Journal of Information Systems Engineering and Management
2025, 10(7s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Spontaneous Micro-facial Expression Detection using Attention-based Convolutional Gated Recurrent Neural Networks with RMSProp Optimization

Ramakrishna Gandi[1], Dr.A.Geetha[2],Dr.B.Ramasubba Reddy[3]

[1]Research Scholar,Computer Science & Engineering Department,Faculty of Engineering & Technology,Annamalai University,Annamalainagar, Tamilnadu,INDIA,Email:gandiramakrishna2@gmail.com

[2]Professor,Computer Science & Engineering Department,Faculty of Engineering & Technology,Annamalai University,Annamalainagar, Tamilnadu,INDIA,Email:aucsegeetha@yahoo.com

[3]Professor,Department of Computer Science and Engineering,School of Computing, Mohan Babu University,Tirupati, A.P,INDIA,Email:rsreddyphd@gmail.com

ARTICLE INFO

Received: 05 Oct 2024

Revised: 05 Dec 2024

Accepted: 22 Dec 2024

ABSTRACT

Introduction: Facial micro expressions have a significant role in communicating hidden emotions, offering profound implications in multiple areas such as psychology, cybersecurity, and the study of human-computer interaction. However, recognizing micro-expressions is tough because of their transient nature. Furthermore,Micro expressions are significantly shaped by the visual characteristics of the face and the interactions among its various sub-regions. Recent advances in computer vision have improved self-supervised learning.

Traditional CNNs for vision disorders learn only from whole photos or videos and cannot consistently distinguish facial micro expressions. Some existing CNN models might lack advanced attention mechanisms to effectively prioritize subtle micro-expression features within the complex facial data. They may leading to potential misinterpretation of expressions.

Objectives: This research offers the Attention Convolutional Gated Recurrent Neural Network with RMSProp (ACGRNN- RMSProp) classifier aims to achieve high accuracy and robustness in micro facial expression recognition. The attention mechanism sharpens the model's emphasis on important facial regions, while convolutional layers gather spatial features, and GRUs effectively capture temporal dynamics. The RMSProp optimizer ensures efficient and stable training, making this approach well-suited for the complex task of recognizing subtle and fleeting micro expressions.. This suggested approach is applied to the FER2013 dataset and compared to conventional micro expression recognition baseline methods.

Methods: The FER2013 dataset is initially loaded and processed. Facial landmarks in this dataset are detected using a Multi-Task Cascaded Convolutional Neural Network, after which features are extracted from the processed data using EfficientNetB3 and InceptionResNetV2. The features obtained from extraction are then fused and subsequently classified using an Attention Convolutional Gated Recurrent Neural Network with RMSProp optimization.

Results:This hybrid model demonstrates exceptional classification performance, achieving precision, recall, and F1-score of 94.5% each, along with an accuracy of 96.43%. Furthermore, it includes additional metrics such as the Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic(ROC) Curve in this field.

Conclusions:This study showcases enhanced recognition accuracy when compared to a range of leading models in the field,eliminated the requirement for system preprocessing procedures such as contrast enhancement, gradient operators, or dimensionality reduction for accurate results.Leverage the inherent relationships between face identification and other face analysis tasks to enhance future performance.

Keywords: micro-expressions, facial detection, neural network, feature extraction, RMSProp.

INTRODUCTION

Microexpressions (MEs) are fleeting facial expressions that reveal emotions a person attempts to hide. They can manifest as either repression (unconscious concealment) or suppression (intentional concealment), typically lasting between 1/25 and 1/5 of a second[1].These brief expressions are vital for understanding human behavior, uncovering hidden emotions, and are employed in various fields, including security and surveillance. For instance, the U.S. Transportation Security Administration's SPOT program[2] trains airport personnel to identify potentially suspicious individuals by analyzing their facial expressions and verbal cues.Although humans can be specifically educated to detect micro-expressions, the results are frequently very poor [3]. Due to human beings limited capacity for effective micro-expression recognition, automated systems that are capable of doing so must be developed. The application of computer-based techniques for the identification of micro-expressions is known as automatic micro-expression recognition. There are seven distinct ways to express oneself, including via macro- and micro-expressions: sorrow, happiness, fear, rage, surprise, disgust, and contempt[4].

