Journal of Signal Processing, Vol.25, No.2, pp.73-85, March 2021

PAPER

Multimodal Emotion Recognition Using Non-Inertial Loss Function

Jargalsaikhan Orgil[1], Stephen Karungaru[1], Kenji Terada[1] and Ganbold Shagdar[2]

[1]Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Tokushima 770-0851, Japan
[2]Graduate School of Information and Communication Technology, Mongolian University of Science Technology
Ulaanbaatar 770-0851, Mongolia
E-mail: orgil@must.edu.mn, karungaru@tokushima-u.ac.jp

Abstract **Automatic understanding of human emotion in a wild setting using audiovisual signals is extremely challenging. Latent continuous dimensions can be used to accomplish the analysis of human emotional states, behaviors, and reactions displayed in real-world settings. Moreover, Valence and Arousal combinations constitute well-known and effective representations of emotions. In this paper, a new Non-inertial loss function is proposed to train emotion recognition deep learning models. It is evaluated in wild settings using four types of candidate networks with different pipelines and sequence lengths. It is then compared to the Concordance Correlation Coefficient (CCC) and Mean Squared Error (MSE) losses commonly used for training. To prove its effectiveness on efficiency and stability in continuous or non-continuous input data, experiments were performed using the Aff-Wild dataset. Encouraging results were obtained.**

Keywords: deep emotion recognition, emotion recognition, emotion, body language, intonation

1. Introduction

To understand each other, emotions have played an important predictive role. Emotions are an extremely complex brain function that reacts to previous actions based on the human brain functional system, called the ”Limbic System”. Recently, there has been an enormous interest in this emotion recognition field.

Emotion can be recognized using many methods including voice intonation, body pose, and more complex ones such as electroencephalography (EEG) [1], etc. Facial emotion can be visualized by the expansion and contraction of the muscles located around the mouth, nose, and eyes [6]. Previous studies of facial emotion recognition concentrated on seven basic categories: anger, fear, disgust, happiness, sadness, surprise, and contempt [5]. Moreover, Action Units (AUs) [2], [3], [4] were proposed to model facial behavior, and the combination of AUs could also be utilized for facial expression recognition. Most studies are based on the seven basic categories [7], [8], [9], [10], [11], [12] with some researchers using triplet expression recognition [13].

Consequently, facial expression recognition has attracted renewed attention owing to recent advanced network architectures. In facial expression recognition, real-time automated analysis of facial expression in video plays an essential role in implementing human-computer interaction interfaces.

The 2-D Emotion Wheel [22], shows valence ranging from extremely positive to negative and arousal ranging from extremely active to passive. Based on the data from the activity a person is engaged in, discriminating information on the valence of a person's emotions may be known. Emotion sensing parameters can be distinguished as to whether they provide information on qualification of the valence aspect or the arousal aspect. There are many related works in this field depending on data sets used and models proposed [20], [21]. There are also previous studies on emotion recognition in videos [18], [25]. Valence and Arousal (V-A) are not separated values; binding these two parameters describe an emotion.

For most emotion recognition methods so far, the Concordance Correlation Coefficient (CCC) loss and the Mean Squared Error (MSE) loss are widely used during training. FATAUVA-Net Chang et al. [25] has provided the best-confirmed results using mean CCC and mean MSE for valence and arousal. The authors concentrated on the connection between V-A estimation and Action Unit such as the face and its parts. Moreover, their research environment was based on a wild setting. Yang et al. [15] concentrated on feature extraction. A network was assembled that extracted features with a Recurrent unit and was trained on MSE loss. Vielzeuf et al. [14] trained audiovisual ensemble network on emotion video classification. MIMAMO Net Deng et al. [18] trained a spatial, temporal network with CCC loss.

Two criteria were measured for evaluating the performance of the networks; Valence and Arousal fractional range is between $[-1, 1]$. The main problem is that no loss function that can quickly train a given network on less data. Moreover, Valence and Arousal are not separated values, and it is usually considered important to train them together in a coordinate system points.

Therefore, in this paper, the unit circle map of the V-A is used. Annotated targets are moving objects in the unit circle map of Valence and Arousal. Predictions of the model should be the small difference value of targets. Moreover, there are 3 dynamical physical calculation formulas used in the proposed method. This is the basis of the proposed loss function that we call ”Non-Inertial”. Experiments were conducted on multiple model architectures and sequence lengths in training.

The contributions of the proposed method are as follows:

1. The new loss function allows for Valence and Arousal to be viewed together
2. Ability to train on less data
3. Better results
4. Faster training times

As shown in Table 1, emotion recognition model input parameters can be linked to whether they provide the Valence or Arousal aspect quantification information [27]. This research uses the following relations: Facial expression, facial muscle activity around the lip, eye, and nose, voice, and intonation. Body pose relation is not engaged.

Table 1 Emotion related parameters

Emotion-related effect	Arousal	Valence
Emotion induced sweating	+	
Breathing rhythm variations	+	+
Heart rate variability	+	+
Blood pressure	+	
Core body temperature	+	
Heart rate	+	
Facial expression		+
Facial muscle activity	+	
Voice intonation	+	+
Questionnaire	+	+

The rest of the paper is organized as follows: Section 2 explains the proposed method. Experimental result and their analysis are presented in Section 3 and finally, in Section 4, the paper is concluded and future works discussed.

2. Proposed Method

Many learning models and training loss functions have already been proposed for emotion recognition. This work introduces an additional new loss function and performs experiments using multiple existing model architectures and sequence lengths in training to prove its effectiveness. The ResNet50 model is selected as the proposed loss function evaluation model. One of the main emotion recognition problems in sequence data is defining interrupted and continuous data. Based on our previous paper [17], in the current work, a new specific loss function is introduced to deal with the issue. Fig.1 shows the flow of the proposed method.

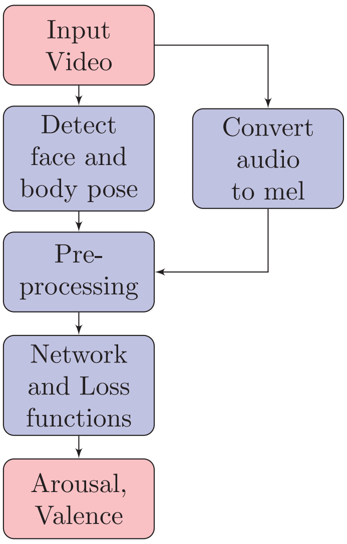

Fig. 1 Flow chart

In this work, after video input, multi-modal features including face landmarks and the mel-spectrum data are extracted. This is followed by pre-processing steps like facial bounding box expansion and feature vector construction. Four selected networks are trained using the proposed loss function and the two other existing ones (CCC and MSE) for comparison. The results of Arousal and Valence obtained provide a means for evaluating the loss functions.

2.1 Loss functions

Before discussing the proposed non-inertial loss function, for comparison purposes, two other largely used loss functions are briefly introduced.

