Signal, Image and Video Processing (2022) 16:1281–1288
https://doi.org/10.1007/s11760-021-02079-x

ORIGINAL PAPER

Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning

P. Sreevidya[1] · S. Veni[1] · O. V. Ramana Murthy[2]

Received: 13 March 2021 / Revised: 16 October 2021 / Accepted: 1 November 2021 / Published online: 18 January 2022

Abstract

The objective of the work is to develop an automated emotion recognition system specifically targeted to elderly people. A multi-modal system is developed which has integrated information from audio and video modalities. The database selected for experiments is ElderReact, which contains 1323 video clips of 3 to 8 s duration of people above the age of 50. Here, all the six available emotions Disgust, Anger, Fear, Happiness, Sadness and Surprise are considered. In order to develop an automated emotion recognition system for aged adults, we attempted different modeling techniques. Features are extracted, and neural network models with backpropagation are attempted for developing the models. Further, for the raw video model, transfer learning from pretrained networks is attempted. Convolutional neural network and long short-time memory-based models were taken by maintaining the continuity in time between the frames while capturing the emotions. For the audio model, cross-model transfer learning is applied. Both the models are combined by fusion of intermediate layers. The layers are selected through a grid-based search algorithm. The accuracy and F1-score show that the proposed approach is outperforming the state-of-the-art results. Classification of all the images shows a minimum relative improvement of 6.5% for happiness to a maximum of 46% increase for sadness over the baseline results.

Keywords Emotion classification · Cross-model transfer learning · CNN · Fusion

1 Introduction

Human affective computation and social cognition are increasingly becoming an important research area. The emotion recognition is an integral part of cognition and non-verbal communication. Especially in post-covid-19 digital scenario, as our social lives are becoming more and more automated, algorithms are deployed for automating the associated tasks in the fields like healthcare, education, advertisement, automated job interviews, interactive voice assistants, human assistive robotics, etc.

The emotions can be identified through different modalities like audio, video, image, gestures/poses, text or from physiological signals. Specifically, multimedia signals are noninvasive, information-rich medium which can be explored using the said methods. Handcrafted features or deep features can be used for machine learning-based methods for emotion classification [1]. Deep learning models which are trained on established datasets like ImageNet [3], CIFAR10 [2], COCO [4] database are available for image classifications, while there are audio models trained with datasets like Audioset. There are tailor-made multimodal-based datasets like IMEOCAP [5], EMOReact [6], AffectNet [7] and EMOTIC [8] for emotion recognition.

When the problem of emotion recognition is addressed, the research results shows that there is a marked deviation in display of emotions in normal people and elderly people [9]. According to the meta-analysis by Hayes et al. [10], the older adults less accurately identify facial expressions of sadness, fear and anger compared to younger people. The effect is lesser in surprise and happiness, and disgust is identified

✉ P. Sreevidya
sreevidyapmenon@gmail.com

S. Veni
s_veni@cb.amrita.edu

O. V. Ramana Murthy
ovr_murthy@cb.amrita.edu

[1] Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

[2] Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

equally as that of the young. It implies that custom-made automated systems are to be developed for addressing the emotional-level requirements of aged adults. The said findings were the motivation behind the work.

The proposed work attempts to address the following issues:

1. To identify suitable methods to develop audio and video models for emotion recognition in aged individuals.
2. To suggest a suitable multimodal fusion technique for emotion recognition system in aged adults.
3. To compare the performance of the results of the various experiments conducted.

The proposed model is a fusion of audio and video modalities. The audio models are developed using cross-modal transfer learning techniques in addition to a feature-based approach. We used pretrained Inception Nets which are trained on ImageNet to transfer knowledge on audio spectrograms. The video signals are sampled, and a convolutional neural network–long short-time memory (CNN-LSTM) network is used for developing the model. Further, the fusion between these two modalities are performed. Experiments were conducted to analyze the significance of customized datasets for the people of age group above 60.

Various sections of this paper are arranged as follows. In Sect. 2, the state of the art in emotion recognition and multimodal approaches is discussed. In Sect. 3, the proposed framework is presented, and Sect. 4 discusses the different experiments conducted by us. The dataset for emotion recognition in elderly people are also discussed here. The analysis of the results is given in section, which is followed by conclusion and future work.

