Automatic Facial Expression Recognition Based on Deep Layered Representation of Convolution Neural Networks

Arun Kumar Dubey and Vanita Jain

Abstract Facial expression is one of the utmost dominant, natural and instant ways for human beings to communicate their sentiments and feelings. Automatic facial expression recognition is an exciting and challenging problem due to its variability and complexity that impacts its importance in human–computer interaction applications. This paper illustrates a novel layered extended convolution neural networks architecture named deep layered representation (DLR). In this paper, we have used Kaggle dataset FER2013 for our layered deep neural network based approach. The implementation of DLR has shown better results. Results are also analyzed and compared using five generalized activation functions: Elu, ReLu, Softplus, Sigmoid and Selu, in the last dense layer. We have also compared our Elu and ReLu-based model with GoogLeNet and VGG16 + SVM-based model on the same dataset.

Keywords CNN · Activation function · Facial expression · FER2013

1 Introduction

Ideas and emotions are easily expressible with verbal communication, but world does not adopted any common language to identify emotion of a person. However, facial expression is one of the utmost dominant, natural and instant ways for human beings to communicate their sentiments and feelings [1–3]. Facial expression is a nonverbal method for sharing expression. Common human expressions which can be easily identified by us are angry, surprise, contempt, neutral, happy, sad and disgust [4–6]. Children are also capable of recognizing these emotions despite their country, race

A. K. Dubey
USICT Guru Gobind Singh Indraprastha University, New Delhi, India
e-mail: arudubey@gmail.comm

V. Jain (✉)
Bharati Vidyapeeth's College of Engineering, New Delhi, India
e-mail: vanita.jain@bharatividyapeeth.eduu

A. Abraham et al. (eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks*, Lecture Notes in Networks and Systems 167,
https://doi.org/10.1007/978-981-15-9712-1_7

knowledge, etc. Same capability can also be developed in machine if properly being modeled [3–10].

Automatic facial expression recognition is an exciting and challenging problem due to its variability and complexity that impacts its importance in human–computer interaction applications [11–13]. Facial expression recognition (FER) has helped a lot to medical science to understand psychological behavior of patient [14]. Human eyes can easily identify emotion of a person if it exists on the face more than half second. To incorporate and mimic this human ability in human–computer-based interaction systems, facial expression recognition(FER) gravity increased in recent years [15–18].

Further, this paper is organized in five sections. Section 1 gives the introduction that covers the problem statement and its motivation. Section 2 describes the related work of facial expression recognition systems. In the Sect. 3, proposed model has been discussed. Section 4 shows the experimental results and its comparison with other models. Section 5 gives the conclusion of the paper.

2 Related Work

Facial expression is one of the favorite topics of experts in computer vision. A radical survey on existing FER has done within last decade [1, 2, 4, 19, 20]. Many extraordinary works are done on static image to detect emotion of human [5–16]. Zeng et al. [19] published a survey on audio, visual and spontaneous expressions and shown their importance and comparisons between voice and facial images. Video-based facial expression recognition (FER) is proposed by Sariyanidi et al. [21], and residue learning [22] has shown some important variants of FER.

In previous works, emotion recognition depends on the traditional two-step machine learning strategy, where in its very first step, a number of unique characteristics or features are extracted from the pictures, and in the second step, a classifier (such as SVM [23], neural network [24] or random forest) is used to identify the emotions [23–25]. Several popular handcrafted capabilities are employed for facial expression recognition [26]. CK, CK+ [23], Fer2013 were popular dataset for facial expression [27, 28]. Typical FERs focused on still face images and analyzed statistical features, while some researchers have explored the different features such as optical flow features [29] and LBP [30]. For video-based facial expression recognition, motion units in faces are identified by Walecki et al. [31]. He emphasized the temporal variation of FER. In 2019, J. Yang et al. performed action unit [32]-based facial expression recognition system which incorporates facial muscle movements that effectively reflect the changes in people's facial expressions.

Due to the inherent importance of deep learning [27, 29], this paper illustrates a novel layered extended convolution neural networks architecture named deep layered representation (DLR). Extended convolution neural networks architecture, compared to other state-of-the-art methods, has demonstrated better result after implementation.