2.1.1 Mean squared error

The first most comparative metric criterion is the Mean Squared Error (MSE), defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{1}$$

where x_i predictions including (valence-arousal) values and y_i annotations (valence-arousal), and N is the total number of samples. The MSE calculates an approximate indication of how the obtained training model is performing. A small value of MSE is desirable.

2.1.2 Concordance correlation coefficient

The second one is the Concordance Correlation Coefficient (CCC) [23]. It is widely used to measure the performance of dimensional emotion recognition methods, e.g., in the Aff-Wild challenge [26]. CCC calculates the similarity between two time-series (e.g., all video annotations and predictions) by scaling their correlation coefficients with their mean square difference. The predictions that are well correlated with the annotations but shifted in value are penalized in proportion to the deviation. CCC values are in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes discordance. The higher the value of the CCC, the better the fit between annotations and predictions. The mean value of CCC for valence and arousal estimation was used as the main evaluation criterion.

CCC is defined as follows:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\hat{x} - \hat{y})^2} = \frac{2s_x s_y \rho_{xy}}{s_x^2 + s_y^2 + (\hat{x} - \hat{y})^2} \tag{2}$$

where ρ_{xy} is the Pearson Correlation Coefficient (Pearson CC), s_x and s_y are the variances of all valence or arousal video annotations and predicted values, respectively and s_{xy} is the corresponding covariance value, $\hat{x}$ and $\hat{y}$ are mean values of predictions and annotations.

2.1.3 Proposed non-inertial loss function

Two criteria are measured for evaluating the performance of networks. Valence and Arousal's fractional range is between $[-1, 1]$. The problem is that there is no loss function to quickly train the networks on less data. Moreover, Valence and Arousal are not separated values, and it is usually considered important to train them together in a coordinate system of points. This work uses the unit circle map of V-A. This new definition is named the "Non-Inertial loss".

The new loss function for the V-A is defined by the following equations.

$$\Delta l_c = \sqrt{x_c^2 - y_c^2} \tag{3}$$

$$\begin{aligned}\Delta l_{x,y} = (&\sqrt{(x_c(v) - x_p(v))^2 + (x_c(a) - x_p(a))^2} \\ &- \sqrt{(y_c(v) - y_p(v))^2 + (y_c(a) - y_p(a))^2})/t\end{aligned} \tag{4}$$

$$\begin{aligned}\Delta\alpha &= tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} \\ &= abs(a_x * v_y - a_y * v_x)\end{aligned} \tag{5}$$

$$loss_t = mean(\Delta l_c + \Delta l_{x,y} + \Delta\alpha) \tag{6}$$

Equation (3) is referred to as norm 2, Euclidean distance or RMSE of the annotations and predictions, which calculates how far away in V-A unit map the prediction is. A small value for the distances is desired. Equation (4), where t is time, $x_c(v)$ is the current prediction valence, and $x_p(a)$ is the previous prediction arousal, calculates the difference of prediction and annotation V-A velocities. To avoid the network parameters exploding, $t = 1$ is assumed. Velocity will be enhanced in the recurrent section of the training model, which in our network is the LSTM layer. Equation (5) is the direction. It calculates the differences of angles. The difference between annotation and prediction angles are expected to become zero; a_x is the projection of prediction arousal and v_y, is the projection of annotation valence. Differences of angles are converted to avoid subtraction because of zero denominators and value more than 1 is overlooked in the arcsin function. The proof of the Equations (5) is in Appendix A.

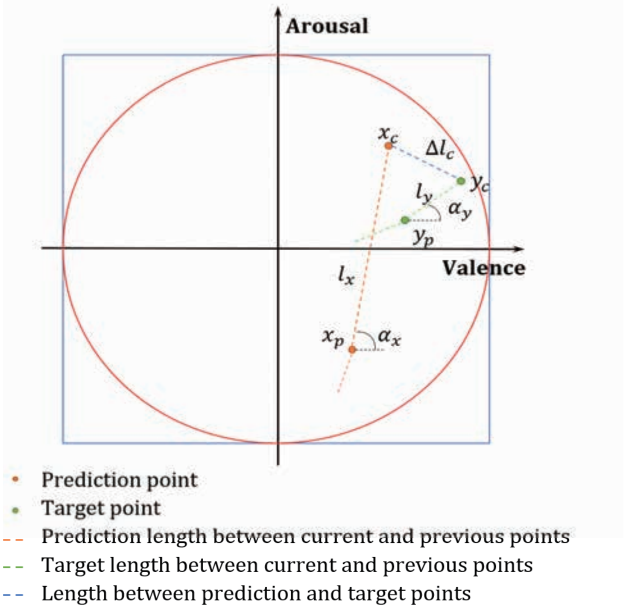
Arousal
x_c
Δl_c
y_c
l_y
α_y
y_p
Valence
l_x
x_p
α_x
● Prediction point
● Target point
-- Prediction length between current and previous points
-- Target length between current and previous points
-- Length between prediction and target points

Fig. 2 Non-inertial loss function illustration

The non-inertial loss function works with the form $[batch, sequence, 2]$, which is our annotation size. If the annotation size is more than 2, then an acceleration parameter can be added and the circle of angle difference can be extended from the first column to the end.

Fig.2 shows an illustration of an example loss function annotation and prediction sequence in arousal and valence map.

2.2 Proposed training model

The proposed model (Full Net) is illustrated in Fig.3. It consists of three stages. The first stage inputs the openPose features. The second stage consists of the Linear layers, and the last stage is a Recurrent stage LSTM, which is widely used and already proven effective in Action Recognition [16]. The first stage extracts a feature representation of a snippet, which consists of a facial RGB image, face landmarks, body pose, and Mel spectrogram features. The RGB image is fed into the ResNet50 network. The sequence of extracted features of openPose outputs is fed directly into the second stage linear layers. The output of the linear layers feeds the last stage of the LSTM network and classifier. For low-cost training, features were extracted using a pre-trained ResNet50 in our previous study [17]. In this work, only openPose and pre-trained ResNet50 are not trainable. All other parameters are trained. Moreover, to test the proposed loss functions in different network parameter sizes, the Full network is split into 3 additional models; ”Without mel Network”, ”Audio Network”, and ”Pose Network”. For interrupted data network, ”Without mel Network” and ”Pose Network”, are used because of input data interruption, that is, when there is no detected pose estimation and face features from openPose. For ”Full Network” and ”Audio Network”, representation of continuous networks is assumed. All these candidate networks are trained using the 3 loss functions.

2.3 Training data

Mollahosseini et. al. [20] deal with the largest database for Emotion Recognition in a Wild setting referred to as Aff-Wild [26]. The database contains more than 30 hours of footage in 298 videos. To satisfy the ”in the wild” concept, data videos specifically not made for any specific task of emotions recognition are required. Hence, the videos were gathered from YouTube using the keyword ”Reaction”. Most contained people who naturally reacted to unexpected plots, exciting situations, etc. The data also contains the Training and Test videos, facial bounding box, facial landmark, and target, in two dimensions, Arousal and Valence. From the training data, the following two features are extracted for training, Mel spectrogram, and openPose. They are briefly described below.

