Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.035356
Article

Facial Emotion Recognition Using Swarm Optimized Multi-Dimensional DeepNets with Losses Calculated by Cross Entropy Function

A. N. Arun[1,*], P. Maheswaravenkatesh[2] and T. Jayasankar[2]

[1]Department of Computer Science & Engineering, Sri Venkateswara Institute of Science and Technology, Tiruvallur, Tamilnadu, India
[2]Department of Electronics and Communication Engineering, Anna University, Trichy, Tamilnadu, India
*Corresponding Author: A. N. Arun. Email: arun.svist.cse@gmail.com
Received: 17 August 2022; Accepted: 28 December 2022

Abstract: The human face forms a canvas wherein various non-verbal expressions are communicated. These expressional cues and verbal communication represent the accurate perception of the actual intent. In many cases, a person may present an outward expression that might differ from the genuine emotion or the feeling that the person experiences. Even when people try to hide these emotions, the real emotions that are internally felt might reflect as facial expressions in the form of micro expressions. These micro expressions cannot be masked and reflect the actual emotional state of a person under study. Such micro expressions are on display for a tiny time frame, making it difficult for a typical person to spot and recognize them. This necessitates a place for Machine Learning, where machines can be trained to look for these micro expressions and categorize them once they are on display. The study's primary purpose is to spot and correctly classify these micro expressions, which are very difficult for a casual observer to identify. This research improves upon the accuracy of the recognition by using a novel learning technique that not only captures and recognizes multimodal facial micro expressions but also has features for aligning, cropping, and superimposing these feature frames to produce highly accurate and consistent results. A modified variant of the deep learning architecture of Convolutional Neural Networks combined with the swarm-based optimality technique of the Artificial Bee Colony Algorithm is proposed to effectively get an accuracy of more than 85% in identifying and classifying these micro expressions in contrast to other algorithms that have relatively less accuracy. One of the main aspects of processing these expressions from video or live feeds is aligning the frames homographically and identifying these concise bursts of micro expressions, which significantly increases the accuracy of the outcomes. The proposed swarm-based technique handles this in the research to precisely align and crop the subsequent frames, resulting in much superior detection rates in identifying the micro expressions when on display.

Keywords: Facial micro expression recognition; deep learning CNN; artificial bee colony

1 Introduction

The face is the index of the mind. As humans, this index is expressed outward by various facial expressions. Some of these expressions reveal our inner emotions, like happiness, sadness, anger, excitement/surprise, fear, disgust, or just being neutral. Emotions thus play a significant role in building a form of a non-verbal cue in society. These expressions and various communication modes give a holistic perspective on the ideas we try to put across. But there are times wherein the actual expression felt by the person may differ from that of what is projected outside. For instance, the person who puts on a happy face on the surface for others to see may internally hide a ton of other contrasting emotions that might be felt within. Even though the brain system tries to conceal these inner feelings, our brains are wired so that, unconsciously, genuine emotions are revealed even without us realizing it. These emotional qualities are known as micro-expressions. The micro-expressions are very brief emotions that might be on display for 0.5 to 1 s at max; hence, only someone specialized and trained to detect them will be able to focus on and find them. Thus, the challenge of finding the micro-expressions falls mainly in Artificial Intelligence and Deep Learning, where research systems train machines to spot these expressions when on display. Micro-expressions knowledge is used in criminology, job recruitment, customer feedback surveys, online learning portals, etc. The model uses Deep Neural Networks to recognize and spot these micro expressional changes. Once the changes are identified, they are mapped to Action Units through a modified swarm-based Artificial Bee Colony algorithm, which effectively categorizes the exhibited micro expressions.

2 Related Works

2.1 Static Facial Expression Analysis

Facial recognition of emotions has long been the field of interest for many research works. The static analysis of facial expression is done by extracting feature sets from images and thereby using algorithms to infer the exhibited expression. Ekman [1], in his work on facial emotions, analyzed the various references on the cross-cultural impacts of different groups of people for similar sets of expression classes. Tang [2], in his work on the FER2013 Kaggle competition, used datasets like the FER2013 Dataset, Japanese Female Facial Expression Dataset, Mobile Web App Dataset, and the Extended Cohn-Kanade Dataset to train his model on static images. The study used a combination of a Neural Network along with a Support Vector Machine model to classify the emotions. Most facial emotion recognition research using static expression analysis used images or discrete video frames captured as images. This leads to a loss of continuity in the expressions being exhibited, where a person might display emotion for less and then change their emotions to an entirely different category of emotions. Thus, analyzing static images helped pick the emotional cues for discrete emotions. Still, it lacked the continuous monitoring of emotions essential for spotting and correctly identifying micro expressional emotions.

2.2 Dynamic Facial Expression Analysis

The dynamic facial expressions are mainly captured from existing video data or acquired from real-time live video feeds. The main objective of this approach is to first extract the apex frames from the video data and then apply techniques to recognize facial expressions. In analyzing micro expressions, this task is made more difficult to data's temporal nature. Micro expressions being very short-lived in expression, capturing the frames that transform from an existing state of expression to a micro expression and then resuming its prior form becomes difficult. Data sets like In the Wild [3] (ITW). The work of Allaert et al. made significant progress in classifying micro expressions by using the elasticity

and deformations produced in the face when emotion is exhibited and created accuracies close to 70% in the CASME II dataset. Introducing the CK+ dataset with video data on micro expression vastly improved the research on micro expression analysis. With the introduction of standardized deep learning architectures, the field of micro expression analysis gained good momentum in which vastly improved techniques and architectures were employed to increase the accuracy of detection and classification of the micro expressions on a real-time basis. Some of the standard methods that are being utilized for the recognition of micro facial expressions are tabulated in Table 1:

