

Emotion Detection Using Deep Learning Algorithm

Shital S. Yadav¹, Anup S. Vibhute²

^{1,2} Department of Electronics and telecommunication Engineering, SVERI's College of Engineering,
Pandharpur, Maharashtra, India.

ABSTRACT: Automatic emotion detection is a key task in human machine interaction, where emotion detection makes system more natural. In this paper, we propose an emotion detection using deep learning algorithm. The proposed algorithm uses end to end CNN. To increase computational efficiency of the deep network, we make use of trained weight parameters of the Mobile Net [1] to initialize the weight parameters of our system. To make our system independent of the input image size, we place global average pooling layer On top of the last convolution layer of it. Proposed system is validated for emotion detection using two benchmark datasets viz. Cohn–Kanade+ (CK+) and Japanese female facial expression (JAFPE). The experimental results show that the proposed method outperforms the other existing methods for emotion detection.

KEYWORDS: Emotion detection, Human machine interaction, deep learning

1. INTRODUCTION

Through outside agents such as speech, gestures, and facial expressions, individuals communicate their real intent and emotions. Facial features provide valuable information on an individual's emotional state, theory of mind, public persona, and psychology [2]. Fully automated Facial Expression Recognition (FER) is a non-intrusive approach to the analysis of human affective behavioral patterns. The FER device plays a key role in human–computer interaction, tracking, deception or lie detection, behavioral profiling, and healthcare applications. Automated emotion study has received increasing attention in recent years because of a very diverse variety and continued development in such applications.

In [3], authors make use of facial action coding system (FACS) for emotion detection. This study witnessed the characterization of emotions is approximately same across the globe. Also, they categorized the human emotions into anger, sad, fear, happy, disgust, and surprise. A traditional emotion detection approach includes (1) image acquisition (2) image pre-processing (3) feature extraction and (4) classification (emotion detection). Accuracy of such traditional emotion detection system is depends upon the robustness of feature extraction and classification stage.

In this paper, we proposed an emotion detection approach using convolution neural network (CNN). The proposed end-to-end CNN is therefore named as ENet. Rest of the manuscript is organized as follow:

Section 1 introduces the automatic emotion detection system, Section 2 enlightens the existing approaches for emotion detection, the proposed approach for emotion detection is discussed in the Section 3. Training details of the proposed ENet are given in the Section 4. The experimental analysis has

been carried out in Section 5. Conclusion is drawn in Section 6.

2. LITERATURE SURVEY

A general FER follows five stages which includes the task of image capturing, the creation of pre-processing techniques, effective feature extraction, recognition and post- processing. The usefulness of such a structure depends to a large degree on the exact mechanism of abstraction and classification of features. Even whilst using the best classification model, insufficient extraction of the function will degrade the efficiency. For a reliable FER system, developing a suitable characteristic descriptor is indeed vital.

Techniques for feature extraction process may be generally grouped into two types: handmade features and [4] learned features. The handmade features are well before-designed to capture specific facial expressions while the learned features are coded utilizing convolution neural networks (CNN). The CNN based methods [5–17] jointly learn to classify the facial expression through the correct attributes and weights. Handmade features proposed in the existing method broadly comes under appearance based features and geometric features. The geometric features [18, 19] encode the face image with the help of geometric properties like deformation, contour, and various other geometric properties. Zhang et al. [20] represented face image by 34 facial points and utilized them as a landmark points. Further, these landmark points are used to extract geometric features. Valstar et al. [21] proposed to track the facial points and detect the AUs (Action Units) in the face image. The facial expressions can be recognized based on the detected AUs in the image. The geometric features fail to

identify the minute characteristics such as ridges and skin texture changes and are dependent on reliable and accurate feature detection and tracking. In addition, preprocessing techniques are required to localize various facial components before the extraction of facial features.

Appearance based methods have been widely used to measure the physical appearance of a facial image. Especially, the local feature descriptor based methods for facial appearance analysis have gained popularity due to their ease of implementation, pose invariance, and robustness to illumination. These methods capture the spatial topology of the image in the local neighborhood. Shan et al. [22] used Local Binary Pattern (LBP) to extract the facial features for expression recognition. Lai et al. [23] employed a two stage feature extraction; first, they retrieved threshold LBP responses and then applied centre symmetric LBP. In local directional patterns (LDP) [24], the local edge responses in eight directions were computed using eight different masks. Extraction of the salient directional responses increases the

discriminative capability of the descriptors. Rivera et al. [25] represented the directional information by encoding a more discriminative Local Directional Number (LDN) patterns. Rivera et al. [26] proposed to extract the texture information by identifying the principal directions and encoded the intensity variation of the principal directions into response numbers called Local Directional Ternary Pattern (LDTP). Ryu et al. [27] used the ternary patterns to extract the directional information and designed a multi-level grid based approach to characterize the coarse and finer features separately. The coarse grids are used for stable codes, which are closely related to non-expressions, whereas the finer grids are used for active codes which are closely related to expressions. Furthermore, the facial image can be divided into non-overlapping regions to extract features, which increase the performance of the system [22]. The selection of salient grids, size, and location of the grids directly affect the recognition accuracy of the FER system.

