Research Article

Facial emotion recognition using convolutional neural networks (FERC)

Ninad Mehendale[1,2]

Received: 16 July 2019 / Accepted: 12 February 2020 / Published online: 18 February 2020

Abstract
Facial expression for emotion detection has always been an easy task for humans, but achieving the same task with a computer algorithm is quite challenging. With the recent advancement in computer vision and machine learning, it is possible to detect emotions from images. In this paper, we propose a novel technique called facial emotion recognition using convolutional neural networks (FERC). The FERC is based on two-part convolutional neural network (CNN): The first-part removes the background from the picture, and the second part concentrates on the facial feature vector extraction. In FERC model, expressional vector (EV) is used to find the five different types of regular facial expression. Supervisory data were obtained from the stored database of 10,000 images (154 persons). It was possible to correctly highlight the emotion with 96% accuracy, using a EV of length 24 values. The two-level CNN works in series, and the last layer of perceptron adjusts the weights and exponent values with each iteration. FERC differs from generally followed strategies with single-level CNN, hence improving the accuracy. Furthermore, a novel background removal procedure applied, before the generation of EV, avoids dealing with multiple problems that may occur (for example distance from the camera). FERC was extensively tested with more than 750K images using extended Cohn–Kanade expression, Caltech faces, CMU and NIST datasets. We expect the FERC emotion detection to be useful in many applications such as predictive learning of students, lie detectors, etc.

Keywords Emotion recognition · Facial expression · CNN

1 Introduction

Facial expressions are the vital identifiers for human feelings, because it corresponds to the emotions. Most of the times (roughly in 55% cases) [1], the facial expression is a nonverbal way of emotional expression, and it can be considered as concrete evidence to uncover whether an individual is speaking the truth or not [2].

The current approaches primarily focus on facial investigation keeping background intact and hence built up a lot of unnecessary and misleading features that confuse CNN training process. The current manuscript focuses on five essential facial expression classes reported, which are displeasure/anger, sad/unhappy, smiling/happy, feared, and surprised/astonished [3]. The FERC algorithm presented in this manuscript aims for expressional examination and to characterize the given image into these five essential emotion classes.

Reported techniques on facial expression detection can be described as two major approaches. The first one is distinguishing expressions [4] that are identified with an explicit classifier, and the second one is making characterization dependent on the extracted facial highlights [5]. In the facial action coding system (FACS) [6], action units are used as expression markers. These AUs were discriminable by facial muscle changes.

✉ Ninad Mehendale, ninad.mehendale@gmail.com | [1]Ninad's Research Lab, Thane, India. [2]K. J. Somaiya College of Engineering, Mumbai, India.

SN Applied Sciences (2020) 2:446 | https://doi.org/10.1007/s42452-020-2234-1

2 Literature review

Facial expression is the common signal for all humans to convey the mood. There are many attempts to make an automatic facial expression analysis tools [7] as it has application in many fields such as robotics, medicine, driving assist systems, and lie detector [8–10]. Since the twentieth century, Ekman et al. [11] defined seven basic emotions, irrespective of culture in which a human grows with the seven expressions (anger, feared, happy, sad, contempt [12], disgust, and surprise). In a recent study on the facial recognition technology (FERET) dataset, Sajid et al. found out the impact of facial asymmetry as a marker of age estimation [13]. Their finding states that right face asymmetry is better compared to the left face asymmetry. Face pose appearance is still a big issue with face detection. Ratyal et al. provided the solution for variability in facial pose appearance. They have used three-dimensional pose invariant approach using subject-specific descriptors [14, 15]. There are many issues like excessive makeup [16] pose and expression [17] which are solved using convolutional networks. Recently, researchers have made extraordinary accomplishment in facial expression detection [18–20], which led to improvements in neuroscience [21] and cognitive science [22] that drive the advancement of research, in the field of facial expression. Also, the development in computer vision [23] and machine learning [24] makes emotion identification much more accurate and accessible to the general population. As a result, facial expression recognition is growing rapidly as a sub-field of image processing. Some of the possible applications are human–computer interaction [25], psychiatric observations [26], drunk driver recognition [27], and the most important is lie detector [28].