Table 1: Survey of popular micro expression analysis methodologies

Reference and Year	Techniques/Methodology	Outcome
2018 [4]	Affective computing	Accuracy-70.2%
2019 [5]	Local motion patterns (LMP) feature & face skin temporal elasticity and face deformations	CASME II-70.2% SMIC-67.68%
2020 [6]	2D landmark feature map (LFM), convolutional neural network (CNN) and Long Short-Term Memory (LSTM)	Accuracy-71% to 74%
2020 [7]	Pixel-level change rates in the frequency domain	CASME-60.82% CASME II-65.02% SAMM-40.9%
2020 [8]	Cross-database micro-expression recognition (CDMER), domain adaptation (DA), spatiotemporal descriptors	Accuracy-64.01%
2020 [9]	3D morphable model (3DMM), LSTM	Accuracy-51% to 57%
2020 [10]	Local motion patterns (LMP)	CASME II-63.27 SAMM-70.66 SMIC2-64.43
2020 [11]	Deep spatio-temporal geometric features, recurrent neural network	Accuracy-85.5%
2020 [12]	Local-region division and the feature selection & relief algorithm	LBP-TOP descriptor-49.39% HOG-TOP descriptor-54.4%
2021 [13]	AI-based FER methodologies	Accuracy-71.64 %
2021 [14]	Transfer learning convolutional neural network	MMEW-69.4%
2021 [15]	Dual temporal scale convolutional neural network	SAMM-69.2%
2022 [16]	CAS(ME)3: Facial spontaneous micro-expression database with depth information	Work in progress

3 Proposed Methodology

3.1 Data Sets for Facial Expressions

The proposed work is implemented using the following publicly available datasets: FER2013 [17], Google Data Set, CK+ [18], FEC, and the In the Wild dataset. Pierre-Luc Carrier and Aaron Courville developed the FER2013 dataset. The dataset consists of 48 × 48 pixel greyscale images with about

28,709 examples for the training set. The emotions are categorized into 6 classes: Angry, Disgust, Fear, Happy [19], Sad, surprised, and Neutral. The Extended Cohn-Kanade (CK+) Dataset was released as an extension to the original CK dataset. The Ck dataset is one of the most widely used datasets for evaluating facial expressions in the research domain. It consists of 97 subjects recorded over 486 sequences. The frames are labeled using the standardized Facial Action Coding System (FACS). In addition, 26 more topics and 107 series were added to create the CK+ dataset. These classify expressions like Anger, Contempt, Disgust, Fear, Happy [20], Sadness, and Surprise.

3.2 Frame Sequencing

The video sequences are collected from various datasets mentioned above and represented as a normalized dataset. The video sequences are fed into a frame grabber that isolates each frame into a series of images depending on the duration of the video file considered. Each of these frames is then analyzed for the presence of facial topologies; if present, those frames are sequenced and preserved. The rest of the frames that contain non-facial information are discarded.

3.3 Image Alignment

The frames obtained from the above process are aligned using a sequence of basic affine transforms like rotation, scaling, and translation. This changes the input coordinates of the image to the desired output coordinate system that is expected from the resultant reference frames. The Harr Cascade Classifier of the OpenCV toolkit is used to detect the facial and eye regions in the presented structure. The facial region is seen from the rest of the image and is marked. Then within the area, the left and right eye are caught and observed, and their coordinates are calculated as Region of Interest. The mid-point of each of these eye regions is calculated. A line is drawn from the mid-points of these two regions, and the line's slope gives the directionality of the rotation of the image, as shown in Figs. 1 and 2.

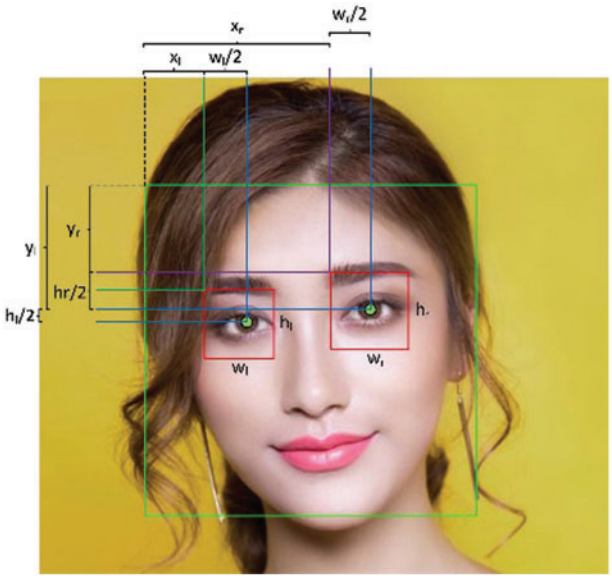

Figure 1: Calculating the alignment of the facial features

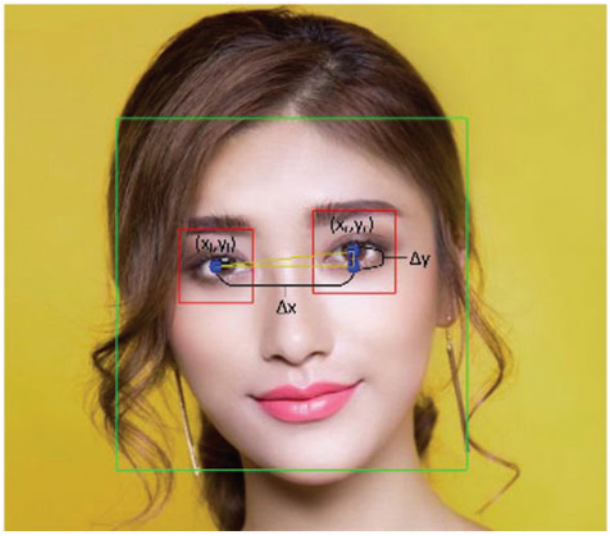

Figure 2: Calculating the rotational coefficient to align frames

For the left eye,

$$x_{cr} = x_r + w_r/2 \tag{1}$$

$$y_{cr} = y_r + h_r/2 \tag{2}$$

And for the right eye,

$$x_{cr} = x_r + w_r/2 \tag{3}$$

$$y_{cr} = y_r + h_r/2 \tag{4}$$

The angle of rotation is computed as

$$\Delta x = xr - x1 \tag{5}$$

$$\Delta y = yr - y1 \tag{6}$$

$$\theta = \arctan\left(\frac{\Delta y}{\Delta x}\right) \tag{7}$$

The angle that is obtained is used in rotating the image for vertical alignment for stacking all related image frames.