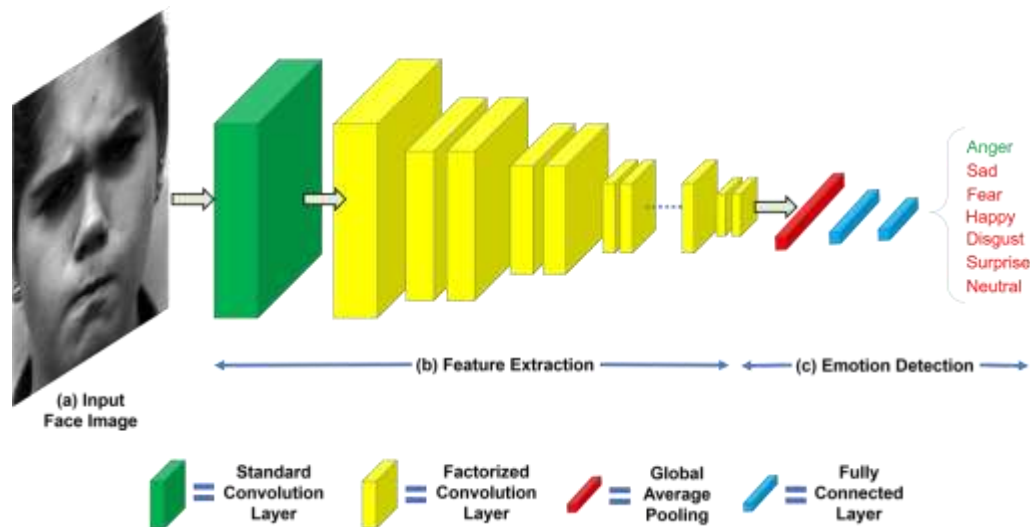


Fig1: End-to-end deep network for emotion detection

Lee et al. [28] proposed to use the sparse representation of images to reduce the intra-class variations of expressions. Mohammadzade and Hatzinakos [29] introduced the concept of expression subspace which represents a particular expression with one subspace and new expressions can be synthesized from an image by applying projections into different expression subspaces. Hybrid methods incorporate various techniques from geometric and appearance-based approaches to attain enhanced performance. Zhang et al. [30] proposed to capture facial movement features based on distance features. These distance features are computed by extracting the salient patch-based Gabor features. Happy et al. [31] extracted the salient patches from various active facial patches during the emotion elicitation. They further extracted LBP features from the salient patches to generate the feature vector and classified the expression images using the SVM classifier. Furthermore, many significant works for the advancement of FER systems were done in [32–

34].

More recent works in facial expression analysis have used deep learning approaches to solve the recognition problem. The deep learning methods learn both the feature extraction and the network weight parameters for accurate classification using the training data. Burkert et al. [5] designed a deep convolutional neural network (CNN) inspired by Google Net and introduced a parallel feature extraction (FeatEx) block to extract features at different scales. Mollahosseini et al. [6] applied the concept of network in network architecture and added inception layers after two traditional CNN layers. The inception layers are concatenated as output and connected to the fully connected layers. Barsoum et al. [7] trained a customized VGG13 network to verify different crowd sourced label distribution techniques and undertook facial expression classification as a case study. Hasani and Mahoor [8] designed a network consisting of 3D inception ResNet layers followed by a Long Short Term Memory (LSTM)

module. These layers extract both the spatial and temporal relations in the face image and frame sequences in the video, respectively. A two step training method was proposed by Ding et al. [9] where, in the first stage, the convolution layer weights are regularized and in the second stage, the fully connected layers are added to the pre-tuned convolution layers and train the complete network to learn the optimal classification parameters. A combination of multiple CNN architectures is yielding better classification accuracy. With this consideration, Pons et al. [10] proposed to improve the FER accuracy by supervised learning of committee of CNNs. Kim et al. [11] combined the decisions from a hierarchical committee of CNNs and hand-crafted hierarchical decision rule. Other CNN based systems such as VGG [12], ResNet [13], DTAGN [14], DTAGN-Joint [15], spatio-temporal [16] and GCNet [17] have also achieved accelerated growth in the field of FER.