The primary techniques for micro-expression identification in the past were those based on machine learning, optical flow, and local binary patterns (LBP).Typically, they extract features using optical flow or LBP-based techniques, and then utilize Support Vector Machine (SVM) classification to categorize them. The predominant baseline feature extractor in the majority of databases now in use [5] is LBP with Three Orthogonal Planes (LBP-TOP) [6,7].A lightweight representation known as LBP-SIP was introduced by Wang et al. [8] and its foundation consisted of three intersecting lines that went through LBP-TOP's center point. Moreover, the optical flow field was used by the LBP-TOP approach as the primary feature to illustrate the micro-expression motion pattern. Following this, it derived additional simplified emotional expressions to create a facial dynamic map (FDM). Both traditional machine learning and deep learning techniques are capable of analyzing micro-expressions, with convolutional neural networks (CNNs) making significant strides in the field of computer vision recently. It has the potential to be used for the purpose of detecting microexpressions.CNN and temporal interpolation were used in the solution provided by Patel et al. [9].Khor et al. [10] introduced a neural network with three streams and an enrichment strategy to merge different data sources. However, deep learning is rarely used to differentiate micro expressions in comparison to standard machine learning. [11].CNN models face challenges related to unbalanced datasets and the necessity for more nuanced input features beyond mere video sequences, optical flows, or grayscale images.This study presents a technique based on deep learning to accurately identify micro expressions, aiming to address the aforementioned challenges. Therefore, the study's contributions may be summarized as:

- A cascaded multi-task architecture optimizes performance by using intrinsic correlations. This technology employs a layered architecture made up of three intricately designed deep convolutional networks.These networks are used to anticipate the positioning of faces and landmarks in a gradual and precise manner.
- From a predetermined reference point, it is crucial to comprehend the differences in the pixel locations of the mouth, eyes, and eyebrows. Thus, the Inception of EfficientNetB3+Geometric feature extraction is aided by ResNetV2.
- The proposed model employs a basic proposal loss to maximize feature discrimination and Combining CGRNN with a constructive but attention method captures micro-expressions' tiny fluctuations.

The paper is structured in the following manner. Section II provides the existing approaches. Section III describes our suggested technique, which is separated into various section. In Section IV, the experimental findings for the algorithm assessment. Finally, Section V brings this work to its conclusion.

RELATED WORKS

This section offers a concise overview of facial expressions and subtle micro-expressions. A geometry-aware conditional network (GACN) designed for continuous pose-invariant facial emotion recognition and modification is outlined in [12]. The network is capable of capturing long-range correlations. The GACN can efficiently perform photo editing tasks that are not affected by changes in posture, thanks to the integration of conditional self-attention processes that can handle long-range dependencies. For posture-invariant facial expression recognition (FER), in [13] , a dual-branch network is developed that employs deep global multi-scale attention and local patch attention to mitigate the impacts of self-occlusion and positional shifts. In [21],Emotions can be expressed through various mediums such as vocal tone, written communication, and facial expressions. Detecting emotions in text represents a content-driven classification task that integrates principles from natural language processing and machine learning.For this research, the GMS-LPA network is comprised of four primary modules: the feature extraction module, which includes the local patches attention (LPA) module and the global multiple-scale (GMS) module, along with the model-level fusion model, constitutes a comprehensive framework. The double-stream 3D convolutional neural network (DS-3DCNN) serves as an advanced deep learning architecture specifically designed for detecting video motion events (ME)[14].The recognition architecture uses two 3D-CNN streams. The first video from the unprocessed ME dataset is used to extract spatiotemporal features. Within the spatiotemporal domain, the second component isolates and captures the changes in facial motions. In order to improve the process of extracting features, the little movement that naturally occurs in a moving entity is amplified. The C3DBed[15] model incorporates an innovative embedding technique utilizing a three-dimensional convolutional neural network within its transformer architecture. A mechanism determines the attention weight of each micro-expression picture area in this model. As a result, it is able to accurately identify subtle variations in the facial image and extract reliable local details. Address the problems caused by the low-intensity placement of facial muscle activity in local regions, such as model complexity and information redundancy. The Lossless Attention Residual Network (LARNet) was compared to many cutting-edge micro facial expression algorithms in [16]. However, the main purpose is to explore the segments needed to distinguish facial microexpressions. Several CNN-based techniques analyze facial pixels for local-level feature extraction. LARNet encodes face spatial and temporal data to extract characteristics from the nose, cheeks, mouth, and eyes. For the classification and detection of micro-expressions, a proposed architecture utilizes a three-stream fusion of 2D and 3D convolutional neural networks (TSNN) [17]. The models TSNN-IF and TSNN-LF are designed to simultaneously automate the learning of spatial and temporal variables.

