SeLF: A Deep Neural Network Based Multimodal Sequential Late Fusion Approach for Human Emotion Recognition

Anitha Modi(✉) and Priyanka Sharma

Department of Computer Science and Engineering,
Nirma University, Ahmedabad, India
{anitha.modi, priyanka.sharma}@nirmauni.ac.in

Abstract. Computer vision domain consists of algorithms and techniques to enhance computers with the ability to see and perceive. Human emotion recognition using computer vision is a challenging research area. Facial expression may not always give accurate judgment of emotion hence needs to be combined with other modalities such as voice, text and physiological signals. Several fusion approaches such as direct, early and late were introduced but the problem still persists. This paper focuses on deep neural network (NN) based sequential late fusion approach to identify emotions from various available modalities. Modalities are integrated into the system sequentially at the decision level. A deep CNN was trained to identify face emotions. Short videos were analyzed to recognize emotions. Further, frames were extracted and the emotions were analyzed. The voice channel was processed and transcripts were generated. Each channel outcome was compared for accuracy. The opinion was recorded manually for conformance of results. The opinion matched with the emotion classified by the system.

Keywords: Emotion recognition · Deep neural network · Multimodal features · Late fusion · Sequential approach

1 Introduction

Emotions are the inherent feature of human being. The ability to express emotions and the intensity of expressing emotions depends on the stimulus given. The key challenge is to recognize the distinguished pattern and develop a robust system to identify the expressed emotions. Further, there is a need towards automating the emotion recognition system which would assist in a situation such as identifying boredom and improvising visual experience required to maintain interestingness in gaming, website and online tutorials [22].

There is a specific pattern involved while expressing emotions. Ekman, Pulchik, Parrot [1–3] concentrated on clustering emotions based on their expressive state, intensity and relationship among them. These were first studied and encoded in the form of AU(Action Units) and FACS [4] for images and FAP's [5] for videos.

The face was primarily studied as a key to recognizing emotions experienced by a human being. Face images were extensively analyzed since FACS was introduced [6].

M. Singh et al. (Eds.): ICACDS 2019, CCIS 1045, pp. 275–283, 2019.
https://doi.org/10.1007/978-981-13-9939-8_25

With the introduction of various face image databases in 2D such as CK, CK+ [7], 3D [8] and 4D [9], the study intensified. Apart from RGB other formats of images such as thermal [10] was also taken into account for studies. It was evident from the research that automatic face emotion recognition system with the highest accuracy failed in real scenarios. It failed due to inaccuracy in the training dataset or other factors such as regional, cultural, gender and age group dependencies. The approach broadened with the introduction of other modalities for studies such as voice [11], text [12, 14] and physiological signals [13]. The methods to recognize human emotions spanned across modalities. The multimodal approach combines different modalities to produce desired efficiency and accuracy. The combination of the modalities was done such as face and voice [13], face and physiological signal [15]. The major drawback in the available dataset is that they are acquired under an experimental environment which is quite unrealistic categorized as a posed expression.

Several works have been carried out on the dataset in wild acquired under realistic environment. Such studies are subjected to practical problems such as non-availability of the frontal face as most of these algorithms work on the frontal face. Gesture-based studies were conducted to eliminate this issue [16]. Further research is carried out towards defining a process to combine the extracted features and produce desired results in less computation time. Combining modalities is compute intensive process as the complexity increases with an increase in features.

2 Related Work

Several feature fusion approaches such as direct, early, late and sequential fusion were introduced based on correlation, synchronous or asynchronous nature of features and their availability in time.

Direct [17] fusion approach is advantageous if the dataset is a rich feature source and are correlated both in the spatial and temporal domain. Feature level fusion before training the system was experimented in early [18, 19] method but required synchronous feature source. There is a higher dimension of features leading to overfitting.

Late fusion [20] is applicable at the decision level either through polling or maximization process and can handle asynchronous data sources. But the decision needs to be taken at the initial level regarding the feature sources that are experimented for the purpose. Integration of features in sequential order is the key feature of sequential fusion [21] approach such as rule-based and is less studied. The details of fusion approaches are described in Table 1.