Results are also analyzed and compared using five generalized activation functions: Elu, ReLu, Softplus, Sigmoid and Selu, in the last dense layer.

3 Proposed model

This paper illustrates a novel layered extended convolution neural networks architecture named deep layered representation (DLR). The architecture of proposed deep layered representation (DLR) model for facial expression is shown in Fig. 1. It is depicting the details of the Input, Filter, Stride (st.), Padding (pad), Output size and Parameters. For each convolutional layer, the output is given by

$$\rho_{xy}^{i} = \sum_{\alpha=0}^{k-1} \sum_{\beta=0}^{k-1} w_{\alpha\beta} q_{(x+\alpha)(y+\beta)}^{i} + \chi i \tag{1}$$

where

ρ_{xy}^{i} the output at position (x, y).
$w_{\alpha\beta}$ is the weight of the kernel and
χi is the bias on layer i.

In each 2D convolution layer, 64 3 × 3 filters are used. Output layer has seven labels with softmax classifier. This layered approach has used batch normalization and dropout apart from convolution and max pooling layer. Distribution of input to the layers has been changed after each batch but batch normalization standardized it and helped to reduce epochs. Dropout is used for reducing hyper-parameter and

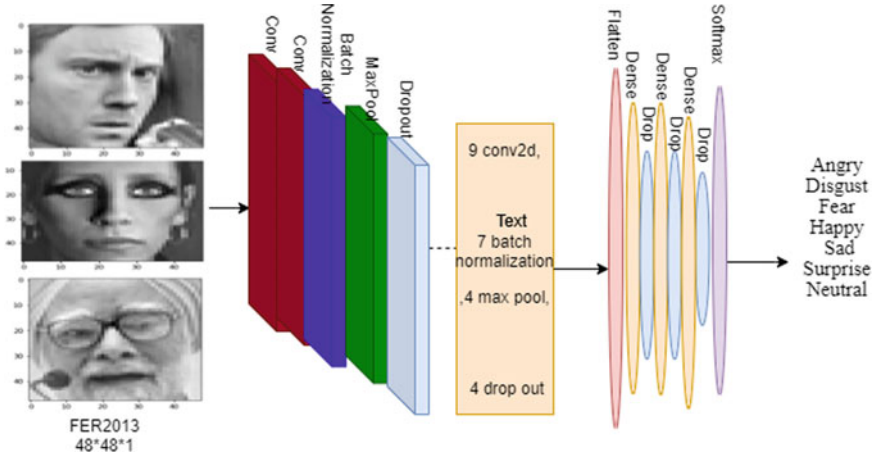

Fig. 1 Proposed deep layered representation (DLR) model for facial expression

helped the model to minimize overfitting. ReLu activation function has been used in each layer except last dense layer.

For the given sample image z, three layered process are followed. First, two normalized convolutional layers plus one fully connected (FC) layer with additional dropout layer are aligned. This generates a feature vector $\nu = \rho(z)$. In the second step, nine convolutional layers with seven batch normalization plus four fully connected layers with additional four dropout layers are aligned which converts $\rho(z)$ into a feature vector $\xi = \Psi(\rho(z) - d)$. In the last step, this combines FC layers with dropouts which finally generate deep layered features (DLF) ϕ.

$$\phi = \delta\big(\Psi(\rho(z) - d) - d'\big) \tag{2}$$

where d and d' are a difference given by dropout layer, and $d \neq d'$. δ, Ψ, ρ are responses of each convolution sets.

Last dense layer is generalized and tested with five activation functions, ReLu, Elu, Selu, Sigmoid and Softplus. Then, each model is designed as DLF + ReLu, DLF + Elu, DLF + Selu, DLF + Sigmoid and DLF + Softplus model. Additionally, zero point four dropout rate has been taken in last three dropouts—dropout 5, dropout 6 and dropout 7. It has dropped some neurons randomly. From the facial expressions, angry, disgust, fear, happy, sad, surprise and neutral emotion have been detected by this model on FER2013 dataset. Kaggle dataset is already processed and converted in to csv file. It is available on Kaggle Web site, and size of each image is 48×48 on single channel. Total parameters, trainable parameters and non-trainable parameters are 5,905,863, 5902, 151 and 3712, respectively.