3 Methodology

Convolutional neural network (CNN) is the most popular way of analyzing images. CNN is different from a multi-layer perceptron (MLP) as they have hidden layers, called convolutional layers. The proposed method is based on a two-level CNN framework. The first level recommended is background removal [29], used to extract emotions from an image, as shown in Fig. 1. Here, the conventional CNN network module is used to extract primary expressional vector (EV). The expressional vector (EV) is generated by tracking down relevant facial points of importance. EV is directly related to changes in expression. The EV is obtained using a basic perceptron unit applied on a background-removed face image. In the proposed FERC model, we also have a non-convolutional perceptron layer as the last stage. Each of the convolutional layers receives the input data (or image), transforms it, and then outputs it to the next level. This transformation is convolution operation, as shown in Fig. 2. All the convolutional layers used are capable of pattern detection. Within each convolutional layer, four filters were used. The input image fed to the first-part CNN (used for background removal) generally consists of shapes, edges, textures, and objects along with the face. The edge detector, circle detector, and corner detector filters are used at the start of the convolutional layer 1. Once the face has been detected, the second-part CNN filter catches facial features, such as eyes, ears, lips, nose, and cheeks. The edge detection filters used in this layer are shown in Fig. 3a. The second-part CNN consists of layers with 3×3 kernel matrix, e.g., [0.25, 0.17, 0.9; 0.89, 0.36, 0.63; 0.7, 0.24, 0.82]. These numbers are selected between 0 and 1 initially. These numbers are optimized for EV detection, based on the ground truth we had, in the supervisory training dataset. Here, we used minimum error decoding to optimize filter values. Once the filter is tuned by supervisory learning, it is then applied to the background-removed face (i.e., on the output image of the first-part CNN), for detection of different facial parts (e.g., eye, lips. nose, ears, etc.)

To generate the EV matrix, in all 24 various facial features are extracted. The EV feature vector is nothing but values of normalized Euclidian distance between each face part, as shown in Fig. 3b.

3.1 Key frame extraction from input video

FERC works with an image as well as video input. In case, when the input to the FERC is video, then the difference between respective frames is computed. The maximally stable frames occur whenever the intra-frame difference is zero. Then for all of these stable frames, a Canny edge detector was applied, and then the aggregated sum of white pixels was calculated. After comparing the aggregated sums for all stable frames, the frame with the maximum aggregated sum is selected because this frame has maximum details as per edges (more edges more details). This frame is then selected as an input to FERC. The logic behind choosing this image is that blurry images have minimum edges or no edges.

3.2 Background removal

Once the input image is obtained, skin tone detection algorithm [30] is applied to extract human body parts from the image. This skin tone-detected output image is

