Video-Audio Emotion Recognition Based on Feature Fusion Deep Learning Method

Yanan Song, Yuanyang Cai, and Lizhe Tan
Department of Electrical and Computer Engineering, Purdue University Northwest, Indiana, USA
Email: song635@pnw.edu; cai250@pnw.edu; lizhetan@pnw.edu

2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) | 978-1-6654-2461-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/MWSCAS47672.2021.9531812

***Abstract*—In this paper, we propose a video-audio based emotion recognition system in order to improve the successive classification rate. The features from audio frames are extracted using Mel frequency Cepstral coefficients (MFCC) while the features from video frames are extracted from VGG16 with pre-trained weights on the ImageNet dataset [17]. Then recurrent neural networks (RNN) are further applied to process the sequence information. The outputs of both RNN are fused into a concatenate layer and then the final classification result is obtained by the softmax layer. Our proposed system achieves 90% accuracy based on the RAVDESS dataset for eight emotion classes.**

***Keywords*— Emotion recognition, MFCC, Transfer learning, Deep learning, RNN.**

I. INTRODUCTION

Emotion recognition applications have been increasing in communication, mobile applications, human-machine interface, healthcare industry for training and diagnosis, emotional care to people, and interactive learning and gaming [1]–[15] due to advances in computing technology and deep learning neural network algorithms. The common emotions for human being are classified as Happy, Sad, Angry, Fear, Surprise, and Disgust in general. As a unimodal social behavior, emotions can be expressed in terms of speech, facial expression, and gestures; and on the other hand, the emotions exhibit in the multimodal form such as audio, video, and physiological signal. Thus, the emotion recognition system has been successfully developed using features from speech, images, videos, text, or physiological information. Emotion recognition recently finds a significant application in mobile phone application [2].

A number of emotion recognition systems exist in the literature and commercial use. Among them, speech, images, and videos are commonly used. Emotion recognition using speech and deep learning neural network has been reported in references [5]–[9]. Emotion recognition using facial images can be found in [12]–[13]. Researchers [20] have developed a deep learning based approach to recognize emotion from video clips and temporal modeling. A recent emotion recognition system using both audio and visual modalities has been proposed in [21]. In addition, the emotion recognition using physiological data such as electroencephalogram (EEG) signals was evidenced to investigate social psychology [14].

Although many emotion recognition systems use single modality via speech/audio, images, videos, or physiological data along, it is expected that using multi-modalities along with data fusion [19], [22] can improve the classification accuracy. Emotion recognition using a feature fusion method can be found in [23]-[24], where the transfer learning is used to extract visual features. The feature extraction for emotion recognition may be considered and adopted in order to achieve the reduction of data dimension and remove noise in data, resulting in an improved performance.

In our study, a video-audio emotion recognition system is developed. For the video frames, the feature-representation-transfer learning [16] is applied to extract the features to correctly represent emotion information with a reduced dimension of input data, that is, instead of using randomly initialized weights to train the model from scratch, a transfer learning technique uses pre-trained models in which the weights are usually trained on a large dataset such as ImageNet [17]. In this paper, the VGG16 with the ImageNet pre-trained weights is proposed to be used as the feature extractor for the visual data [12]. On the other hand, the features from audio frames are extracted using Mel frequency cepstral coefficients (MFCC). Next, a recurrent neural network (RNN) is proposed to be used, since the RNN has been proven to be very effective and widely used in video and audio processing [6]-[10], and natural language processing [9] to process sequential information of the extracted audio and video features. In our study, a particular type of RNN, that is, the gated recurrent unit (GRU) [18] has been adopted. Our developed system conducts emotion recognition by applying the CNN-GRU subsystem for video data and the MFCC-GRU subsystem for speech/audio data simultaneously, and then concatenating two models via the feature fusion [19] in order to achieve the improved emotion classification rate.