3. PROPOSED METHOD FOR EMOTION DETECTION

In this section, the proposed approach for emotion detection is discussed. We propose an end-to-end convolution neural network named as ENet for emotion detection from images. Keeping in mind the computational efficiency of the deep network, we make use of pre- trained weight parameters of the existing MobileNet [1] to initialize the weight parameters of the ENet. MobileNet is a deep network designed for object recognition. Due to the following facts we choose pre-trained weight parameters of the MobileNet:

- It is trained on large-scale dataset for object recognition task. It is compact and computationally less expensive deep network as compared to the other existing deep networks [6, 12, 13].

The proposed ENet is divided into two parts *viz.* (1) feature extraction (2) emotion detection.

3.1 Feature Extraction

The basic building blocks of CNN to extract features are

convolution, pooling and rectified linear unit (ReLU) layers. A standard convolution layer both filters and combines input feature maps into a new set of features in one step. This turns into the increase in number of multiplications and additions in performing convolution operation. In [1], authors showed an effective way to reduce the number of computations by splitting the standard convolution layer operation into two layers without much affecting the accuracy. Thus, in ENet, instead of using standard convolution filters, we use factorized convolution filters [1] which factorize a standard convolution into a depth-wise convolution and a point-wise convolution.

Here, depth-wise convolution applies a single filter to each input feature map while the point-wise convolution applies a 1 1 convolution to combine the outputs of the depth-wise convolution. This factorization has the effect of drastically reducing computation and model size. Feature maps of both the layers are passed through the batch norm and ReLU nonlinearities. To reduce the number of computations further, each convolution layer with stride 2 is used instead of convolution followed by pooling layer. These building blocks are used to make feature extraction module of ENet for emotion detection. The detailed network architecture is given in Table 1.

3.2 Emotion Detection

To detect the human emotion from the extracted features, we employed global feature average layer followed by the classification layer on top of the extracted feature maps.

3.2.1 Global Average Pooling (GAP)

is an operation that calculates the average output of each feature map in the previous layer. This fairly simple operation reduces the data significantly and prepares the model for the final classification layer. Also, due to the averaging operation over the feature maps, the model becomes more robust to spatial translations in the data.

3.2.2 Classification layer
Comprise of seven neurons belongs to seven emotions *viz.* anger, sad, fear, happy, disgust and surprise. SoftMax layer is employed as an activation function

Table 1: Network architecture details of the proposed ENet for emotion recognition.

Feature Extraction	Input Size	Type/Stride	Filter Shape	Output Size
	128x128x3	Conv/2	3x3x3x32	64x64x32
	64x64x3	Conv dw/1	3x3x32 dw	64x64x32
	64x64x3	Conv pw/1	1x1x32x64	64x64x64
	64x64x64	Conv dw/2	3x3x64 dw	32x32x64
	32x32x64	Conv pw/1	1x1x64x128	32x32x128
	32x32x128	Conv dw/1	3x3x128 dw	32x32x128
	32x32x128	Conv pw/1	1x1x128x128	32x32x128
	32x32x128	Conv dw/2	3x3x128 dw	16x16x128
	16x16x128	Conv pw/1	1x1x128x256	16x16x256
	16x16x256	Conv dw/1	3x3x256 dw	16x16x256
	16x16x256	Conv pw/1	1x1x256x256	16x16x256

“Emotion Detection Using Deep Learning Algorithm”

Emotion Detection	16x16x256	Conv dw/2	3x3x256 dw	8x8x256
	8x8x256	Conv pw/1	1x1x256x512	8x8x512
	8x8x512	Conv dw/1	3x3x512 dw	8x8x512
	8x8x512	Conv pw/1	1x1x512x512	8x8x512
	8x8x512	Conv dw/2	3x3x512 dw	4x4x512
	4x4x512	Conv pw/1	1x1x512x1024	4x4x1024
	4x4x1024	Conv dw/1	3x3x1024 dw	4x4x1024
	4x4x1024	Conv pw/1	1x1x1024x1024	4x4x1024
	4x4x1024	GAP	-	1024
	1x1x1024	FC	1024x512	1x1x512
	1x1x512	FC	512x128	1x1x128
	1x1x128	FC	128x7	1x1x7
	1x1x7	Classifier	SoftMax	

