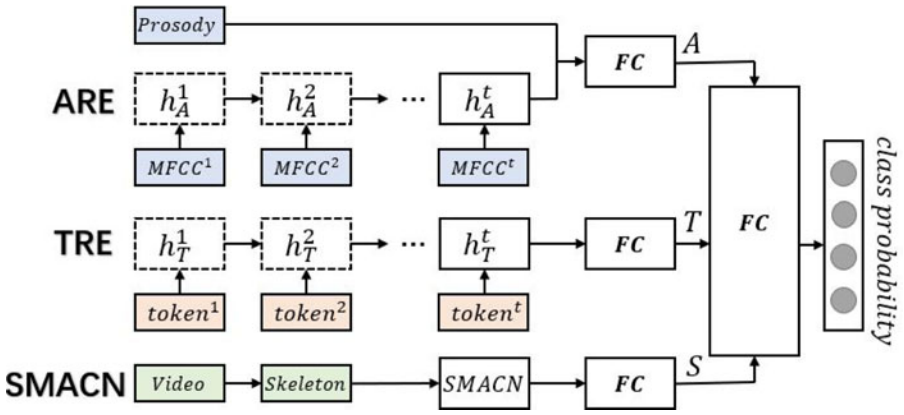

Fig. 2. SMERN framework where audio, text, and gesture are used for emotion classification simultaneously.

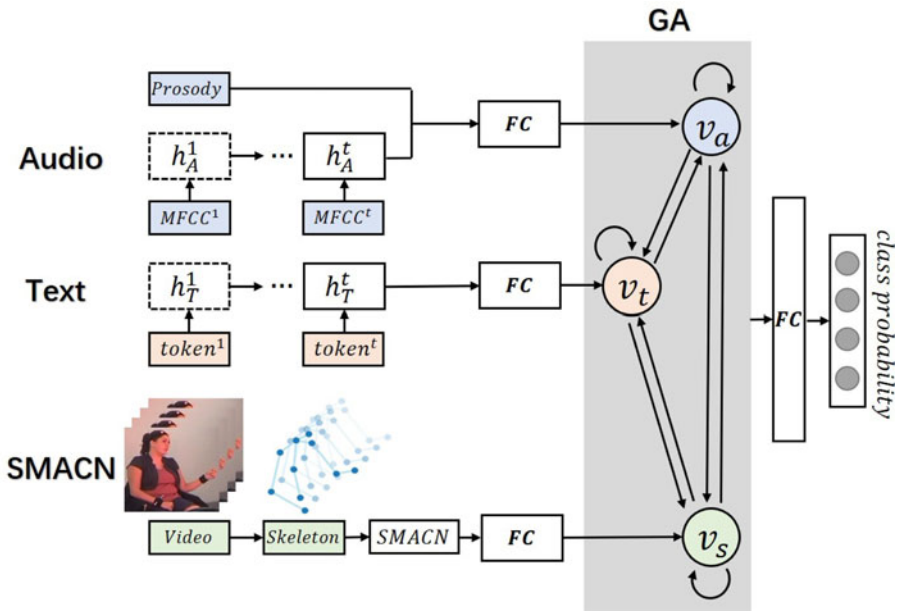

Fig. 3. Illustration of the multimodal model with graph attention. GA represents the graph attention module.

OpenSMILE toolkit [47] is employed to extract MFCC features and prosodic features. The MFCC features consist of 39 features, whose frame size is set to 25 ms at a rate of 10 ms with the Hamming window. The prosodic features include 35 features, comprising the Fo fundamental frequency, the voicing probability, and the loudness contours. For the textual transcript, we use a pretrained 300-dimensional GloVe vector [48] to initialize each token.

C) Experiment setting

In our experiments, five-fold cross-validation is applied to evaluate the performance of the model. The samples of each fold are divided into the training set, development set, and test set, with a ratio of 8:0.5:1.5. This process is repeated for five iterations, and then the prediction results are integrated to calculate the final value. Crossentropy loss is employed as the loss function for the outputs of all networks after passing the softmax function. We trained the models with the size of mini batch 128 and Adam as optimizer. The learning rate was set to 0.001. All models were implemented by Pytorch framework.

