Facial emotion recognition using convolutional neural networks (FERC)

Ninad Mehendale[1]

[1]Affiliation not available

January 12, 2019

Abstract

Facial expression for emotion detection has always been an easy task for humans (especially parents) but doing the same thing using the computer algorithm is a challenging task. With the advancement of computer vision and machine learning in the recent decade, 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 networks (CNN) where first-half removes the background from the picture while the second part concentrates on the facial feature vector extraction. In FERC model expressional vector (EV) is used to find the different five types of normal facial expression. Supervisory data obtained from the stored database of 10000 images (154 persons). With total 24 value long EV it was possible to accuratly highlighting the emotion with 96% accuracy. The two-level CNN works within serise and last layer of perceptron adusts the weights and exponents values with each iteration and improves accuracy per stages. FERC contrast, generally followed strategies with single level CNN and hence improving the accuracy. Furthermore, a novel background removal procedure before EV avoids dealing with multiple problems that may occur, such as, distance from the camera. FERC was tested with extended Cohn-Kanade expression datasets. We expect the FERC emotion detection to be useful in many applications such as predictive learning of students, lie detectors etc.

Introduction

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

Recently, researchers have made extraordinary accomplishment in facial expression detection[3][4][5]. Improvements in neuroscience[6] and cognitive science[7] drive the advancement of research in the field of facial expression. Also, the development in computer vision[8] and machine learning[9]makes emotion identification much more accurate and accessible to the general population. As a result, facial expression recognition is growing rapidly as a subfield of image processing. Some of the possible applications are human-PC interaction[10], Mental patient observation[11], drunk driver recognition[12] and most important is lie detector[13].

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. There are five essential facial expression classification classes reported which are displeasure/anger, sad/unhappy, smiling/happy, feared, and surprised/astonished[14]. The current FERC algorithm presented in this manuscript aim for current expressional examination and to characterize given image into these five essential emotion classes. Reported techniques on facial expression detection can be characterized as two major approaches. First is distinguishing[15] that are identified with an explicit classifier and second is making characterization dependent on the extracted facial highlights[16][16][17]. In the Facial Action Coding System (FACS) [18], action units are used as expression markers. These AUs were discriminable by facial muscle changes.

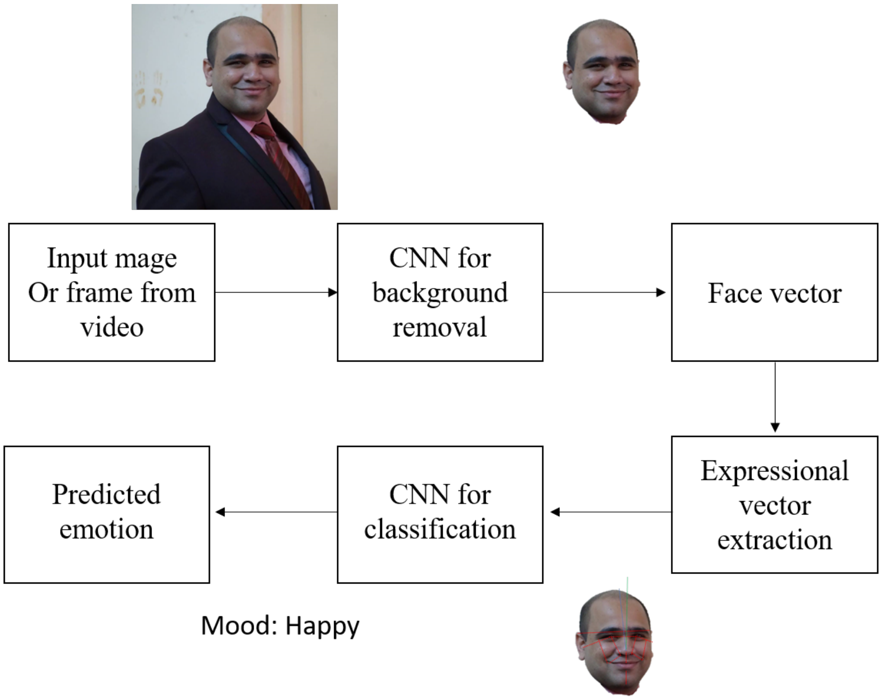

Input mage Or frame from video
CNN for background removal
Face vector
Expressional vector extraction
CNN for classification
Predicted emotion
Mood: Happy

Figure 1: Block diagram of FERC. The input image is either taken from camera or extracted from the video. The input image is then passed to CNN for background removal. After background removal facial expressional vector (EV) is generated. CNN is applied with the supervisory model obtained from the database. Finally, emotion from the image is detected.

