3D Skeletal Movement enhanced Emotion Recognition Network

Jiaqi Shi*†, Chaoran Liu†, Carlos Toshinori Ishi† and Hiroshi Ishiguro*†
* Osaka University, Japan
E-mail: {shi.jiaqi, ishiguro}@irl.sys.es.osaka-u.ac.jp
† Advanced Telecommunications Research Institute International, Japan
E-mail: {chaoran.liu, carlos}@atr.jp

***Abstract*—Automatic emotion recognition has become an important trend in the field 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, motion modalities are rarely considered in the study of affective computing. An important reason is the lack of large open emotion databases containing skeletal movement data. In this paper, we extract 3D skeleton information from video, 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 speaker's emotion state. We also combine our model with models using other modalities to provide complementary information in the emotion classification task. The combined model utilizes audio signals, text information and skeletal data simultaneously. The performance of the model significantly outperforms the bimodal model, proving the effectiveness of the method.**

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 human and computers, so that the machine can understand the user's intention and generate natural feedback 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 system, mental health care and so on [1]. For example, an interactive system with the ability of affective 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 [2], [3].

Humans recognize emotions through a variety of different modalities during natural interaction, e. g. facial expressions, voice tone and body movement [4]. Through the acquisition of information from different modalities, humans can obtain multiple related but different aspects of emotional information, so as 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 [5], [6], [7], [8].

Gesture is one of the most important forms of nonverbal communication, which plays an extremely important role in the recognition of emotions [9]. 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 [10], [11]. On the other hand, skeletal movement data is the most natural and intuitive depiction of body movements, which can represent the interrelationships of body parts and joint movements [12], [13]. 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 emotional databases containing 3D skeletal movement data.

There are some multimodal emotion databases containing 3D skeleton data, such as emoFBVP database [14] and Multimodal Database of Emotional Speech, Video and Gestures [15]. 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 dialogue and interaction between people. IEMOCAP [16] 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 work, we add a new modality of skeletal data for IEMOCAP database, and propose a skeletal motion 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 through adding a new modality. The extracted data is a representation of body movements, in the form of 3D joints positions sequence.

2) We propose an attention-based CNN network for obtain-

ing informative representations of the skeleton data to identify emotional classes.

3) We fuse the motion representations extracted by the proposed model with the audio and text features extracted by the existing methods, to utilize audio, text and skeleton data simultaneously. The performance of the model exceeds the prior model significantly, which proves the effectiveness of the extracted modality.

The rest of the paper is organized as follows: Section 2 describes the method of extracting skeleton data and our unimodal and multi-modal models; the experiment and results are described in Section 3; finally, we conclude this paper with a brief summary and mention some future work.

II. Related works

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 [17], [18]. 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 [19], [20], but also the dynamic characteristics of body movement, e. g. amount, speed, force, fluency and size, can help to accurately identify emotions [21].

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 [22], [23]. 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 [12], [24]. However, these studies are limited by the relatively small amount of data and the lack of interaction and dialogue 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.

III. Methodology

This section describes the method of extracting gesture modality from the video and the structure of the proposed model. We extract skeleton data from the original video files and perform data cleaning aimed at removing noise. The preprocessed data is 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 body movement information 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 video. 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 requires some preprocessing operations in order to be fed into 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.

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 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 work, AlphaPose is used as a 2D pose detector. AlphaPose [25], [26], [27] is an open-source pose estimation system with extremely high accuracy. The AlphaPose detector pretrained on the COCO dataset [28] 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 [29], which is proved to be effective in 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 video.

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 noise and get clean data, a low-pass filter is applied to the detection. After some test, we selected 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.