D) Results

Consistent with the basic model, the weighted average precision (WAP) is calculated as the indicator of the model performance. Since WAP is rarely used in emotion recognition tasks, we also list the unweighted and weighted average recall (UAR/WAR) of the models in our experiment. Precision and recall are defined as:

$$Precision = \frac{tp}{tp + fp}, \tag{9}$$

$$Recall = \frac{tp}{tp + fn}, \tag{10}$$

where tp is the number of true-positive samples, fp is the number of false-positive samples, and fn is the number of

Table 1. Comparison for unimodal and multimodal

Model	Modality	WAP	UAR	WAR
Mocap$_$Model [49]	M	51.1	–	–
ARE [30] (reported)	A	54.6 ± 0.9	–	–
TRE [30] (reported)	T	63.5 ± 1.8	–	–
ARE [30]	A	54.7 ± 0.9	59.7	57.1
TRE [30]	T	63.9 ± 1.0	65.9	64.5
SMACN	S	64.8 ± 0.6	65.6	65.9
MDRE [30] (reported)	A + T	71.8 ± 1.9	–	–
MDRE [30]	A + T	71.8 ± 0.9	73.4	72.7
CMA [32]	A + T + S	76.1 ± 0.4	77.2	76.9
Xu *et al.* [33]	A + T + S	76.7 ± 0.4	77.6	77.2
SMERN	A + T + S	76.6 ± 0.6	78.3	77.7
SMERN-GA	A + T + S	**77.9 ± 0.6**	**79.0**	**78.6**

false negative. And the weighted score is calculated as:

$$weighted = \frac{1}{\sum_{l\in L}|\hat{y}_l|}\sum_{l\in L}|\hat{y}_l|\phi(y_l,\hat{y}_l), \tag{11}$$

where L is the set of labels, $\hat{y}$ is the true labels, y is the predicted labels, $|\hat{y}_l|$ is the number of true labels that have the label l, $|y_l|$ is the number of predicted labels that have the label l, and ϕ is the function that computes the precision or recall.

Table 1 shows the performance of the models in the emotion recognition task, which is shown in the form of the mean and standard deviation for the results in the 10 experiments. For [30], both the reported results in the paper and the implemented results on our data are listed. In order to verify the effectiveness of the extracted skeleton data, we compare it with the MoCap data of the head, hands, and face contained in the database which also represents body movement of the actor. The MoCap-based emotion detection model in [49] is used as the baseline model of motion data, which uses five 2D convolutional layers along with ReLu activation function followed by a dense layer. Evaluation results of uni-modal models that utilize audio signals (A), textual transcription (T), skeletal movement (S), and MoCap data (M) respectively, are listed in Table 1. From the results, it can be seen that the proposed SMACN largely outperforms the model based on MoCap, which indicates that the collected motion data contain more informative features related to emotion.

In this study, we also apply different types of neural networks to the extracted skeleton data, to compare their performance in gesture-based emotion recognition. We use CNN, long short-term memory network (LSTM), and long short-term memory network with attention (LSTM+Att) to analyze the skeleton sequence. For each of the models, the following structures are tested: (1) LSTMs contain from 1 to 2 layers, and each layer has from 128 to 512 hidden states. (2) LSTM with attention networks also include from 1 to 2 layers, and each layer contains from 128 to 512 hidden states. The attention mechanism is applied to the output of the last layer of LSTM and the weighted sum of the output of each timestep is calculated. (3) CNNs and 3DCNNs contain from 4 to 5 layers followed by maxpooling layers, each layer includes from 64 to 512 channels. Table 2 shows the best performance of each model for the extracted skeleton data.