3.4 Optical Flow and Homography Using Ensemble Learning

Homography of the different image planes is done after stacking the various facial images one behind the other and looking for points of similarity to precisely align points of interest on the different apex frames. The points or regions of interest are mainly chosen over the eye and nasal regions for defining the homography matrix, as shown in Fig. 3. For instance, if 'x' is a point of interest in one image and 'x'' is another pointin the next image to be aligned, they can be represented as:

$$x = (u, v, 1) \tag{8}$$

$$x' = (u', v', 1) \tag{9}$$

Then the homography matrix 'M' can be represented as

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} \tag{10}$$

where every pixel in the second image x′ is wrapped around the original image x by the homographic transform

$$x' = M.x \tag{11}$$

Figure 3: Rotation and crop of required facial landmark features

The ensemble learning algorithm being used is the Artificial Bee Colony algorithm. This acts as an evolutionary particle swarm optimizer in the way that there are three sets of bees that enable the detection and learning of the optical flow patterns between the consecutive frames of the apex images chosen and homographically aligned [21]. An additional feedback loop is added to the swarm intelligence layer, which acts as a reinforcement loop to increase the efficacy of the inputs supplied for the training based on the losses calculated by a cross-entropy function. Micro-expressions mainly focus on facial landmarks like eyebrows, eyes, and lips. These regions are marked and segmented, and variance in the general optical flow vector is computed.

The Artificial Bee Colony algorithm employs three groups of bees: the employed bees, the onlooker bees, and the scout bees. This algorithm is used to train the bees in detecting the apex frames from the video segments that display the micro expressions and then to relate the associated structures that are temporally linked with the critical micro expression. The nectar, the food source, represents the video segment's individual frames. The employed bees and the onlooker bees continuously exploit these frames or food sources until all the video frames are processed. The employed bees associate themselves with the food source, and the onlooker bees watch the dance of the employed bees. Thus, they find and align the optimal food source or the video frames that maximize the chance of finding the related micro expression being displayed. Once all the food sources are found, the employed bee becomes a scout bee and waits for the following video frames showing the next micro expression category. Based on the variance of the optical flow vector calculated between consecutive frames, losses are calculated based on the cross entropy function and again fed back to the loop. This reinforces the learning process and thereby trains the bees to become competent by associating themselves with important food sources to optimize the solution's output. The scout bees are initialized with the optical vectors of the population of food sources, $\vec{x}_m$'s , (m = 1...SN , SN: frame count). Each food source, $\vec{x}_m$, is a solution vector to the optimization problem, each $\vec{x}_m$ vector holds n variables, ($\vec{x}_{mi}$, I = 1...n), which are to be optimized based on the nature of the received video frames. The employed bees search to find new frames $\vec{v}_m$, calculates

the food sources that have more nectar within the available neighboring structures or food sources, $\vec{x}_m$ and eventually evaluates their fitness values. The fitness values of the frames can be calculated as

$$fit_m\left(\vec{x}_m\right) = \begin{cases} \dfrac{1}{1+f_m\left(\vec{x}_m\right)} & if\ f_m\left(\vec{x}_m\right) \geq 0 \\ 1+\text{abs}\left(f_m\left(\vec{x}_m\right)\right) & if\ f_m\left(\vec{x}_m\right) < 0 \end{cases} \tag{12}$$

where $f_m\left(\vec{x}_m\right)$ is the objective function that is to be optimized based on the value obtained from the solution obtained $\vec{x}_m$. Based on the probability value p_m of the fitness calculated by the employed bee, the onlooker bee chooses the best frame from the given input of video frame sequences taken as inputs.

$$p_m = \frac{fit_m\left(\vec{x}_m\right)}{\sum_{m=1}^{SN} fit_m\left(\vec{x}_m\right)} \tag{13}$$

As more onlooker bees join the search, the resulting solution's quality keeps increasing, and a positive reinforcement behavior is established. The scout bees randomly check for the food sources till all the frames have been exhausted and wait for the following sequence of new structures to be supplied. Solutions with poor fitness value are rejected and removed from the source. This technique dramatically improves the classification accuracy of the result when contrasted with other research solutions because only frames with high fitness constraints that genuinely contribute to the solution's usefulness are retained. The frames with high fitness values are aligned, cropped, and stacked to be given as inputs to the CNN for further categorizing the micro expressions into their respective emotional classes.

The images are aligned and resized to 127 × 127 pixels in dimension, and the optical flow between each consecutive apex frame is determined, as shown in Fig. 4. The optical flow is computed using the Lucas Kanade feature tracker method [22]. The motion estimation is calculated as displacing the pixels on one frame of the apex image to the next successive frame [23]. The magnitude of the vector implies the rate of change of the facial features as a function of time and is modeled as a partial differential equation in Fig. 5.

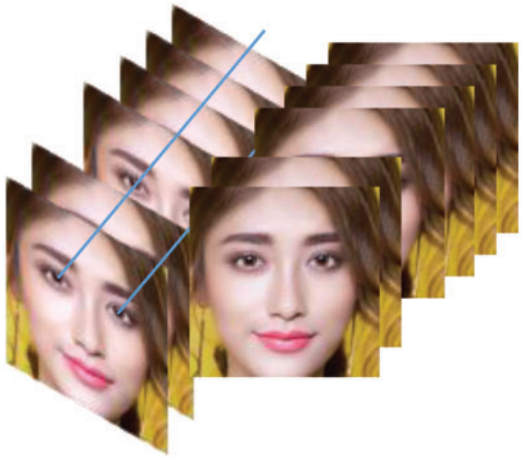

Figure 4: Homographic alignment of consecutive frames

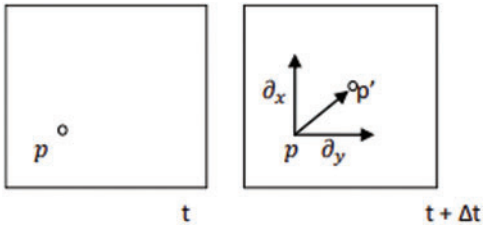

Figure 5: Calculating the vector of rate of the change of facial features

The matrix form of the transform can be described as

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_1) \end{bmatrix} v = \begin{bmatrix} V_x \\ V_y \end{bmatrix} b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(qn) \end{bmatrix} \tag{14}$$

where the partial derivatives of image I, are $I_x(q_1)$, $I_y(q_1)$, and $I_t(q_1)$ concerning the positions (x,y) and the time t, of the current frame. The translation of points in the image is denoted as

$$x' = x + t, x' = \begin{bmatrix} I & t \end{bmatrix} \bar{x} \tag{15}$$

where I is the identity matrix.