Further, with the introduction of different deep neural network architectures, there was a change in choice of deep neural network architecture to increase the accuracy of the system. A bimodal (video and voice) late fusion was applied on videos in which the voice channel was extracted and processed [23]. A similar study was done using 3D CNN for video and 2D CNN for voice [24]. Text and voice correlations in expressing emotions were studied using CNN architecture [25]. Feature level fusion approach was explored using LSTM architecture [26]. Hardware acceleration was used to speed up the process for reduced computation time [27].

Table 1. Fusion approaches with different modalities and the number of emotions detected.

Author	Fusion approach	Modalities	Dataset	No. of Emotions	Model description	Results	Open issues
Ranganathan et al. [17]	Direct	Audio, face video, body video, physiological signals	emoFBVP	23	DBN (Deep Belief Networks) and CDBN (Convolutional Deep Belief Networks) were used to study	SVM baseline: 75.67% DBN: 76.54% CDBN: 81.41%	Requires rich feature source with the high temporal and spatial correlation between data sources
Huang et al. [18]	Early with Plain fusion	Audio and face	Author collected dataset	6	Prosodic feature for voice, feature based study for video	Audio: 75% Video: 69.4% Audio + video: 91.7%	Requires synchronous feature source
Gunes et al. [19]	Early and late	Face and body	Face and upper body gesture dataset generated by the author	6	C4.5 and BayesNet were used for classification. Feature fusion approach for early and decision level fusion for late was used	Early fusion overall: 96% Late fusion Sum rule: 86% Product rule: 80% Weight rule: 82%	High dimensional feature sources were used
Yoshitomi et al. [20]	Late fusion	Voice and face	Author-generated voice, thermal (IR) and visible images (VR)	5	HMM for voice, Neural Network for image classification with a weighted sum for multimodal late fusion approach	VR: 85% IR: 75% VR + IR: 95% Voice: 84% Voice + VR: 92.5%	Availability of IR data source for classification as there is a change in the initial decision regarding data source for experimentation
Chen et al. [21]	Sequential rule-based approach	Video images and voice	An author-generated dataset of video clips and voice samples	4	F0 contours for voice and Fourier Transforms (FT) features fed into HMM for videos. A rule-based approach which exploits the relationship between the two modalities	Low accuracy due to insufficient data and features	Well defined rules need to be established well in advance before integration of features at the initial level

The earlier work requires a fixed and predefined set of input sources towards building a highly accurate system. Further there no scope for inclusion of any other available data sources with rich features in the existing system. The main focus of our work is to build a dynamic system which can incorporate a classification model for various available data sources with different modalities.

3 Proposed Approach

The proposed approach provides a framework to recognize emotions based on the devices and modality of data available during the data gathering process. Initially, the available modality is used to classify the emotion. Based on the output class probability we sequentially integrate the next available data channel from a different source into the model. Then the output class probability of the modalities is compared. The process is repeated till the same class labels are acquired with output probability greater than the desired threshold.

Currently, videos recorded during conversation such as project review meeting are used to build and test the model. The selected videos contain interactions that are conducted in a realistic environment without any specialized lab setup or devices. The recorded video clips are fed to the system and the emotion is recognized and further subjected to emotional analysis. The proposed system flow diagram is depicted in Fig. 1.

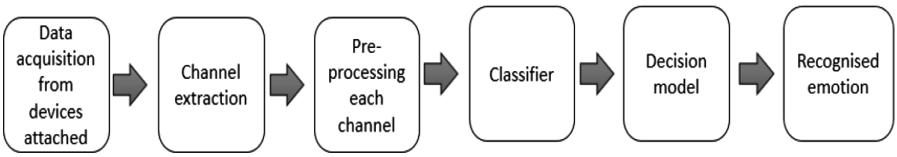

Fig. 1. Flow diagram of the proposed system.

