Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 4, 2022, 529-537.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Audio-visual emotion recognition based on a deep convolutional neural network

Khadijeh Aghajani*

Department of Computer Engineering, University of Mazandaran, Babolsar, Iran.

Article Info

Article History:
Received 09 April 2022
Revised 25 July 2022
Accepted 25 August 2022

DOI:10.22044/jadm.2022.11809.2331

Keywords:
Speech emotion recognition, Facial emotion recognition, Deep learning, Transfer learning.

**Corresponding author: kh.aghajani@umz.ac.ir (Kh. Aghajani).*

Abstract

Emotion recognition has several applications in various fields, including human-computer interactions. In the recent years, various methods have been proposed to recognize emotion using facial or speech information, while the fusion of these two has been paid less attention in emotion recognition. In this work, first of all, the use of only face or speech information in emotion recognition is examined. For emotion recognition through speech, a pre-trained network called YAMNet is used to extract the features. After passing through a convolutional neural network (CNN), the extracted features are then fed into a bi-LSTM with an attention mechanism to perform the recognition. For emotion recognition through facial information, a deep CNN-based model is proposed. Finally, after reviewing these two approaches, an emotion detection framework based on the fusion of these two models is proposed. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) containing videos taken from 24 actors (12 men and 12 women) with 8 categories is used to evaluate the proposed model. The results of the implementation show that a combination of the face and speech information improves the performance of the emotion recognizer.

1. Introduction

Automatic recognition of emotion plays an important role in several applications including computer games, human-computer interactions, educational software, robots, automobile safety, as well as treatment of mental illnesses such as depression. The humans express their emotions through sound, face, gesture, and body posture. However, it should be noted that in some applications, such as call centers, some of these resources may not be available. Each of these sources contains important information related to emotion recognition. Research in this regard has applied one or a combination of these modalities to recognize emotion. Among these sources, speech and facial information have received more attention than the other modalities.

In the traditional approaches, first, hand-crafted features have been extracted from the input signal, and then recognition has been performed using an appropriate classification method such as SVM (support vector machine), and RF (random forest). With the significant progress of computers and the applicability of end-to-end CNN-based models, emotion recognition has been done with a higher accuracy and a faster speed.

In this research work, the goal is to recognize the emotion of a person based on a video of his face. Due to the availability of speech and facial information in this problem and the recent research in these two fields, these sources are examined separately, and finally, the audio-visual framework is proposed by combining information from these two sources.

Here, three approaches (speech-based, face-based, and audio-visual-based) have been proposed for emotion recognition. To recognize emotion from speech, the output of the third convolutional block ("activation-3" layer) of a pre-trained model called YAMNet has been used as a feature vector. This feature vector is then fed as an input to a model consisting of CNN followed by a bi-LSTM with an attention layer. For the face-based

approach, after extracting faces from video frames, a CNN-based model has been used to recognize emotion. Finally, by combining these two methods, a multi-modal model for recognizing emotion with the help of facial and speech information has been presented. The proposed model reached an accuracy of 81.04% on the RAVDESS database.

In summary, the contributions of this study are as follows:

- Two approaches based on only facial or speech information for emotion recognition have been proposed. For each source (speech or face), the proposed models are evaluated. For the speech signal, a pre-trained network is used to extract the low-level features.
- Based on the configuration of the models obtained in the speech-based and facial-based approaches, the final audio-visual model is proposed and evaluated.

The rest of the paper is organized as what follows. According to the research done in the fields of speech-based, facial-based, and audio-visual emotion recognition, in the second section, some recent research works done in all these three approaches has been reviewed. In the third section, three approaches are presented to recognize emotion, one solely based on face, the other solely based on speech, and the third based on a combination of the two. The fourth section is dedicated to implementation and observations. The final section contains the conclusion.

2. Related Works

In some studies, speech signals have been used alone to recognize emotion. In most of them, hand-crafted features such as Mel Frequency Magnitude Coefficients (MFCCs) [1], Mel-Energy Spectrum Dynamic Coefficients [2], Linear Predictive Spectrum Coding [3], Discrete Wavelet Transform [4], speaking rate, and pitch [5]. have been extracted from the speech signal. Then a machine learning method such as Bayesian Networks (BN) or SVM is used to detect emotion. In [6], considering MFCCs and their spectral centroids as the input features and using SVM as a classifier in the RAVDESS database, an accuracy of around 71% was achieved.