2 Related work

Here, the state-of-the-art techniques for the emotion recognition in the wild are discussed. Since the work addresses the problem of identifying the emotions in elderly people, the available datasets for this purpose is investigated. The ElderReact [11] dataset proposes an emotion reaction video dataset which has only elderly adults as actors in it. FACES and Lifespan are some of the datasets that contain emotion annotations for elderly people.

For emotion recognition from audio signals, [12] suggested a cross-modal transfer learning framework [13], which transfers knowledge from AlexNet, which is trained on ImageNet. Thus, it can be concluded that large-scale image classification benchmarks can help audio classification.

Similarly in [14], spectrogram-based CNN models for speech emotion recognition are implemented on Berlin Dataset [15]. According to [16,17], the techniques in neural style transfer [18] can be applied for spectrograms as it is the two-dimensional representations of audio frequencies with respect to time. In [19], Poorna et al. applied a multistage learning network for classifying speech emotions in Arabic speaking community. Kown et al. [20] proposed an artificial intelligence-assisted deep stride convolution neural network architecture using the plain nets strategy to learn discriminative and salient features from spectrogram of speech signals that are enhanced in prior steps to perform better.

Boateng et al. [21] applied a transfer learning technique using YAMNet CNN for classifying emotions in elder individuals. YAMNet is a pretrained network with 1024 embeddings that are based on MobileNet.

Experiments on FER-13 and AffectNet were done by [22] to show the combined effect of handcrafted and deep features. In Emotiw2019, Zhou and his team [23] taken a feature fusion strategy for classification of emotions.

Zadeh et al. [24] introduced a multimodal dictionary to understand the interaction between facial gestures and spoken words for expressing sentiment, which is basically taking positive, negative and neutral expressions in a better manner. In [25], late fusion network was used for sentiment classification using MOUD dataset, and [26] used fusion of text and speech for emotion classification on eNTERFACE dataset.

The multimodal clues from videos were taken into different modalities and explored in [27,28]. The hybrid deep learning framework introduced through this work include static spatial appearance information, motion patterns within a short-time window, audio information, as well as long-range temporal dynamics. Three CNN models were operating on static frames, and temporal relations were extracted through two LSTM networks.

Hunag et al. [29] used a transformer model and a LSTM model for classifying the audio and video modalities. It had a multi-head attention mechanism by which multimodal emotional intermediate representations from common semantic feature space were used after encoding audio and visual modalities.

3 Methodology

The proposed frame work incorporates a multimodal interaction between audio and video modalities. The audio model has been developed by combining the spectrogram features as well as the handcrafted features. In the video modality, CNN-based networks are incorporated to learn the information from videos. We also tried a CNN network and LSTM network for modeling the raw video data. The input data were given to the network after performing necessary preprocessing steps. Further, feature-level and hybrid-type approaches are adapted to develop the final model as shown in Fig. 1.

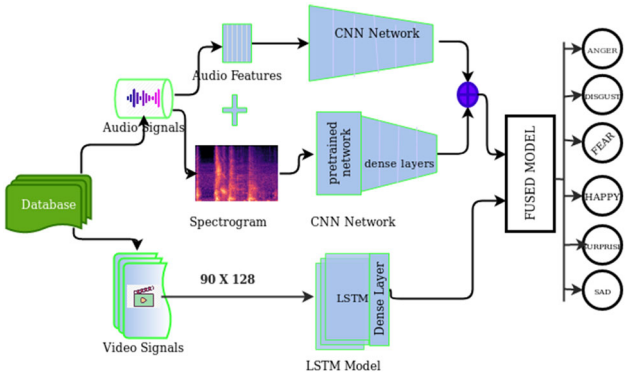

Fig. 1 Structure of the proposed model

3.1 Dataset

In order to classify the emotions in elderly people, a major limitation is the lack of suitable datasets conducting the experiments. The ElderReact, a dataset which has description of emotion of old age people above fifty only, is selected for the experimentation purpose. This is one of the largest dataset available for emotion recognition in aged individuals. It contains 1323 video clips of from 46 elderly people, which are divided into 615 clips for training, 353 clips for testing and 355 clips for validations [11]. These videos are collected from YouTube channels.