4 System Architecture

A deep convolution neural network (CNN) was used to train FER2013 face emotion dataset. The dataset comprises of 35887 pre-cropped, 48 × 48 size grayscale images of faces. Each face image was labeled with one of the seven emotion classes: anger, disgust, fear, happiness, sadness, surprise and neutral. A small snapshot of images is shown in Fig. 2. Deep CNN model was trained on NVIDIA GPU system with adadelta optimizer and softmax classifier and achieved an accuracy of 61%.

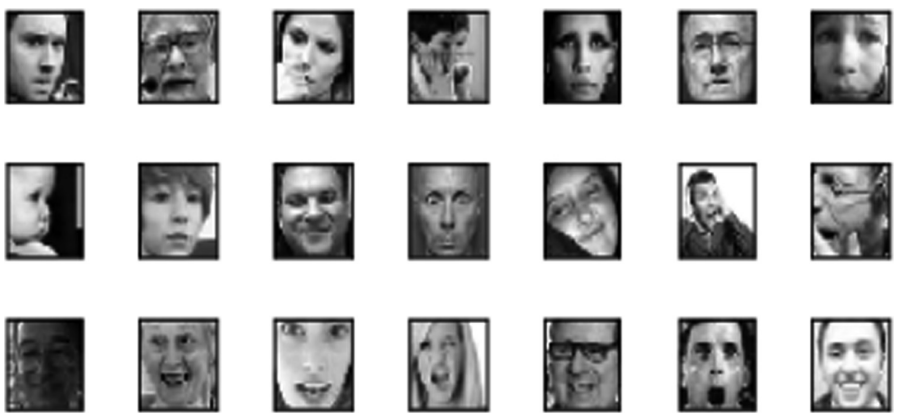

Fig. 2. FER2013 dataset

The voice component is extracted from the video using open source audio extractor. The extracted audio was pre-processed using open source software Audacity. Noise and silence were removed. The transcript of the pre-processed voice was generated. The video clippings were fed to the system. The entire video summarized to one

emotion. The system extracted frames from a video containing a face and fed to a trained deep CNN model. The output is a class probability representing six basic emotion classes. The detailed architecture is shown in Fig. 3. Frame wise detailed study was conducted to analyze the recognized emotion.

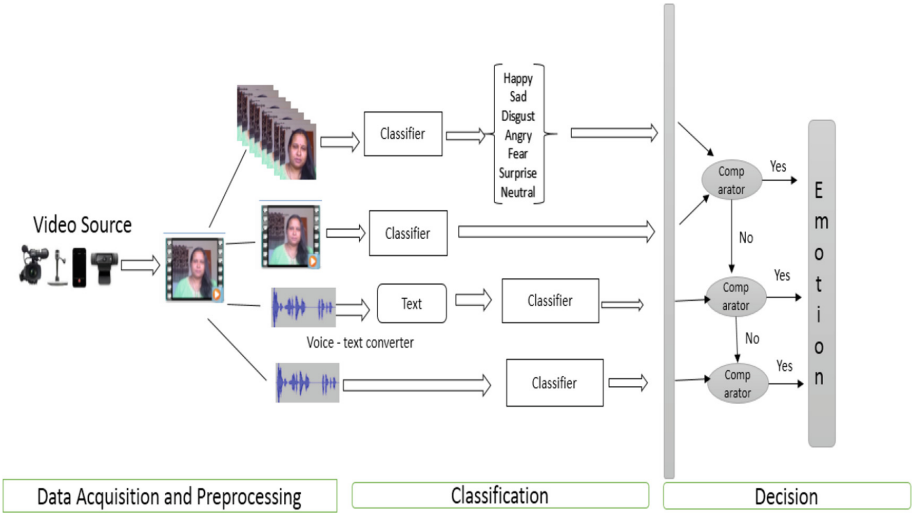

Fig. 3. System architecture

5 Results

The proposed architecture focuses on sequential approach towards a fusion of modalities. Table 2 summarizes the results. For experimental purpose short video of a few minutes were taken.

The numerical value depicted in Table 2 indicates the following results:

- 0 - 'Angry'
- 1 - 'Disgust'
- 2 - 'Fear'
- 3 - 'Happy'
- 4 - 'Sad'
- 5 - 'Surprise
- 6 - 'Neutral'

Figure 4 gives a frame-wise classification for better analysis of the results. Further short sentences extracted from the transcript were summarized and analyzed manually and observation was included. At decision level, the frame outcome and video outcome is matched based on a max count on frames.