3.5 FACS Coding System

The Facial Action Coding System (FACS) was initially developed by Carl-Herman Hjortsjö [24] as a template to taxonomize the movements of the facial muscles in human beings. The Facial Action Coding System (FACS) is a standardized marker for analyzing facial emotion features. This constantly evolved to be one of the standards to categorize the expressions exhibited by the face and facial structures. The coding of the muscle groups is used as an index to represent the facial expressions called the Action Units (AU). This index can identify specific facial regions of interest, which could be tagged as particular expressions. The minute changes in the facial features of the micro expressions are captured and aggregated to form a model organized into one or more of 44 Action Units and supplied later as inputs to the CNN for training. The emotion-related action units can be tabulated, as shown in Table 2:

Table 2: Macro and micro emotion action units

Emotion	Macro action units	Micro action units
Happiness	6 + 12	AU6/AU12/AU6 + AU12
Sadness	1 + 4 + 15	AU14/AU17/AU14 + 17/AU1 + 2
Surprise	1 + 2 + 5B + 26	AU5/AU26/AU1 + AU2/AU18
Fear	1 + 2 + 4 + 5 + 7 + 20 + 26	AU4 + 5/AU20/AU4 + 11/AU14
Anger	4 + 5 + 7 + 23	AU17 + AU13/AU7/AU9
Disgust	9 + 15 + 17	AU9/AU10/AU4 + 7/AU4 + 9

A total of 16 AU are used to map the facial expressions that belong to a particular class of emotions. In addition, relaxed facial features like brows, eyelids, nose, lips, cheeks, and jaw are considered to be the baseline for a neutral emotion. All these 22 emotional classes are represented

in Table 3, of an Image 'I' are encoded as an attribute vector 'a' and are computed as a sparse matrix representing the image 'I' dimensionality.

Table 3: FACS for mapping the action units

Index	Description	Action Unit	Angry	Disgust	Fear	Happiness	Sadness	Surprised	Neutral
1	Inner Brow Raiser	AU1			✓		✓	✓	
2	Outer Brow Raiser	AU2			✓			✓	
3	Brow Lowerer	AU4	✓	✓	✓		✓		
4	Upper Lid Raiser	AU5			✓			✓	
5	Cheek Raiser	AU6				✓	✓		
6	Lid Tightner	AU7	✓		✓				
7	Nose Wrinkler	AU9		✓					
8	Upper Lid Raiser	AU10	✓	✓					
9	Lip Corner Puller	AI12				✓			
10	Lip Corner Depressor	AU15					✓		
11	Chin Raiser	AU17	✓	✓					
12	Lip Stretcher	AU20			✓				
13	Lip Tightner	AU23	✓						
14	Lip Pressor	AU24	✓	✓					
15	Lip Part	AU25			✓	✓	✓	✓	
16	Jaw Drop	AU26			✓			✓	
17	Relaxed Brows	–							✓
18	Relaxed Lids	–							✓
19	Relaxed Nose	–							✓
20	Relaxed Lips	–							✓
21	Relaxed Cheeks	–							✓
22	Relaxed Jaws	–							✓

3.6 Modelling and Training the Convolutional Neural Network

The flow map is supplied as an input to the modified Convolutional Neural Network, wherein the inputs are trained by tagging them based on their respective Action Units. The whole architecture of the system is shown below in Fig. 6:

The internal architecture of the Convolutional Neural Network is designed based on the Inception V3 architecture [25]. This architecture is modified to have a factorized 3×3 convolutional filter for improved granularity of convergence and also has features like label smoothing and auxiliary classifiers to propagate the information obtained from the ensemble learners down the pipeline along the network. A feedback loop is added to the CNN Layer which provides the swarm module with constant reinforcement control to reward and punish behaviors that eventually converge to highly accurate modules for identifying the specific micro-expressional features. With machines getting faster, this considerably reduces the training time of the model as the multi-dimensional boosting and swarm model takes up the heavy load of convolutions that are mapped by using the modified Convolutional Neural Network.