The dataset was annotated manually for six basic emotions along with valance and gender using Amazon Mechanical Turk. The emotions considered in this work are anger, fear, disgust, happiness, surprise and sadness along with valence and gender information. The cropped samples of faces of aged people from the database are shown in Fig. 2. The annotations in the train, validation and test segments are distributed as shown in Fig. 3. The two other datasets considered here for comparison purposes are EmoReact and RAVDESS [30].

3.2 Audio model

The emotion classifications based on the audio signals are carried out by both 1-D and 2-D approaches. At first, the audio features like prosody, spectral coefficients and voice quality features like tenseness, creakiness, etc., are extracted. There are 72 selected features. The features are extracted using the open-source tool, COVARAP [31] with frame length 10 ms. The model based on the handcrafted features is developed by forming a deep neural network with two 1-D CNN layers and two dense layers. The number of filters are 64 and 32, respectively. The dense layers have 256 and 128 neurons in it. Mean square error is monitored for convergence. The network was optimized with Adam optimizer with a learning rate of 0.001.

Fig. 2 Sample elder images in the dataset

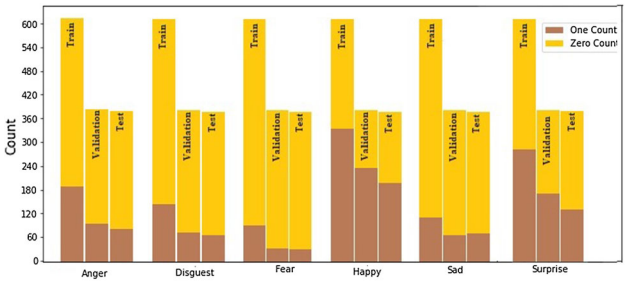

Fig. 3 Distribution of the presence of six emotions

Further, a spectrogram model was developed. The spectrogram is a 2-D representation of the audio signal [32]. It appertains instantaneous frequency information of the audio feeds. The amplitudes are mapped into the intensity levels. In order to generate spectrograms, audio files are segmented uniformly. Each audio sample is sampled at 44.1 KHz frequency. The images constituted patches of 20 ms with an overlap of 75%. The short-time Fourier transform (STFT) was applied on the original signal. The Hanning window of length 10 ms was selected which hopes on the spectrum for adjusting the weighing factor. The spectrogram images obtained was pseudo-color-mapped. These images were standardized, before applying on the model proposed. The selection of the window type decides the sidelobe suppression, and the hop size determines the time–frequency smearing. By using Hanning window, we could ensure that there is smooth transition from main lobes to sidelobes, and there is no discontinuities due to windowing. The 25% hop-size was suitable for improving the time resolution.

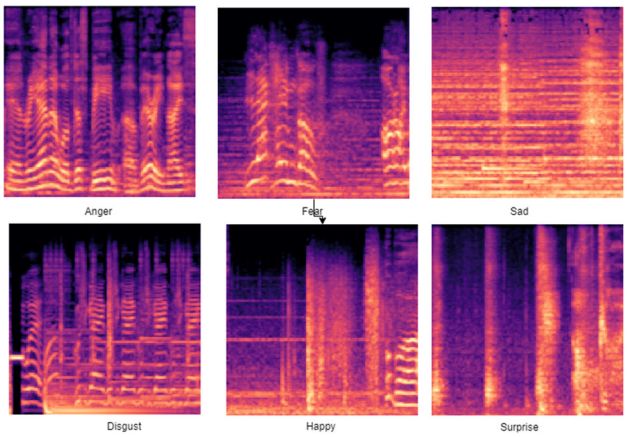

Fig. 4 Spectrogram obtained for different emotions

A cross-model transfer learning technique was applied on the spectrogram images as shown in Fig. 4. The idea here was to utilize the rich set of weights of the pretrained models. The pretrained models trained on ImageNet are retrained with training data of the dataset for the learning purpose. The Inception-V2 is identified as the suitable pretrained network [33]. This is because of the separable convolutions in the inception units. The filterbanks in the network are making the modules wider than just deeper, and there is an internal regularization that prevents overfitting.