Methodology

The proposed method is based on a two-level CNN framework. As shown in figure 1, in order to extract emotions from an image we propose firstly background removal[19][20]. We further modify the typical CNN network module to extract basic expressional vector (EV) using perceptron basic unit from a face image with background removed. The expressional vector is generated by tracking down relevant facial points of importance and EV is directly related to changes in expression.

Convolutional neural networks (CNN) is the most popular way of analyzing images. CNN is different from multilayer perceptron (MLP), such that they have hidden layers called convolutional layers. In the FERC model, we do also have a non-convolutional perceptron layer at the end. Each of the convolutional layers receives the input image, transforms it, and then outputs it to the next level. This transformation is convolution operation as given by figure 2.

All the convolutional layers used are capable of detection patterns. Within each convolutional layer, we used

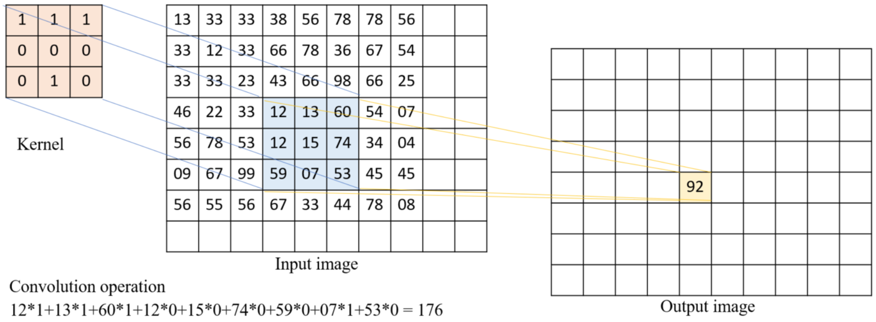
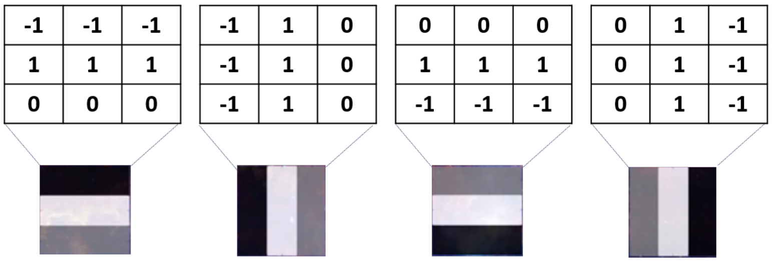


Figure 2: Convolution filter operation with the 3x3 kernel. Each pixel from the input image and its eight neighboring pixels are multiplied with the corresponding value in the kernel matrix and finally, all multiplied values are added together to achieve final output value.

4 filters. The input image to 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 here at the start of the convolutional layer 1 of this first CNN. Once the face is detected, in the second layer CNN filters detect eyes, ears, lips, nose, and cheeks. The edge detection filters in the layer is as shown in figure 3 (a). The second-part CNN consists of layers with 3x3 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 to 1 initially and later optimized for EV detection based on ground-truth in the supervisory training dataset. Here we used minimum error decoding to optimize filter values. once the filter is tuned it is applied to background removed face, for detection of different facial parts (e.g. eye, lips. nose, ears etc.) To generate EV matrix we take 24 different features. This EV feature vector is nothing but values of normalized euclidian distance between each of face part as shown in figure 3 (b).