This paper is organized as follows. Section II illustrates the framework of the proposed system, including a feature fusion model, GRU based video subsystem and audio subsystem, respectively. In Section III, the experiments are described and results are presented. Finally, the conclusions are given in Section V.

II. The Proposed System

A. Feature Fusion Model

Our proposed system is depicted in Fig.1.

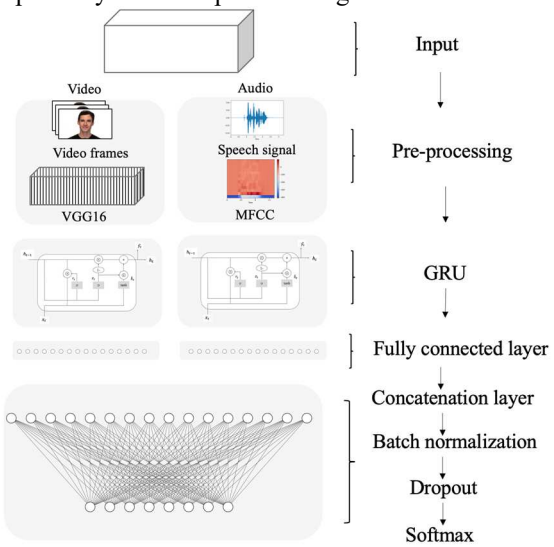

Fig. 1. Feature fusion model.

As shown in Fig. 1, the system contains two subsystems: VGG16-GRU for video frames and MFCC-GRU for audio frames. VGG16-GRU subsystem takes video input and generates the VGG16 features using the pre-trained VGG16 weights and then the achieved features are fed to the gated recurrent unit (GRU) layer, which is a particular type of the recurrent neural network (RNN). Meanwhile, the MFCC-GRU subsystem takes audio input to produce the Mel frequency Cepstral coefficients (MFCC). The obtained MFCC features are fed to three GRU layers. The video feature vectors from the last fully connected layer in VGG16-GRU model and the audio features vectors from the last fully connected layer in the MFCC-GRU model are fused via concatenation. The finally, the system performs batch normalization and classification by the softmax operation.

B. VGG16-GRU Subsystem

The VGG16-GRU subsystem shown in Fig. 1 is detailed in Fig. 2.

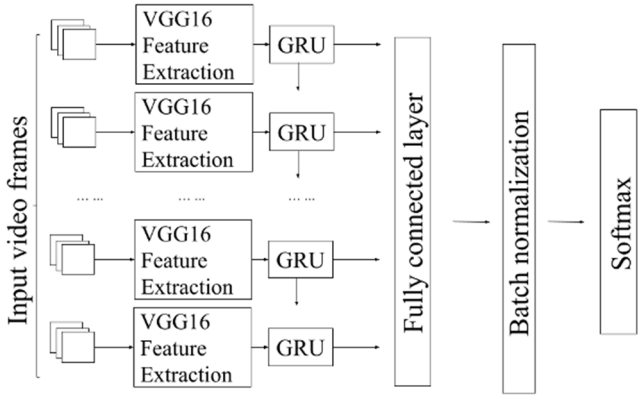

Fig. 2. VGG16-1 layer GRU model.

As shown Fig. 2, the VGG16-GRU subsystem possesses a VGG16 network [25], cascaded with one GRU layer followed by a fully connected layer. Each VGG16 unit consists of five blocks as described in Fig. 3, where each block has two convolutional layers and a max-pooling layer. Note that the pre-trained weights on the ImageNet dataset from the first block to the fifth block are frozen during the training progress.

Gated recurrent unit (GRU) is a type of recurrent neural network (RNN) introduced in 2014 by Cho et al. [18]. A GRU operates using an update gate and reset gate. The update gate controls how much past information pass to the next states while the reset gate controls how much past information to forget.