3.3 Video model

For the video model, at first, feature extraction was done from the visual modality through OPENFACE [34], and face only frames were selected from the extracted frames. The selection of 178 features was done based on gaze, head pose detection, facial action units and non-rigid shape parameters as these features are the prominent visual indicators of the emotion that the participant is displaying [11]. A CNN model was developed and trained for classifying the emotions from these handcrafted features. There is batch normalization layer [35] incorporated which will prevent the model from overfitting by giving internal co-variance shift among minibatches.

The raw video data were sampled and frames with only face images were selected for the purpose. During the preprocessing steps, the successful face recognition algorithms were executed to get the cropped face only images [36]. The multi-task cascaded convolution neural network (MTCNN) algorithm was applied for selecting face only frames from the video data. As a result, there are 90 face only images of size 160×160 is stacked into a single folder. The CNN network was developed. The model is pretrained with FER-2013 database. This dataset has a versatile set of images which has complex and subtle emotions. This dataset contains 35,685 examples of 48×48 pixel gray-scale images of faces. The image array is passed through this network, and 128 embeddings were taken for each emotion. Subsequently, this features are passed through another CNN network for classification of the emotions. The diagrammatic representation of the process is shown in Fig. 5.

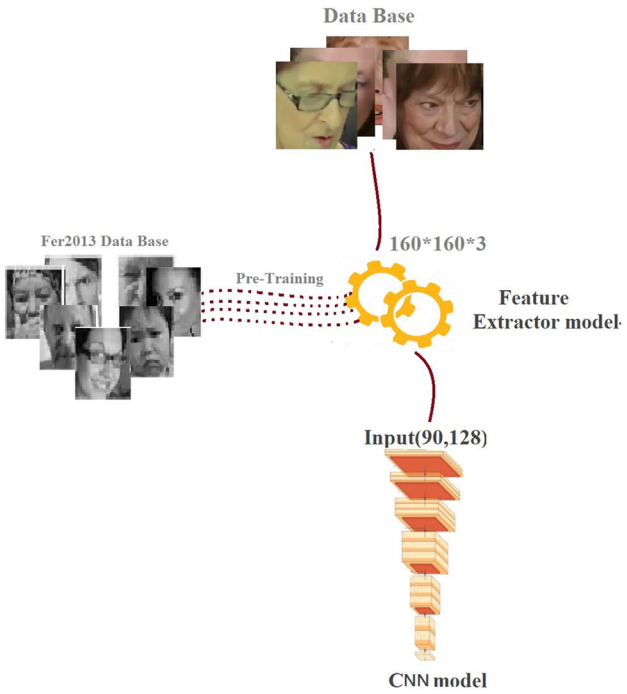

Fig. 5 Diagrammatic representation of video model

3.4 Combined model

Further, for classifying the emotions, both visual and audio modalities are important. Therefore, fusion methods are attempted. The embedding from experimentally selected intermediate layers is taken for both audio and video modalities. The layers are selected through a grid-based search algorithm.

The intermediate layers are fused as in

$$f_{\text{fusion}} = opt_{i,j}(x_{\text{video}}, x_{\text{audio}})$$

where $[f_{\text{fusion}}]$ indicates the fusion layer, $[x_{\text{video}}]$ indicates the $[i\text{th}]$ layer in video model, and $[x_{\text{audio}}]$ indicates the $[j\text{th}]$ layer in audio model.

The embeddings are collected from the selected layers and applied as the input to the combined model. The results shows that features from both the modalities are contributing to the problem under consideration. There were 512 features from the video model, and 384 features were from the sound model.

Table 1 Accuracy and F1-score of extracted feature-based sound model and visual model

Emotion	Accuracy	F1-score	Accuracy	F1-score
Anger	55.4	57.14	57.4	61.9
Disgust	57.0	51.0	63.0	67.0
Fear	65.3	67.7	58.0	65.1
Happy	65.9	70.0	59.0	67.0
Sad	57.9	57.7	53.9	58.1
Surprise	55.4	61.1	61.3	66.0

Table 2 Accuracy and F1-score of the extracted feature of combined visual and audio model

Emotion	Accuracy	F1-score
Anger	61.4	61.7
Disgust	65.0	61.5
Fear	65.4	73.0
Happy	66.9	76.8
Sad	56.0	65.0
Surprise	65.3	70.1

4 Experiments

Figure 3 shows that the dataset is imbalanced except for the emotions happy and surprise. So at the preprocessing stage, the dataset is preprocessed with some resampling and subsampling methods. The dataset is normalized between the minimum and maximum value of the available feature values.