Table 2. Performances of different models for skeleton movement-based emotion recognition

Model	Number of parameters	UAR	WAR
LSTM	3217.4k	54.0	52.7
LSTM+Att	1126.9k	54.8	53.9
CNN	940.8k	61.7	61.0
3DCNN	1350.4k	60.6	61.9
SMACN	1143.8k	**65.6**	**65.9**

For LSTM, the network with two layers of 512 hidden states achieves the best results. For LSTM with attention, the network containing one layer of 512 hidden states obtains the best results. For CNN models, the best results are obtained for the network of four convolutional layers with 64, 128, 256, and 512 channels respectively. For the 3DCNN models, the best results are obtained for the network of four convolutional layers with 256 channels. However, in fact, the performance of each type of model does not change much for different hyperparameters. In addition, SMACN significantly outperforms other networks in this task, which shows that our network structure, especially the spatial multi-head attention mechanism, can effectively extract the emotional features from skeleton sequences.

We also compared the performance of SMERN with MDRE, which is used as the basic model of audio and text feature extraction. As shown in Table 1, for the IEMOCAP dataset, our model outperforms the best baseline model (MDRE) that only uses audio and text information by the WAP of 4.8%, the unweighted average recall of 4.9%, and the weighted average recall of 5.0%. All metrics pass the Student's t-test with $p < 0.001$. The results verify that the performance of the model using gesture information is much better than that of the model without using gesture information. This demonstrates that the new modality provides information that is not contained in the original modalities, and the multimodal fusion with the extracted gesture information contains more plentiful related features to enhance the ability of emotion analysis.

For trimodal (audio, text, and skeleton modalities) emotion prediction, the cross-model attention (CMA) model

[32] and attention-based alignment model [33] are compared as baseline models. CMA contains two modules. One of the CMA modules takes the hidden features from the audio encoder as query vectors and takes the hidden features from the text encoder as key and value vectors and employs multi-head scaled dot product attention. The other CMA module takes the hidden features from the text encoder as query vectors and takes the hidden features from the audio encoder as key and value vectors and employs multi-head scaled dot product attention. And the statistics pooling layer calculates the mean and standard deviation as output feature representations. The model proposed by Xu *et al.* calculates the attention weighted sum of the speech hidden features based on the text hidden features and concatenates the weighted sum with the text hidden features. The concatenated vectors are fed into a multimodal BiLSTM for feature fusion. The outputs are applied an maxpooling to get output feature representations. In our experiment, we concatenate the output feature representations of the CMA or the model proposed by Xu *et al.* with prosodic features and the skeleton feature representations to make prediction. The result is shown in Table 1. Our multimodal model with graph attention obtains better performance compared to the model proposed by Xu *et al.*, CMA, and SMERN models, and there is significant difference between the SMERN-GT with the SMERN model (WAP: $p = 0.0048 < 0.01$, UAR: $p = 0.0129 < 0.05$, WAR: $p = 0.0046 < 0.01$). The results show that the graph attention module can effectively find the inter-modal relationship among skeleton, audio, and text features and make better use of multimodal information.

E) Multimodal analysis

Different modalities reflect distinct aspects of human emotional expression and recognition. The different form and amount of information of each modality may cause dissimilar characteristics for emotion recognition. Figure 4 presents the confusion matrices of each model in our experiment. As seen from Fig. 4(a), in gesture-based emotion detection, the performance of neutral emotion is relatively high among all classes, while in the false recognition samples of the other three categories, the ones that are misclassified as neutral emotion are the most. We speculate that this may be because, in many utterances, the range of actor's movement is relatively small, even almost zero, which is closer to the features of the neutral expression for the model, thus these samples will be recognized as neutral labels, even if the true labels of the samples are not neutral. For audio information (Fig. 4(b)), samples of sadness are well detected, but in addition to the confusion between neutral and other emotions, anger is frequently confused with happiness. Both of the emotions have high arousal in the emotion space, which may lead to similar acoustic features that cause the misrecognition. For the text model (Fig. 4(c)), the difference between anger and happiness is identified, while it is difficult to distinguish the neutral category from other categories. The multimodal model (Fig. 4(d)) reduces the defect of text and audio model for recognizing neutral emotion to a certain extent by synthesizing multi-modal information, and integrates the strengths of uni-modal models, achieving a balanced and high performance for each emotional category.