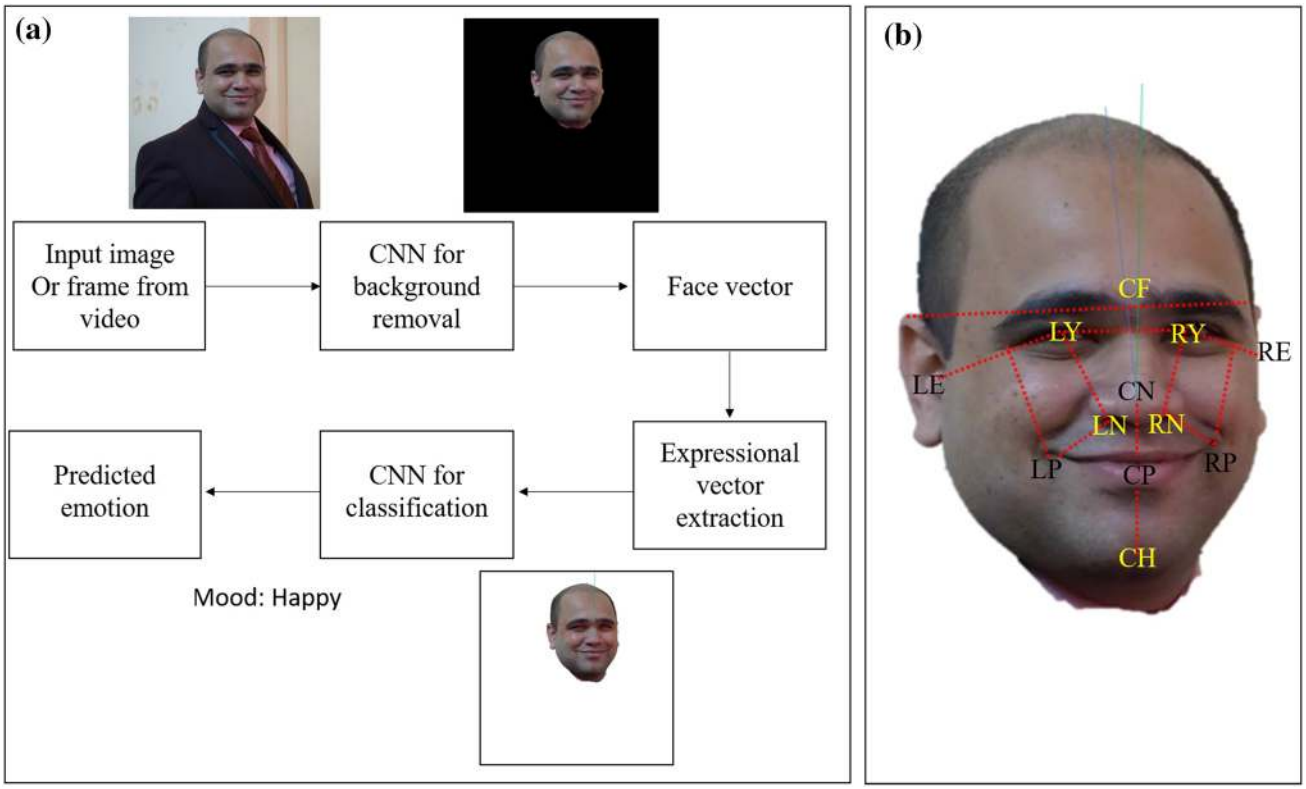

Fig. 1 **a** Block diagram of FERC. The input image is (taken from camera or) extracted from the video. The input image is then passed to the first-part CNN for background removal. After background removal, facial expressional vector (EV) is generated. Another CNN (the second-part CNN) is applied with the supervisory model obtained from the ground-truth database. Finally, emotion from the current input image is detected. **b** Facial vectors marked on the background-removed face. Here, nose (N), lip (P), forehead (F), eyes (Y) are marked using edge detection and nearest cluster mapping. The position left, right, and center are represented using L, R, and C, respectively

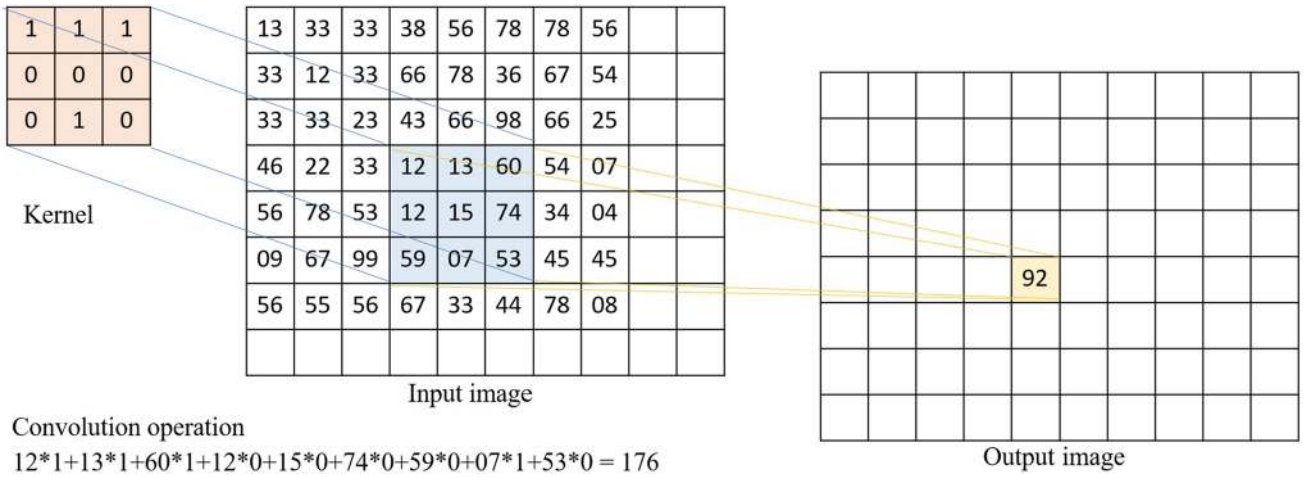

Fig. 2 Convolution filter operation with the 3×3 kernel. Each pixel from the input image and its eight neighboring pixels are multiplied with the corresponding value in the kernel matrix, and finally, all multiplied values are added together to achieve the final output value

a binary image and used as the feature, for the first layer of background removal CNN (also referred to as the first-part CNN in this manuscript). This skin tone detection depends on the type of input image. If the image is the colored image, then YCbCr color threshold can be used. For skin tome, the *Y*-value should be greater than 80, Cb should