2.3.1 Mel spectrogram

Frequency Mel-scale describes the perceptual distance between pitches of different frequencies. A classical approximation is to define the frequency-to-mel transform function for a frequency f as

$$m = 2595 * log_{10}(1 + \frac{f}{700}) \tag{7}$$

Our method uses Mel-frequency Spectrogram as input to the Linear layers. Spectrograms are generated when Short Term Fourier Transform (STFT) is applied on windowed audio or speech signal. The audio is sampled at 22050Hz. Windowing is then carried out on each audio frame using a ”Hann” window of length 2048. Fast Fourier Transform (FFT) windows of length 2048 are then applied on the said windowed audio samples with an STFT hop-length equal to 512. The obtained Spectrogram magnitudes are then mapped to the Mel-scale to get Mel-spectrograms. 128 Spectrogram coefficients per window are used in this model. The Mel-frequency scale emphasizes the lower end of the frequency spectrum over the higher ones, thus, imitating the perceptual hearing capabilities of humans. The ”librosa” python package, along with the above-mentioned parameters are used to compute the Mel-spectrograms. In speech emotion recognition, Chan et al. [19] previously researched this topic. A sample Spectrogram corresponding to Aff-Wild training 110.avi audio is shown below in Fig.4.

2.3.2 OpenPose

Human 2D pose estimation is a method for localizing human body parts such as the shoulders, elbows, and ankles from an input image or video. openPose [24], developed by researchers at the Carnegie Mellon University can be considered as the state of the art approach for real-time human pose estimation. A sample Body and Face key points corresponding to Aff-Wild training video number 110.avi and frame 5389 is shown below in Fig.5 and 6.

The two green crosses in Fig.5 and 6 are the nose key points, at same location. A previous study also used body pose in video emotion recognition [15].

3. Experiments and Results

The four training networks used for evaluation of the proposed Non-inertial loss function are implemented as follows:

3.1 Training network details

1. Full Network: The network architecture is shown in Fig.3. The flow is as follows.

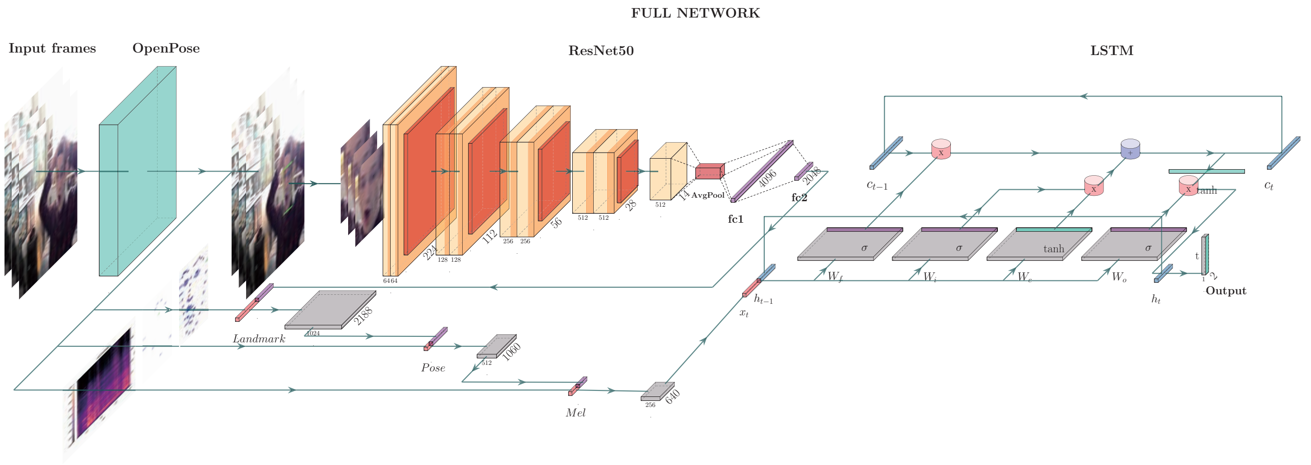

Fig. 3 Network architecture

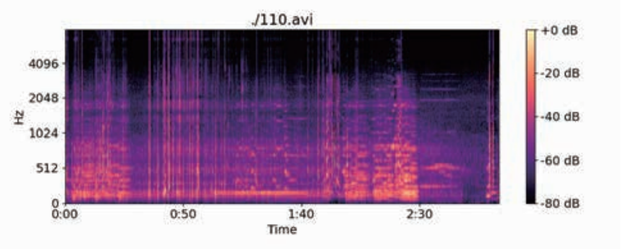

Fig. 4 Aff-Wild training 110.avi audio spectrogram

Fig. 5 OpenPose face key points 110.avi frame 5389

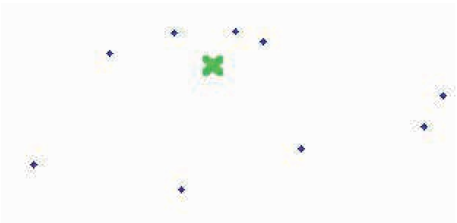

Fig. 6 OpenPose body key points 110.avi frame 5389

The Input(Batch(B), Sequence(S), Height(H), Width(W), Channel(C)) feed the openPose to get pose estimation and face features. Face feature coordination is used to calculate and preprocess the face with a bounding box and then feed into ResNet50 network. The previous layer output is concatenated with the face landmarks $(B \times S, 2048 + 140)$ and fed into the linear layer named Landmark. The landmark layer directly feeds into the Pose layer $(B \times S, 1024 + 36)$. After the Pose layer, Mel layer continuous input is $(B \times S, 512 + 128)$. Mel layer output is $(B \times S, 256)$ which is directed as input into the LSTM layer $(B, S, 256)$. The LSTM layer requires additional Sequence dimension, the dimension of Batch, Sequence, and output of previous layers. At the end of the classifier layer $(B, S, 2)$ the final outputs are calculated.

2. Without mel Network: This network does not include continuous input Mel spectrogram and Mel layer. Other parameters are the same as the previous Full Network.

3. Audio Network: $Input(B \times S, 128)$ continuous audio data is fed into Mel layer and output is $(B \times S, 256)$. Before the LSTM layer, a new dimension of sequence $(B, S, 256)$ is added. The end Classifier linear layer $(B, S, 2)$ classifies arousal and valence.

4. Pose Network: As non-continuous data, Pose data $Input(B \times S, 36)$ is fed into the Pose layer. Also, the dimension of sequence $(B, S, 256)$ is added and fed into LSTM $(B, S, 256)$ layer. This is the same as the previous networks. The classifier layer $(B, S, 2)$ also classifies arousal and valence.