In addition, the lower body of the actors in the IEMOCAP dataset is basically 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 motion data in each sample is a variable-length sequence, which cannot be directly used as the input of CNN. In order 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 Figure 1. It takes the time sequence of skeletal movement as input, extracts emotion-related features through the convolution layers and

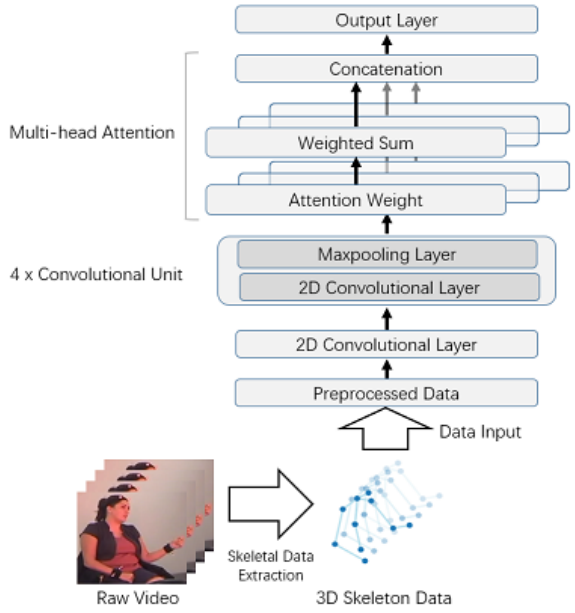


Fig. 1. Architecture of the proposed SMACN.

the attention layer, and predicts the emotion class. The convolutional layers are trained to detect affective 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 classes.

In recent years, convolutional neural networks have achieved excellent performance in many tasks related to digital image processing, e. g., target detection[30], [31], [32] and human pose estimation[29], [33]. CNN is 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, motion data can also be regarded as a special kind of image data, CNN is 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 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 4 convolutional units to abstract 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'* 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.

1) Multi-head Spatial Attention Layer: In the spectrogram representation based speech emotion recognition task of [34] and [35], 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 of 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 $L = T' \times V'$:

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

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

$$y_i = \tanh(Wa_i + b), \tag{2}$$

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

$$E_i = Uy_i, \tag{3}$$

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

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 of the weights on each attention head form a two-dimensional 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 $G \in \mathbb{R}^R, R = H \times L$:

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

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

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 motion information (see Figure 2). For text and audio data, Multimodal Dual Recurrent Encoder (MDRE)[36] 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

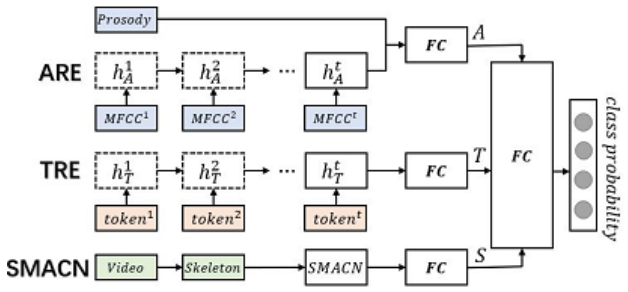


Fig. 2. The SMERN framework where audio, text and gesture are used for emotion classification simultaneously.

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*. On the other hand, 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, and then the final hidden state after a fully connected layer is the vector representation *T* of the text. The concatenation of vectors *A* and *T* contain both audio and text information and is used for emotion prediction.

In our work, 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 network, respectively, to obtain the uni-modal feature representations in the first phase. In the second phase, the feature representations are passed through the fully connected layer to change dimensions, and then concatenated to pass to the fully connected layer to form a multimodal feature representation that is ultimately used for emotion prediction.

IV. Experiments and Results

A. Dataset Description

We use Interactive emotional dyadic motion capture database (IEMOCAP), which contains more than 10 hours of audio and video data from ten actors. For simulating natural binary interaction between people, the dialogues 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, 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 work of [36]. For speech data, OpenSMILE toolkit [37] 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 F0 fundamental frequency, the voicing probability, and the loudness contours. For the textual transcript, we use a pretrained 300-dimensional GloVe vector [38] to initialize each token.