Table 3. Comparison between noisy data and clean data

Data	UAR	WAR
Preprocessed	65.6	65.9
Unpreprocessed	61.9	62.8

V. ABLATION STUDY AND DISCUSSION

A) Effect of data cleaning

In Section III-A), we applied the low-pass filter for the detection results of estimated key points to get clean data. In order to confirm the effectiveness of the data cleaning strategy, we perform an ablation study for the use of the preprocessing process. The unpreprocessed noisy data and the preprocessed clean data are fed into two SMACNs and are trained separately. The results are listed in Table 3.

The results show that the use of the low-pass filter can effectively remove the noise information of the skeleton data and improve the performance.

B) Cross-speaker emotion recognition

Some studies have achieved considerable performance in emotion recognition based on electroencephalography (EEG) and speech features across different subjects, indicating generality and wide applicability of EEG and speech emotional features for different people [50, 51].

However, a lot of studies have shown that there are individual differences in the perception and expression of gestures due to cultures, personal characteristics, etc. [52, 53]. For example, the thumb-up symbol might express different meanings and emotions in different cultures [10]. It is widely regarded as a positive gesture and is used to express agreement, consent, or interest, while it may be considered insulting in some cultures. In recent years, due to the increasing frequency of cultural exchanges, there is a trend of globalization of gestures and the influence of cultural background on the meaning of gestures gradually decreases [54]. However, although people are in the same cultural background, they also differ in the frequency of using gestures, the purposes of producing gestures, the generated gesture spaces, and so on [55].

IEMOCAP dataset contains audio and video data from 10 speakers. In Section IV, the data of different speakers were not explicitly separated when we divided the samples into training set, development set, and test set. To evaluate the performance of the skeleton modality on cross-speaker emotion recognition, we redivide the dataset such that there is no speaker sharing between the training set and the test