All the network configurations used batch normalization 1D and dropout 0.5 after every fully connected layer.

3.2 Data pre-processing

Initially, for all the train and test sets, all body pose and facial landmarks are extracted using openPose [24]. From this output, objective personal data is obtained, using landmarks, by calculating a new facial bounding box expanded 5 pixels and then resizing it to 224x224 pixels. This facial bounding box was applied to extract features on the sequential image using the

ResNet50 network. The ResNet50 network obtained the final linear layer output and saved 2048x1 features with facial landmarks (70x2) and body pose (18x2). Moreover, from the video, the audio signal was separated and converted to the Mel-Spectrogram features. All training and validation data are normalized to between 0 and 1.

3.3 Experimental configuration

Training: All the following configurations were similar in the 4 candidate networks. The Aff-Wild dataset was split into 2 Train and Validation sets as 70%:30% and 90%:10% respectively to make sure that our loss function and models are accurate for both standard or less training and validation data proportions. During training, the 3 loss functions and the standard stochastic gradient descent optimizer, with momentum (0.9) and weight decay ($5e^{-4}$) were used. The number of epochs was set to 70. Early stopping (15 epochs) was used to prevent over-fitting. The Learning rate was 0.01 with the scheduler (patience 10 epochs). Batch size was set at 64, with the Sequence length varied in Validation:30% $[32, 64, 128, 256, 384]$ and Validation:10% $[128, 256, 512]$. However, because different experiments had different GPU consumption, the largest batch size that fits in Train and Validation set video length and our GPU memory (11GB) was chosen. However, some of Validation:30% video sequences were less than 512, the max sequence length was reduced to 384. Iteration of data-loader, drop last, and shuffle parameters were set to False. Train and validation video sets were shuffled at the beginning of each epoch. The pre-trained ResNet50 model weights were fixed during training; all the other layers were trainable. Batch normalization and dropout were used after every fully connected layer.

Validation: The losses of MSE, Non-Inertial, and CCC were calculated in each video and sample output of arousal and valence saved using TensorBoard SummaryWriter. Additionally, all the best losses of the model were saved for the rest of the test set.

3.4 Results

Due to comparison difficulty, the base unit in the CCC loss score was set as the base, which in training, $VA = -2$ is the best loss, and in validation, 2 is the best score. The 3 loss functions trained on the Aff-Wild dataset was observed in 2 ways; 5 times trained for stability check-in 4 networks and split/trained candidates of the 4 networks in training and validation data set proportions 70% : 30% and 90% : 10% among the configuration sequence length.

3.4.1 Validation 30%, 10%

In the train and validation, pipeline batch iteration took around $0.01 sec$, which as expected, was fast enough even though it was the same when max sequence length 512 was used.

Fig.7 shows the comparison of our loss function, MSE, and CCC in various network parameter size space. From this result, the new loss function was sensitive for network parameter size like the MSE function. In our candidate four networks, "Pose Net" input size is $[B, S, 36]$ normalized pose coordinates, which is small enough. The relation between pose and emotion is 0.163, which shows a very weak correlation.

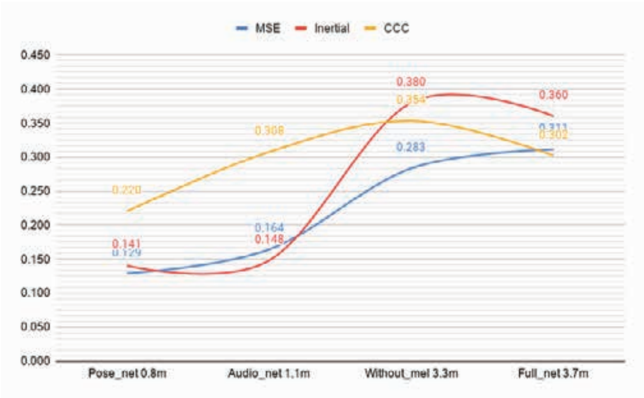

Fig. 7 Parameter size vs loss (average sequence length val 10%)

Another expectation of the result in the "Audio Net" mean was 0.206, which shows intonation and emotion as weakly correlated. "Without Mel" and "Full Net" did not produce any notable difference in the CCC correlation. The results of these two networks' were 0.339 and 0.324, respectively. On the other hand, in these networks' training, our loss function shows the best result.

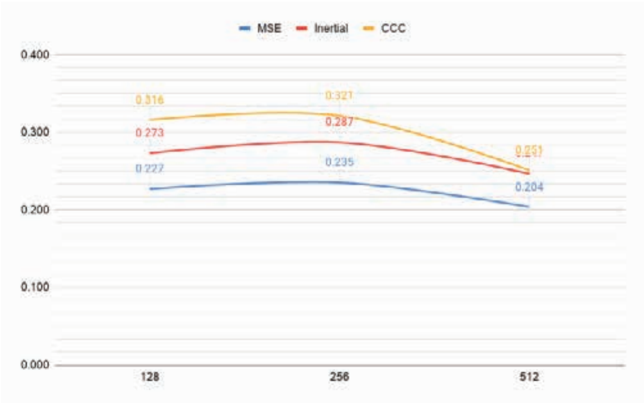

Fig. 8 Sequence length vs loss (full networks val 10%)

In Fig.8, the four networks were trained using various sequence length $128, 256, 512$ and validation 10% proportion with the 3 loss functions. The result is the average of four candidate networks. From the result, the CCC loss shows the top result, because the CCC calculation uses $[std, mean]$ values of prediction, which is more advanced. On the other hand, these values showed some disadvantage if sequence length were not comfortable, as shown in Fig. 8 and 9.

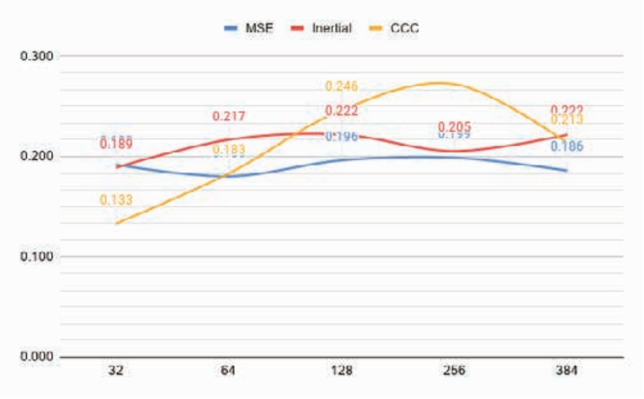

Fig. 9 Sequence length vs loss (full networks Val 30%)