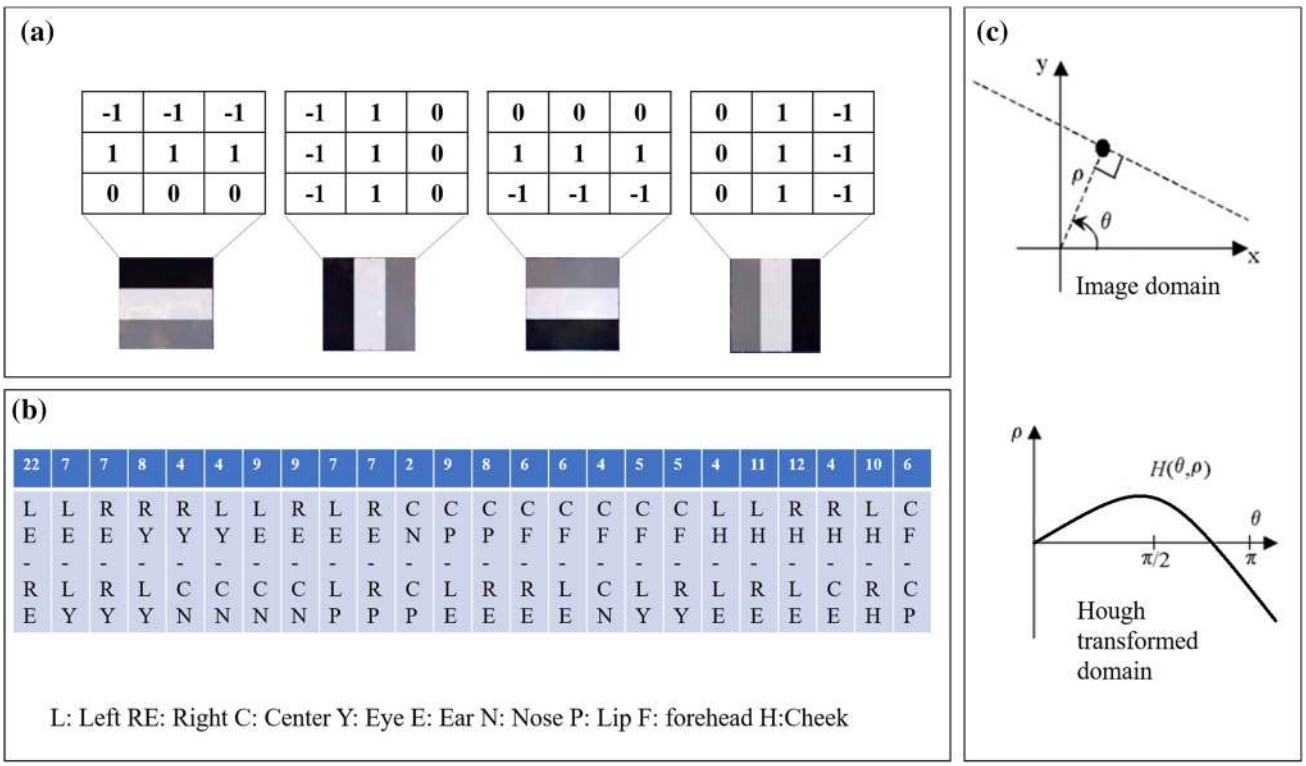

Fig. 3 **a** Vertical and horizontal edge detector filter matrix used at layer 1 of background removal CNN (first-part CNN). **b** Sample EV matrix showing all 24 values in the pixel in top and parameter measured at bottom. **c** Representation of point in Image domain (top panel) to Hough transform domain (bottom panel) using Hough transform

range between 85 and 140, Cr value should be between 135 and 200. The set of values mentioned in the above line was chosen by trial-and-error method and worked for almost all of the skin tones available. We found that if the input image is grayscale, then skin tone detection algorithm has very low accuracy. To improve accuracy during background removal, CNN also uses the circles-in-circle filter. This filter operation uses Hough transform values for each circle detection. To maintain uniformity irrespective of the type of input image, Hough transform (Fig. 3c) was always used as the second input feature to background removal CNN. The formula used for Hough transform is as shown in Eq. 1

$$H(\theta,\rho)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}A(x,y)\delta(\rho-x\cos\theta-y\sin\theta)\mathrm{d}x\mathrm{d}y \quad (1)$$