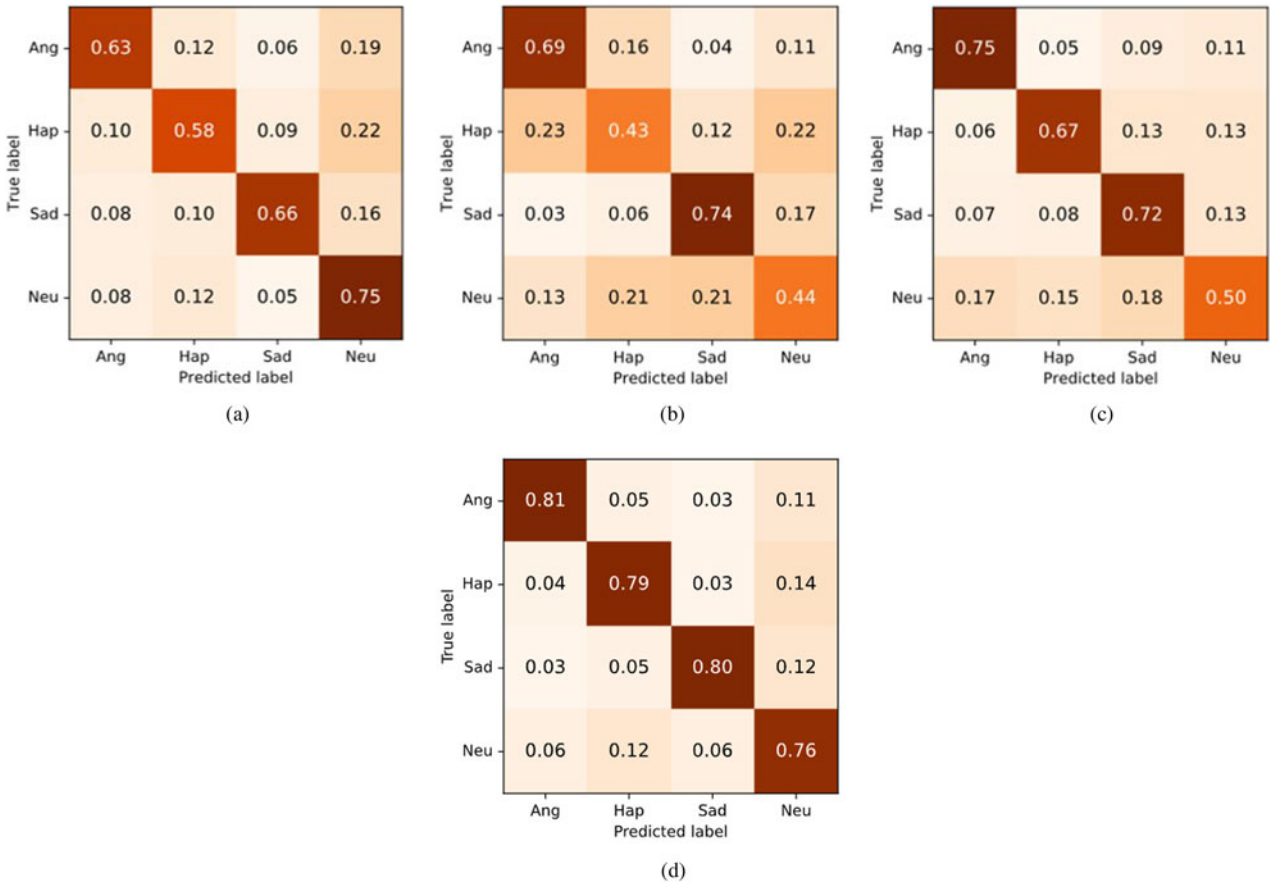

Fig. 4. Confusion matrices of each model in our experiment: (a) gesture, (b) audio, (c) text, and (d) multimodal.

Table 4. Comparison of division methods of the dataset

Division method	Ratio	UAR	WAR
Division without considering speakers	8:0.5:1.5	65.6	65.9
No speaker sharing	8:0.5:1.5	33.1	39.3

set. We compare the performance on the re-divided dataset against the performance of the previously divided dataset in Table 4.

The result shows that the obtained recognition performance in a cross-speaker manner is far lower than the performance on the previously divided data. It indicates that gestures have variation across individuals, e.g. different gesture spaces, which makes it difficult for the model to extract effective generic emotional features from skeleton data in an end-to-end manner, and for skeleton it may be necessary to learn the approximate gestural distribution of the individual for achieving an accurate emotion prediction.

C) Intra-speaker emotion recognition with pretrained model

Although the above experiment shows that the skeleton modality does not perform well across speakers, it is still an interesting topic whether the emotion recognition model pretrained on other speakers' data can be helpful in predicting the emotions of a new speaker. To do this, we divide the dataset into two parts with eight speakers and two speakers, respectively. We train an SMACN using the data of the former part and then divide the latter part with different ratios. The proportion of data of the latter part used for training ranges from 2 to 50%. The training data are fed into the pretrained model to fine-tune and the rest of the data is used for validating and testing. The performance of the model changes with the proportion of training data as shown in Fig. 5. The performance of the model increases rapidly as the proportion of training data increases and tends to be stable after 40%.

In order to verify the effect of the pretrained model, we also compare it with the unpretrained model that is only trained on the subset with two speakers. The ratio of data used for training is set to be the same. As shown in Fig. 5, the performance of the pretrained model significantly outperforms the unpretrained model when the proportion of training data is low. The gap gradually narrows as the ratio increases. The result shows that the pretraining on other speakers' data brings a significant improvement on the prediction performance especially when the size of training data is small, and the pretrained model can better extract emotional gestural features and infer the individual characteristics of gesture when the available data of the speaker is not enough.