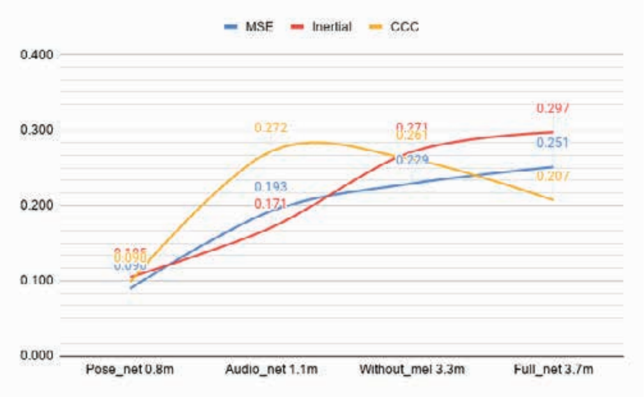

Fig. 10 Parameter size vs loss (average sequence length val 30%)

Next, the Validation:30% proportion was experimented on to verify that the result is the same as the previous experiment with additional sequence lengths. In Fig.10, for the training mean sequence length in "Full Net" and "Without mel", our loss function shows the best result and is 0.07 lower than 10% proportion. In Fig.9 the lower sequence length was extended to prove that the CCC loss has a disadvantage if sequence length in a small range is chosen. The loss would be worse than the other losses, which shows that in GPU memory a trade-off between sequence length with parameter size and accuracy may be necessary. Fig.11 shows the average of the previous two training sessions in different proportions of the Validation set. In the results of "Without Mel" and "Full Net" there is no big difference. In addition, the other two loss functions especially our new loss was shown to be stable and produced the best results among different sequence lengths.

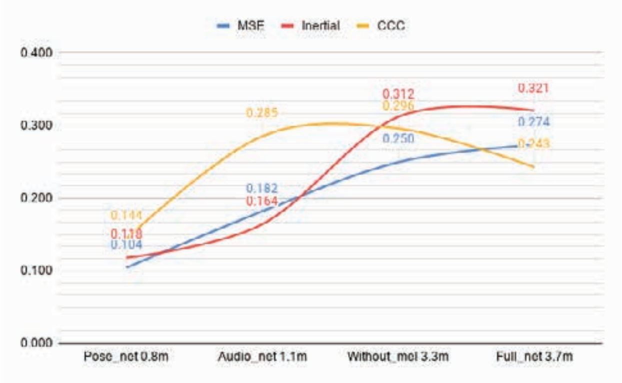

Fig. 11 Parameter size vs loss (total)

3.4.2 5-fold training

The Full network was trained five times, with a sequence length of 256, which was shown as optimal in previous training and the result of this training are shown in Table 2, where rows represent validation's main loss function and column shows parallel calculated losses. In Fig.12, the 5 training Average Line of Non-Inertial and MSE losses were nearly constant. It should be noted that the CCC loss average rapidly increased, which shows it is less stable than our new loss function. The other 3 networks training results are also shown in Table 4 Appendix B.

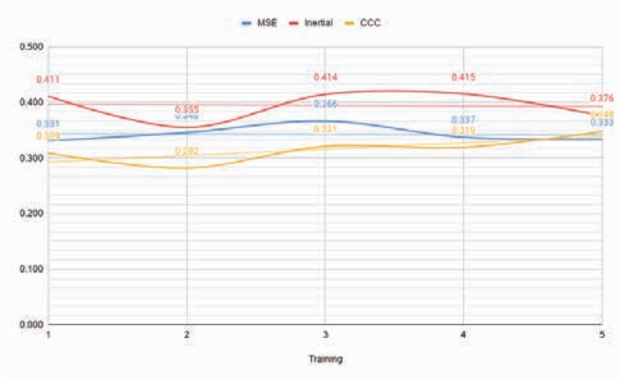

Fig. 12 Full net 5 training CCC loss comparison (sequence 256 val 10%)

3.4.3 Example valence and arousal of full network

Figures 13, 14, and 15 show our new loss function training and sample outputs for arousal and valence. Figures 16, 17 and 18 are the CCC loss training and sample output of the Validation video 158.avi. The best loss on "Full Net" in Validation was 0.86 for Arousal and 0.404 for Valence, for which backpropagation was done separately. Also, in Fig.17 Arousal was positively correlated but missed the constant value. In Figures 13, and 16 training epochs was

Table 2 Full network loss of 5 training (sequence length 256)

Train Loss	Train	MSE	CCC	Non-Inertial	Average	Differences	Sum
MSE	1	0.209	0.331	0.105	0.210	0.0000014	
	2	0.211	0.346	0.107	0.210	0.0000005	
	3	0.207	0.366	0.104	0.210	0.0000116	
	4	0.225	0.337	0.113	0.210	0.0002326	
	5	0.199	0.333	0.100	0.210	0.0001291	0.000375
CCC	1	0.348	0.309	0.176	0.316	0.0000454	
	2	0.364	0.282	0.191	0.316	0.0011459	
	3	0.335	0.321	0.170	0.316	0.0000262	
	4	0.415	0.319	0.213	0.316	0.0000107	
	5	0.271	0.348	0.136	0.316	0.0010367	0.002265
Non-Inertial	1	0.202	0.411	0.102	0.103	0.0000025	
	2	0.206	0.355	0.104	0.103	0.0000006	
	3	0.204	0.414	0.103	0.103	0.0000000	
	4	0.202	0.415	0.102	0.103	0.0000016	
	5	0.209	0.376	0.105	0.103	0.0000047	0.000009

70 but the training sessions did not reach 40 epochs in some models because of early stopping.

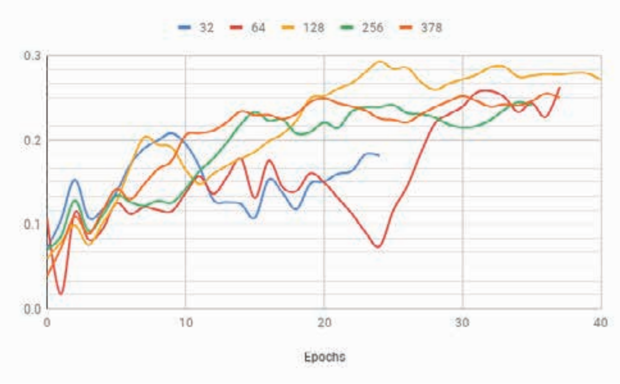

Fig. 13 Full network validation CCC loss (Non-inertial training)

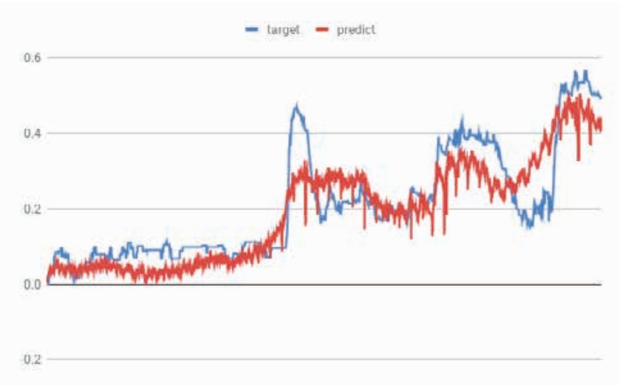

Fig. 14 Non-inertial loss example validation of arousal (158.avi)