3.3 Convolution filter

As shown in Fig. 2 for each convolution operation, the entire image is divided into overlapping 3×3 matrices, and then the corresponding 3×3 filter is convolved over each 3×3 matrix obtained from the image. The sliding and taking dot product operation is called 'convolution' and hence the name 'convolutional filter.' During the convolution, dot product of both 3×3 matrix is computed and stored at a corresponding location, e.g., (1,1) at the output, as shown in Fig. 2. Once the entire output matrix is calculated, then this output is passed to the next layer of CNN for another round of convolution. The last layer of face feature extracting CNN is a simple perceptron, which tries to optimize values of scale factor and exponent depending upon deviation from the ground truth.

3.4 Hardware and software details

All the programs were executed on Lenovo Yoga 530 model laptop with Intel i5 8th generation CPU and 8 GB RAM with 512 GB SSD hard disk. Software used to run the experiment were Python (Using Thonny IDE), MATLAB 2018a, and ImageJ.

4 Results and discussions

To analyze the performance of the algorithm, extended Cohn–Kanade expression dataset [31] was used initially. Dataset had only 486 sequences with 97 posers, causing accuracy to reach up to 45% maximum. To overcome the

problem of low efficiency, multiple datasets were downloaded from the Internet [32, 33], and also author's own pictures at different expressions were included. As the number of images in dataset increases, the accuracy also increased. We kept 70% of 10K dataset images as training and 30% dataset images as testing images. In all 25 iterations were carried out, with the different sets of 70% training data each time. Finally, the error bar was computed as the standard deviation. Figure 4a shows the optimization of the number of layers for CNN. For simplicity, we kept the number of layers and the number of filters, for background removal CNN (first-part CNN) as well as face feature extraction CNN (the second-part CNN) to be the same. In this study, we varied the number of layers from 1 to 8. We found out that maximum accuracy was obtained around 4. It was not very intuitive, as we assume the number of layers is directly proportional to accuracy and inversely proportional to execution time. Hence due to maximum accuracy obtained with 4 layers, we selected the number of layers to be 4. The execution time was increasing with the number of layers, and it was not adding significant value to our study, hence not reported in the current manuscript. Figure 4b shows the number of filters optimization for both layers. Again, 1–8 filters were tried for each of the four-layer CNN networks. We found that four filters were giving good accuracy. Hence, FERC was designed with four layers and four filters. As a future scope of this study, researchers can try varying the number of layers for both CNN independently. Also, the vast amount of work can be done if each layer is fed with a different number of filters. This could be automated using servers. Due to computational power limitation of the author, we did not carry out this study, but it will be highly appreciated if other researchers come out with a better number than 4 (layers), 4 (filters) and increase the accuracy beyond 96%, which we could achieve. Figure 4c and e shows regular front-facing cases with angry and surprise emotions, and the algorithm could easily detect them (Fig. 4d, f). The only challenging part in these images was skin tone detection, because of the grayscale nature of these images. With color images, background removal with the help of skin tone detection was straightforward, but with grayscale images, we observed false face detection in many cases. Image, as shown in Fig. 4g, was challenging because of the orientation. Fortunately, with 24 dimensions EV feature vector, we could correctly classify 30° oriented faces using FERC. We do accept the method has some limitations such as high computing power during CNN tuning, and also, facial hair causes a lot of issues. But other than these problems, the accuracy of our algorithm is very high (i.e., 96%), which