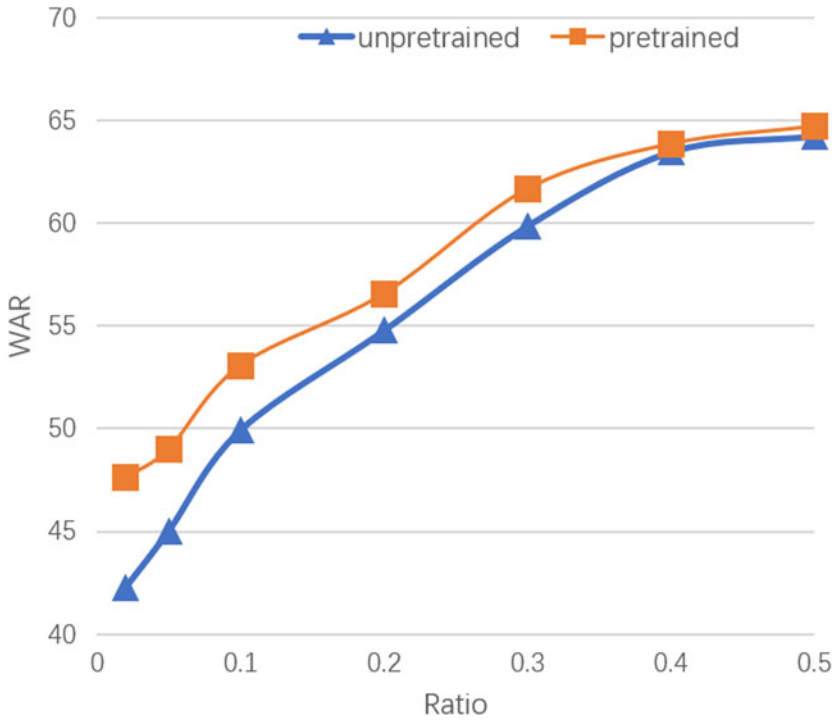

Fig. 5. Performance of the model changes with the proportion of training data using and not using pretrained model.

VI. CONCLUSIONS

In this study, we applied a method to the IEMOCAP database for extracting skeleton data from videos. We proposed a multi-head attention-based convolutional network for gesture emotion recognition and a graph attention-based multimodal emotion recognition network for integrating information from speech signals, text data, and body movements. Our experimental results indicated that skeletal movement can serve as an effective source of emotional information and the multimodal network can effectively fuse the information from multiple modalities. In future research, we plan to further explore more effective methods to make better use of the skeleton data and fuse multimodal information.

FINANCIAL SUPPORT

This study was partly supported by JSPS, Grant Number JP20H05576, by JST, ERATO, Grant Number JPMJER1401, and by JST, Moonshot R&D Grant Number JPMJMS2011.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations, *arXiv preprint arXiv:1810.02508*, 2018.

[2] Zhang, S.; Zhang, S.; Huang, T.; Gao, W.: Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control*, **59** (2020), 101894.

[3] Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B.: Emotional chatting machine: Emotional conversation generation with internal and external memory, in *Thirty-Second AAAI Conf. Artif. Intell.*, 2018.

[4] Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L.: I know the feeling: Learning to converse with empathy, *arXiv preprint arXiv:1811.00207v1*, 2018.

[5] Jaimes, A.; Sebe, N.: Multimodal human computer interaction: A survey, in *International Workshop on Human–Computer Interaction*, Springer, 2005, 1–15.

[6] Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.*, **11** (2017), 1301–1309.

[7] Li, J.L.; Lee, C.C.: Attentive to individual: A multimodal emotion recognition network with personalized attention profile, in *Proc. Interspeech 2019*, 2019, 211–215.

[8] Yoon, S.; Dey, S.; Lee, H.; Jung, K.: Attentive modality hopping mechanism for speech emotion recognition, *arXiv preprint arXiv:1912.00846*, 2019.

[9] Heusser, V.; Freymuth, N.; Constantin, S.; Waibel, A.: Bimodal speech emotion recognition using pre-trained language models, *arXiv preprint arXiv:1912.02610*, 2019.

[10] Noroozi, F.; Kaminska, D.; Corneanu, C.; Sapinski, T.; Escalera, S.; Anbarjafari, G.: Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput*, **12**(2018), 505–523.