Upon examining a few of the most current studies on facial expression identification, it is clear that emotion recognition remains a difficult computer vision task with several obstacles. It is evident from recent surveys that many novel models have been presented that can even identify emotions from photographs with varied face angles, although the majority of classic emotion identification systems can only detect the expression from an image with front facial angles. Furthermore, the surveys mentioned above make it abundantly evident that, when examining the last five years of data, publishing in the field of emotion recognition is fast growing.

METHODOLOGY

Initially, the FER2013 as Data Input are loaded as shown in Fig 1. Facial landmarks from this dataset are identified using a Multi-Task Cascaded Convolutional Neural Network. After detection, EfficientNetB3 and InceptionResNetV2are adopted for feature extraction. The extracted features are fused and classified using Attention Convolutional Gated Recurrent Neural Network(ACGRNN) With RMSProp.

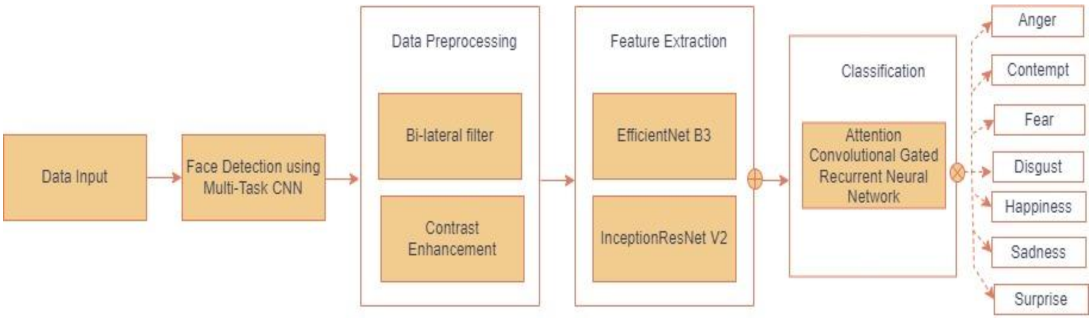


Fig 1: block diagram for micro facial expression detection

DATASET DESCRIPTION

Pierre-Luc Carrier and Aaron Courvill showcased the FER2013 dataset, released by Kaggle, during the 2013 International Conference on Machine Learning (ICML)[18]. FER-2013 grayscale images are 48 pixels by 48 pixels and every face has been classed by emotion. The FER-2013 dataset contains 35,887 micro expressions grouped into seven groupings from index 0 to 6. The collection contains 48x48 pixel grayscale face photographs. Faces are automatically centered and occupy similar space in each image. Classify each face by emotion: angry, disgusted, afraid, happy, sad, surprised, neutral. Public test set has 3,589 samples, training set 28,709 occurrences.Samples of the dataset with different types of expressions is shown in below Fig 2