*GAP: Global Average Pooling Layer

*FC: Fully Connected Layer

4. EXPERIMENTAL ANALYSIS

In this Section, we have validated the propose ENet for emotion detection from images. We have considered two benchmark datasets for the evaluation *viz.* Cohn–Kanade+ (CK+) [35] and Japanese female facial expression (JAFPE) . The considered evaluation measure is emotion recognition accuracy which is formulated as:

$$ERA = \frac{N1}{N} \dots\dots\dots [1]$$

Where ERA - Emotion Recognition Accuracy

N1 - Total No of Correctly detected samples

N - Total no of Samples

In this study, image sets are randomly divided into N parts for getting better classification accuracy. The N-1 parts are used as a training set and rest is used as a testing set. We have randomly divided the dataset into a ratio of 80:20 and selected the training and testing set images, respectively. The final recognition accuracy is calculated by taking the average of the accuracy produced after five iterations.



Fig 2: Sample images from each category of ck+ and JAFPE dataset

4.1 Results on Jaffe Database

The (JAFPE) [36] dataset contains 213 facial images of ten Japanese females. The subjects posed for neutral and six basic facial expressions. Each expression set consists of an almost same number of images. The facial images were captured from the frontal view and hair of the female subjects were tied back to expose all the expressive zones of facial region. Sample image from each emotion category are shown in the Figure 2. We have given the average recognition accuracy for

six class and seven class problem over JAFPE dataset in Table 2. The proposed method achieves better recognition accuracy as compared to state of the art handcrafted approaches as well as some of the deep learning techniques. More specifically, the proposed method attain recognition rate improvement of 10.8 and 9.3% for the six class problem and 11.1 and 9.2% for the seven class problem over LBP [22] and LDN [25], respectively.

Table 2: Emotion recognition accuracy of the proposed ENet and existing approaches on CK+ and JAFFE dataset

Method	CK+		JAFFE	
	6EX	7EX	6EX	7EX
LBP [22]	93.50	89.00	85.20	84.30
two-phase [23]	88.20	79.50	83.30	80.50
LDP [24]	96.20	92.90	85.00	80.50
LDN [25]	94.80	91.70	86.70	86.20
LDTP [26]	95.30	91.90	83.30	82.90
LDTerP [27]	95.70	91.50		
VGG16 [12]	96.70	95.20	61.10	73.80
VGG19 [12]	97.20	81.20	66.70	73.80
ResNet50 [13]	94.00	91.80	75.00	64.30
RADAP [37]	96.20	94.70	95.00	88.10
XRADAP [37]	96.60	93.80	88.40	88.10
ARADAP [37]	96.20	93.80	91.10	88.10
DRADAP [37]	96.00	95.40	93.90	90.50
ENet	98.44	97.22	96.00	95.40

4.2 Results on CK+ Database

The CK+ [35] [38] database includes 593 image sequences of 123 different subjects. The subjects are of American, African–American, Asian and Latin origin. Each sequence starts with the neutral state and ends at the apex of an expression. The dataset provides frontal facial images of six expressions: anger, disgust, fear, happy, sad, and surprise. We selected three apex frames from each sequence to prepare the image set for an expression class. We also collected the neutral state images from the onset of the image sequences to create a neutral image set. Finally, we augmented a total of 1043 images: 132-anger, 180-disgust, 75-fear, 204-happy, 87-sad, 249-surprise, and 116-neutral. Sample image from each emotion category are shown in the Figure 2.

In our experiments, in order to compare the performance of the proposed ENet with other deep learning techniques, we have trained the VGG16, VGG19 and ResNet50 networks over the CK+ dataset. The pre-trained weights from ImageNet were used as initial weight parameters while fine-tuning these networks. We have measured the performance of the FER system in terms of average recognition accuracy. In Table 2, we have shown a comprehensive comparative analysis of the proposed methods with state-of-the-art approaches including recent deep learning methods. From Table 2, it is evident that the proposed methods achieve superior recognition accuracy as compared to the existing feature descriptors. It also outperforms ResNet50 for both 6-class and 7-class recognition problem for emotion recognition

5. CONCLUSION

In this paper, we have proposed an emotion recognition network named as ENet. To make the ENet computationally efficient, we make use of factorized convolution layer. The global average pooling layer is employed on top of the

extracted feature maps to make the proposed ENet invariant to the spatial translations in the data. Experimental analysis has been carried out on two benchmark datasets used in the emotion recognition studies viz. CK+ and JAFFE. The performance of the proposed ENet compared with the existing deep learning as well as non-deep learning based approaches. Comparative analysis shows that the proposed ENet outperforms the other existing approaches for emotion recognition.