Fig. 4 shows the standard architecture of a GRU unit [18], [26], used in our research. Equations (1)-(4) define the signal flows:

$$h_t^j = \left(1 - z_t^j\right)h_{t-1}^j + z_t^j \tilde{h}_t^j \tag{1}$$

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \tag{2}$$

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

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \tag{4}$$

where σ is the logistic sigmoid function while $\odot$ denotes the element-wise multiplication. The number of GRU layers, number of neurons in the GRU, and batch normalization can be specified in the developed system.

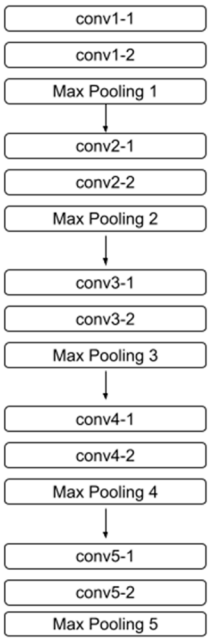

Fig. 3. VGG16 model in VGG16-GRU subsystem.

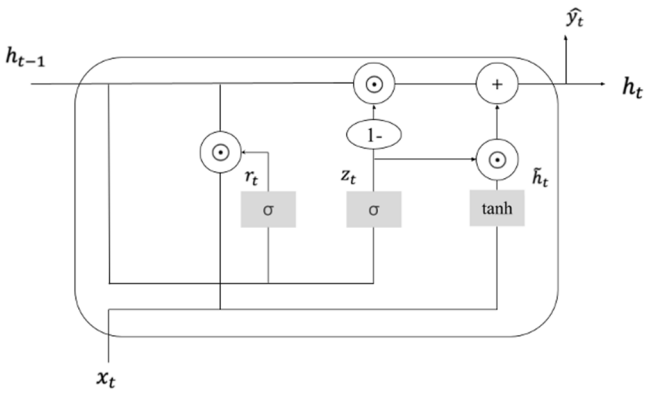

Fig. 4. GRU architecture [18].

Note that based on the results of our experiments, one layer of GRU units is proposed for our VGG16-GRU subsystem.

C. MFCC-GRU Subsystem

Fig. 5 displays the MFCC-3 layers subsystem. As illustrated, the Mel-frequency cepstral coefficients (MFCC) are extracted at the first stage and adopted as the audio features. MFCC has been widely found in speech recognition and music information retrieval applications [7]. Acquiring MFCC features requires the following steps: pre-emphasis, frame blocking, hamming windowing operation, fast Fourier transform (FFT), Mel filter bank processing, discrete cosine transform (DCT), delta energy, and delta spectrum. The obtained MFCC features are fed to GRU units at the input layer. After a variety of experiments by adopting various GRU layers, neurons in the GRU, and batch normalization, three layers of GRUs are employed.

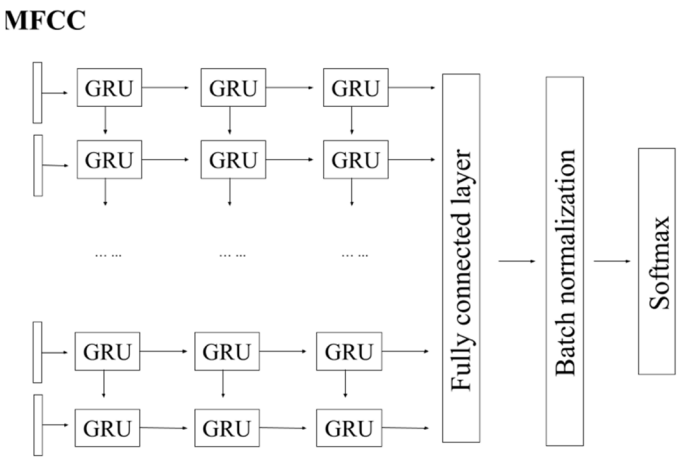

Fig. 5. MFCC-3 layer GRU model.