[11] Ofodile, I. *et al.*: Action recognition using single-pixel time-of-flight detection. *Entropy*. **21** (2019), 414.

[12] Kipp, M.; Martin, J.C.: Gesture and emotion: Can basic gestural form features discriminate emotions?, in *2009 3rd Int. Conf. Affective Comput. Intell. Interact. Workshops*, IEEE, 2009, 1–8.

[13] Sapiński, T.; Kamińska, D.; Pelikant, A.; Anbarjafari, G.: Emotion recognition from skeletal movements. *Entropy*, **21** (2019), 646.

[14] Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R.: Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. *J. Vis. Commun. Image. Represent.*, **25** (2014), 24–38.

[15] Ranganathan, H.; Chakraborty, S.; Panchanathan, S.: Multimodal emotion recognition using deep learning architectures, in *2016 IEEE Winter Con. Appl. Comput. Vis. (WACV)*, IEEE, 2016, 1–9.

[16] Sapiński, T.; Kamińska, D.; Pelikant, A.; Ozcinar, C.; Avots, E.; Anbarjafari, G.: Multimodal database of emotional speech, video and gestures, in *Int. Conf. Pattern Recognit.*, Springer, 2018, 153–163.

[17] Busso, C. *et al.*: IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, **42** (2008), 335.

[18] Ekman, P.: Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, **3**(1992), 34–38.

[19] Russell, J.A.; Bachorowski, J.A.; Fernández-Dols, J.M.: Facial and vocal expressions of emotion. *Annu. Rev. Psychol.*, **54** (2003), 329–349.

[20] Coulson, M.: Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. *J. Nonverbal. Behav.*, **28** (2004), 117–139.

[21] Tracy, J.L.; Robins, R.W.: Show your pride: evidence for a discrete emotion expression. *Psychol. Sci.*, **15** (2004), 194–197.

[22] Dael, N.; Goudbeek, M.; Scherer, K.R: Perceived gesture dynamics in nonverbal expression of emotion. *Perception*, **42** (2013), 642–657.

[23] Kaza, K. *et al.*: Body motion analysis for emotion recognition in serious games, in *Int. Conf. on Universal Access in Human–Computer Interaction*, Springer, 2016, 33–42.

[24] Piana, S.; Stagliano, A.; Odone, F.; Verri, A.; Camurri, A.: Real-time automatic emotion recognition from body gestures, *arXiv preprint arXiv:1402.5047*, 2014.

[25] Barros, P.; Jirak, D.; Weber, C.; Wermter, S: Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural. Netw.*, **72** (2015), 140–151.

[26] Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S.: A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (2008), 39–58.

[27] Chen, S.; Jin, Q.: Multi-modal conditional attention fusion for dimensional emotion prediction, in *Proc. of the 24th ACM Int. Conf. Multimedia*, 2016, 571–575.

[28] Wu, C.H.; Lin, J.C.; Wei, W.L.: Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process*, **3**(2014), 12.

[29] Lin, J.C.; Wu, C.H.; Wei, W.L.: Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Trans. Multimedia*, **14** (2011), 142–156.

[30] Yoon, S.; Byun, S.; Jung, K.: Multimodal speech emotion recognition using audio and text, in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, 112–118.

[31] Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P: Tensor fusion network for multimodal sentiment analysis, *arXiv preprint arXiv:1707.07250*, 2017.

[32] Krishna, D.; Patil, A.: Multimodal Emotion Recognition using Cross-Modal Attention and 1D Convolutional Neural Networks, in *Interspeech*, 2020.

[33] Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; Li, X.: Learning alignment for multimodal emotion recognition from speech, *arXiv preprint arXiv:1909.05645*, 2019.

[34] Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S.: Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, **8** (2020), 176274–176285.

[35] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.: Graph attention networks, *arXiv preprint arXiv:1710.10903*, 2017.

[36] Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C.: RMPE: Regional multi-person pose estimation, in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, 2334–2343.