In the recent years, like many other applications, the use of deep learning networks for emotion recognition has become widespread [7-9]. In [10], the features consisting of Mel-frequency cepstral coefficient, Mel-scale-spectrogram, chromatogram, and Tonnez representation were extracted from the speech signal. The extracted features were then fed into a 1D CNN-based network. The accuracy reported in the RAVDESS database was 71.6%. In [8], the 3-D log Mel spectrum features were extracted from speech signal, and then these features were fed into a CNN-based model. Zhao *et al.* have proposed two frameworks [11]. The first one is a combination of 1D CNN and an LSTM conducted on speech and the second one is a combination of 2D CNN and LSTM conducted on log-Mel-spectrum. The results show that the second approach outperforms the other one. Tzirakis *et al.* proposed a model consisting of a 1D CNN followed by a 2-layer LSTM conducted on a speech signal [12]. In [13], the 3D scalograms extracted from speech signals are fed into the CNN layers followed by an LSTM with an attention mechanism to recognize the emotion. Li *et al.* have proposed a CNN-BiLSTM model to classify both gender and emotion [14].

Facial information whether it is a single image or part of a video in the form of frames contains important information helping to recognize emotion. One of the common approaches for emotion recognition is to check the coordinates of some key points allocated around the eyes, nose, mouth, jaw, and eyebrows [15]. Bagheri *et al.* have proposed a method based on extracting action units [16]. In this work, the extracted features are applied to an auto-encoder, and the recognition operation is performed. Wang *et al.* have used stationary wavelet entropy for feature extraction [17]. The extracted features have been fed into a neural network with a hidden layer to recognize emotion.

In [18], a two-stage CNN-based framework is proposed for facial emotion recognition. In the first stage, the background is removed from the whole picture. Then a CNN is used to extract a facial feature vector for emotion recognition. In [19], Gabor filters were applied to the original images, and then the obtained images were fed into a CNN-based model for emotion recognition. The performance of GoogLNet and AlexNet on facial emotion recognition has been examined in [20]. In [21, 22], the CNN-based models have been proposed and learned for facial emotion recognition. Falahzadeh *et al.* have utilized the horizontal and the vertical gradients of the original image as the second and third layers of the AlexNet-DCNN input. After utilizing transfer learning the presented model is fine-tuned on the training samples [23].

In the case of using video frames, emotion recognition can be done with the help of two approaches. A) Using the information of each frame as separate images, making the initial recognition, and then using a voting mechanism to

make the final decision [24]; B) Using the sequence of information belonging to each frame and examining the process of face changes in the video. For this purpose, the pre-trained networks to extract face embedding in each frame can be used. By applying these features to a temporal network such as LSTM or bi-LSTM, recognition operations can be performed [25, 26]. In the recent years, several studies have used a combination of audio-visual data to recognize emotion. Data combination from different sources can be done at the feature level, decision level, and model level. In the feature level approach, the features extracted from each source are combined and then applied as input to the model. This method is often less popular due to differences in the nature of different sources (e.g. video is three-dimensional, while audio is one-dimensional). However, it is probably more effective due to the consideration of the relationship between the features belonging to different sources. In the second approach, the relationship between different sources is not considered, and the final output is determined from the outputs of the models related to each source. The third approach mediates the other two approaches. In this approach, the output of the hidden layers of the models related to each source are combined and after passing the final hidden layers, the result is obtained. In the following, some studies in this field are reviewed.

In [27], a combination of audio-visual data was used to distinguish 6 senses. For sound, features such as zero-crossing, MFCC, LPC, and PLP as well as their first-order derivatives were first extracted from the speech signal. For the face, the temporal variations for some face landmarks in consecutive frames were used to extract the feature. By joining these features, the final feature vector was created. In this work, the use of a combination of features led to an increased accuracy. Ren *et al.* have recognized emotion from a combination of audio-visual data [28]. They extracted the desired features by passing the Mel-spectrograms through several convolutional layers followed by an LSTM and used triplet loss to further distinguish the class features. Then they made classifications. Song *et al.* have used the EdNet network for facial feature extraction and a speech analysis engine for audio feature extraction [29]. An ANN network was then used to recognize emotion according to the extracted vectors. The accuracy reported in this work based on SAVEE audio-visual data related to 7 senses was reported to be 43%. Ortega *et al.* have used a combination of audio, video, and text data to recognize emotion [30]. In this paper, a feature vector including 23 speech descriptors such as energy, spectral, and pith per frame was used. For the face, the feature vector was extracted from frames with the time step of 20 milliseconds. The facial features included some information such as the coordinates of 59 specific face landmarks and the normalized face orientation in degrees. The bag of word feature was extracted for the text. The applied dictionary contained 512 words. The vectors obtained from all three sources were then applied to three separate ANN networks. The outputs were combined. By applying the combined data to an ANN, the recognition operation was performed.