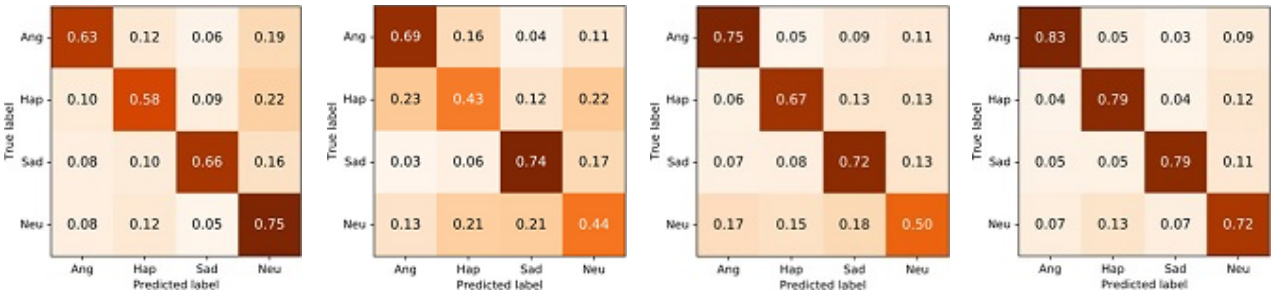
C. Experiment Setting

In our experiments, 5-fold cross validation is applied to evaluate the performance of the model. The samples of each fold is divided into training set, development set and test set, with a ratio of 8: 0.5: 1.5. This process is repeated for 5 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.

D. Results and Discussion

Consistent with previous work, the weighted average precision (WAP) is calculated as the indicator of the model performance. Table I 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. 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 gestures of the actor. The MoCap based emotion detection model in [39] 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 I. 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 contains more informative features related to emotion.

We also compared the performance of SMERN with Multimodal Dual Recurrent Encoder (MDRE), that is used as the basic model of audio and text feature extraction, and Multimodal Dual Recurrent Encoder with Attention (MDREA)[36], that applies the output of ARE to generate attention weights for text features. As shown in Table I, for the IEMOCAP dataset, our model outperforms the best baseline model (MDRE) that only uses audio and text information by the weighted average precision of 4.8%. Since WAP is rarely used in emotion recognition tasks, we also list the unweighted and weighted



(a) Gesture (b) Audio (c) Text (d) Multimodal

Fig. 3. Confusion matrices of each model in our experiment

TABLE I
COMPARISON FOR UNIMODAL AND MULTIMODAL

Model	Modality	WAP	UAR	WAR
Mocap_Model[39]	M	0.511	-	-
ARE[36]	A	0.546 ±0.009	0.597	0.571
TRE[36]	T	0.635 ±0.018	0.659	0.645
SMACN	S	0.648 ±0.006	0.656	0.659
MDRE[36]	A + T	0.718 ±0.019	0.720	0.714
MDREA[36]	A + T	0.690 ±0.019	-	-
SMERN	A + T + S	**0.766 ±0.006**	**0.783**	**0.777**

average recall (UAR/WAR) of the models in our experiment. Precision and recall are defined as:

$$Precision = \frac{tp}{tp + fp}, \tag{7}$$

$$Recall = \frac{tp}{tp + fn}, \tag{8}$$

where tp is the number of true positive samples, fp is the number of false positive samples, and fn the number of is false negative. And the weighted score is calculated as:

$$weighted = \frac{1}{\sum_{l \in L} |\hat{y}_l|} \sum_{l \in L} |\hat{y}_l| \phi(y_l, \hat{y}_l), \tag{9}$$

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, ϕ is the function that computes the precision or recall. The results also 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.

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 the emotion recognition. Figure 3 presents the confusion matrices of each model in our experiment. As can be seen from Figure 3(a), in gesture based emotion detection, the accuracy 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 true label of the sample is not neutral. For audio information (Figure 3(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 a high arousal in the emotion space, which may lead to rather similar acoustic features that cause the misrecognition. For the text model (Figure 3(c)), the difference between anger and happiness is identified, while it is difficult to distinguish the neutral category with other categories. SMERN (Figure 3(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 recognition of each emotion category.