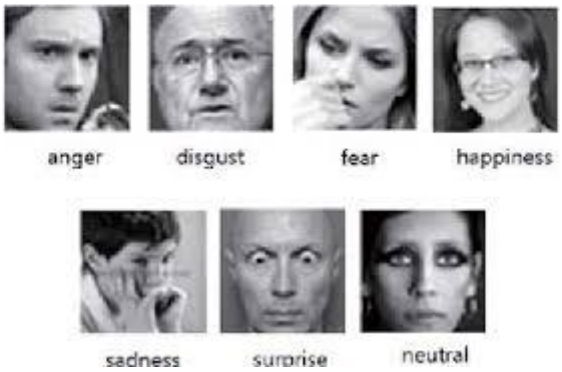


Fig 2: Samples of the dataset of different types of expressions

FACE IDENTIFICATION USING A MULTI-TASK CASCADED CONVOLUTIONAL NEURAL NETWORK

Multi-task Cascaded Convolutional Networks (MTCNN) is a framework based on deep learning that simultaneously addresses both face detection and the localization of facial landmarks.The system utilizes a three-tiered cascaded structure of meticulously crafted convolutional neural networks to predict face bounding boxes and the locations of facial keypoints, such as the eyes, nose, and mouth, in a progressively refined approach.

MTCNN is structured as a sophisticated deep learning framework that integrates three convolutional neural networks in a cascaded manner, as illustrated in Figure 3.The Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net) work in tandem to enhance face detection. P-Net efficiently produces a multitude of potential face bounding boxes from the input image,which are then refined using Non-Maximum Suppression (NMS) to combine highly overlapping areas.R-Net takes the candidate bounding boxes provided by P-Net, discards a large portion of non-face boxes, and then applies Non-Maximum Suppression (NMS) to combine the remaining bounding boxes.Finally, O-Net outputs the locations of five facial keypoints. The output from each layer is adjusted to incorporate face classification, bounding box regression, facial landmark localization, and head pose estimation, allowing MTCNN to simultaneously execute face detection, face alignment, and head pose estimation.

This model feature descriptor is mainly composed of three key elements: a classifier for distinguishing between faces and non-faces, a bounding box regression mechanism, and a system for localizing landmarks.

$$L_i^{det} = -(y_i^{det} \, log(p_i) + (1 - y_i^{det})(1 - \log(\mathrm{p_i}))) \text{, where } y_i^{det} \in \{0,1\} \; (1)$$

The formula outlined above defines a cross-entropy loss function utilized in face classification, where p_i signifies the probability of a face, and y_i^{det}represents the actual label for the background.

$$L_i^{box} = \left\|\bar{y}_i{}^{box} - y_i^{box}\right\|_2^2, where \; y_i^{box} \in \mathbb{R}^4 \; (2)$$

The equation above represents the regression loss computed using Euclidean distance, where $\bar{y}_i$ is the predicted output from the network, and y refers to the actual coordinates of the background, represented as a quadrilateral (upper left x, upper left y, width, height).

The equation presented above illustrates the regression loss calculated using Euclidean distance, with $\bar{y}_i$ denoting the network's predicted output and y indicating the actual coordinates of the background, which are represented as a quadrilateral defined by the upper left corner (x, y), width, and height.

$$L_i^{landmark} = \left\|\bar{y}_i{}^{landmark} - y_i^{landmark}\right\|_2^2 \text{,where } y_i^{landmark} \in \mathbb{R}^{10} \; (3)$$

Just like in bounding box regression, we calculate and minimize the Euclidean distance between the predicted landmark positions and their actual coordinates. In this context, $\bar{y}$denotes the coordinates predicted by the network, whereas y represents the true landmark coordinates. Given that there are five landmarks, each with two coordinates, the ground truth y forms a ten-element tuple. The training process incorporates multiple input sources, which are configured as follows:

$$\min \sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{j} \in \{\mathrm{det,box.landmark}\}} \alpha_j \beta_i^j L_i^j \qquad (4)$$

$$\beta_i^j \in \{0,1\} \qquad (5)$$

$$\mathrm{P-\ Net}\ R-\mathrm{Net}(\alpha_{\mathrm{det}}=1,\alpha_{\mathrm{box}}=0.5,\alpha_{\mathrm{landmark}}=0.5)\ (6)$$

$$\mathrm{O-Net}\ (\alpha_{\mathrm{det}}=1,\alpha_{\mathrm{box}}=0.5,\alpha_{\mathrm{landmark}}=1) \qquad (7)$$

The primary goal of the training process is to reduce the objective function, where NN signifies the number of training samples, α_j represents the importance weight assigned to each task, β_i^j denotes the sample label, and L_i^j corresponds to the specific loss function previously defined for each task.