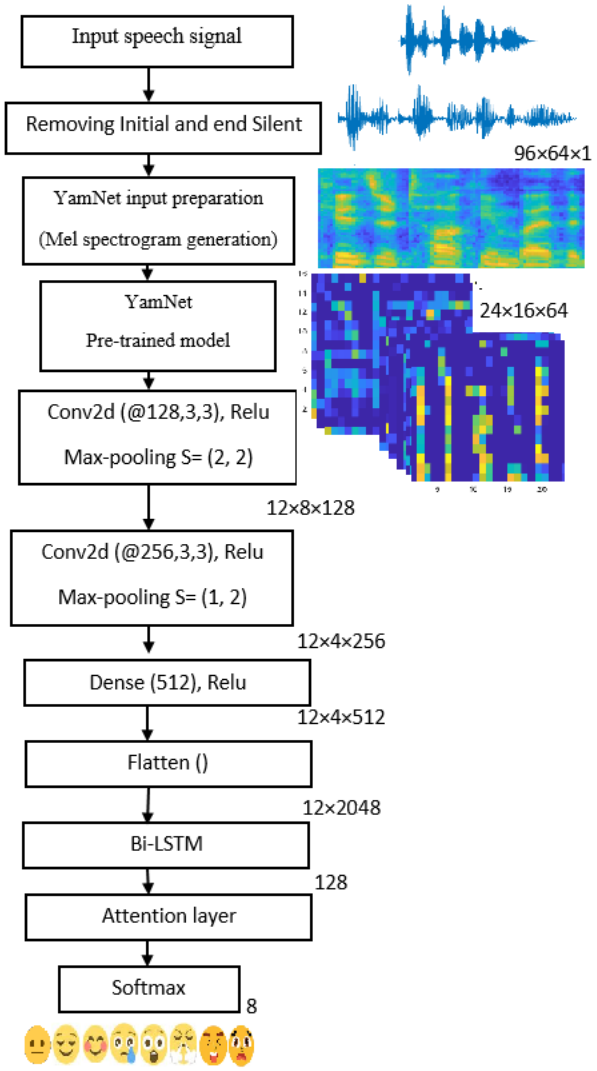

Figure 1. Proposed framework for speech-based emotion recognition.

In [31], after examining the strengths and weaknesses of the use of only audio or video data in emotion recognition, a multi-modal approach was presented based on the combination of these data in emotion recognition. In this study, for each video, four emotions including sadness, anger, happiness, and neutral were considered. The

emotions were recognized by considering the motions of some landmarks on the face combined with the features extracted from the speech signals. In this work, it was revealed that the method based on the video information gave more accurate results compared to the audio data, and the combination of the two increased the accuracy of the recognition. In [32], the 1D CNN network was used to extract the speech feature from the speech signals, and the ResNet model was used to extract the facial features. The extracted features were then joined together, and finally, using a 2-layer LSTM, recognition was done.

3. Proposed Method

In this research work, the main goal is to present a model that can be used to recognize the person's feelings based on the information of the person's voice and face in the form of a video. The nature of these two sources is different. Accordingly, the mechanisms that can extract features from these two are also different. So, at first, we considered this problem as two sub-problems to examine the configuration of the models that could extract suitable features from each source for emotion recognition. Then according to these sub-models, the final audio-visual model is proposed.

In the following, first, two approaches based on the use of only speech or facial information for emotion recognition are examined. Then a multi-modal model is proposed.

3.1. Speech-based emotion recognition

Here, the aim is to use speech information to recognize emotion. A pre-trained network called YAMNet has been used to extract the low-level features from the speech signal. This model consists of 27 convolutional blocks followed by an average pooling, a fully connected, and a softmax layer to classify the sound signals into more than 521 categories. Each block is a convolution layer followed by the relu activation function and batch normalization. It has been trained with a large amount of data. Here, the feature maps of the third block are used as the low-level features for further processing.

Before using the YAMNet model, the initial and end silent parts of the speech signal are eliminated by using a voice activity detection (VAD) method. The YAMNet model expects a $96 \times 64 \times 1 \times n_s$ size Mel spectrogram as the input. Here, 96, 64, and n_s are the number of 25ms frames, the number of Mel bands, and the number of spectrograms, respectively. n_s depends on the length of the speech signal, and the percentage of the frames overlap. The generated spectrograms are fed into the YAMNet model, and then the output of the "activation-3" layer is used as the extracted low-level features to recognize emotion. The output of this layer is a tensor with dimensions of $24 \times 16 \times 64 \times n_s$. In the training phase, each spectrogram is considered as a sample with a label that is the same as the speech label. The proposed model for emotion recognition includes two convolutional layers followed by a Bi-LSTM with an attention layer and finally a softmax layer. The proposed framework is presented in Figure 1.