D. Feature Fusion

The feature fusion displayed in Fig. 1 is illustrated as follows. First, the video and audio feature vectors extracted from the VGG16-GRU network (Fig. 2) and MFCC-GRU network (Fig. 5) are fed to the respective GRU networks. The fully connected layer output from the VGG16-GRU network and the fully connected layer output from the MFCC-GRU network are concatenated into a new feature vector. In our proposed system, the sizes of the fully connected layers for both VGG16-GRU and MFCC-GRU networks are experimentally set to 256 and 256, respectively. Therefore, the concatenate layer has 512 units, composing of the audio and video feature vectors.

E. Batch normalization

In order to reduce the internal covariate shift during the training procedure, we adopt the batch normalization described in [27] between the fully connected layer and the concatenation layer. Based on our experiments shown in Tables II and III listed in the next section, the batch normalization not only can improve the classification accuracy but also reduce the training time. Consider a mini-batch B of size m, where $B = \{x_1 \cdots x_m\}$. The batch normalization is conducted using Equations (5)-(8) [27] listed below:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{5}$$

$$\sigma_B{}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{6}$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B{}^2 + \epsilon}} \tag{7}$$

$$y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i) \tag{8}$$

where m denotes the batch size, ϵ denotes a small constant for numerical stability, γ and β are a pair of parameters that scale and shift the normalized value. $BN_{\gamma,\beta}(\ )$ denotes the batch normalization transform.

F. Softmax Classifier

The features from previous steps are further be processed by the softmax layer. Softmax can calculate the probabilities of each class by mapping the value of neuron= output from 0 to 1, the final classification result can be calculated by arguments of the maxima (argmax) which refers to select the largest probability in the softmax output. The calculation of softmax is shown in Equation (9).

$$\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{9}$$

where *z* denotes input vectors, *e* denotes the natural base, and *K* denotes the number of classes.

III. Experiments And Results

A. Dataset

To evaluate our developed emotion recognition system, we adopted the Ryerson Audio-Visual Database [1] of Emotional Speech and Song (RAVDESS), which contains 7356 files, including audio and visual data from 24 actors consisting of 12 males and 12 females. Speech data contains eight emotions, that is, neutral, calm, happy, sad, angry, fearful, disgust, and surprise whereas song data contains calm, happy, sad, angry, and fearful emotions. In our study, we only used the speech-

video data [1], which consists of 1440 audio-visual files (96 files for neutral, 192 calm, 192 happy, 192 sad, 192 angry, 192 fearful, 192 disgust, and 192 surprise). Each audio-visual file has video recording format with a scan resolution of 1920x1080 pixels at a frame rate of 30 frames per second (fps) and speech recording format at a sampling rate of 48 kHz at 16-bit resolution. Figs. 6 and 7 display samples of the video and speech data, respectively.

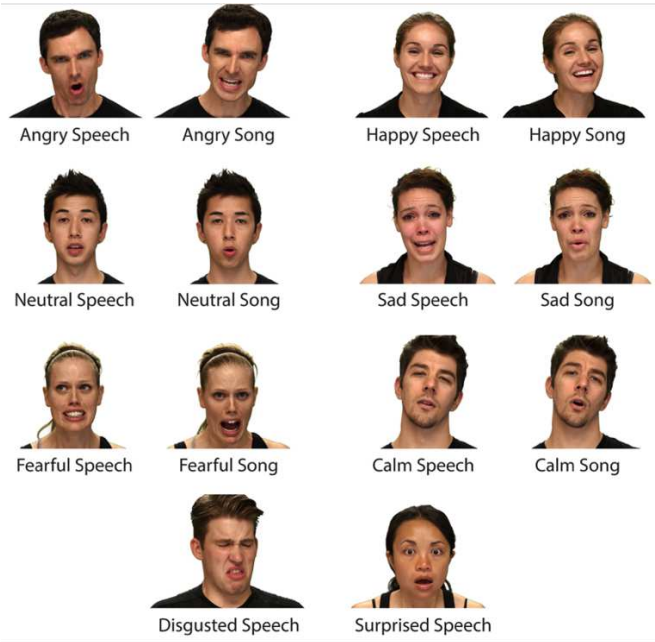

Fig. 6. Example of RAVDESS database.