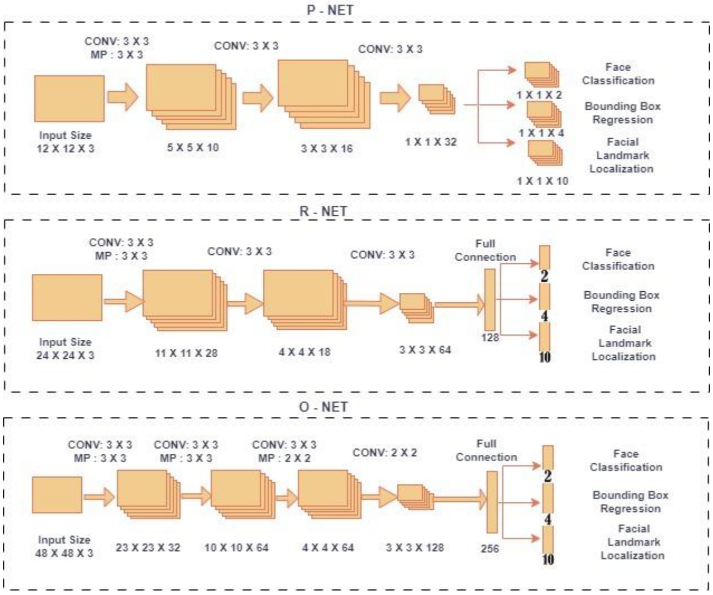


Fig 3:MTCNN-based face feature detection

In Fig 3, the MTCNN establishes a mapping function that transforms face images into a lower-dimensional Euclidean feature space, where the spatial distances between feature vectors directly correlate with a measure of facial similarity. In this space, images of the same person are mapped to feature vectors that are close together, while images of different individuals are mapped to distant feature vectors. Once this mapping function is learned, the subsequent face recognition task becomes straightforward, as the similarity between faces can be directly inferred from their distances in the feature space.

FEATURE EXTRACTION

This research analyzes geometric characteristics using two methods: EfficientNetB3 and InceptionResNetV2[20]. Both methods employed facial landmark detection. As shown in Fig 4, this application uses the histogram of oriented gradients (HOG) face detector to create a 68-point model of a face's form, brow, eyes, nose, and mouth. In compared to previous face detection methods, this facial landmark detection achieves great speed and accuracy.The steps involved in feature extraction is shown in below Fig 4.

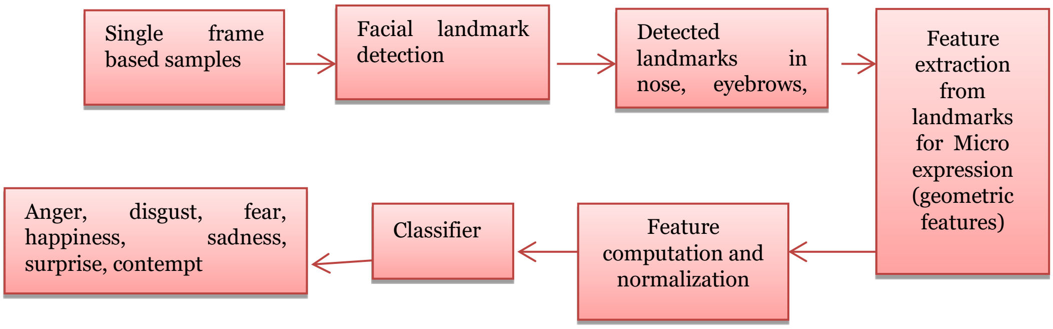

Single frame based samples
Facial landmark detection
Detected landmarks in nose, eyebrows,
Feature extraction from landmarks for Micro expression (geometric features)
Feature computation and normalization
Classifier
Anger, disgust, fear, happiness, sadness, surprise, contempt