[37] Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.S.; Lu, C.: Crowdpose: efficient crowded scenes pose estimation and a new benchmark, *arXiv preprint arXiv:1812.00324*, 2018.

[38] Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C.: Pose flow: efficient online pose tracking, in *BMVC*, 2018.

[39] Lin, T.Y. *et al.*: Microsoft Coco: Common objects in context, in *Eur. Con. Comput. Vis.*, Springer, 2014, 740–755.

[40] Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *Proc. IEEE Con. Comput. Vis. Pattern Recognit.*, 2019, 7753–7762.

[41] Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X.: Object detection with deep learning: a review. *IEEE Trans. Neural. Netw. Learn. Syst.*, **30** (2019), 3212–3232.

[42] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.: Mask r-CNN, in *Proc. IEEE Int. Con. Comput. Vis.*, 2017, 2961–2969.

[43] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A.: You only look once: Unified, real-time object detection, in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, 2016, 779–788.

[44] Bulat, A.; Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression, in *Eur. Conf. Comput. Vis.*, Springer, 2016, 717–732.

[45] Karim, F.; Majumdar, S.; Darabi, H.; Chen, S.: LSTM fully convolutional networks for time series classification. *IEEE Access*, **6** (2017), 1662–1669.

[46] Zhao, Z. *et al.*: Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, **7** (2019), 97515–97525.

[47] Eyben, F.; Weninger, F.; Gross, F.; Schuller, B.: Recent developments in OpenSMILE, the Munich open-source multimedia feature extractor, in *Proc. of the 21st ACM Int. Conf. Multimedia*, 2013, 835–838.

[48] Pennington, J.; Socher, R.; Manning, C.D.: Glove: Global vectors for word representation, in *Proc. of the 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543.

[49] Tripathi, S.; Tripathi, S.; Beigi, H.: Multi-modal emotion recognition on IEMOCAP with neural networks. *arXiv preprint arXiv:1804.05788*, 2018.

[50] Li, H.; Jin, Y.M.; Zheng, W.L.; Lu, B.L.: Cross-subject emotion recognition using deep adaptation networks, in *Int. Conf. Neural Inf. Process.*, Springer, 2018, 403–413.

[51] Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E.: Dialoguernn: An attentive RNN for emotion detection in conversations, in *Proc. AAAI Conf. Artif. Intell.*, vol. **33**, 2019, 6818–6825.

[52] Kendon, A.: The study of gesture: Some remarks on its history, in *Semiotics 1981*, Springer, 1983, 153–164.

[53] Marstaller, L.; Burianová, H.: Individual differences in the gesture effect on working memory. *Psychon. Bull. Rev.*, **20** (2013), 496–500.

[54] Pease, B.; Pease, A.: The definitive book of body language: The hidden meaning behind people's gestures and expressions, Bantam Books, New York, 2006.

[55] Özer, D.; Göksun, T.: Gesture use and processing: a review on individual differences in cognitive resources, *Front. Psychol.*, **11**, 2020.

Jiaqi Shi is a graduate student at the Graduate School of Engineering Science, Osaka University, Japan, and works as an intern student in the Institute of Physical and Chemical Research (RIKEN), Japan.

Chaoran Liu obtained his Ph.D. degree in engineering from Osaka University, Japan. He is currently a researcher in the Hiroshi Ishiguro Laboratories, Advanced Telecommunications Research Institute International (ATR), Japan.

Carlos T. Ishi received his Ph.D. degree in engineering from The University of Tokyo, Japan. He joined ATR Intelligent Robotics and Communication Labs in 2005, and became the group leader of the Department of Sound Environment Intelligence since 2013. He joined the Guardian Robot Project, RIKEN, in 2020. His research topics include speech and gestures applied for human–robot interaction.

Hiroshi Ishiguro received his D.Eng. in systems engineering from Osaka University, Japan. He is currently a professor in the Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University and a distinguished professor of Osaka University. He is also a visiting director of the ATR Hiroshi Ishiguro Laboratories and an ATR fellow. His research interests include sensor networks, interactive robotics, and android science.