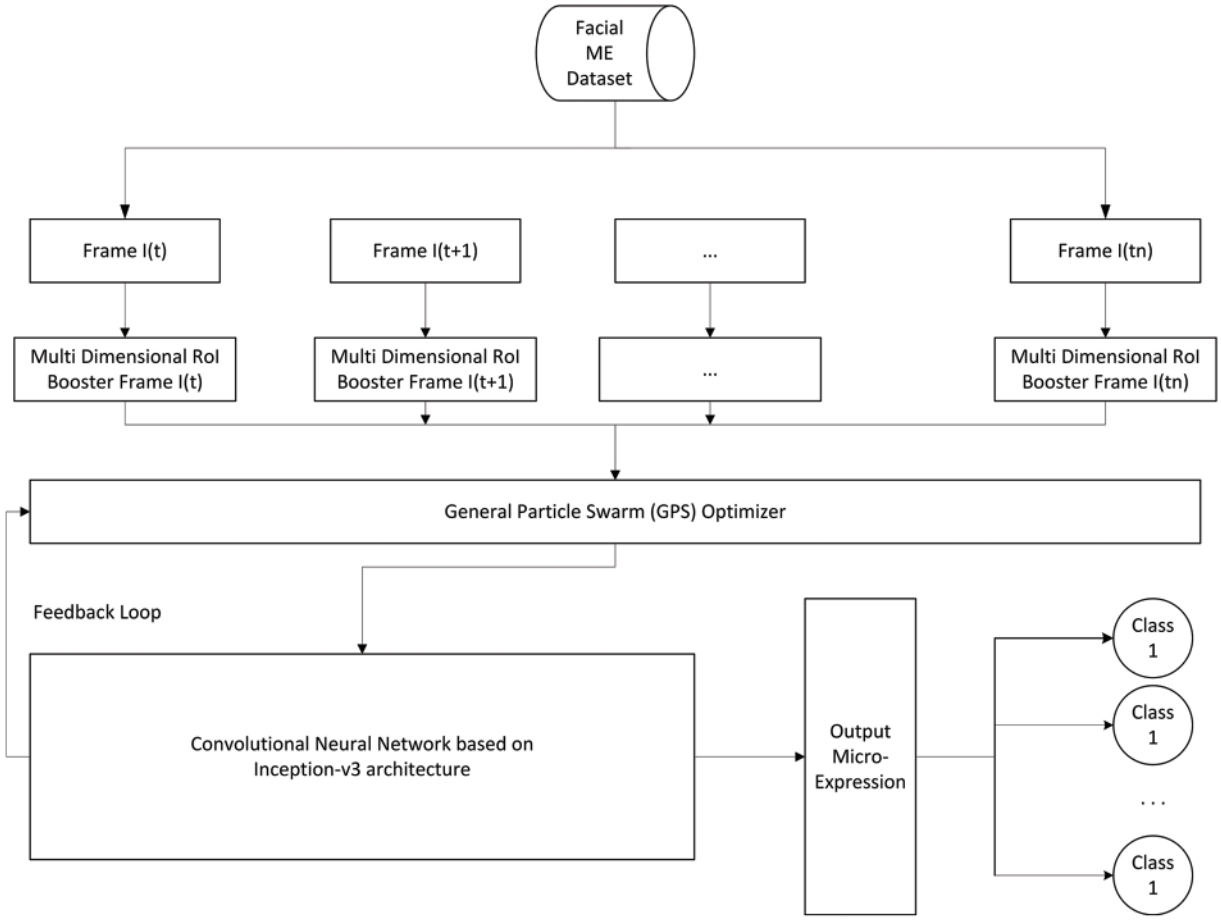

Figure 6: Architecture of the proposed system

The Inception v3 model is chosen for the deep learning model because it consumes less memory in the order of around 92 MB and can have a topological depth of 159 layers and 23 M parameters [26]. The modified CNN uses a 3×3 standard convolutional kernel to train the network and has a Rectified Linear Unit (ReLU) and a pooling layer to flatten the model to be supplied for the fully connected classification layer. Hence the proposed DeepNet model takes less time to train as the taught inputs are already presented by the pre-processing stage, with weightage given to certain special features to look for in the inputs. The modified architecture sub-modules A, C, and E help factorize the parameter values, and sub-modules B and D enable reducing the grid size, as shown below in Fig. 7.

The Inception V3 architecture is a base model because it comprises different filters and stacks up on the different layers after adding nonlinearity to the network. The model typically uses a 1×1 2D convolution filter which results in faster training of the model by shrinking the number of channels, thereby reducing the computational cost and increasing the performance of the design as represented in Fig. 8. The neural network processes the Spatio-temporal feature vectors encoded to produce classes categorized into mapping that is supplied during the training of the network.

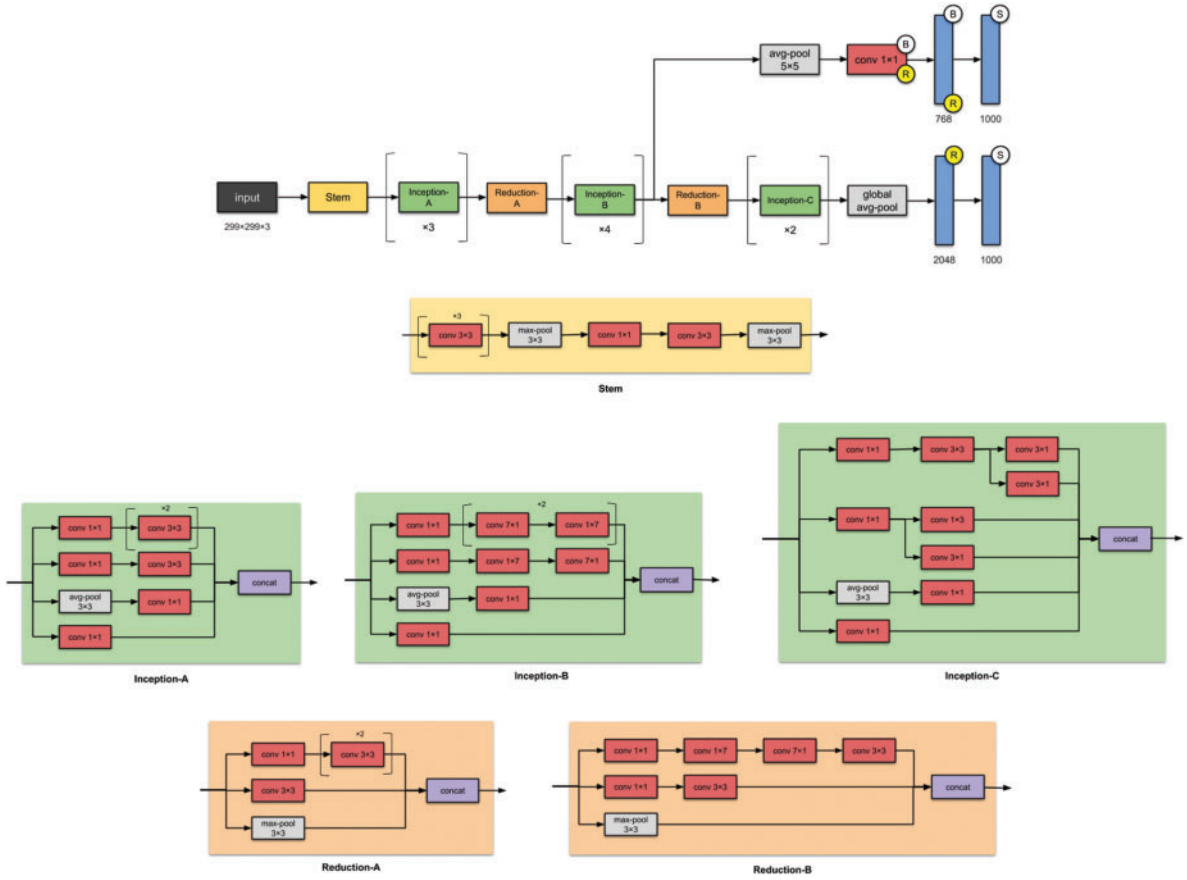

Figure 7: Inception v3–Higher-level architecture

4 Experiments and Results

4.1 Experimental Setup

Experiments were conducted on the following datasets: CK+, In the Wild, and FER2013. These data sets have more than 800 video and image sequences from which a randomized sample of 80%–20% split is used for training and testing. The output of the training was classified into one among the 6 emotions: Angry (A), Disgust (D), Fear (F), Happy (H), Sad (S), Surprise (Su), and Neutral (N). The optical flow characteristics of the images are supplied as inputs to the CNN for training. The video dataset is split into frames and these images are analyzed for the micro expressions on the above categorical classes. The testing was done using the composite data files from In the Wild (ITW), FER 2013, and CK+ datasets. The network can correctly identify the apex frame, as shown in Fig. 9 for the corresponding facial emotion shown in Fig. 10, in which the micro expression is exhibited most of the time with higher levels of accuracy. Table 4 shown the magnitude for all the emotional values being registered.