4.1 Feature-based models

The models are developed using the extracted features, and the results are tabulated in Table 1. The accuracy and F1-score are both presented in the said table. It shows good F1-score above 70% for happiness. Further, for the feature-based model in visual environment, again a 1-D model has been developed. The 1-D CNN layers are selected with the number of filters 256 and 128, respectively. Each layer was followed by drop out layers of 0.2 and 0.5, respectively. No regularization parameter was used except the batch normalization. Here again, Adam optimizer is used with a learning rate 0f 0.001. The results are tabulated in Table 2. It is observed that while developing the model, increasing the depth of the model resulted in overfitting.

The features are then concatenated, and fusion of two modalities was tried. The embeddings from intermediate layers are taken and fused together. The fused model has three dense layers with decreasing number of neurons applied. Each layer was followed by dropouts carefully chosen to avoid overfitting. Batch normalization was applied to normalize the minibatches before applying the classifier. The softmax activation function was used in the last layer. The hyperparameter tuning was done using Adam optimizer, with a learning rate of 0.001. The early stopping technique was applied with a patience level of 10. The results are tabulated in Table 3.

It can be observed that there is a significant improvement in all results except for disgust. This is especially true while considering the F1-score. The F1-score of happiness was increased by 6.8%, and the F1-score of surprise was found to increase by 4%. The F1-score of fear class was improved by 5.3%. It again reemphasizes the requirement for customized models for emotion classification in aged people.

Table 3 State-of-the-art results on ElderReact dataset for emotion classification

Model	Angr	Disg	Fear	Hap	Sad	Surpr
Random	30	26	14	**51**	27	41
Naive Bayes	34	27	25	**56**	33	45
SVM	41	35	16	**70**	34	54
XGBOOST	43	36	17	**71**	35	14

Bold values indicate the best value

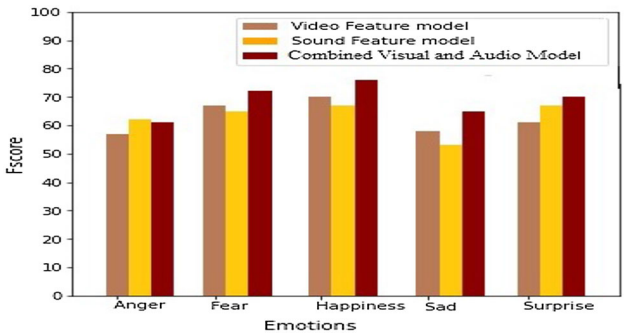

Fig. 6 Comparison of audio, video and fusion models

While comparing the results obtained with the state-of-the-art results, it can be concluded that the proposed model has outperformed the existing results. The state-of-the-art results are tabulated in Table 4. The algorithms like random forest, Naive Bayes, SVM and XGBOOST were applied on the dataset. It is found that except for happiness none of the classes could perform well using these algorithms. There was a strong consistency in the results for our proposed model. The comparison between the combined model and unimodels is shown in Fig. 6. It shows that the fusion model is better performing on classification of all images, and there is a slight decrease of 0.3%. in F1-score for anger.

Table 4 Accuracy and F1-score of raw video model

Emotion	Accuracy	F1-score
Anger	55.9	56.1
Disgust	57.0	61.2
Fear	62.4	66.1
Happy	73.7	77.1
Sad	56.1	64.1
Surprise	56.0	64.0

Table 5 Accuracy and F1-score of spectrogram model

Emotion	Accuracy	F1-score
Anger	50.7	62.1
Disgust	57.1	51.2
Fear	60.4	62.5
Happy	58.6	61.4
Sad	60.5	60.5
Surprise	62.0	61.0

Table 6 Accuracy and F1-score of combined model using CNN model as well as LSTM model