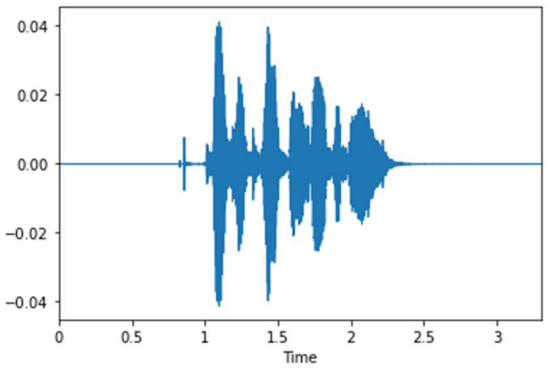

Fig. 7. Speech signal samples (Actor01).

B. Pre-processing

The video frame rate was dropped by a factor 4 for incoming video sequence. In the VGG16-GRU subsystem, a face detector was applied to preprocess each video frame. After face detection, all detected facial images were resized to $48\times48\times3$. As illustrated in Fig. 2, the facial features were produced from the pooling layer from block 5 using the pre-trained weights from the ImageNet dataset. From the MFCC-GRU subsystem, thirteen MFCC coefficients were calculated according to the Mel frequency scale specification for each speech frame with a length of 4096 and used for the network inputs. Fig. 8 displays the MFCC features in terms of frequency (vertical axis) and temporal information (horizontal axis).

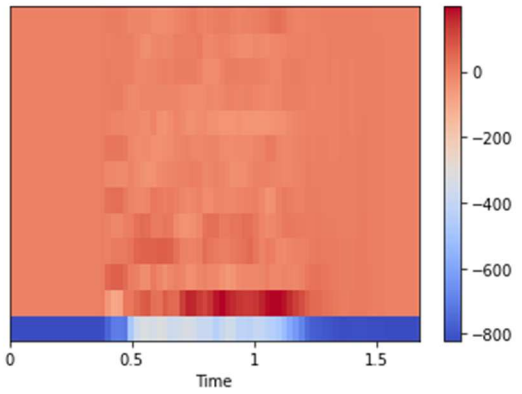

Fig. 8. MFCC features extracted from speech.

For the VGG16-GRU and MFCC-GRU networks, respectively, we randomly chose 80% data for training and 20% data for validation. In the audio-video feature fusion network, we performed 10-fold cross-validation.

C. Results from VGG16-GRU Classifier (Video)

For a comparison purpose, we used the VGG16-GRU network alone to validate the performance. Table I shows the validation accuracy according to the number of layers for GRUs used, the number of units in each GRU, and the option for batch normalization.

TABLE I VGG16-GRU MODELS ACCURACY

Model	Validation accuracy
VGG16 and 1 layer GRU (256 units) with batch normalization	78%
VGG16 and 3 layer GRU (256 units) with batch normalization	74%
VGG16 and 1 layer GRU (512 units) without batch normalization	76%
VGG16 and 1 layer GRU (512 units) with batch normalization	**79%**
VGG16 and 3 layer GRU (512 units) without batch normalization	74%
VGG16 and 3 layer GRU (512 units) with batch normalization	76%
VGG16 and 1 layer GRU (1024 units) without batch normalization	77%
VGG16 and 1 layer GRU (1024 units) With batch normalization	78%

From the experiments shown in Table I, the VGG16 1-layer GRU with 512 units using batch normalization obtains 79% accuracy. The performance confusion matrix is illustrated in Fig. 9, where the emotion of "calm" has the highest rate of 95% while there is only 53% of accuracy for "fearful", which is the lowest accuracy.