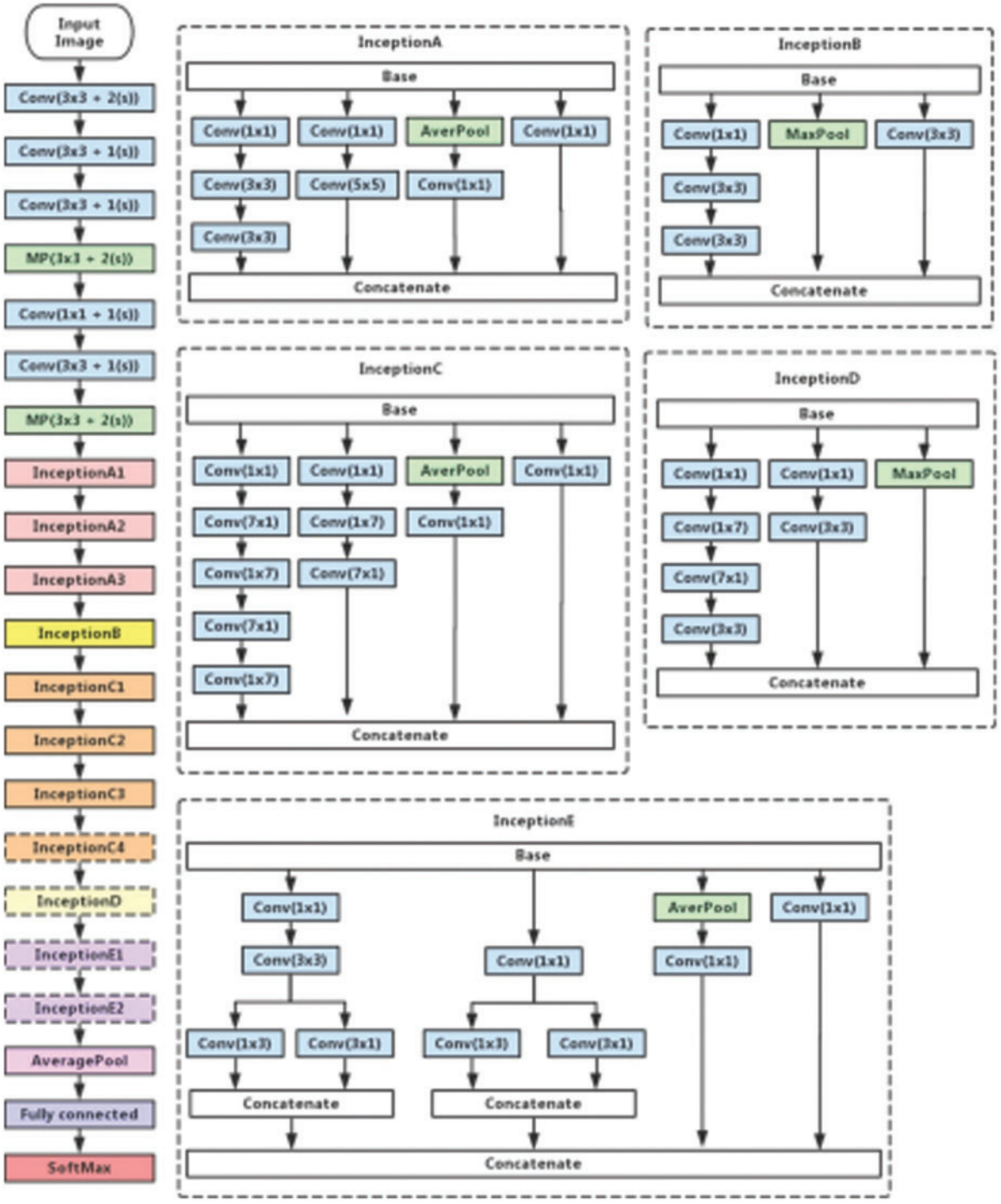

Figure 8: Inception v3–Detailed layer architecture

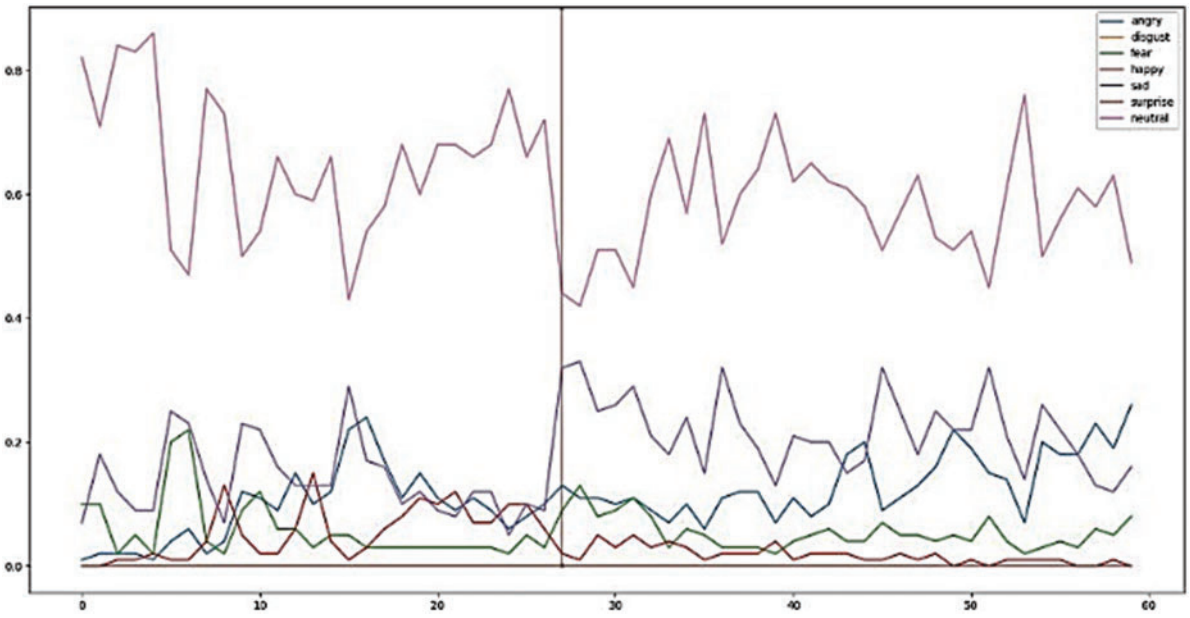

Figure 9: Various micro expressions on display at time, T_n

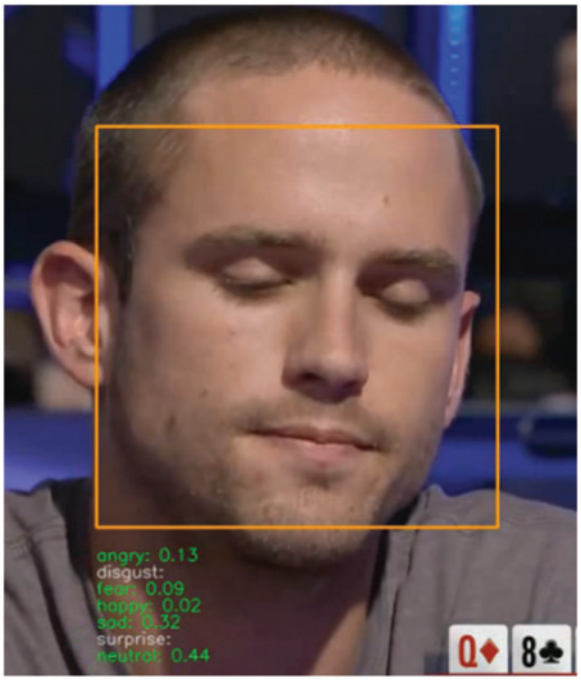

Figure 10: Micro expression on display and various magnitudes identified

Table 4: Emotional classes and relatively scaled magnitudes

	Human emotions	Emotion value from the video
0	Angry	6.93
1	Disgust	0.00
2	Fear	3.36
3	Happy	2.05
4	Sad	10.94
5	Surprise	0.00
6	Neutral	36.74

4.2 Time Constraints

The proposed approach was trained and evaluated on Nvidia GTX 1060 processor and also tested on the TPU architecture of the Google Cloud. The training and validation took approximately 28–30 s for each epoch on the local cluster. There were around 28,709 files that were trained and 7178 files that were tested for accuracy.

4.3 Performance

The model resulted in an overall training accuracy of more than 90% and an average validation accuracy of around 85% for the different classes of micro emotions being recognized after 20 epochs of training as depicted in Table 5. This is a significant improvement over the rest of the models being evaluated.

Table 5: Performance of various techniques on different datasets

	Databases			
Approach	ITW	FER 2013	CK+	CASME II
Proposed method	**83.2**	**85.4**	**84.1**	**87.7**
LBP-TOP	57	52.1	48	38
STCLQ	–	–	64	59.5