4 Experiment Result

We have used google co-lab GPU platform for implementation. Model is executed on hundred epoch. Last dense layer's activation function has been kept generalized. This last fully connected layer's activation function is replaced by Elu, ReLu, Softplu, Selu and Sigmoid for hundred epoch. Python Keras framework has helped us to implement our approach.

4.1 Database

We have tested our proposed model on benchmark FER2013 dataset. There are seven emotions: angry, disgust, fear, happy, sad, surprise and neutral labeled as 0–6 in the database. The highest information available in dataset is of happy, whereas disgust is least. Table 1 presents database description as sample per emotion and sample dataset.

Table 1 Database description as sample per emotion and sample dataset

	S. No.	Emotion	Usage	Pixels
Samples per emotion: 3 8989 Happy 6 6198 Neutral 4 6077 Sad 2 5121 Fear 0 4953 Angry 5 4002 Surprise 1 547 Disgust Name: emotion, dtype: int64 Number of Labels: 7	0	0	Training	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121…
	1	0	Training	151 150 147 155 148 133 111 140 170 174 182 15…
	2	2	Training	231 212 156 164 174 138 161 173 182 200 106 38…
	Number of pixel per emotion is: 2304 Number of examples in dataset: 35887			

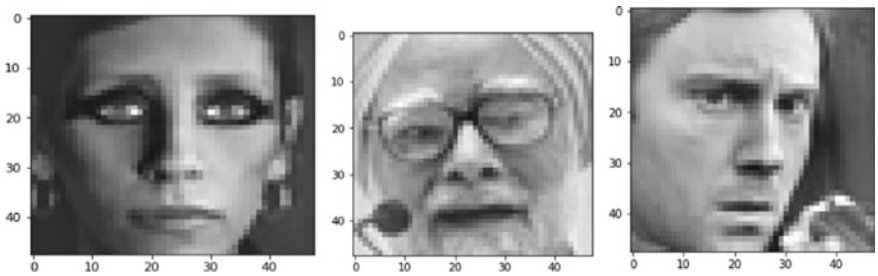

Fig. 2 Faces of dataset consist of 48 × 48 gray image

This dataset consists of set of real-world images like Man Mohan Singh, etc. These are static gray images of size 48 × 48 and processed for single channel. Sample of faces is shown in Fig. 2.

To evaluate accuracy and loss in each epoch, we have trained proposed model on above dataset as per following distribution.

We have calculated the performance of proposed model for 100 epochs using accuracy and loss graph. In each epoch, DLF + Elu model, DLF + Softplus model and DLF + ReLu model-based layered approach have performed better than DLF + Sigmoid model and DLF + Selu model. DLF + Selu model has shown 25% accuracy, and however, it is stable throughout training and validation. Figure 3 represents accuracy of each model in duration of 100 epochs.

In Fig. 4, the loss of the training and validation over 100 epochs has been shown. It is seen that Elu, ReLu and Softplus-based layered models have shown good performance for expression detection. Elu, ReLu and Softplus-based models have shown almost similar performance and outperformed the other two. All three models do not have early stopping point for overfitting.

To analyze the performance of DLR model, receiver operative characteristics (ROC) curves on different activation functions are also presented. Figure 5 shows