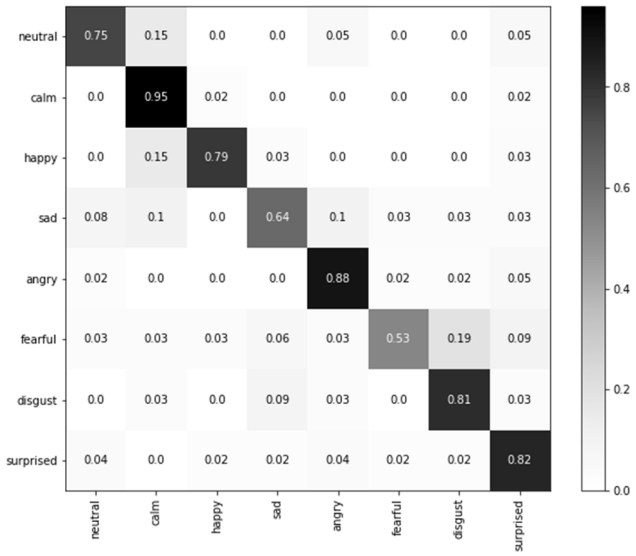

Fig. 9. Confusion matrix VGG16-1 layer GRU with 512 units and batch normalization.

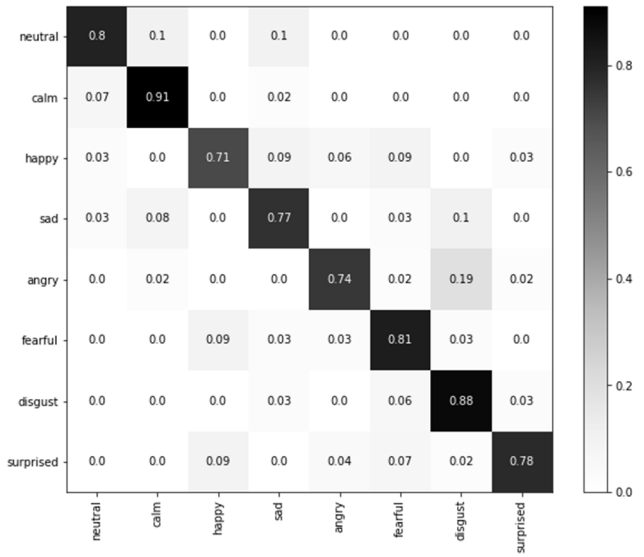

Fig. 10. Confusion matrix MFCC-3 layer GRU with 256 units and batch normalization.

D. Results from MFCC-GRU Classifier (Audio)

The performance results from the MFCC-GRU subsystem using a various number of GRU layers and GRU units with the option of using batch normalization are listed in Table II.

TABLE II MFCC-GRU MODELS ACCURACY

Model	Validation accuracy
MFCC and 1 layer GRU (256 units) without batch normalization	69%
MFCC and 1 layer GRU (256 units) with batch normalization	75%
MFCC and 3 layer GRU (256 units) without batch normalization	75%
MFCC and 3 layer GRU (256 units) with batch normalization	**80%**
MFCC and 1 layer GRU (512 units) with batch normalization	75%
MFCC and 3 layer GRU (512 units) with batch normalization	79%

From the experiments shown in Table II, the MFCC-3 layer GRU with 256 units and the use of batch normalization obtains 80% of accuracy. The performance confusion matrix is illustrated in Fig. 10. Note that the classification of "calm" obtains the highest accuracy of 91% and the recognition of "happy" has the lowest accuracy of 71%.

E. Results from Audio-Video Feature Fusion System

By using one VGG16-1 layer GRU with 512 units for video frames, three VGG16-3 layer GRU with 256 units for audio frames, and the feature fusion model as depicted in Fig. 1, the proposed video-audio feature fusion system achieved the improved performance with the accuracy of 90%. Fig. 11 shows the confusion matrix. Note that the accuracy is the average over 10 cross-validation folds.