If the input to the FERC is video, then the difference between individual frames is computed. Whenever the difference is zero maximally stable frames occurs. Then out of these stable frames aggregated edge detection output is computed. After comparing this aggregated sum for all stable frames, the frame with maximum sum is selected because in a way that frame has maximum details. This frame is then selected as an input to FERC. The logic behind selecting image with more edges is that blurry images have minimum edges or no edges. Once the input image is obtained skin tone detection algorithm[21] is applied to extract human body parts from the image. This skin-tone detected image is a binary image and used as one of the feature vectors for the first layer of background removal CNN. The other feature is Hough transformed image. If the input image is gray-scale then skin tone detection algorithm is low on accuracy. To overcome this problem second level of background removal CNN, uses circles-in-circle filter. This filter uses Hough transform values for each circle detection[22]. As shown in figure 2 for each convolution operation, the entire image is divided into overlapping 3x3 matrix and then corresponding 3x3 filer, is convolved over each 3x3 matrix from the image. The sliding and taking dot product operation is called convolution and hence the name convolutional. During convolution dot product of both 3x3 matrix is computed and stored at corresponding location e.g.(1,1) at the output (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 simple perceptron, which tries to optimize values of scale factor and exponent depending upon ground truth.[23]

(a)



$$\begin{bmatrix} -1 & -1 & -1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} -1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

(b)

22	7	7	8	4	4	9	9	7	7	2	9	8	6	6	4	5	5	4	11	12	4	10	6
LE-RE	LE-LY	RE-RY	RY-LY	RY-CN	LY-CN	LE-CN	RE-CN	LE-LP	RE-RP	CN-CP	CP-LE	CP-RE	CF-RE	CF-LE	CF-CN	CF-LY	CF-RY	LH-LE	LH-RE	RH-LE	RH-CE	LH-RH	CF-CP

L: Left RE: Right C: Center Y: Eye E: Ear N: Nose P: Lip F: forehead H:Cheek

Figure 3: (a)Vertical and horizontal edge detector filter matrix used at layer 1 of background removal CNN (b) sample EV matrix showing all 24 values in pixel in top and parameter measured at bottom.

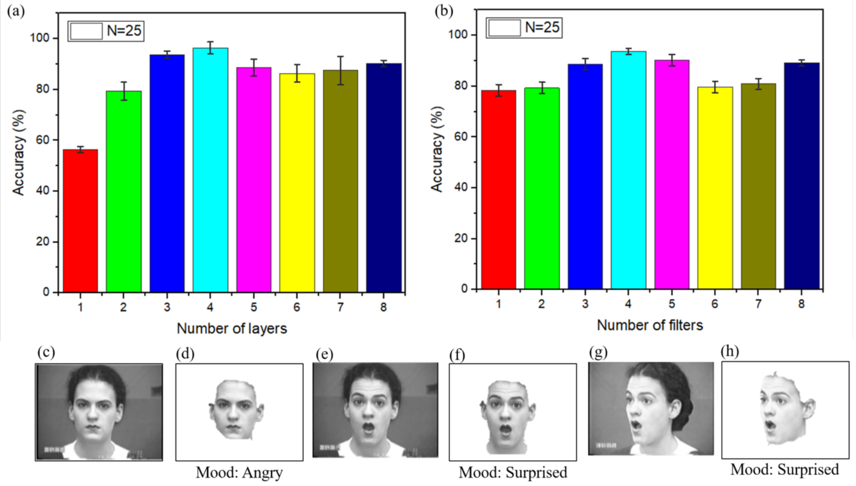


Figure 4: (a) Optimization for number of CNN layers. Maximum accuracy was achieved for 4 layer CNN (b) Optimization for number of filters. 4 filters per layer gave maximum accuracy. (c, e, g) different input images from data set. (d,f,h) Output of background removal with final predicted output of emotion.