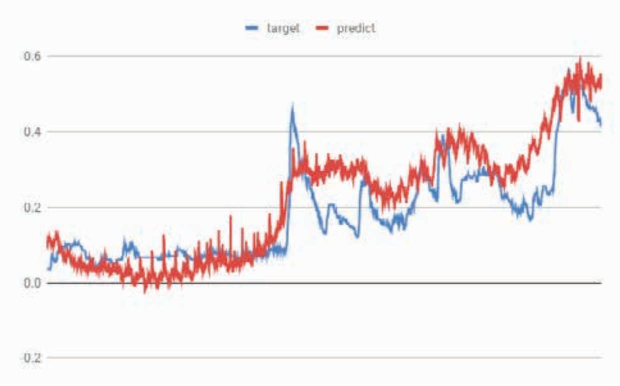

Fig. 15 Non-inertial loss example validation of valence (158.avi)

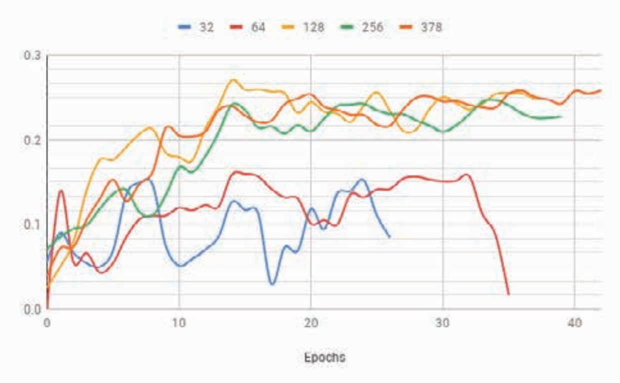

Fig. 16 Full network validation CCC loss (CCC training)

4. Discussion

This work shows that the proposed new loss function has an advantage if we choose the correct network

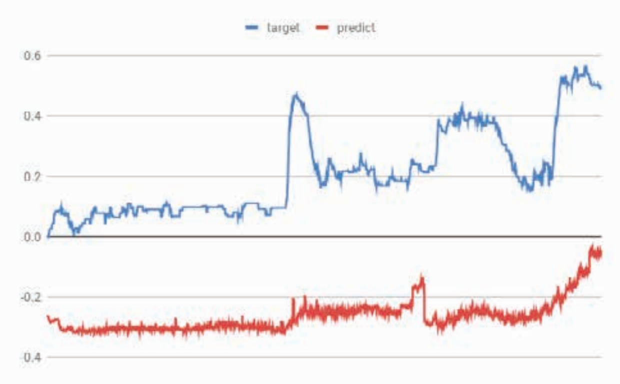

Fig. 17 CCC loss example validation of arousal (158.avi)

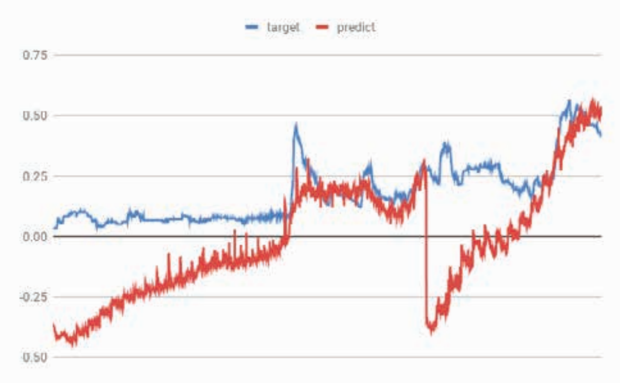

Fig. 18 CCC loss example validation of valence (158.avi)

pipeline. Fig.11 shows the average of the previous two training sessions in different proportions of the validation set. In the results of ”Without Mel” and ”Full Net” there is no big difference. This is because the Aff-Wild set audio was used directly without any preprocessing and transcribing.

From Table 1, the emotion parameters relation used is ”Facial expression”, ”Facial muscle activity” and ”Voice intonation”. In addition, new relations for ”Body pose” and emotion were added. Consequently, the representations of interrupted networks ”Without mel Network” and ”Pose Network” accuracy was lower than the continuous networks, which shows the data interruption effect. Moreover, our loss function training on ”Full net” was acceptable, excluding fitting annotation explosion in the vertical axis. It could be easily fixed by another network design, that is more deep or wide.

Our promised contributions in Section 1., are as follows:

1. In the proposed loss function, the angle difference calculation allowed for more accurate Arousal and Valence through back-propagation compared to the other loss functions. The results are shown in Figures 14, 15, 17, and 18
2. Figures 8 to 9 shows our proposed loss function trained more accurately on less data.
3. Additionally, as shown in Table 2, our loss function provides better accuracy training on ”5 training” data.
4. In training and validation, pipeline batch iteration took around $0.01sec$, that as expected, was fast enough even though it was the same for the max sequence length 512.

5. Conclusion

In this work, we proposed a new loss function named ”Non-Inertial Loss” and proved its stability and effectiveness using four networks designed with different sequence lengths, and validation proportion. We observed that, in data relation for continuous and interrupted data, if the network and data were low, the interruption effect was high. Conversely, if data and networks were large, the interruption effect was low.

Our loss function was limited by a smaller network parameter size, just like the other existing losses (CCC and MSE). However, in the sequence range, our loss function showed better results in the lower rank, which allows better trade-off cost between the network size and sequence length.

In the future, we will continue to test the new loss function with acceleration detailed in Appendix A. We will also evaluate this new loss function using other states of the art networks to see how it improves emotion recognition accuracy.

Appendix

Extended proposed loss function

$$Distance = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{8}$$

$$Velocity = |\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\partial x_{i,j}}{\partial t} - \frac{\partial y_{i,j}}{\partial t}| \tag{9}$$

$$Alpha = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} = |\frac{1}{N(M+1)} \sum_{i=1}^{N}\sum_{j=1}^{M+1}(\frac{\partial x_{i,j}}{\partial t}\frac{\partial y_{i,j+1}}{\partial t} - \frac{\partial x_{i,j+1}}{\partial t}\frac{\partial y_{i,j}}{\partial t})| \tag{10}$$

$$Acceleration = |\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\partial^2 x_{i,j}}{\partial^2 t} - \frac{\partial^2 y_{i,j}}{\partial^2 t}| \tag{11}$$

$$loss = \frac{D + V + A + Ac}{4} \tag{12}$$

Proof of Equation 5:

$$\begin{gathered}
\Delta\alpha = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} => \\
\lim_{\Delta\alpha \to 0} \Delta\alpha = 0 \quad => \\
\Delta\alpha = 0 \quad => \\
0 = tan^{-1}\frac{a_x}{v_x} - tan^{-1}\frac{a_y}{v_y} \quad */tan \\
tan(0) = \frac{a_x}{v_x} - \frac{a_y}{v_y} \quad => \\
0 = \frac{a_x * v_y - a_y * v_x}{v_x * v_y} \quad => \\
0 = a_x * v_y - a_y * v_x \quad => \\
\Delta\alpha = a_x * v_y - a_y * v_x
\end{gathered} \tag{13}$$