Fig 4:steps involved in feature extraction

EFFICIENTNET-B3

In 2019, Google introduced the EfficientNet architecture at the International Conference on Machine Learning (ICML), which is a convolutional neural network (CNN) that utilizes an innovative structural methodology (Tan & Le, 2019). Convolutional Neural Networks (CNNs) are widely acknowledged as a highly effective subset of deep neural networks, especially renowned for their capabilities in processing and analyzing images.The structure of a CNN consists of input and output layers, with several intermediate layers, often called hidden layers, located in between.A CNN typically consists of three main types of intermediate layers: convolutional layers, pooling layers, and fully connected layers.

The convolutional layer is responsible for generating an activation map through the application of filters to the input data via addition and multiplication operations. In this stage, key features are extracted from the data, and then a pooling layer reduces the data's dimensionality by subsampling it nonlinearly, thereby simplifying the output (Stanford.edu, 2018). The fully connected layer converts the output from the previous layer into a one-dimensional array, ensuring complete interconnectivity among all neurons in the subsequent layer.A fully connected layer typically comes before the classification layer, which serves as the final layer in a convolutional neural network (CNN).

To enhance computational performance, the depth of CNN architectures has been progressively increased; however, improvements in accuracy plateaued after a certain threshold. The EfficientNet-B3 model significantly outperforms other CNN architectures by achieving higher success rates without necessitating increased depth. A notable aspect of this architecture is its compound scaling method, which improves not only the depth but also the breadth and resolution metrics (Tan & Le, 2019).

Figures that compare the proposed architecture with various CNN models demonstrate that EfficientNet (Tan & Le, 2019) outperforms its counterparts while utilizing a smaller number of parameters. The EfficientNet family consists of eight models, ranging from B0 to B7, with each successive model demonstrating enhanced accuracy. The chosen model for this research is EfficientNet-B3 (300x300) (Tan & Le, 2019), due to its input size being closely aligned with that of the Inception-ResNet-V2 architecture (299x299) (Szegedy et al., 2017), allowing for a resolution effect comparison to be disregarded.

Distinct from the B0 model, the EfficientNet-B3 model, which has an input size and parameter count of 12M compared to B0's 5.3M, features seven blocks, each with varying numbers of mobile inverted bottleneck convolutions (MBConV).Throughout the training phase of the EfficientNet-B3 model, a range of parameters was employed, ultimately identifying the most effective values. The Adam optimization algorithm was utilized with a learning rate of 0.0001, a batch size of 16, and binary cross-entropy designated as the loss function,the sigmoid function was utilized as the activation function in the final layer of classification. Additionally, 10-fold cross-validation was implemented throughout the training process.

INCEPTION-RESNET-V2

The Inception-ResNet-V2 architecture, introduced by Szegedy et al. in 2017, integrates the Inception architecture with residual connections and has been trained on more than one million images sourced from the ImageNet database(n.d.). This CNN model features a depth of 164 layers and benefits from rich feature representations across various images, attributed to the diverse training dataset.The model processes input images with dimensions of 299x299 pixels. Figure 5 depicts the core architecture of Inception-ResNet-V2 (Szegedy et al., 2017) prior to the fusion stage.In the Inception-ResNet-V2 architecture, multi-dimensional convolution filters are combined with residual connections, effectively minimizing the distortions typical of deep networks while also shortening the training duration.

The training parameters selected for the Inception-ResNet-V2 model were similar to those employed in the EfficientNet-B3 model. The Adam optimization algorithm was utilized with a learning rate set to 0.0001 and a batch size of 32.The loss function utilized was binary cross-entropy, and the final classification layer employed the sigmoid function as its activation function. Additionally, 10-fold cross-validation was implemented throughout the training process.