Table 2. Summarization of results

Video	Duration	Emotion identified	Frame extracted	Emotion per frame	Opinion	Text analysis from transcript
Vid_gen_1	65 s	4	16	3,4,3,2,6,4,2,3,6,4,4,2,3,4,3,6	Frustrated and sad	"Not happening ": indicates sad and frustrated
Vid_stud_2	90 s	2	20	4,0,2,0,0,0,2,2,2,4,2,2,2,2,2,2,2,4,2	Scared	"Sorry, I couldn't finish": indicates sad and fearful for comments
Vid_stud_3	30 s	3	27	2,4,2,3,3,4,2,2,4,2,3,3,3,3,4,4,4,4,3,3,3,3,4,4,4,3,3	Smiling and Happy	"I finished the module" Yes.: indicated a sense of happiness
Vid_stud_4	60 s	4	13	4,4,4,4,3,4,3,4,4,4,4,4,4	Confused and sad	"I don't know what to do": indicated frustration
Vid_gen_2	52 s	4	38	4,4,4,4,4,4,4,4,3,4,3,3,3,4,4,4,4,3,3,4,4,4,4,4,4,4,4,4,4,3,4,4,4,4,4,4,4,3,4,4,4,4,4,4,4,3	Not so happy	"The results will be improved in next meet". Indicated unhappiness

It was noted that frame extraction count varies due to non-availability of face region for recognition. Experiments were repeated on Vid_gen_2 with different frame count to study the effect. Table 3 and Fig. 4 shows the results for random count value.

Table 3. Frame count analysis.

Count	Emotion for extracted frames	No. of frames extracted
10	4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 3, 3, 3, 4, 4, 4, 4, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4	37
20	4, 4, 4, 4, 3, 3, 3, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4	19
30	4, 4, 4, 4, 3, 3, 3, 4, 4, 3, 4, 4, 3	13

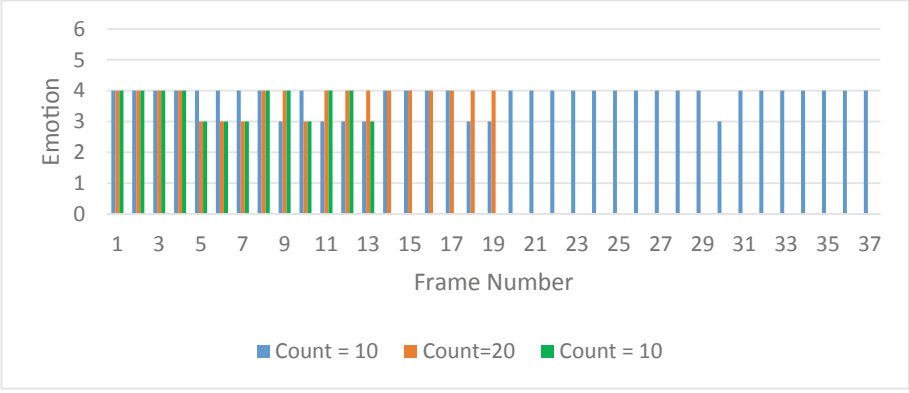

Fig. 4. Frame analysis with varied count values for a particular experimental video Vid_gen_2.

6 Conclusion

Multimodal emotion is compute intensive. The purpose of this paper is to provide a framework to integrate modalities at a later stage only if there is a difference in the outcome of any two available modalities. In our current study, the outcome video emotions matched with the frame outcome and further matched with the manual opinion. Choice of modality plays a vital role and depends on the situation and device attached for data gathering. The availability of channels with required data such as face region might not be present always in a real scenario. Under such circumstances, our approach assists in proceeding towards the next available modality. Further, the experiments can be conducted with various illumination, orientation, camera quality and an initial selection of modality.

However, our work required manual conformance of results for text and audio channel. Hence it is a semi-automated system. This partial automation can be further extended to a fully automated system with minimal manual observation for conformance of results.