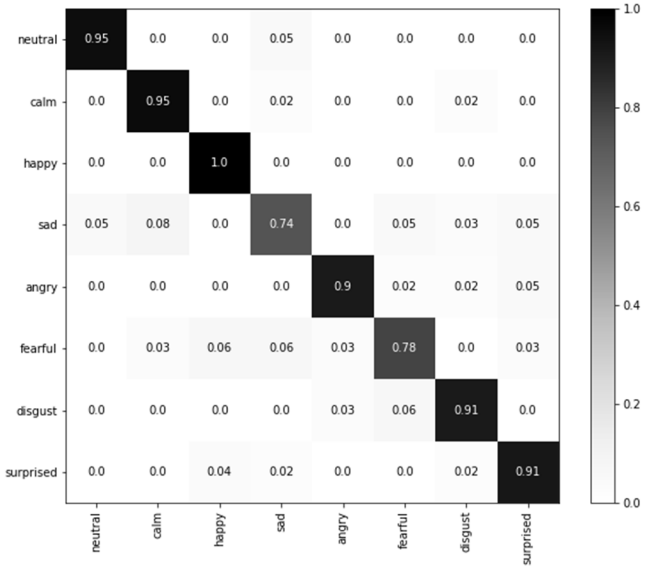

Fig. 11. Feature fusion model confusion matrix.

Table III summarizes the comparison results with the ones from the other published methods. As shown in Table III, our MFCC-GRU subsystem using audio information only achieves 80% accuracy versus 62% accuracy by human volunteers listed in the database [1]. Our VGG16-GRU subsystem using video information only has the accuracy of 79% while the human volunteers listed in the database [1] only achieve 72%. Using both audio and video information, our audio-video (AV) feature fusion system finally reaches the accuracy of 90%, which is 10% above the result from human volunteers in the database. According to the results, the proposed AV feature fusion system outperforms the subsystems using video or audio alone, as well as outperforms the models proposed by Ghaleb et al. [28], Mustaqeem et al. [29], and Issa et al. [30].

TABLE III COMPARISON WITH THE PREVIOUS WORKS USING RAVDESS DATABASE

Method	Accuracy
Multimodal and temporal perception (AV)[28]	67.70%
Incorporating learned features and deep BiLSTM (AV)[29]	77.02%
Deep convolutional neural networks (AO)[30]	71.61%
Human volunteers (AO)[1]	62%
Ours (AO)	80%
Human volunteers (VO)[1]	72%
Ours (VO)	79%
Human volunteers (AV)[1]	80%
Ours (AV)	**90%**

IV. CONCLUSIONS

In this paper, we propose a video-audio based emotion recognition system. The system has two subsystems. The first one uses VGG16 to extract video features followed by the single layer GRU network and the second one applies the MFCC audio features to the multiple layer GRU network. The outputs of both networks are fused into a concatenate layer and the final classification result is calculated by the softmax operation. Our proposed AV fusion system can achieve 90% accuracy based on the RAVDESS dataset with eight different emotions. Furthermore, our system training and testing are accelerated by Google Cloud TPU. It takes about 2.5s to perform emotion recognition given a 3.0s video clip The improvement on computation reduction, preprocessing speed, and effectiveness of the network structure will be our further work.

REFERENCES

[1] S. R. Livingstone and F. A. Russo, "The RYERSON audio-visual database of emotional speech and song (ravdess): A dynamic multimodal set of facial and vocal expressions in north American English," PloS one, vol. 13, no. 5, pp. e0196391, 2018.

[2] M. S. Hossain, G. Muhammad, "Emotion recognition system for mobile applications," IEEE Access, pp. 2281-2287, February 2017.