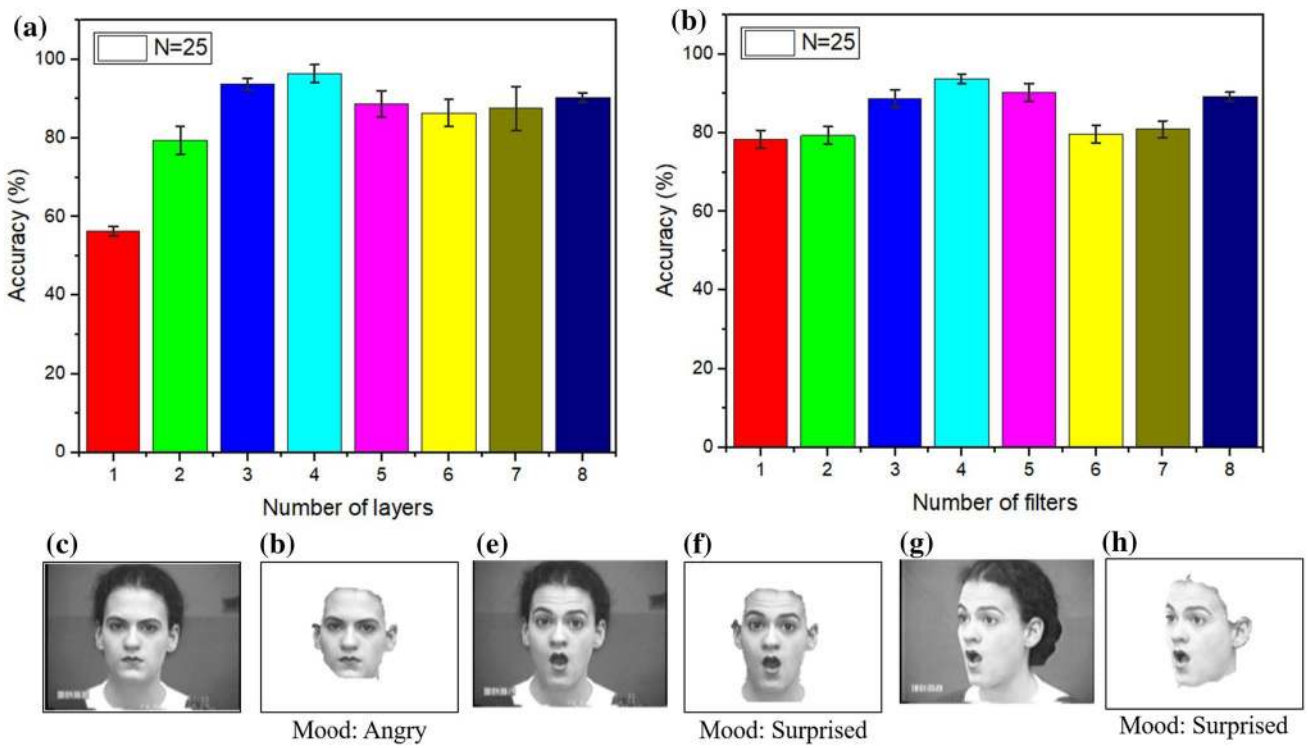

Fig. 4 **a** Optimization for the number of CNN layers. Maximum accuracy was achieved for four-layer CNN. **b** Optimization for the number of filters. Four filters per layer gave maximum accuracy. **c**, **e**, **g** Different input images from the dataset. **d**, **f**, **h** The output of background removal with a final predicted output of emotion

Table 1 Results obtained with different databases

Database	Images (people)	Accuracy (%)
Caltech faces [34]	450 (27)	85
The CMU database [35]	750,000 (337)	78
NIST database [36]	3248 (1573)	96

is comparable to most of the reported studies (Table 2). One of the major limitations of this method is when all 24 features in EV vector are not obtained due to orientation or shadow on the face. Authors are trying to overcome shadow limitation by automated gamma correction on images (manuscript under preparation). For orientation, we could not find any strong solution, other than assuming facial symmetry. Due to facial symmetry, we are generating missing feature parameters by copying the same 12 values for missing entries in the EV matrix (e.g., the distance between the left eye to the left ear (LY–LE) is assumed the same as a right eye to the right ear (RY–RE), etc.) The algorithm also failed when multiple faces were present in the same image, with equal distance from the camera. For testing data selection, the same dataset with 30% data which was not used for training was used. For each pre-processing epoch, all the 100 % data were taken as new fresh sample data in all 25 folds of training. To find the performance of FERC with large datasets Caltech faces, CMU database and NIST database were used (Table 1). It was found that Accuracy goes down with an increasing number of images because of the over-fitting. Also, accuracy remained low, when the number of training images is less. The ideal number of images was found out to be in the range of 2000–10,000 for FERC to work properly.