Proof of Equation 5 in an example:

$$\begin{gathered}
assume \\
\angle annotation = 90° \\
\angle prediction = 30° \\
\Delta\alpha = 90° - 30° = 60° => \\
a_x = \sin(\pi/2) = 1 \quad v_x = \cos(\pi/2) = 0 \\
a_y = \sin(\pi/6) = 1/2 \quad v_y = \cos(\pi/6) = \sqrt{3}/2 \\
\sin(A - B) = \sin(A) * \cos(B) - \cos(A) * \sin(B) \\
\Delta\alpha = \arcsin(a_x * v_y - a_y * v_x) \\
\Delta\alpha = \arcsin(1 * \sqrt{3}/2 - 0 * 1/2) \\
\Delta\alpha = \arcsin(\sqrt{3}/2) = 60° \\
60° = 60°
\end{gathered} \tag{14}$$

Training tables

Table 3 Validation: 10% training

Network	Train Loss	Sequence Length	MSE	CCC	Non-Inertial
Full Network	MSE	128	0.233	0.334	0.117
		256	0.245	0.312	0.123
		512	0.229	0.288	0.115
	CCC	128	0.293	0.307	0.155
		256	0.303	0.338	0.161
		512	0.380	0.262	0.195
	Non-Inertial	128	0.216	0.404	0.109
		256	0.214	0.334	0.108
		512	0.217	0.343	0.109
Without mel Spectrogram	MSE	128	0.217	0.319	0.109
		256	0.226	0.277	0.115
		512	0.223	0.254	0.112
	CCC	128	0.287	0.426	0.147
		256	0.321	0.342	0.163
		512	0.330	0.292	0.182
	Non-Inertial	128	0.214	0.430	0.108
		256	0.210	0.375	0.106
		512	0.214	0.335	0.108
Audio Network	MSE	128	0.213	0.151	0.107
		256	0.199	0.202	0.100
		512	0.218	0.139	0.110
	CCC	128	0.231	0.300	0.116
		256	0.252	0.349	0.126
		512	0.256	0.274	0.129
	Non Inertial	128	0.208	0.131	0.105
		512	0.213	0.165	0.107
Pose Network	MSE	128	0.234	0.103	0.118
		256	0.232	0.150	0.117
		512	0.222	0.134	0.112
	CCC	128	0.335	0.230	0.180
		256	0.329	0.255	0.179
		512	0.457	0.176	0.247
	Non-Inertial	128	0.222	0.126	0.112
		256	0.229	0.151	0.115
		512	0.225	0.144	0.113

Table 4 5 Training (sequence 256 val 10%)

Network	Train Loss	Fold	MSE	CCC	Non-Inertial
Full Network	MSE	1	0.209	0.331	0.105
		2	0.211	0.346	0.107
		3	0.207	0.366	0.104
		4	0.225	0.337	0.113
		5	0.199	0.333	0.100
	CCC	1	0.348	0.309	0.176
		2	0.364	0.282	0.191
		3	0.335	0.321	0.170
		4	0.415	0.319	0.213
		5	0.271	0.348	0.136
	Non-Inertial	1	0.202	0.411	0.102
		2	0.206	0.355	0.104
		3	0.204	0.414	0.103
		4	0.202	0.415	0.102
		5	0.209	0.376	0.105
Without mel Spectrogram	MSE	1	0.209	0.328	0.105
		2	0.225	0.354	0.115
		3	0.227	0.304	0.115
		4	0.235	0.326	0.118
		5	0.227	0.304	0.114
	CCC	1	0.332	0.342	0.176
		2	0.344	0.310	0.176
		3	0.300	0.353	0.153
		4	0.383	0.339	0.199
		5	0.292	0.330	0.156
	Non-Inertial	1	0.209	0.398	0.105
		2	0.214	0.370	0.108
		3	0.206	0.402	0.104
		4	0.206	0.379	0.104
		5	0.205	0.375	0.103
Audio Network	MSE	1	0.214	0.180	0.108
		2	0.216	0.250	0.109
		3	0.219	0.193	0.110
		4	0.219	0.172	0.110
		5	0.217	0.211	0.109
	CCC	1	0.246	0.300	0.124
		2	0.239	0.288	0.120
		3	0.246	0.314	0.124
		4	0.239	0.321	0.120
		5	0.239	0.319	0.120
	Non-Inertial	1	0.215	0.221	0.108
		2	0.214	0.175	0.108
		3	0.212	0.209	0.107
		4	0.214	0.187	0.108
		5	0.215	0.182	0.108
Pose Network	MSE	1	0.226	0.097	0.114
		2	0.226	0.075	0.114
		3	0.228	0.071	0.115
		4	0.223	0.071	0.112
		5	0.233	0.077	0.117
	CCC	1	0.466	0.210	0.239
		2	0.389	0.205	0.203
		3	0.453	0.143	0.235
		4	0.360	0.173	0.187
		5	0.360	0.176	0.191
	Non-Inertial	1	0.228	0.069	0.115
		2	0.233	0.084	0.117
		3	0.227	0.075	0.114
		4	0.228	0.091	0.115
		5	0.226	0.085	0.113

Table 5 Validation: 30% training

Network	Train Loss	Sequence Length	MSE	CCC	Non-Inertial
Full Network	MSE	32	0.203	0.258	0.102
		64	0.201	0.251	0.101
		128	0.204	0.265	0.103
		256	0.198	0.251	0.100
		384	0.206	0.230	0.104
	CCC	32	2.470	0.000	1.235
		64	0.371	0.177	0.192
		128	0.291	0.289	0.148
		256	0.274	0.287	0.144
		384	0.281	0.283	0.149
	Non-Inertial	32	0.202	0.284	0.101
		64	0.204	0.346	0.103
		128	0.193	0.322	0.097
		256	0.194	0.260	0.098
		384	0.203	0.271	0.102
Without mel Spectrogram	MSE	32	0.203	0.249	0.102
		64	0.217	0.233	0.110
		128	0.202	0.241	0.102
		256	0.202	0.223	0.102
		384	0.208	0.201	0.105
	CCC	32	0.270	0.250	0.137
		64	0.359	0.229	0.180
		128	0.325	0.262	0.168
		256	0.268	0.290	0.140
		384	0.287	0.275	0.152
	Non-Inertial	32	0.208	0.243	0.105
		64	0.202	0.305	0.102
		128	0.191	0.285	0.096
		256	0.194	0.234	0.097
		384	0.195	0.290	0.098
Audio Network	MSE	32	0.213	0.202	0.107
		64	0.205	0.205	0.103
		128	0.201	0.184	0.101
		256	0.194	0.192	0.098
		384	0.200	0.180	0.101
	CCC	32	0.249	0.238	0.125
		64	0.244	0.252	0.123
		128	0.231	0.278	0.116
		256	0.251	0.300	0.126
		384	0.234	0.293	0.118
	Non-Inertial	32	0.202	0.184	0.102
		64	0.201	0.160	0.101
		128	0.199	0.162	0.100
		256	0.191	0.177	0.096
		384	0.195	0.171	0.098
Pose Network	MSE	32	0.229	0.059	0.115
		64	0.213	0.032	0.107
		128	0.209	0.094	0.105
		256	0.204	0.129	0.103
		384	0.207	0.134	0.104
	CCC	32	0.373	0.045	0.194
		64	0.432	0.076	0.225
		128	0.365	0.157	0.193
		256	0.350	0.213	0.179
		384	2.265	0.001	1.133
	Non-Inertial	32	0.212	0.046	0.106
		64	0.207	0.056	0.104
		128	0.205	0.118	0.103
		256	0.204	0.151	0.103
		384	0.202	0.154	0.101