References

1. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3-4), 169–200 (1992)
2. Plutchik, R., Kellerman, H.: Emotion, Theory, Research, and Experience, vol. 1. Academic Press, London (1980)
3. Parrott, W.G. (eds.): Emotions in Social Psychology: Essential Readings. Psychology Press, New York (2001)
4. Eckman, P., Friesen, W.V.: Manual for the Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1977)
5. MPEG Video and SNHC, Text of ISO/IEC FDIS 14 496-3: Audio, Atlantic City MPEG Mtg (1998)
6. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Mountain View (1978)
7. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, pp. 46–53 (2000)
8. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, pp. 211–216 (2006)
9. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: Proceedings of 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, pp. 1–6 (2008)
10. Nguyen, H., Kotani, K., Chen, F., Le, B.: A thermal facial emotion database and its analysis. In: Klette, R., Rivera, M., Satoh, S. (eds.) PSIVT 2013. LNCS, vol. 8333, pp. 397–408. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-53842-1_34
11. Paeschke, A., Kienast, M., Sendlmeier, W.F.: F0-contours in emotional speech. In: Proceedings of 14th International Congress of Phonetic Sciences, vol. 2 (1999)
12. Binali, H., Wu, C., Potdar, V.: Computational approaches for emotion detection in text. In: 4th IEEE International Conference on Digital Ecosystems and Technologies, Dubai, pp. 172–177 (2010)
13. Thushara, S., Veni, S.: A multimodal emotion recognition system from video. In: International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, pp. 1–5 (2016)
14. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing. ACM (2008)
15. Huang, Y., Yang, J., Liao, P., Pan, J.: Fusion of facial expressions and EEG for multimodal emotion recognition. Comput. Intell. Neurosci. **2017**, 8 (2017)
16. Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., Driessen, P.F.: Gesture-based affective computing on motion capture data. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 1–7. Springer, Heidelberg (2005). https://doi.org/10.1007/11573548_1
17. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, New York, pp. 1–9 (2016)
18. Huang, T.S., Chen, L.S., Tao, H., Miyasato, T., Nakatsu, R.: Bimodal emotion recognition by man and machine. In: ATR Workshop on Virtual Communication Environments, vol. 31 (1998)

19. Gunes, H., Piccardi, M.: Affect recognition from face and body: early fusion vs. late fusion. In: 2005 IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, HI, vol. 4, pp. 3437–3443 (2005)
20. Yoshitomi, Y., Kim, S.-I., Kawano, T., Kilazoe, T.: Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In: Proceedings of the 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000, Osaka, Japan, pp. 178–183 (2000)
21. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, pp. 366–371 (1998)
22. Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Emotion recognition and its applications. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T., Wtorek, J. (eds.) Human-Computer Systems Interaction: Backgrounds and Applications 3. AISC, vol. 300, pp. 51–62. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08491-6_5
23. Song, K., Nho, Y., Seo, J., Kwon, D.: Decision-level fusion method for emotion recognition using multimodal emotion recognition information. In: 15th International Conference on Ubiquitous Robots (UR), Honolulu, HI, pp. 472–476 (2018)
24. Hossain, M.S., Muhammad, G.: Emotion recognition using deep learning approach from audio-visual emotional big data. Inf. Fusion **49**, 69–78 (2019)
25. Choi, W.Y., Song, K.Y., Lee, C.W.: Convolutional attention networks for multimodal emotion recognition from speech and text data. In: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pp. 28–34 (2018)
26. Tan, Z.X., Goel, A., Nguyen, T.-S., Ong, D.C.: A multimodal LSTM for predicting listener empathic responses over time. arXiv preprint arXiv:1812.04891 (2018)
27. Sonawane, B., Sharma, P.: Acceleration of CNN-based facial emotion detection using NVIDIA GPU. In: Bhalla, S., Bhateja, V., Chandavale, A.A., Hiwale, A.S., Satapathy, S.C. (eds.) Intelligent Computing and Information and Communication. AISC, vol. 673, pp. 257–264. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-7245-1_26