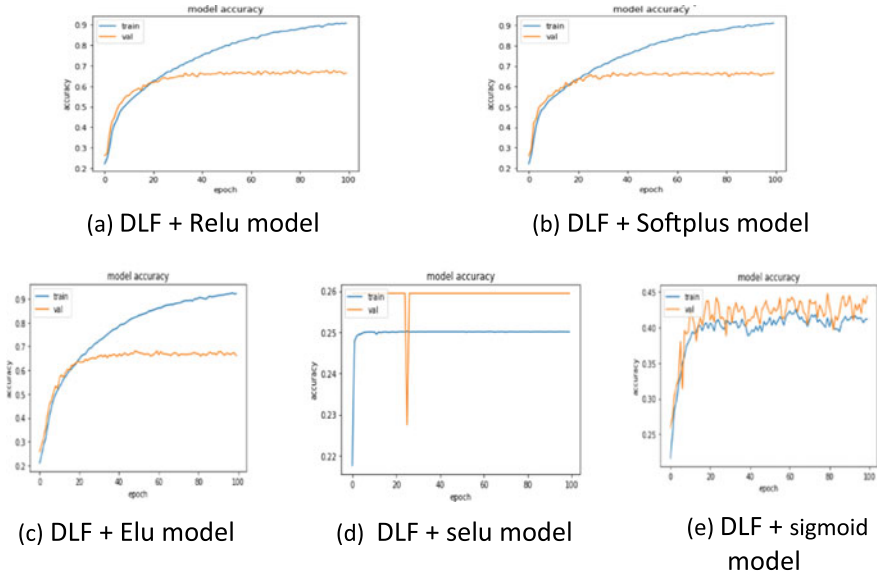

Fig. 3 Training and validation accuracy of each model over 100 epochs

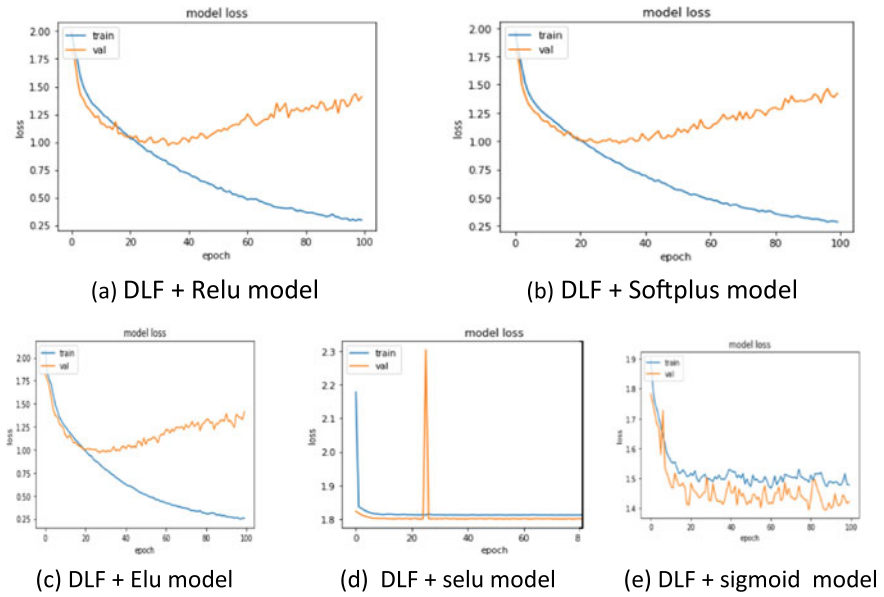

Fig. 4 Loss versus epoch graph on five activation function

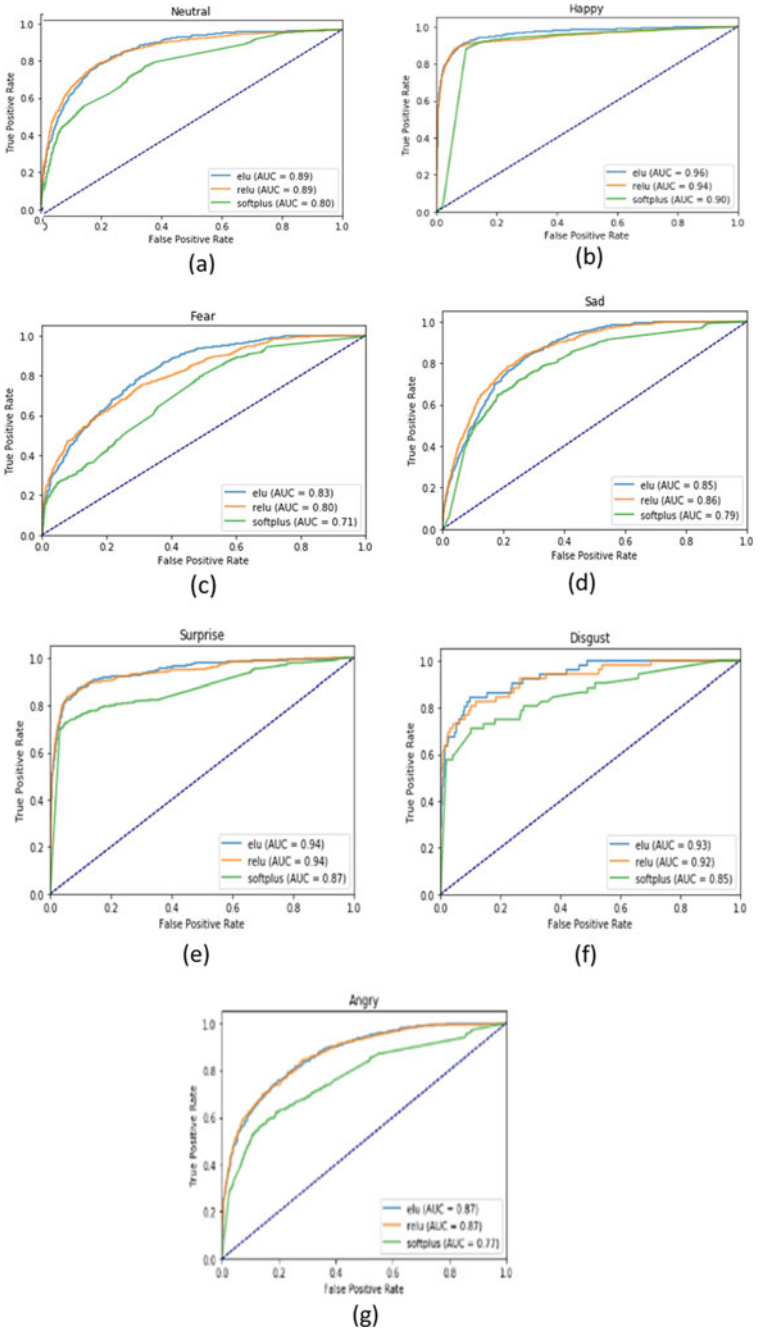

Fig. 5 ROC curve of top three activation function-based proposed approach

the ROC curve on predicting different emotion on FER2013 dataset. From Fig. 5, it is clear that area under curve (AUC) of the DLF + Elu model and DLF + ReLu model has better than others. It is found that happy facial expression prediction modeling is better than ROC of other emotion prediction.

We can also analyze the system performance in the terms of precision, recall, $F1$-score and support which shows the correct predicted values in dataset. Precision is the ratio between true observation and total positive observation. Recall is the sensitivity which shows the ratio between truly predicted positive observations and all observation in actual class, marked as Yes. $F1$-score shows the weighted average of precision and recall. Supports describe the number of occurrence of each label in tested dataset. Table 2 shows the precision, recall, $F1$-score and support-based comparison of DLF + Elu, DLF + ReLu and DLF + Softplus models.

Confusion matrix shows the matrix between true value and predicted value. Highest accurate prediction in DLF + ReLu model is happy where true label and predicted label are 735. DLF + Softplus model, DLF + Elu model and DLF + ReLu model are better in expression prediction in comparison with other two models. It is found that models-based ReLu, Elu, Softplus and Sigmoid are best for happy expression. DLF + Selu models have shown worst performance, and all the expressions are detected as happy. Accuracy and loss graph have depicted that proposed model is the best for Elu, ReLu and Softplus activation function. Figure 6 depicts confusion matrix of top three confusion matrix.

4.2 Accuracy Comparison with Other Research Work on Same Dataset

We have compared our model with other existing and latest models on fer2013 dataset. It is found that our layered approach has performed well on Elu activation function in comparison with the previous model. It has 68.11 accuracies on sixty epoch and 66.50 on a hundred epoch, which is shown in table. ReLu and Softplus models have also achieved significant accuracy result 67.99 and 67.09 on eighty-one and ninety-one epoch, respectively. Table 3 shows the comparative result with the existing model.