[3] M. Chen, Y. Hao, Y. Li, D. Wu, and D. Huang, "Demo: LIVES: Learning through interactive video and emotion-aware system," in Proc. ACM Mobihoc, pp. 399-400, Hangzhou, China, Jun. 2015.

[4] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio_visual emotion-aware cloud gaming framework," IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 12, pp. 2105_2118, Dec. 2015.

[5] W. Minker, J. Pittermann, A. Pittermann, P. Strauss and D. Bühler, "Challenges in speech-based human–computer interfaces," Int. J. Speech Technol., vol. 10, no. 2-3, pp. 109-119, 2007.

[6] W. Zhang, D. Zhao, X. Chen, and Y. Zhang, "Deep learning based emotion recognition from Chinese speech," in Proc. 14th Int. Conf. Inclusive Smart Cities Digit. Health (ICOST), vol. 9677, pp. 49-58, New York, NY, USA, 2016.

[7] C. S. Kumar and P. R. Mallikarjuna, "Design of an automatic speaker recognition system using MFCC vector quantization and LBG algorithm," Int. J. Comput. Sci. Eng., vol. 3, no. 8, 2011.

[8] A. Graves, A.-R. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013 IEEE international conference on acoustics speech and signal processing, pp. 6645-6649, 2013.

[9] T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent trends in deep learning based natural language processing," IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55-75, Aug. 2018.

[10] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: Audio visual and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39-58, 2009.

[11] J. W. Ragsdale, R. Van Deusen, D. Rubio, and C. Spagnoletti, "Recognizing patients emotions: Teaching health care providers to interpret facial expressions," Acad. Med., vol. 91, no. 9, pp. 1270-1275, Sep. 2016.

[12] B. Yang, J. Cao, R. Ni and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," IEEE Access, vol. 6, pp. 4630-4640, 2017.

[13] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," Sensors, vol. 11, no. 10, pp. 9573_9588, 2011.

[14] M. Ali, A. H. Mosa, F. A. Machot, and K.Kyamakya, "EEG-based emotion recognition approach for e-healthcare applications," in Proc. 8th Int. Conf. Ubiquitous Future Netw. (ICUFN), pp. 946-950, Vienna, Austria, 2016.

[15] Y. Fan, X. Lu, D. Li and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," Proc. 18th ACM Int. Conf. Multimodal Interaction, pp. 445-450, 2016.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. FeiFei, "ImageNet: A large-scale hierarchical image database," Conference on Computer Vision and Pattern Recognition, pp.248-255, June 2009.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473, 2014. 1, 2

[19] U. G. Mangai, S. Samanta, S. Das and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," ETE Tech. Rev., vol. 27, no. 4, pp. 293-307, 2010.

[20] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Multi-scale temporal modeling for dimensional emotion recognition in video," in Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge (AVEC), pp. 11-18, Orlando, FL, USA, Nov. 2014.

[21] S. E. Kahou et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," J. Multimodal User Interfaces, vol. 10, no. 2, pp. 99-111, Jun. 2016.

[22] S. Khalid, T. Khalil and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," Science and Information Conference (SAI) 2014, pp. 372-378, 2014.

[23] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimedia Tools Appl, vol. 80, (2), pp. 2887-2905, 2021.

[24] J. D. S. Ortega, P. Cardinal and A. L. Koerich, "Emotion Recognition Using Fusion of Audio and Video Features", 2019 IEEE International Conference on Systems Man and Cybernetics (SMC), 2019.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[26] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action recognition in video sequences using deep Bi-directional LSTM with CNN features," IEEE Access, vol. 6, pp. 1155-1166, 2017.

[27] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[28] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII), Cambridge, U.K., Sep. 2019, pp. 552–558.

[29] M. Mustaqeem,M.Sajjad,andS.Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," IEEE Access, vol. 8, pp. 79861–79875, 2020.

[30] D. Issa, M. F. Demirci and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", Biomedical Signal Processing and Control, vol. 59, pp. 101894, 2020.