FUSIONING OF FEATURES

In Fig 5,once the features have been extracted, the issue of how to combine the K feature representations remains unanswered. Inspired by the feature fusion approach described in reference [19],we employ element-wise summation to incorporate all properties specific to the expression. Alternatively, individual qualities of the expression may be combined to create a unified feature vector. In addition, we analyze concatenation as a fusion function. Based on our testing results, element-wise summing outperforms concatenation in this scenario. The rationale for this outcome may be elucidated as follows: (1) Concatenation adds features but not distinctness for all classes. (2) Performing summation on an element-wise basis accelerates training and enhances the flow of gradients. (3) Summarization enhances important information in each dimension. Hence, the consolidated feature representation z' for x is defined in the following manner

$$z' = \sum_{k=1}^{K} z_k^*$$

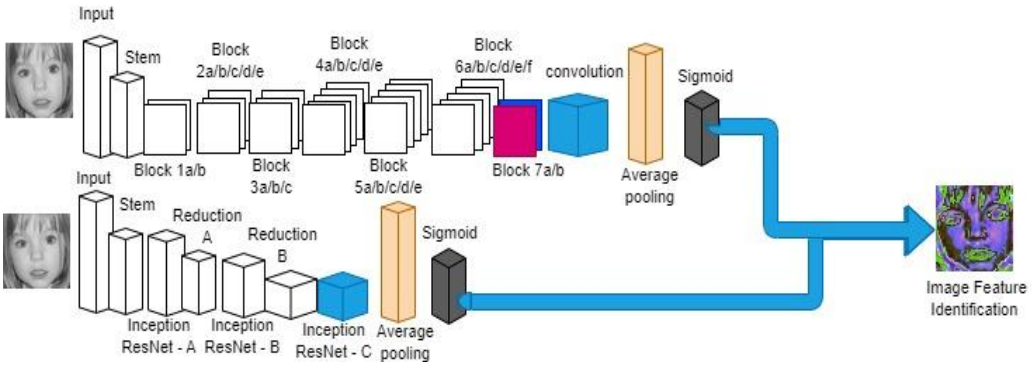


Fig 5:Utilizing a combination of EfficientNet-B3 and Inception-ResNet-V2 for enhanced feature extraction

The kth expression-specific detector's sample expression is represented by z_k^*. The polished feature, obtained from the expression, is forwarded to the final classification module, composed of two interconnected layers. An initial fully connected layer is succeeded by a dropout layer with a probability of 0.5 to mitigate the risk of overfitting. The output of the last fully connected layer is activated by a softmax function, and we utilize L_{cls} the standard softmax cross-entropy loss for classification purposes.

$$L_{cls}(\varphi; y) = -y\log(y^{\wedge})$$

In this context, the variable y denotes the ground truth label vector, which corresponds to the one-hot vector for the sample x.The variable $y^{\wedge}$ reflects the prediction probabilities associated with the sample x, denoted as $y^{\wedge} = [y^{\wedge}1, y^{\wedge}2, \ldots, y^{\wedge}K]$, where K is the number of micro-expression categories.

The variable $y^{\wedge}$ represents the predicted probabilities linked to the sample x, indicated as $y^{\wedge} = [y^{\wedge}1, y^{\wedge}2, \ldots, y^{\wedge}K]$, with K signifying the number of micro-expression categories.

ATTENTION CONVOLUTIONAL GATED RECURRENT NEURAL NETWORK FOR CLASSIFICATION

In Fig 6,this hybrid model incorporates an attention-gated layer, a max-over-time pooling layer, and a convolutional layer to effectively process the input matrix, along with a fully connected layer that incorporates dropout and produces softmax outputs. Let's use a 7-by-7-pixel, 4-dimensional image for illustration.The first convolutional layer using 2 or 3 window size convolution kernels. However, the attention gated layer convolution layer uses convolution kernels with 1 or 3 feature windows.