4.1 Comparison with other methods

As shown in Table 2, FERC method is a unique method developed with two 4-layer networks with an accuracy of 96%, where others have just gone for a combined approach of solving background removal and face expression detection in a single CNN network. Addressing both issues separately reduces complexity and also the tuning time. Although we only have considered five moods to classify, the sixth and seventh mood cases were misclassified, adding to the error. Zao et al. [37] have achieved maximum accuracy up to 99.3% but at the cost of 22 layers neural network. Training such a large network is a time-consuming job. Compared to existing methods, only FERC has keyframe extraction method, whereas others have only gone for the last frame. Jung et al. [38] tried to work with fixed frames which make the system not so efficient with video input. The number of folds of training in most of the other cases was ten only, whereas we could go up to 25-fold training because of small network size.

Table 3 Comparison table of FERC with standard networks

Algorithm	Accuracy (%)	Computational complexity
Alexnet [40]	57–87	O^4
VGG [41]	67–68	O^9
GoogleNet [42]	83–87	O^5
Resnet [41]	73.30	O^{16}
FERC	78–96	O^4

As shown in Table 3, FERC has similar complexity as that of Alexnet. FERC is much faster, compared to VGG, GoogleNet, and Resnet. In terms of accuracy, FERC outperforms existing standard networks. However, in some cases we found GoogleNet out-performs FERC, especially when the iteration of GoogleNet reaches in the range of 5000 and above.

Another unique contribution of FERC is skin tone-based feature and Hough transform for circles-in-circle filters. The skin tone is a pretty fast and robust method of pre-processing the input data. We expect that with these new functionalities, FERC will be the most preferred method for mood detection in the upcoming years.

Table 2 Comparison table with similar methods reported in the literature

	No. of mood	Key frame	N/W size	Accuracy	No. fold
FERC	5	Edge based	8	96	25
Zao et al. [37]	6	Last frame	22	99.3	10
Jung et al. [38]	7	Fixed frame	4	91.44	10
Zang et al. [39]	7	Last frame	7	97.78	10

5 Conclusions

FERC is a novel way of facial emotion detection that uses the advantages of CNN and supervised learning (feasible due to big data). The main advantage of the FERC algorithm is that it works with different orientations (less than 30°) due to the unique 24 digit long EV feature matrix. The background removal added a great advantage in accurately determining the emotions. FERC could be the starting step, for many of the emotion-based applications such as lie detector and also mood-based learning for students, etc.