Result and discussions

To analyze the performance of the algorithm Extended Cohn-Kanade expression dataset[24]; [25] was used initially. Data set had only 486 sequences with 97 posers and causing accuracy just to reach up to 45% maximum. To overcome this problem of low accuracy, multiple datasets were downloaded from the internet[26]; [27]; [28]; [29]; [30]; [31] and also authors own pictures at different expressions were included. As the number of images in dataset increase, 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 set of 70% training data and then error bar was computed as standard deviation. Figure 4 (a) shows optimization of the number of layers for CNN. For simplicity, we kept the number of layers and number of filters for background removal CNN and face feature extraction 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 we selected the number of layers to be 4. The execution time was increasing with the number of layers, and it was not adding great value to our study hence not reported in the current manuscript. Figure 4 (b) shows the number of filters optimization for both layers. Again 1 to 8 filters were tried for each of 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 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 appriciated if other researchers to come out with a better number than 4 (layers), 4(filters) and increase the accuracy beyond 96%, which we could achieve.

Figure 4 (c and e) were normal front facing cases with angry and surprised emotions and the algorithm could easily detect them (fig 4 d and 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 really easy, but with grayscale images we observed false face detection in many cases. Image, such as, figure 4 g was challenging because of the orientation. Fortunately, with 24 dimension EV feature vector, we could correctly classify 30 degree 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 problems. But other than these problems the accuracy of our algorithm is very high (i.e. 96%) which is comparable to most of reported literature[5]; [4]; [32]; [33]; [34]; [35]. One of the major limitations of this method is when all 24 features in EV vector is 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. Distance between left eye to left ear (LY-LE) is assumes same as right eye to right ear(RY-RE) etc.) Algorithm also failed, when mutiple faces were present in same image, with equal distance from camera.

Conclusions

FERC is a novel way of facial emotion detection that uses 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 degrees) due to unique 24 dimensions EV feature matrix. The background removal added a great advantage, in accurately determining the emotions. FERC could be starting step for many of the emotion-based applications such as lie detector and also mood based study for students etc.

Acknowledgements

Author would like to thank Dr. Madhura Mehendale for her constant support on database generation and corresponding ground truths cross validation.

References

1. Mehrabian A (2017) Nonverbal Communication. doi: 10.4324/9781351308724

2. Ekman P, Rosenberg EL (2005) What the Face RevealsBasic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). doi: 10.1093/acprof:oso/9780195179644.001.0001

3. Xie S, Hu H (2019) Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks. IEEE Transactions on Multimedia 21:211–220. doi: 10.1109/tmm.2018.2844085

4. (2010) AUTOMATIC FACIAL FEATURE DETECTION FOR FACIAL EXPRESSION RECOGNITION. Proceedings of the International Conference on Computer Vision Theory and Applications. doi: 10.5220/0002838404070412

5. Mal HP, Swarnalatha P (2017) Facial expression detection using facial expression model. 2017 International Conference on Energy Communication, Data Analytics and Soft Computing (ICECDS). doi: 10.1109/icecds.2017.8389644

6. Parr LA (2009) Facial Expression in Primate Communication. In: Encyclopedia of Neuroscience. Elsevier, pp 193–200

7. Russell JA, Dols JMF (2017) The Science of Facial Expression. doi: 10.1093/acprof:oso/9780190613501.001.0001

8. Application: Facial Expression Recognition. In: Computational Imaging and Vision. Springer-Verlag, pp 187–209

9. Xue Y-li, Mao X, Zhang F (2006) Beihang University Facial Expression Database and Multiple Facial Expression Recognition. 2006 International Conference on Machine Learning and Cybernetics. doi: 10.1109/icmlc.2006.258460

10. Hyoung D, Ho K, Geol Y, Jin M (2007) A Facial Expression Imitation System for the Primitive of Intuitive Human-Robot Interaction. Human Robot Interaction. doi: 10.5772/5194

11. Ernst H (1934) EVOLUTION OF FACIAL MUSCULATURE AND FACIAL EXPRESSION. The Journal of Nervous and Mental Disease 79:109. doi: 10.1097/00005053-193401000-00073

12. S.ChidanandKumar K (2012) Morphology based Facial Feature Extraction and Facial Expression Recognition for Driver Vigilance. International Journal of Computer Applications 51:17–24. doi: 10.5120/8014-1142

13. (2013) Expression Detector System based on Facial Images. Proceedings of the International Conference on Bio-inspired Systems and Signal Processing. doi: 10.5220/0004322504110418