Result shows that our proposed layered model has achieved better results in comparison with the previous existing model on FER2013.

5 Conclusion and Future Work

Our deep layered approach has demonstrated a significant result on facial expression. Different variants are also examined based on activation function. Best, 68.11%,

Table 2 Precision, recall, $F1$-score and support-based comparison with DLF + Elu, DLF + ReLu and DLF + Softplus models

Model	Label	precision	Recall	$F1$-score	Support
DLF + Elu	0	0.569	0.556	0.562	498
	1	0.647	0.635	0.641	52
	2	0.581	0.479	0.525	545
	3	0.838	0.852	0.845	881
	4	0.529	0.549	0.539	588
	5	0.766	0.773	0.769	414
	6	0.620	0.687	0.652	611
Accuracy		–	–	0.665	3589
Macro avg.		0.650	0.647	0.648	3589
Weighted avg.		0.663	0.665	0.663	3589
DLF + softplus	0	0.604	0.488	0.540	498
	1	0.653	0.615	0.634	52
	2	0.589	0.486	0.533	545
	3	0.831	0.863	0.846	881
	4	0.517	0.590	0.551	588
	5	0.801	0.698	0.746	414
	6	0.564	0.684	0.618	611
Accuracy				0.656	3589
Macro avg.		0.651	0.632	0.638	3589
Weighted avg.		0.660	0.656	0.654	3589
DLF + Relu	0	0.593	0.536	0.563	498
	1	0.655	0.692	0.673	52
	2	0.550	0.505	0.526	545
	3	0.854	0.834	0.844	881
	4	0.521	0.580	0.549	588
	5	0.738	0.771	0.754	414
	6	0.589	0.614	0.601	611
Accuracy				0.654	3589
Macro avg.		0.643	0.647	0.644	3589
Weighted avg.		0.656	0.654	0.654	3589

accuracy achieved by DLF + Elu model on 60th epoch and showed the correct prediction of facial expression on FER2013 dataset. Elu and ReLu-based models have outperformed the others with 66.50% and 65.40% accuracy, respectively. Different quantitative and qualitative performance measures have supported our proposed

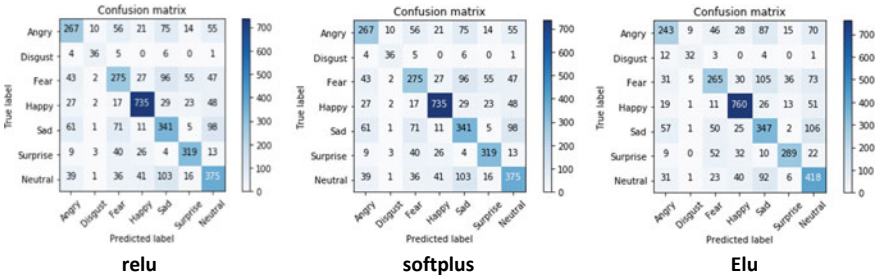

Fig. 6 Confusion matrix of proposed model on ReLu, Elu and Softplus

Table 3 Accuracy comparison with other existing model

Model/Research work	Accuracy
VGG + SVM [9]	66.31
GoogLeNet [10]	65.20
Ergen et al. [11]	57.10
Proposed layered approach on ReLu	67.99
Proposed layered approach on Elu	68.11
Proposed layered approach on Softplus	67.09

model. Our extended deep neural network approach with Elu, ReLu activation function has shown modest performance than GoogLeNet and VGG + SVM-based model.

Fusion of model is being used in the latest research era. FER accuracy can be increased with this method. This model may be fused with one or more than one model, and it can deliver some significant results. In future work, we will also try to combine two or more than two models and analyze its effect on facial expression recognition.