The maximum accuracy on the prior approach tends to be in the range of around 38%–64% [27,28]. The training accuracy of the model consistently increases over each epoch, and the validation accuracy closely follows the trend depicting that the model does not overfit the data, as shown in Figs. 11 and 12.

At each epoch, the training & validation losses have also decreased as expected and result in the convergence of weights to best suit the learning parameters of the modeled neural network. The classification accuracy plot is created on the test data comparing the existing models to the proposed model to identify the instances of correctly classifying the different class assignments of the various categories of micro expressions, as displayed in Fig. 13.

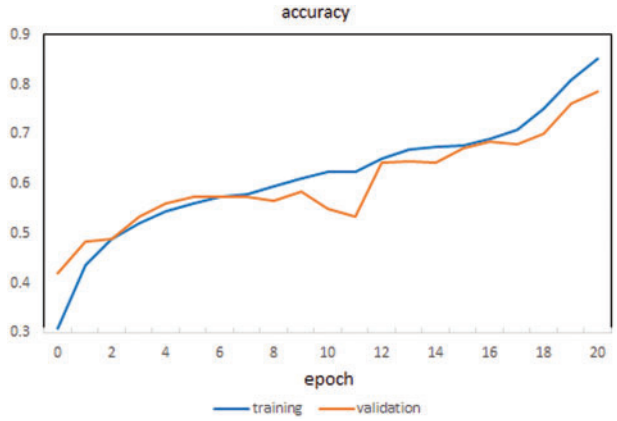

Figure 11: Accuracy of the proposed method

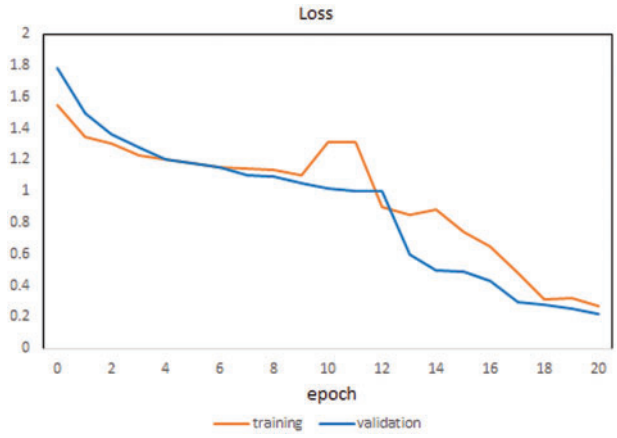

Figure 12: Loss of the proposed method for each epoch

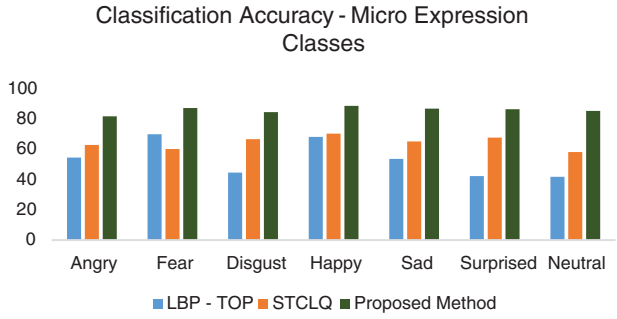

Figure 13: Classification accuracy of different methods

5 Conclusion

The main objective of the research is to improve the recognition and classification capacities of the Neural Network in the field of micro expression analysis. The challenges arise due to the very short duration of capturing the apex frames and the limited number of data sets available for training the network. The proposed model addresses these issues by providing a pre-processing stage by using