REFERENCES

1. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.
2. G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Classifying facial actions,” IEEE Transactions on pattern analysis and machine intelligence, vol. 21, no. 10, pp. 974–989, 1999.
3. P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” Journal of personality and social psychology, vol. 17, no. 2, p. 124, 1971.
4. C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 8, pp. 1548–1568, 2016.
5. P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, “Dexpres- sion: Deep convolutional neural network for expression recognition,” arXiv preprint arXiv:1509.05371, 2015

7. A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in 2016 IEEE Winter conference on applications of computer vision (WACV). IEEE, 2016, pp. 1–10.
8. E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.
9. B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3d convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 30–40.
10. H. Ding, S. K. Zhou, and R. Chellappa, “Facenet2expnet: Regularizing a deep face recognition net for expression recognition,” in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 118–126.
11. G. Pons and D. Masip, “Supervised committee of convolutional neural networks in automated facial expression analysis,” IEEE Transactions on Affective Computing, vol. 9, no. 3, pp. 343–350, 2017.
12. B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition,” Journal on Multimodal User Interfaces, vol. 10, no. 2, pp. 173–189, 2016.
13. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
14. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
15. H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, “Deep temporal appearance-geometry network for facial expression recognition,” arXiv preprint arXiv:1503.01532, 2015.
16. H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.
17. K. Zhang, Y. Huang, Y. Du, and L. Wang, “Facial expression recognition based on deep evolutionary spatial-temporal networks,” IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4193–4203, 2017.
18. Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, “Deep generative-contrastive networks for facial expression recognition,” arXiv preprint arXiv:1703.07140, 2017.
19. M. Pantic and I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 2, pp. 433–449, 2006.
20. N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, “Authentic facial expression analysis,” Image and Vision Computing, vol. 25, no. 12, pp. 1856–1863, 2007.
21. Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron,” 05 1998, pp. 454 – 459.
22. M. F. Valstar, I. Patras, and M. Pantic, “Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)-Workshops. IEEE, 2005, pp. 76–76.
23. C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image and vision Computing, vol. 27, no. 6, pp. 803–816, 2009.
24. C.-C. Lai and C.-H. Ko, “Facial expression recognition based on two-stage features extraction,” Optik, vol. 125, no. 22, pp. 6678–6680, 2014.
25. T. Jabid, M. H. Kabir, and O. Chae, “Robust facial expression recognition based on local directional pattern,” ETRI journal, vol. 32, no. 5, pp. 784–794, 2010.
26. A. R. Rivera, J. R. Castillo, and O. O. Chae, “Local directional number pattern for face analysis: Face and expression recognition,” IEEE transactions on image processing, vol. 22, no. 5, pp. 1740–1752, 2012.
27. A. R. Rivera, J. R. Castillo, and O. Chae, “Local directional texture pattern image descriptor,” Pattern Recognition Letters, vol. 51, pp. 94–100, 2015.
28. B. Ryu, A. R. Rivera, J. Kim, and O. Chae, “Local directional ternary pattern for facial expression recognition,” IEEE Transactions on Image Processing, vol. 26, no. 12, pp. 6006–6018, 2017.
29. S. H. Lee, K. N. K. Plataniotis, and Y. M. Ro, “Intra-class variation reduction using training expression images for sparse representation based facial expression recognition,” IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 340–351, 2014.
30. H. Mohammadzade and D. Hatzinakos, “Projection into expression subspaces for face recognition from single sample per person,” IEEE Transactions on Affective Computing, vol. 4, no. 1, pp. 69–82, 2012.

31. L. Zhang and D. Tjondronegoro, “Facial expression recognition using facial movement features,” *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 219–229, 2011.
32. S. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2014.
33. T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, “A deep neural network-driven feature learning method for multi-view facial expression recognition,” *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.
34. W. Zheng, Y. Zong, X. Zhou, and M. Xin, “Cross-domain color facial expression recognition using transductive transfer subspace learning,” *IEEE transactions on Affective Computing*, vol. 9, no. 1, pp. 21–37, 2016.
35. O. Rudovic, M. Pantic, and I. Patras, “Coupled gaussian processes for pose-invariant facial expression recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1357–1369, 2012.
36. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “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*, 2010, pp. 94–101.
37. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.
38. M. Mandal, M. Verma, S. Mathur, S. K. Vipparthi, S. Murala, and D. K. Kumar, “Regional adaptive affinitive patterns (radap) with logical operators for facial expression recognition,” *IET Image Processing*, vol. 13, no. 5, pp. 850–861, 2019.
39. T. Kanade, J. F. Cohn, and Yingli Tian, “Comprehensive database for facial expression analysis,” in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 46–53.