References

1. Pantic M, Rothkrantz L (2000a) Automatic analysis of facial expressions: the state of art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
2. Fasel JL (2003) Automatic facial expression analysis: a survey. Pattern Recogn 36:259–275
3. Pantic M, Rothkrantz L (2003) Toward an affect-sensitive multimodal human–computer interaction. In: Proceeding of the IEEE, vol 91, pp 1370–1390
4. Tian Y, Kanade T, Cohn J (2005) Handbook of face recognition. Springer (Chap. 11. Facial expression analysis)
5. Yacoob Y, Davis LS (1996) Recognizing human facial expression from long image sequences using optical flow. IEEE Trans Pattern Anal Mach Intell 18(6):636–642
6. Essa I, Pentland A (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7):757–763

7. Lyons MJ, Budynek J, Akamatsu S (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Intell 21(12):1357–1362
8. Donato G, Bartlett M, Hager J, Ekman P, Sejnowski T (1999) Classifying facial actions. IEEE Trans Pattern Anal Mach Intell 21(10):974–989
9. Pantic M, Rothkrantz L (2000b) Expert system for automatic analysis of facial expression. Image Vis Comput 18(11):881–905
10. Tian Y, Kanade T, Cohn J (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
11. Cohen I, Sebe N, Garg A, Chen L, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst 91:160–187
12. Yin L, Loi J, Xiong W (2005) Facial expression representation and recognition based on texture augmentation and topographic masking. ACM Multimedia
13. Yeasin M, Bullot B, Sharma R (2004) From facial expression to level of interests: a spatio-temporal approach. IEEE conference on computer vision and pattern recognition (CVPR)
14. Burrows AM, Waller BM, Parr LA, Bonar CJ (2006) Muscles of facial expression in the chimpanzee (pan troglodytes): descriptive, comparative, and phylogenetic contexts. J Anat 208:153–167
15. Hoey J, Little JJ (2004) Value directed learning of gestures and facial displays. In: IEEE conference on computer vision and pattern recognition (CVPR)
16. Chang Y, Hu C, Turk M (2004) Probabilistic expression analysis on manifolds. In: IEEE conference on computer vision and pattern recognition (CVPR)
17. Hu J, Yu B, Yang Y, Feng B (2019) Towards facial de-expression and expression recognition in the wild. In: 2019 8th international conference on affective computing and intelligent interaction (ACII), Cambridge, United Kingdom, pp 157–163
18. Zhong L, Liu Q, Yang P, Huang J, Metaxas DN (2015) Learning multiscale active facial patches for expression analysis. IEEE Trans Cybern 45(8):1499–1510
19. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58
20. Bengio, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
21. Sariyanidi, Gunes H, Cavallaro A (2017) Learning bases of activity for facial expression recognition. IEEE Trans Image Process 26(4):1965–1978
22. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2168–2177
23. Abdulrahman M, Eleyan A (2015) Facial expression recognition using Support Vector Machines. In: 2015 23nd signal processing and communications applications conference (SIU), Malatya, pp 276–279
24. Tian Y (2004) Evaluation of face resolution for expression analysis. In: CVPR workshop on face processing in video
25. Zhang Z, Lyons MJ, Schuster M, Akamatsu S (1998) Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: IEEE international conference on automatic face and gesture recognition (FG)
26. Georgescu M-I, Ionescu RT, Popescu M (2018) Local learning with deep and handcrafted features for facial expression recognition. arXiv:1804.10892
27. Giannopoulos P, Perikos I, Hatzilygeroudis I (2018) Deep learning approaches for facial emotion recognition: a case study on FER-2013. In: Advances in hybridization of intelligent methods. Springer, Cham, pp 1–16
28. Tumen V, Soylemez OF, Ergen B (2017) Facial emotion recognition on a dataset using convolutional neural network. In: Proceedings of the artificial intelligence and data processing symposium, pp 1–5
29. Mase K (1991) Recognition of facial expression from optical flow. IEICE Trans Inf Syst E74-D(10):3474–3483

30. Zhao G, Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
31. Walecki R, Rudovic O, Pavlovic V, Pantic M (2015) Variable-state latent conditional random-elds for facial expression recognition and action unit detection. In: Proceedings of 11th IEEE international conference on workshops automat. Face Gesture Recognit. (FG), May 2015, pp 1–8
32. Yang J, Zhang F, Chen B, Khan SU (2019) Facial expression recognition based on facial action unit. In: 2019 tenth international green and sustainable computing conference (IGSC), Alexandria, VA, USA, pp 1–6