Acknowledgements The author would like to thank Dr. Madhura Mehendale for her constant support on database generation and corresponding ground truths cross-validation. Also, the author would like to thank all the colleagues at K. J. Somaiya College of Engineering.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Mehrabian A (2017) Nonverbal communication. Routledge, London
2. Bartlett M, Littlewort G, Vural E, Lee K, Cetin M, Ercil A, Movellan J (2008) Data mining spontaneous facial behavior with automatic expression coding. In: Esposito A, Bourbakis NG, Avouris N, Hatzilygeroudis I (eds) Verbal and nonverbal features of human–human and human–machine interaction. Springer, Berlin, pp 1–20
3. Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull 115(1):102
4. Gizatdinova Y, Surakka V (2007) Automatic detection of facial landmarks from AU-coded expressive facial images. In: 14th International conference on image analysis and processing (ICIAP). IEEE, pp 419–424
5. Liu Y, Li Y, Ma X, Song R (2017) Facial expression recognition with fusion features extracted from salient facial areas. Sensors 17(4):712
6. Ekman R (1997) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press, New York
7. Zafar B, Ashraf R, Ali N, Iqbal M, Sajid M, Dar S, Ratyal N (2018) A novel discriminating and relative global spatial image representation with applications in CBIR. Appl Sci 8(11):2242
8. Ali N, Zafar B, Riaz F, Dar SH, Ratyal NI, Bajwa KB, Iqbal MK, Sajid M (2018) A hybrid geometric spatial image representation for scene classification. PLoS ONE 13(9):e0203339
9. Ali N, Zafar B, Iqbal MK, Sajid M, Younis MY, Dar SH, Mahmood MT, Lee IH (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. PLoS ONE 14:7
10. Ali N, Bajwa KB, Sablatnig R, Chatzichristofis SA, Iqbal Z, Rashid M, Habib HA (2016) A novel image retrieval based on visual words integration of SIFT and SURF. PLoS ONE 11(6):e0157428
11. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124
12. Matsumoto D (1992) More evidence for the universality of a contempt expression. Motiv Emot 16(4):363
13. Sajid M, Iqbal Ratyal N, Ali N, Zafar B, Dar SH, Mahmood MT, Joo YB (2019) The impact of asymmetric left and asymmetric right face images on accurate age estimation. Math Probl Eng 2019:1–10
14. Ratyal NI, Taj IA, Sajid M, Ali N, Mahmood A, Razzaq S (2019) Three-dimensional face recognition using variance-based registration and subject-specific descriptors. Int J Adv Robot Syst 16(3):1729881419851716
15. Ratyal N, Taj IA, Sajid M, Mahmood A, Razzaq S, Dar SH, Ali N, Usman M, Baig MJA, Mussadiq U (2019) Deeply learned pose invariant image analysis with applications in 3D face recognition. Math Probl Eng 2019:1–21
16. Sajid M, Ali N, Dar SH, Iqbal Ratyal N, Butt AR, Zafar B, Shafique T, Baig MJA, Riaz I, Baig S (2018) Data augmentation-assisted makeup-invariant face recognition. Math Probl Eng 2018:1–10
17. Ratyal N, Taj I, Bajwa U, Sajid M (2018) Pose and expression invariant alignment based multi-view 3D face recognition. KSII Trans Internet Inf Syst 12:10
18. Xie S, Hu H (2018) Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans Multimedia 21(1):211
19. Danisman T, Bilasco M, Ihaddadene N, Djeraba C (2010) Automatic facial feature detection for facial expression recognition. In: Proceedings of the International conference on computer vision theory and applications, pp 407–412. https://doi.org/10.5220/0002838404070412
20. Mal HP, Swarnalatha P (2017) Facial expression detection using facial expression model. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE, pp 1259–1262
21. Parr LA, Waller BM (2006) Understanding chimpanzee facial expression: insights into the evolution of communication. Soc Cogn Affect Neurosci 1(3):221
22. Dols JMF, Russell JA (2017) The science of facial expression. Oxford University Press, Oxford
23. Kong SG, Heo J, Abidi BR, Paik J, Abidi MA (2005) Recent advances in visual and infrared face recognition—a review. Comput Vis Image Underst 97(1):103
24. Xue YI, Mao X, Zhang F (2006) Beihang university facial expression database and multiple facial expression recognition. In: 2006 International conference on machine learning and cybernetics. IEEE, pp 3282–3287
25. Kim DH, An KH, Ryu YG, Chung MJ (2007) A facial expression imitation system for the primitive of intuitive human-robot interaction. In: Sarkar N (ed) Human robot interaction. IntechOpen, London
26. Ernst H (1934) Evolution of facial musculature and facial expression. J Nerv Ment Dis 79(1):109
27. Kumar KC (2012) Morphology based facial feature extraction and facial expression recognition for driver vigilance. Int J Comput Appl 51:2
28. Hernández-Travieso JG, Travieso CM, Pozo-Baños D, Alonso JB et al (2013) Expression detector system based on facial images. In: BIOSIGNALS 2013-proceedings of the international conference on bio-inspired systems and signal processing
29. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1):32
30. Hsu RL, Abdel-Mottaleb M, Jain AK (2002) Face detection in color images. IEEE Trans Pattern Anal Mach Intell 24(5):696

31. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 94–101
32. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (2011) The computer expression recognition toolbox (CERT). In: Face and gesture 2011. IEEE, pp 298–305
33. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803
34. Caltech Faces (2020) http://www.vision.caltech.edu/html-files/archive.html. Accessed 05 Jan 2020
35. The CMU multi-pie face database (2020) http://ww1.multipie.org/. Accessed 05 Jan 2020
36. NIST mugshot identification database (2020) https://www.nist.gov/itl/iad/image-group/resources/biometric-special-databases-and-software. Accessed 05 Jan 2020
37. Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, Yan S (2016) Peak-piloted deep network for facial expression recognition. In: European conference on computer vision. Springer, pp 425–442
38. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision. pp 2983–2991
39. Zhang K, Huang Y, Du Y, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193
40. Wu YL, Tsai HY, Huang YC, Chen BH (2018) Accurate emotion recognition for driving risk prevention in driver monitoring system. In: 2018 IEEE 7th global conference on consumer electronics (GCCE). IEEE, pp 796–797
41. Gajarla V, Gupta A (2015) Emotion detection and sentiment analysis of images. Georgia Institute of Technology, Atlanta
42. Giannopoulos P, Perikos I, Hatzilygeroudis I (2018) Deep learning approaches for facial emotion recognition: a case study on FER-2013. In: Hatzilygeroudis I, Palade V (eds) Advances in hybridization of intelligent methods. Springer, Berlin, pp 1–16