Emotion	Accuracy	F1-score	Accuracy	F1-Score
Anger	55.1	62.3	55.6	65,3
Disgust	51.2	67.1	60.5	66.7
Fear	57.0	64.1	60.5	70.0
Happy	54.9	69.1	66.5	76.0
Sad	54.1	67.1	57.8	67.0
Surprise	56.1	62.3	59.5	69.5

4.2 Raw video and audio models

4.2.1 Raw video model

The raw video models were developed by convolutional neural network-based approaches. The model has two convolutional neural network(CNN) layers followed by a flattening layer. The 2-D CNN layer selected has 32 filters with 3×3 filter size. Then, three more dense layers of neurons 512, 256 and 128, respectively, are added. There are dropouts of 0.1 and batch normalization for normalizing the features extracted. Instead of the rectified linear unit (ReLU) activation function, leaky ReLU was used as the activation function. The small negative slope in the activation function helped to incorporate the negative values also.

The input to this model is embeddings of size 90×128. These embeddings are extracted from a CNN model. The model was trained with FER-2013 database. Now, the model has initial weights learned from the database. The embeddings of the preprocessed image frames are predicted from this pretrained CNN model. The result obtained through this novel approach is tabulated in Table 5. The results show that the video model is giving the best result for happiness, and the failure to distinguish between disgust and anger is the reason for decrease in (56.1%) F1-score for anger.

4.2.2 Spectrogram model

The next experiment was conducted to develop a model based on the raw spectrogram images collected through librosa [37]. The cross-model transfer learning technique was adopted. It was experimentally determined to take the output from the 'mixed9' layer of the pretrained inception model. The layer which is selected is high-dimensional layer of 2048 embeddings with 2D settings. Three Conv2D layers with 512 filters of size 3×3 are applied further. Then, the layers were flattened and two more dense layers are added.

The model was optimized through hyperparameter tuning. Nadam [38] was found to give the best results. The learning rate selected was 0.00001. Table 6 shows the accuracy and F1-score obtained for the customized model for elder emotion recognition task.

4.2.3 Fusion model

Once the raw video model and the spectrogram model are developed, the next step was to observe the performance of the fusion model. The structure of the model is shown in Fig. 1. The embeddings are taken from both audio model and video models from intermediate layers. The layers are selected through a grid-based search. These embeddings are concatenated. Table 7 shows the accuracy and F1-score of the fusion model. The model selected has 1-D CNN layers followed by dense layers. One-dimensional CNN layers has 256 filters in it. Instead of the CNN model, an LSTM model was applied. It has two LSTM layers of 128 units, and a dropout of 0.1 is applied, and finally, a dense layer is added. The results were better than the CNN model and are given in Table 7.

4.3 Comparison of performances of the models

A comparison between spectrogram-based audio model and feature-based audio model is given in Fig. 7. It can be observed that the feature-based model is giving better performance than the spectrogram-based model for all the emotions with respect to F1-score. But accuracy is more for sad and surprise in spectrogram model. But for video model, F1-sore of disgust and anger has come out well than the feature-based model.

Table 7 Comparison of performance of spectrogram model on ElderReact and EmoReact dataset

Dataset	Angr	Disg	Hap	Surpr
ElderReact	62.1	51.2	**61.4**	61
EmoReact	17	10	**77**	64

Bold values indicate the best value

Table 8 Comparison of performance of feature-based model on ElderReact and RAVEDESS dataset

Model	Angr	Disg	Hap	Surpr
ElderReact	57.14	51.0	**70.0**	61.1
RAVEDESS	17	29.0	**67.0**	44.9

Bold values indicate the best value

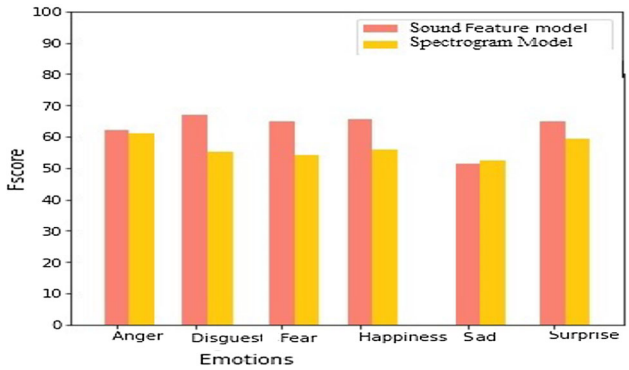

Fig. 7 Comparison of F-score between two audio models developed for classification of emotions