3.2 Facial-based emotion recognition

After extracting the video frames, Multi-Task Cascaded Convolutional Neural Network (MTCNN) is used to detect faces in each frame. The extracted faces are resized to 60 × 60 pixels. In order to minimize the redundant information and reduce the time and space complexity, an optimal selection of faces as the key faces has to be performed. For this purpose, consider a sliding window of length $2r + 1$ faces initialized at position $i = 0$. From this window, the face with the least distance from the other faces is selected as a key face. Then the window shifts by r frames. This process is repeated until the end of the video. All the selected faces are considered for training. The label of each video is assigned to all its key faces. The block diagram of the data generation is depicted in Figure 2.a. To detect emotion according to the facial information, the model presented in Figure 2.b has been used. This model includes 6 convolutional layers and 3 Max-pooling layers. For testing, the extracted key faces belonging to the test video are applied to the trained model, and finally to make the final decision the voting mechanism is utilized.

3.3. Multi-modal emotion recognition

In this section, the proposed multi-modal emotion recognizer is presented. Here, the model-level combination approach is used to fuse information from the two sources. For this aim, the proposed models in Figures 1 and 2.b are combined, as shown in Figure 3. Generating the training and testing samples is done as follows.

For each video triples (S_i, F_i, L_i) are generated. S_i is a feature vector extracted from each speech spectrogram, F_i is a key face, and L_i is the label. In the previous section, we saw that depending on the speech length, n_s consecutive spectrograms are extracted from the speech signal. Also for each video, the number of selected key faces is n_f. According to the length of the videos, the average values of n_s and n_f are 3 and 25, respectively. Depending on the value of these two parameters

in total, up to $n_f \times n_s$ triple samples can be generated for each video. If we want the spectrograms and the faces to be properly synchronized in terms of time, for each spectrogram S_p $(p \epsilon \{1,2,\ldots, n_s\})$ 5 faces are randomly selected from the range $((p-1)*m+1: p*m)$, in which $m = \lfloor n_f / n_s \rfloor$. In this way, for each video with n_s spectrograms and n_f faces, $5 \times n_s$ samples are generated with a label equal to the label of the whole video.

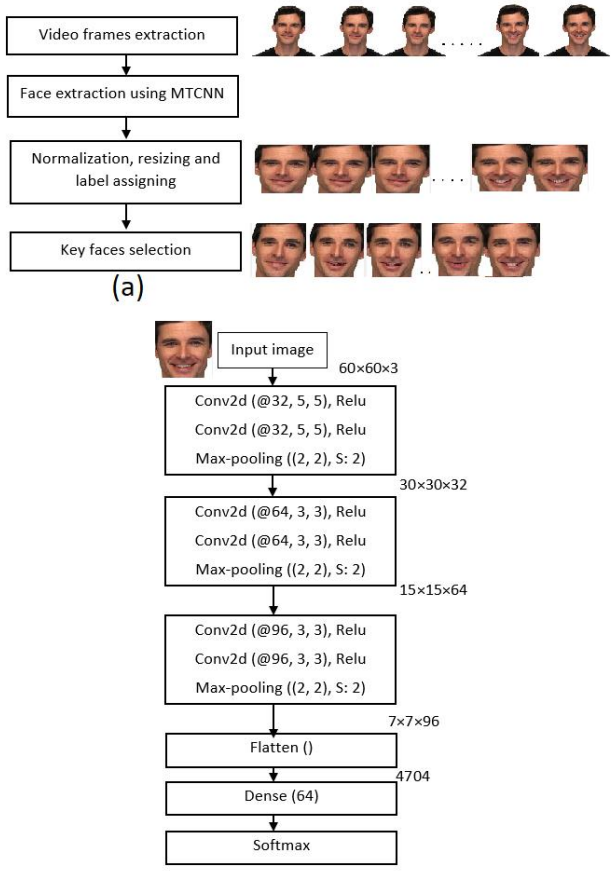

Figure 2. (a) Generating training and test samples. (b) Proposed architecture for facial-based emotion recognition.

4. Experimental results

In this section, the models presented in the previous section are evaluated. All implementations have been performed in the Python environment. In all experiments, 18 actors were used for training and 6 actors were used for testing. Here, the results are evaluated utilizing a 4-fold cross-validation mechanism. The average accuracy obtained by folds is reported at the video level. In all experiments, categorical cross-entropy was used as a cost function. For optimization, the Adam optimization algorithm with a learning rate of 0.001 has been utilized.

Database: In this research work, the RAVDESS database has been used to evaluate the proposed approaches. In this database, the files are modally divided into three categories of audio-only, video-only, and full AV. Moreover, vocally, there are two categories of speech and song in this database. Each file has a label representing an emotion. Eight categories are considered for senses including calm, happy, neutral, sad, angry, fearful, disgusted, and surprised.