additional learners in the form of boosting and swarm-based flow vector detection to simplify the functioning of the modified CNN kernel. This ultimately results in compensating the poor lighting conditions, variations in the orientation and topological morphologies in the facial structures, color and contrast variations, etc. The model is designed with low memory requirements and training time needed to optimize the CNN. The learning and convergence rates of the proposed architecture are pretty high, and the output in recognition rates is more accurate than the currently existing methods. The hyperparameters of the Deep Learning ConvNet can be tuned for good performance so that even real-time micro-expression analysis can be performed on live video feeds for real-time implementations. This presents a lot of applications for facial micro-emotion recognition in fields like detection of emotional intelligence, deep fake detections, medical research, market research surveys, job recruitments, and so on, and it is estimated that the technology has a market potential of $56 billion by the year 2024. The research can be further extended by involving audio elements in the training so that a more accurate model can be formed based on the context of the spoken word, correlated with the associated micro expressions displayed.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[2] Y. Tang, "Deep learning using linear support vector machines," in *Proc. Int. Conf. on Machine Learning 2013: Challenges in Representation Learning Workshop*, Atlanta, Georgia, USA, pp. 1–18, 2013.

[3] P. Husák, J. Cech and J. Matas, "Spotting facial micro-expressions in the wild," in *22nd Computer Vision Winter Workshop*, Retz, Austria, pp. 1–9, 2017.

[4] K. Kulkarni, C. A. Corneanu, I. Ofodile, S. Escalera, X. Baro *et al.,* "Automatic recognition of facial displays of unfelt emotions," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 377–390, 2018.

[5] B. Allaert, I. M. Bilasco and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 147–158, 2019.

[6] D. Y. Choi and B. C. Song, "Facial micro-expression recognition using two-dimensional landmark feature maps," *IEEE Access*, vol. 8, pp. 121549–121563, 2020.

[7] Y. Li, X. Huang and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 249–263, 2020.

[8] T. Zhang, Y. Zong, W. Zheng, C. P. Chen, X. Hong *et al.,* "Cross-database micro-expression recognition: A benchmark," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 2, pp. 544–559, 2020.

[9] E. Pei, M. C. Oveneke, Y. Zhao, D. Jiang and H. Sahli, "Monocular 3D facial expression features for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3540–3550, 2020.

[10] B. Sun, S. Cao, D. Li, J. He and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1037–1043, 2020.

[11] S. K. Jarraya, M. Masmoudi and M. Hammami, "Compound emotion recognition of autistic children during meltdown crisis based on deep spatio-temporal analysis of facial geometric features," *IEEE Access*, vol. 8, pp. 69311–69326, 2020.

[12] Y. Zhang, H. Jiang, X. Li, B. Lu, K. M. Rabie *et al.,* "A new framework combining local-region division and feature selection for micro-expressions recognition," *IEEE Access*, vol. 8, pp. 94499–94509, 2020.

[13] C. Dalvi, M. Rathod, S. Patil, S. Gite and K. Kotecha, "A survey of AI-based facial emotion recognition: Features, ML & DL techniques, age-wise datasets and future directions," *IEEE Access*, vol. 9, pp. 165806–165840, 2021.

[14] X. Ben, Y. Ren, J. Zhang, S. J. Wang and K. Kpalam, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 1–16, 2021.

[15] S. J. Wang, Y. He, J. Li and X. Fu, "MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 3956–3969, 2021.

[16] J. Li, Z. Dong, S. Lu, S. J. Wang, W. J. Yan *et al.,* "CAS(ME)3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Early Access, pp. 1–18, 2022.

[17] J. Shi, S. Zhu and Z. Liang, "Amending facial expression representation via de-albino," in *Proc. 41st Chinese Control Conf. (CCC)*, Hefei, China, pp. 6267–6272, 2022.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar *et al.,* "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 94–101, 2010.

[19] L. Zhang and O. Arandjelović, "Review of automatic microexpression recognition in the past decade," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 414–434, 2021.

[20] O. A. Hassen, N. Azman Abu, Z. Zainal Abidin and S. M. Darwish, "Realistic smile expression recognition approach using ensemble classifier with enhanced bagging," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2453–2469, 2022.

[21] O. A. Hassen, N. A. Abu, Z. Z. Abidin and S. M. Darwish, "A new descriptor for smile classification based on cascade classifier in unconstrained scenarios," *Symmetry*, vol. 13, no. 5, pp. 1–18, 2021.

[22] H. M. Hang, Y. M. Chou and S. C. Cheng, "Motion estimation for video coding standards," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 17, no. 2, pp. 113–136, 1997.

[23] H. Pan, L. Xie, Z. Wang, B. Liu, M. Yang *et al.,* "Review of micro-expression spotting and recognition in video sequences," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 1–17, 2021.

[24] E. A. Clark, J. N. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne *et al.,* "The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review," *Frontiers in Psychology*, vol. 11, pp. 920, 2020.

[25] Y. Sato, Y. Horaguchi, L. Vanel and S. Shioiri, "Prediction of image preferences from spontaneous facial expressions," *Interdisciplinary Information Sciences*, vol. 28, no. 1, pp. 45–53, 2022.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 2818–2826, 2016.

[27] Q. Ji, J. Huang, W. He and Y. Sun, "Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images," *Algorithms*, vol. 12, no. 3, pp. 51, 2019.

[28] D. Patel, X. Hong and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, Cancun, Mexico, pp. 2258–2263, 2016.