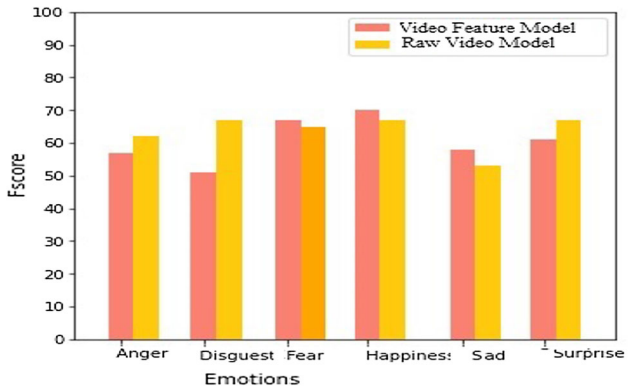

Fig. 8 Comparison of F1-score of video models for classification of emotions

In the next step, spectrogram model was applied on the EmoReact database. It is found that the results were comparable for positive emotions only. Then, we applied audio feature model on a generalized audio emotion classification model, which was trained on RAVDESS dataset. The same pattern could be observed in this case also. The results are tabulated in Tables 7 and 8.

Further, the video models are compared in Fig. 8. For anger and disgust, the model based on the deep feature is better, and for the emotions like fear, happiness, sad and surprise, first model is performing slightly better.

5 Conclusion

With the rapid developments in the field of machine intelligence and deep learning techniques, automated emotion recognition systems are getting developed. But there is the requirement of customized systems for emotion recognition based on age or sex. The work here proposes automated emotion classification in aged people above 60. Various unimodal and fusion modals are tried here. The feature-based fusion model is found to give the best results. Compared to the generalized datasets, the ElderReact dataset is giving better consistency in all results for the proposed models.

There has to be more databases tailor made for elder emotions so that more sophisticated systems can be developed. The results are also in accordance with the meta-analysis on the effect of age on expression of emotions.

References

1. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **16**(7), 64827–36 (2019). https://doi.org/10.1109/ACCESS.2019.2917266
2. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. **25**, 740–755 (2014)
4. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Lawrence Zitnick, C.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer, Cham (2014)
5. Busso, C., Bulut, M., Lee, C.C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335 (2008). https://doi.org/10.1007/s10579-008-9076-6
6. Nojavanasghari, B., Baltrušaitis, T., Hughes, C.E., Morency, L.P.: Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In: ACM International Conference on Multimodal Interaction, vol. 18, pp. 137–144 (2016). https://doi.org/10.1145/2993148.2993168
7. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2019). https://doi.org/10.1109/TAFFC.2017.2740923
8. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using EMOTIC dataset. IEEE Trans. Pattern Anal. Mach. Intell. **42**(11), 2755–2766 (2020). https://doi.org/10.1109/TPAMI.2019.2916866
9. Gonçalves, A.R., Fernandes, C., Pasion, R., Ferreira-Santos, F., Barbosa, F., Marques-Teixeira, J.: Effects of age on the identification of emotions in facial expressions: a meta-analysis. PeerJ (2018). https://doi.org/10.7717/peerj.5278