In this study, data belonging to the category of full AV and speech channels are used as the audio-visual data. The videos belong to 24 actors (12 women and 12 men). Each actor utters two sentences with different emotions in two levels of intensity (except neutral) in two repetitions. One of the important advantages of this database is the approximate balance of labels. However, it should be noted that due to the ignorance of the strong expression of the neutral sense, the number of samples belonging to this sense is slightly less than the rest of the senses. In terms of time, the length of the videos varies from 3 s to 5.5 s. Here, for a more accurate evaluation, the 4-fold-CV technique is used. In each fold, the data belonging to 6 actors are used for the test, and the data belonging to the rest of the actors are used for training phases.

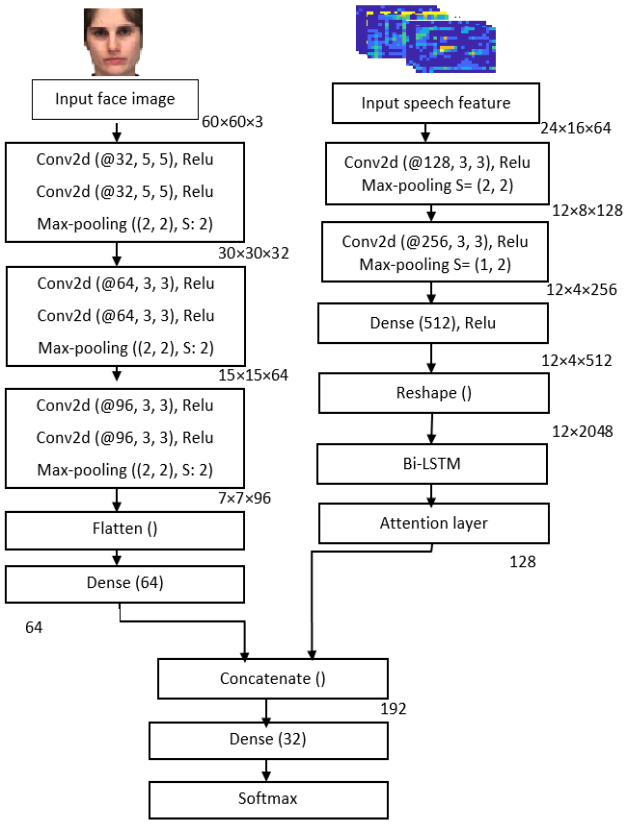

Figure 3. The proposed audio-visual emotion recognition.

One of the problems in comparing the contribution of the existing methods in this field is the lack of a common approach for splitting the data in training and test sets. Some studies did not

specify whether each person was used simultaneously in the test and training sets or not. Due to the repetition of the expression of each sentence in each sense by each actor, the selection of the training and test set using the conventional splitting mechanism leads to a relatively high difference in the evaluation parameters. Therefore, at the end of this section, a comparison is made with studies in which the method of selecting the training set is similar to our work. In the following, the details of the implementation of each approach and the obtained results are studied on the RAVDESS data.

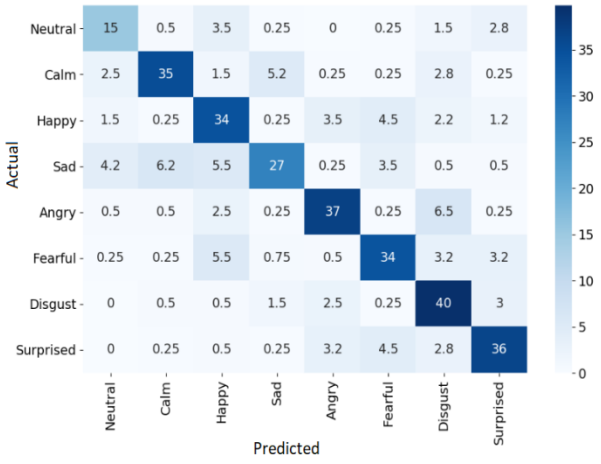

Figure 4. Confusion matrix for emotion recognition based on speech information.

4.1. Speech-based emotion detection

Here, to extract low-level features from the speech signal, a pre-trained network called YAMNet is used. For this aim, for each spectrogram, the output of layer "activation_3", which is a 24 × 16 × 64 tensor, has been considered. In the training phase, each spectrogram is considered a sample. For testing, all the spectrograms of a speech signal are fed into the proposed model. The final decision for the speech signal is obtained through majority voting. The average confusion matrix of this experiment from the folds of 4-CV is displayed in Figure 4. In this figure, the vertical axis is the actual label and the horizontal axis is the label detected by the model. The average accuracy achieved with this method is 72.22%.

The efficiency of the proposed model in recognizing some senses such as disgust and anger is relatively good. However, the error rate of sad recognition is the highest. This feeling is confused with calm, happy and neutral emotions.