14. FRIDLUND ALANJ (1994) FACIAL EXPRESSION AND THE METHODS OF CONTEMPORARY EVOLUTIONARY RESEARCH. In: Human Facial Expression. Elsevier, pp 28–54

15. Gizatdinova Y, Surakka V (2007) Automatic Detection of Facial Landmarks from AU-coded Expressive Facial Images. 14th International Conference on Image Analysis and Processing (ICIAP 2007). doi: 10.1109/iciap.2007.4362814

16. Liu Y, Li Y, Ma X, Song R (2017) Facial Expression Recognition with Fusion Features Extracted from Salient Facial Areas. doi: 10.20944/preprints201701.0102.v1

17. Fasel B Facial expression analysis using shape and motion information extracted by convolutional neural networks. Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing. doi: 10.1109/nnsp.2002.1030072

18. Gavrilescu M (2014) Proposed architecture of a fully integrated modular neural network-based automatic facial emotion recognition system based on Facial Action Coding System. 2014 10th International Conference on Communications (COMM). doi: 10.1109/iccomm.2014.6866754

19. Kong K-H, Kang D-S (2016) A Study of Face Detection Using CNN and Cascade Based on Symmetry-LGP & Uniform-LGP and the Skin Color. doi: 10.14257/astl.2016.139.30

20. Matsugu M, Mori K, Suzuki T (2004) Face Recognition Using SVM Combined with CNN for Face Detection. In: Neural Information Processing. Springer Berlin Heidelberg, pp 356–361

21. Muhammad B, Abu-Bakar SAR (2015) A hybrid skin color detection using HSV and YCgCr color space for face detection. 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). doi: 10.1109/icsipa.2015.7412170

22. Djekoune AO, Messaoudi K, Amara K (2017) Incremental circle hough transform: An improved method for circle detection. Optik 133:17–31. doi: 10.1016/j.ijleo.2016.12.064

23. Arena P, Fortuna L, Graziani S, Muscato G (1991) A real-time implementation of a multi-layer perceptron with automatic tuning of learning parameters. IFAC Proceedings Volumes 24:21–25. doi: 10.1016/b978-0-08-041699-1.50008-2

24. Lucey P, Cohn JF, Kanade T, et al. (2010) The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. doi: 10.1109/cvprw.2010.5543262

25. Tian Y, Kanade T, Cohn JF (2011) Facial Expression Recognition. In: Handbook of Face Recognition. Springer London, pp 487–519

26. Widener A (2010) Language ambiguity and facial expression. doi: 10.1037/e741372011-259

27. Kring AM, Sloan DM (1991) Facial Expression Coding System. doi: 10.1037/t03675-000

28. Russell J (2011) Children's recognition of emotion from facial expression. doi: 10.1037/e634112013-128

29. Schlosberg H (1960) The dimensions of facial expression. doi: 10.1037/e627282012-142

30. (2017) Particularly Exciting Experiments in Psychology: Facial Expression Recognition. doi: 10.1037/e507752018-001

31. Simon L, Csukly G, Takacs B (2005) Facial expression recognition in psychiatric disorders using animated 3D emotional facial expressions. doi: 10.1037/e705572011-024

32. Martinez B, Valstar MF (2016) Advances Challenges, and Opportunities in Automatic Facial Expression Recognition. In: Advances in Face Detection and Facial Image Analysis. Springer International Publishing, pp 63–100

33. Saha A, Das H, Kar N, Pal MC (2013) An Approach of Extracting Facial Components for Facial Expression Detection using Fiducial Point Detection. International Journal of Computer Applications 80:49–53. doi: 10.5120/13901-1923

34. OuYang Y, Sang N (2013) Robust Automatic Facial Expression Detection Method. Journal of Software. doi: 10.4304/jsw.8.7.1759-1764

35. Dantes G, Suarni N, Suputra P, et al. (2017) Face-expression detection: Detection of facial expression for optimizing the role of the e-learning system. Regionalization and Harmonization in TVET. doi: 10.1201/9781315166568-67