10. Hayes, G.S., McLennan, S.N., Henry, J.D., Phillips, L.H., Terrett, G., Rendell, P.G., Pelly, R.M., Labuschagne, I.: Task characteristics influence facial emotion recognition age-effects: a meta-analytic review. Psychol. Aging **35**(2), 295–315 (2020). https://doi.org/10.1037/pag0000441
11. Ma, K., Wang, X., Yang, X., Zhang, M., Girard, J.M., Morency, L.P.: ElderReact: a multimodal dataset for recognizing emotional response in aging adults. In: International Conference on Multimodal Interaction, pp. 349–357 (2019). https://doi.org/10.1145/3340555.3353747
12. Nagarajan, B., Oruganti, V.R.: Cross-domain transfer learning for complex emotion recognition. In: TENSYMP, pp. 649–653 (2019). https://doi.org/10.1109/TENSYMP46218.2019.8971023
13. Liang, P.P., Wu, P., Ziyin, L., Morency, L.P., Salakhutdinov, R.: Cross-modal generalization: learning in low resource modalities via meta-alignment, pp. 2012.02813 (2020)
14. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5 (2017). https://doi.org/10.1109/PlatCon.2017.7883728
15. Burkhardt, F., Paeschke, A., Rolfes, M.W., Sendlmeier, F., Weiss: A database of German emotional speech. In: Interspeech, pp. 1517–1520 (2005)
16. Gatys,L.A., Ecker, L.A., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
17. Li,Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer (2017). https://doi.org/10.24963/ijcai.2017/310
18. Verma, P., Smith, J.O.: Neural style transfer for audio spectrograms (2018)
19. Poorna, S.S., Nair, G.J.: Multistage classification scheme to enhance speech emotion recognition. Int. J. Speech Technol. **22**, 327–340 (2019). https://doi.org/10.1007/s10772-019-09605-w
20. Mustaqeem, Kwon S.: A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**(1), 183 (2020). https://doi.org/10.3390/s20010183
21. Boateng, G., Kowatsch, T.: Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning. In: International Conference on Multimodal Interaction, pp. 12–16 (2020). https://doi.org/10.1145/3395035.3425255
22. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **16**(7), 64827–36 (2019). https://doi.org/10.1109/ACCESS.2019.2917266
23. Hengshun, Z., et al.: Exploring emotion features and fusion strategies for audio-video emotion recognition. In: International Conference on Multimodal Interaction, pp. 562–566 (2019). https://doi.org/10.1145/3340555.3355713
24. Zadeh, A., et al.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016). https://doi.org/10.1109/MIS.2016.94
25. Sreevidya, P., Murthy, O.V., Veni, S.: Sentiment analysis by deep learning approaches. Telkomnika (2020). https://doi.org/10.12928/telkomnika.v18i2.13912
26. Bhaskar, Jasmine, Sruthi, K., Nedungadi, P.: Hybrid approach for emotion classification of audio conversation based on text and speech mining. Procedia Comput. Sci. **46**, 635–643 (2015). https://doi.org/10.1016/j.procs.2015.02.112
27. Jiang, Y.G., Wu, Z., Tang, J., Li, Z., Xue, X., Chang, S.F.: Modeling multimodal clues in a hybrid deep learning framework for video classification. IEEE Trans. Multimed. **20**(11), 3137–3147 (2018). https://doi.org/10.1109/TMM.2018.2823900
28. Jaouedi, N., Boujnah, N., Bouhlel, M.S.: A new hybrid deep learning model for human action recognition. J. King Saud Univ. Comput. Inf. Sci. **32**(4), 447–53 (2020). https://doi.org/10.1016/j.asoc.2015.08.025
29. Huang, J., Tao, J., Liu, B., Lian, Z., Niu, M.: Multimodal transformer fusion for continuous emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3507–3511 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053762
30. de Pinto, M.G., Polignano, M., Lops, P., Semeraro, G.: Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In: IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 1–5 (2020). https://doi.org/10.1109/EAIS48028.2020.9122698
31. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 960–964 (2014). https://doi.org/10.1109/ICASSP.2014.6853739
32. Lech, M., Stolar, M., Bolia, R., Skinner, M.: Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images. Adv. Sci. Technol. Eng. Syst. J **3**(4), 363–371 (2018)
33. Szegedy. et.al.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence, vol. 31, no. 1 (2017)
34. Baltrusaitis, T., Robinson, P., Morency, P.: OpenFace: an open source facial behavior analysis toolkit. In: WACV. IEEE Computer Society, pp. 1–10 (2016). https://doi.org/10.1109/WACV.2016.7477553
35. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
36. Boyko, N., Basystiuk, O., Shakhovska, N.: Performance evaluation and comparison of software for face recognition, based on Dlib and Opencv Library. In: IEEE Second International Conference on Data Stream Mining & Processing, pp. 478–482 (2018). https://doi.org/10.1109/DSMP.2018.8478556
37. Librosa development team. LibROSA. https://librosa.github.io/librosa (2019)
38. Dozat, T.: Incorporating Nesterov momentum into Adam. In: ICLR Workshop (1), 2013–2016