4.2. Facial-based emotion recognition

First, the frames are extracted from each video. Each frame is an image with dimensions 720*1280*3. Using MTCNN, faces are extracted from each frame and then the extracted faces are converted to 60×60 images. Depending on the length of the video, which is 3 to 5.5 seconds, approximately 90 to 160 frames are extracted from each video. By considering shifting 9-frame windows, the number of selected key faces for each video is approximately 25 to 35. By assigning the label of the video to these faces, sufficient samples are produced for training. In the test phase, the key faces for a sample video are fed into the model, and then the final result is obtained using the majority voting mechanism. The average confusion matrix for this experiment for the 4 folds is shown in Figure 5. The average accuracy achieved with this method is 67.85%. The experiments show that the speech-based emotion recognizer achieves more accurate results in comparison with the facial-based emotion recognizer.

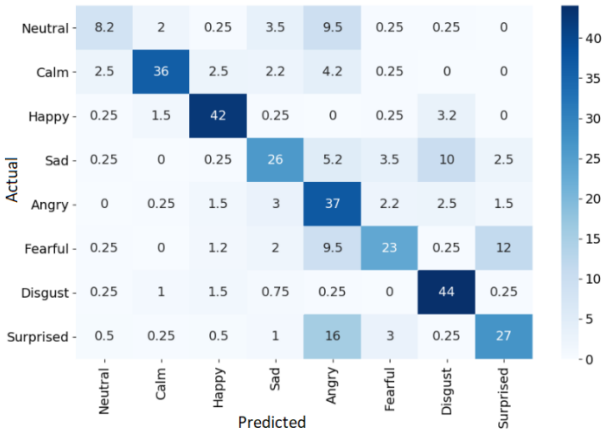

Figure 5. Confusion matrix for emotion recognition based on facial information.

4.3. Audio-visual emotion recognizer

Here, the proposed multi-modal framework depicted in Figure 3 is validated using the used database. The average confusion matrix for this experiment for the 4 folds is shown in Figure 6. The average accuracy achieved with this method is 81.04%. It can be seen that considering image information along with speech has made it possible to better recognize emotions.

As previously mentioned, one of the problems in comparing the performance of different algorithms is how to choose the training and testing set. Considering this problem, a comparison between different methods is presented in Table 1.

This table compares the accuracy of the proposed approaches (speech-based, facial-based, and audio-visual-based emotion recognition) and the results reported in some recent works on the RAVDESS database.

For the speech-based approach, the results reported in [1, 10, 32] are reported. Ancilin and Milton have proposed a modification to the extraction of Mel frequency cepstral coefficients [1]. By considering the modified coefficients along with some other spectral features as the feature vector, they utilized a multi-class support vector machine to classify the emotion of the speech signal. In [10], a 1D convolutional neural network has been proposed to classify the speech emotion. The input of the model is a 193-dimensional vector consisting MFCCs, Mel scaled Scalogram, Chromagram, and Spectral contrast features. In [33], the proposed model is comprised of two BLSTM layers followed by a 1D Conv-capsule layer. The model's input is a 70-dimensional vector consisting of the fundamental frequency, 23-dimensional MFCC, the 23-dimensional delta MFCC, and the 23-dimensional delta-delta MFCC coefficients.

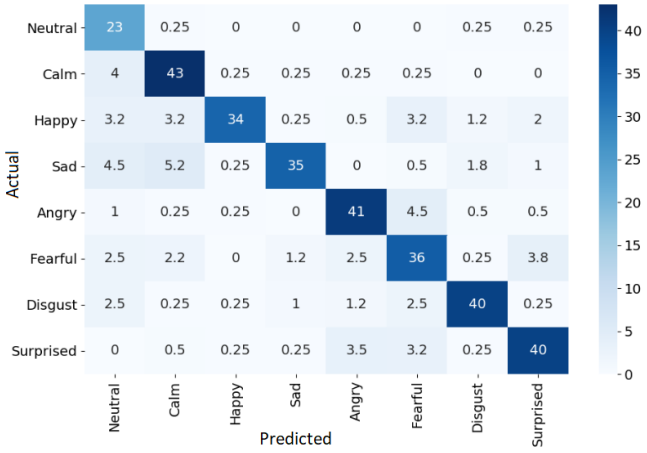

Figure 6. Confusion matrix for emotion recognition based on speech and facial information.

Table 1. Comparison between accuracy of proposed method and reported results in recent works on RAVDESS database. Here, results of all three approaches (speech-based, facial-based, and audio-visual approaches) are reported separately.