References

[1] P. Abhang, S. Rao, B. W. Gawali, and P. Rokade: Emotion recognition using speech and EEG signal a review, International Journal of Computer Applications, Vol. 15, pp. 37-40, 2011.

[2] P. Ekman and E. L. Rosenberg: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), 2nd Edition, Oxford University Press, 2015.

[3] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N.Metaxas: Learning active facial patches for expression analysis, IEEE CVPR, 2012.

[4] A. Yao, J. Shao, N. Ma, and Y. Chen: Capturing AU-Aware facial features and their latent relations for emotion recognition in the wild, Proceedings of the 2015 A C Mon International Conference on Multimodal Interaction, ICMI 15, (New York, NY, USA), pp. 451-458, ACM, 2015.

[5] P. Ekman: Universals and Cultural Differences in Facial Expressions of Emotion, Nebraska, USA: Lincoln University of Nebraska Press, 1971.

[6] D. Al. Chanti and A. Caplier: Spontaneous Facial Expression Recognition Using Sparse Representation, Arxiv.org, 2018.

[7] D. Duncan, G. Shine and C. English: Facial Emotion Recognition in Real Time, 2016.

[8] I. Tautkute and T. Trzcinski: Classifying and Visualizing Emotions with Emotional DAN, Arxiv.org, 2018.

[9] C. Shan, S. Gong, and P. W. McOwan: Facial expression recognition based on local binary patterns: A comprehensive study, Image and Vision Computing, Vol. 27, No. 6, pp. 803-816, 2009.

[10] E. Barsoum, C. Zhang, C. C. Ferrer and Zhengyou Zhang: Training deep networks for facial expression recognition with crowd-sourced label distribution, ACM International Conference on Multimodal Interaction (ICMI), 2016.

[11] A. Dhall, R. Goecke, S. Lucey and T. Gedeon: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark, Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 2106-2112, 2011.

[12] S. Thuseethan and S. Kuhanesan: Eigenface based recognition of emotion variant faces, Computer Engineering and Intelligent, Systems www.iiste.org ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online), Vol.5, No.7, 2014.

[13] R. Vemulapalli and A. Agarwala: A Compact Embedding for Facial Expression Similarity, Arxiv.org, 2019.

[14] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy and F. Jurie: An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets, Arxiv.org, 2018.

[15] J. Yang, K. Wang, X. Peng and Y. Qiao: Deep Recurrent Multi-Instance Learning with Spatio-temporal Features for Engagement Intensity Prediction, EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction, 2018.

[16] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell: Long-term recurrent convolutional networks for visual recognition and description, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, pp. 677-691, 2017.

[17] O. Jargalsaikhan and S. Karungaru: Facial emotion recognition accuracy improvement using deep learning, ACEAT Dec 17-19, pp. 88-93, Kyoto, Japan, 2019.

[18] D. Deng, Z. Chen, Y. Zhou and B. Shi: MIMAMO Net: Integrating Micro- and Macro-motion for Video Emotion Recognition, Arxiv.org, 2018.

[19] C. W. Lee, K. Y. Song, J. Jeong and W. Y. Choi: Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data, Arxiv.org, 2018.

[20] A. Mollahosseini, B. Hasani and M. H. Mahoor: Affect-Net: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE, Transactions on Affective Computing, Vol. 10, pp. 18-31, 2019.

[21] S. Li and W. Deng: Deep Facial Expression Recognition: A Survey, Arxiv.org, 2018.

[22] R. Plutchik: Emotion: A Psychoevolutionary Synthesis, New York, NY: Harpercollins College Division, 1980.

[23] I. Lawrence and K. Lin: A concordance correlation coefcient to evaluate reproducibility, Biometrics, Vol.45, pp. 255-268, 1989.

[24] Z. Cao, G. H. Martinez, T. Simon, S. Wei and Y. A. Sheikh: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No.1, pp. 172-186, 2019.

[25] W. Y. Chang, S. H. Hsu, and J. H. Chien: FATAUVA-Net : An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-8, 2018.

[26] D. Kollias et al.: Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond, International Journal of Computer Vision, Vol.127, pp. 907-929, 2019.

[27] J. Westerink, M. Krans and Martin: Sensing Emotions: The Impact of Context on Experience Measurements Table 2.1. ISSN 1571-5671 26, 2011.

Jargalsaikhan Orgil is a Lecturer at the School of Information and Communication Technology of Mongolian University of Science and Technology. Since 2018, he has been a doctoral course student at the Graduate School of Technology, Industrial and Social Sciences, Tokushima University. His research interests include emotion recognition, image processing and deep learning.

Stephen Karungaru received his B.S degree from the Department of Electrical and Electronics, Moi University, Kenya. He received his master's and doctoral degrees from the Department of Information Science and Intelligent Systems, Faculty of Engineering, Tokushima University, in 2001 and 2004, respectively. He became an Associate Professor, Graduate School of Technology, Industrial and Science, Tokushima University in 2004. His research interests are in face detection, recognition, neural networks, and genetic algorithms.

Kenji Terada received his B.E., M.E., and Dr.Eng. degrees in electrical engineering, all from Keio University, Japan in 1990,1992 and 1995, respectively. He is currently a Professor at Tokushima University. His research interests include the image processing, and computer vision.

Ganbold Shagdar is an Associate Professor in the School of Information and Communication Technology of Mongolian University of Science and Technology (MUST). He is a member of IEEE. He received his B.S. and M.S degrees in electronic engineering from MUST, Ulaanbaatar, Mongolia, in 1991 and 1997, respectively, and Ph.D. degree from Gwangju Institute of Science and Technology, Gwangju, Korea, in 2016. His current research interests include dynamic wavelength and bandwidth allocation, energy efficiency and resilient architecture in optical access networks, optical cross connection architectures, and optical transport network.

(Received July 29, 2020; revised October 9, 2020)