In this work, we also applied different types of neural networks to the extracted skeleton data, so as to compare their performance in motion emotion recognition. We use Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM) and Long Short-Term Memory Network with attention (LSTM+Att) to analyze the motion sequence. For each of the models, the following structures are tested: 1) LSTM contains from 1 to 2 layers, and each layer has from 128 to 512 hidden states. 2) LSTM with attention networks also includes from 1 to 2 layers, each layer contains from 128 to 512 hidden states. 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) CNN contains from 4 to 5 layers followed by maxpooling layers, each layer includes from 64 to 512 channels. Table II shows the performances of the models for the extracted skeleton data.

For LSTM, the network with 2 layer of 512 hidden states achieved the best results. For LSTM with attention, the network containing 1 layer of 512 hidden states obtained the best results. For CNN models, the best results were obtained for the network of 4 convolutional layers with 64, 128, 256 and 512 channels respectively. However in fact, the performance

TABLE II
PERFORMANCES OF DIFFERENT MODELS FOR SKELETON MOVEMENT BASED EMOTION RECOGNITION

Model	#Para	UAR	WAR
LSTM	3217.4K	0.540	0.527
LSTM+Att	1126.9K	0.548	0.539
CNN	940.8K	0.617	0.610
SMACN	1143.8K	**0.656**	**0.659**

of each type of model does not change much for different hyperparameters. In addition, SMACN network 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 motion sequence.

V. CONCLUSIONS

In this work, we applied a method to the IEMOCAP database for extracting skeleton data from videos. We proposed a multi-head attention based CNN network for gesture emotion recognition and skeletal motion enhanced multimodal network for integrating information from speech signals, text data, and body movements, aiming to verify the usefulness of the extracted data in the emotion recognition task. Our experimental results indicated that skeletal movement can serve as an effective source of emotional information, which means that using the method, we may no longer be limited by the lack of skeleton data in the research of emotional motion interaction. In future work, we plan to further explore a more effective fusion strategy for multimodal information aiming to make better use of the features from every modality.

ACKNOWLEDGMENT

This work was partly supported by Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576, and JST, ERATO, Grant Number JPMJER1401.

REFERENCES

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

[2] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "I know the feeling: Learning to converse with empathy," 2018.

[4] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," in *International Workshop on Human-Computer Interaction*. Springer, 2005, pp. 1–15.

[5] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[6] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile," *Proc. Interspeech 2019*, pp. 211–215, 2019.

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

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

[9] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[10] I. Ofodile, A. Helmi, A. Clapés, E. Avots, K. M. Peensoo, S.-M. Valdma, A. Valdmann, H. Valtna-Lukner, S. Omelkov, S. Escalera *et al.*, "Action recognition using single-pixel time-of-flight detection," *Entropy*, vol. 21, no. 4, p. 414, 2019.

[11] M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?" in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–8.

[12] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari, "Emotion recognition from skeletal movements," *Entropy*, vol. 21, no. 7, p. 646, 2019.

[13] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[14] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[15] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *International Conference on Pattern Recognition*. Springer, 2018, pp. 153–163.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[17] P. Ekman, "Facial expressions of emotion: New findings, new questions," 1992.

[18] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual review of psychology*, vol. 54, no. 1, pp. 329–349, 2003.

[19] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[20] J. L. Tracy and R. W. Robins, "Show your pride: Evidence for a discrete emotion expression," *Psychological Science*, vol. 15, no. 3, pp. 194–197, 2004.

[21] N. Dael, M. Goudbeek, and K. R. Scherer, "Perceived gesture dynamics in nonverbal expression of emotion," *Perception*, vol. 42, no. 6, pp. 642–657, 2013.

[22] K. Kaza, A. Psaltis, K. Stefanidis, K. C. Apostolakis, S. Thermos, K. Dimitropoulos, and P. Daras, "Body motion analysis for emotion recognition in serious games," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2016, pp. 33–42.

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

[24] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, pp. 140–151, 2015.

[25] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

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

[27] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *BMVC*, 2018.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[29] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7753–7762.

[30] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in

Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[33] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*. Springer, 2016, pp. 717–732.

[34] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE access*, vol. 6, pp. 1662–1669, 2017.

[35] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019.

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

[37] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[38] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[39] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap with neural networks." *arXiv preprint arXiv:1804.05788*, 2018.