	Method	Accuracy (%)
Speech-based Emotion recognition	[33]	61.8
	[1]	64.31
	[10]	71.61
	P. M.	72.22
Facial-based Emotion recognition	[34]	57.08
	P. M.	67.85
Multimodal Emotion recognition	[34]	80.08
	P. M.	**81.04**

For the facial and audio-visual approaches, the results reported in [34] are considered. For facial emotion recognition, a pre-trained STN (Spatial Transformer Network) model has been used to extract the facial features. Then a two-layer bi-LSTM model has been proposed for emotion recognition. Moreover, in this work using two pre-trained models the extracted embeddings from the speech-based model and facial-based model are concatenated and fed into a two-layer perceptron to recognize emotion.

By studying Table 1, it can be observed that the proposed two sub-systems and the final audio-visual framework have a higher accuracy compared to some similar works done in these fields.

5. Conclusion

In this work, the aim was to recognize emotion based on a video of a person's face. Due to the different nature of audio and image signals, at first, the use of only facial or speech information in emotion recognition was examined. The aim was to determine the configuration of the sub-systems of the final system. For speech-based emotion recognition, a pre-trained network called YAMNet was used to extract the speech features. The extracted feature was fed into a CNN followed by a bi-LSTM with an attention mechanism. For facial-based emotion recognition, a deep CNN-based model was proposed. Finally, an end-to-end framework based on the fusion of these two modalities was proposed to recognize the emotion. The RAVDESS database was used to evaluate the proposed framework. The results showed that a combination of the face and speech information improved the performance of the emotion recognizer. Also in the comparison of the accuracy obtained by the proposed method with similar works of the recent years, its improvement was observed.

References

[1] J. Ancilin, and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl Acoust*, Vol. 179, pp. 108046, 2021.

[2] Y.D. Chavhan, B. S. Yelure, and K. N. Tayade, "Speech emotion recognition using RBF kernel of LIBSVM", *2nd international conference on electronics and communication systems (ICECS)*, pp. 1132-1135, 2015.

[3] A. Chamoli, A. Semwal, and N. Saikia, "Detection of emotion in analysis of speech using linear predictive coding techniques (LPC)", *In 2017 International Conference on Inventive Systems and Control (ICISC)*, pp. 1-4, 2017.

[4] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate", *International Journal of Speech Technology*, Vol. 23(1), pp. 45-55, 2020.

[5] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, and R. K. Muthu, "Speech emotion recognition using support vector machine", arXiv preprint arXiv: 2002.07590, 2020.

[6] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech", *Knowl. Based Syst.*, Vol. 184, pp.104886, 2019.

[7]R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M.H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", *IEEE Access*, Vol. 7, pp.117327-117345, 2019.

[8] H. Meng, T. Yan, F. Yuan, H. and Wei, "Speech emotion recognition from 3D log-Mel spectrograms with deep learning network", *IEEE Access*, Vol. 7, pp.125868-125881, 2019.

[9] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention-based fully convolutional network for speech emotion recognition", *In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* pp. 1771-1775, 2018.

[10] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomed. Signal Process. Control*, Vol. 59,pp. 101894, 2020.

[11] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", *Biomed. Signal Process. Control*, Vol. 47, pp. 312-323, 2019.

[12] P. Tzirakis, J. Zhang, and B. W. Schuller, " End-to-end speech emotion recognition using deep neural networks", *In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5089-5093, 2018, IEEE.

[13] K. Aghajani and I. Esmaili Paeen Afrakoti, " Speech emotion recognition using scalogram-based deep structure", *International Journal of Engineering*, Vol. 33(2), pp. 285-292, 2020.

[14] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition using Self-attention Mechanism and Multitask Learning", *In Interspeech* pp. 2803-2807, 2019.

[15] B. T. Nguyen, M. H. Trinh, T. V. Phan, and H. D. Nguyen, "An efficient real-time emotion detection using camera and facial landmarks", *In 2017 seventh international conference on information science and technology (ICIST)*, pp. 251-255, IEEE, 2017.

[16] E. Bagheri, P. G. Esteban, H. L. Cao, A. D. Beir, D. Lefeber, and B. Vanderborght, "An autonomous cognitive empathy model responsive to users' facial emotion expressions", *ACM Transactions on Interactive Intelligent Systems* (TIIS), Vol. 10(3), pp. 1-23, 2020.

[17] S. H. Wang, P. Phillips, Z. C. Dong, and Y. D. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm", *Neurocomputing*, 272, pp. 668-676, 2018.

[18] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)", *SN Applied Sciences*, Vol. 2(3), pp. 1-8, 2020.

[19] M. M. T. Zadeh, M. Imani, and B. Majidi, "Fast facial emotion recognition using convolutional neural networks and Gabor filters", *In 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pp. 577-581, 2019.

[20] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013", *In Advances in hybridization of intelligent methods*, pp. 1-16, Springer, Cham, 2018.

[21] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)", *SN Applied Sciences*, Vol. 2(3), pp. 1-8, 2020.

[22] I. Lasri, A. R. Solh, and M. El Belkacemi, " Facial emotion recognition of students using convolutional neural network", *In 2019 third international conference on intelligent computing in data sciences (ICDS)*, pp. 1-6, IEEE, 2019.

[23] M. R. Fallahzadeh, F. Farokhi, A. Harimi, and R. Sabbaghi-Nadooshan. "Facial Expression Recognition based on Image Gradient and Deep Convolutional Neural Network." *Journal of AI and Data Mining* , Vol. 9, pp. 259-268 2021.

[24] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild", *Mach. Vis. Appl.*, Vol. 30(5), pp. 975-985, 2019.

[25] M. C. Sun, S. H. Hsu, M. C. Yang, and J. H. Chien, "Context-aware cascade attention-based RNN for video emotion recognition", *In 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1-6, IEEE, 2018.

[26] M. Hu, H. Wang, X. Wang, J. Yang, and R. Wang, "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks", *J. Vis. Commun. Image Represent.*, Vol. 59, pp. 176-185, 2019.

[27] F. Rahdari, E. Rashedi, and M. Eftekhari, "A multimodal emotion recognition system using facial landmark analysis", *Iran. J. Sci. Technol. - Trans. Electr. Eng.*, Vol. 43(1), pp. 171-189, 2019.

[28] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal Correlated Network for emotion recognition in speech", *Vis. Inform.*, Vol. 3(3), pp. 150-155, 2019.

[29] K. S. Song, Y. H. Nho, J. H. Seo, and D. S. Kwon, "Decision-level fusion method for emotion recognition

using multimodal emotion recognition information", *In 2018 15th International Conference on Ubiquitous Robots (UR)*, pp. 472-476, IEEE, 2018.

[30] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition", *arXiv preprint* arXiv: 1907.03196, 2019.

[31] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information", *In Proceedings of the 6th international conference on Multimodal interfaces* , pp. 205-211, 2004.

[32] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks", *IEEE J. Sel. Top. Signal Process.*, Vol. 11(8), pp. 1301-1309, 2017.

[33] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition", *In Proceedings of Interspeech*, pp. 1701-1705, 2019 ISCA.

[34] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS Dataset using Transfer Learning", *Sensors*, Vol. 21(22), p. 7665, 2021.

تشخیص احساسات از روی چهره و صدا بر اساس یک شبکه عصبی کانولوشنال عمیق

خدیجه آقاجانی*

گروه مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه مازندران، بابلسر، ایران.

ارسال ۲۰۲۲/۰۴/۰۹؛ بازنگری ۲۰۲۲/۰۷/۲۵؛ پذیرش ۲۰۲۲/۰۸/۲۵

چکیده:

تشخیص احساسات کاربردهای متعددی در زمینه های مختلف از جمله تعامل انسان و کامپیوتر دارد. در سال‌های اخیر روش‌های مختلفی برای شناخت احساسات با استفاده از اطلاعات چهره یا گفتار ارائه شده است، در حالی که تلفیق این دو در تشخیص احساسات کمتر مورد توجه قرار گرفته است. در این اثر ابتدا استفاده از اطلاعات چهره یا گفتار در تشخیص احساسات مورد بررسی قرار گرفته است. برای تشخیص احساسات از طریق گفتار، از یک شبکه از پیش آموزش دیده به نام YAMNet برای استخراج ویژگی ها استفاده می‌شود. پس از عبور از یک شبکه عصبی کانولوشن (CNN)، ویژگی‌های استخراج‌شده به یک bi-LSTM با مکانیزم attention برای انجام تشخیص وارد می‌شوند. برای تشخیص احساسات از طریق اطلاعات چهره، یک مدل عمیق مبتنی بر CNN پیشنهاد شده است. در نهایت، پس از بررسی این دو رویکرد، یک چارچوب تشخیص احساسات بر اساس ادغام این دو مدل پیشنهاد شده است. پایگاه داده شنیداری- بصری رایرسون از گفتار و آهنگ حاوی احساس (RAVDESS) حاوی ویدئوهای ضبط شده از ۲۴ بازیگر (۱۲ مرد و ۱۲ زن) با ۸ دسته برای ارزیابی مدل پیشنهادی استفاده شده است. نتایج پیاده سازی نشان می‌دهد که ترکیبی از اطلاعات چهره و گفتار باعث بهبود عملکرد تشخیص دهنده احساسات می‌شود.

کلمات کلیدی: تشخیص عواطف گفتار، تشخیص احساسات چهره، یادگیری عمیق، انتقال یادگیری.