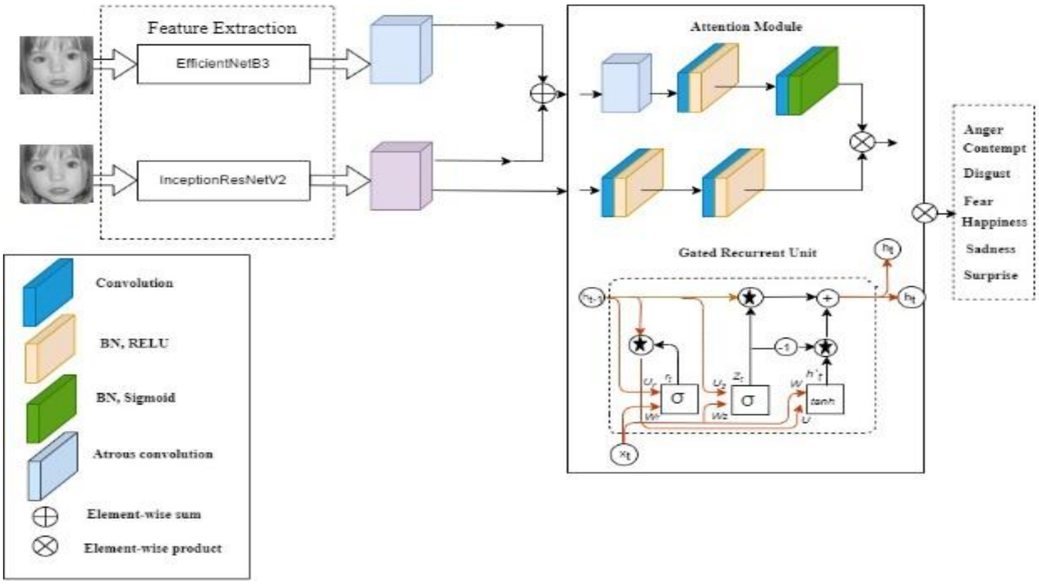


Fig 6: Architecture of ACGRNN

The i-th pixel's d-dimensional word vector is $e_i \in R^d$. A pixel-n input embedding matrix is

$$E_{1:n} = [e_1, e_2, \dots, e_n]^T$$

Where, $E_{1:n} \in R^{nxd}$

A window of h pixels is subjected to a convolution kernel $W \in R^{hxd}$ in the first convolutional layer in order to generate a new feature. A totally abstract characteristic $c_i \in R$ is derived from a set of pixels $E_{i:i+h-1}$'

$$c_i = f(g(W \times E_{i:i+h-1}) + b)$$

where f(.) represent a non-linear activation function, g(.) denote the summation of all elements in a matrix, b∈R represent a bias term, and × indicate the element-wise product of matrices. The convolution kernel produces multiple kernels by moving across the input sentence matrix and applying them to each word window within the phrase $\{E_{1:h}, E_{2:h+1,\dots}E_{n-h+1:n}\}$.An attention-gated layer comprises both a convolutional layer and a gating mechanism.The attention weight matrix is formed by applying a convolution kernel $V \epsilon R^{k \times 1}$ to each context feature $c_j(j = 1,2, \dots n - h + 1)$ in the feature map, using a window size of k.

$$A = [a_1, a_2, \dots a_{n-h+1}]^T$$

Theorem 1. The attention weight element $a_k(j = 1, . . n - h + 1)$ in matrix A is represented as:

$$a_j = \begin{cases} g\left[V \times C_{j-\frac{k-1}{2}.j+\frac{k-1}{2}}\right] (j = 1,2 \dots . n - h + 1 \text{ when k is odd} \\ g\left[V \times C_{j-\frac{k}{2}+1.j+\frac{k}{2}}\right] (j = 1,2 \dots . n - h + 1 \text{ when k is even} \end{cases}$$

Where $a_j \in R$ and $A \in R^{(n-h+1)x1}$. $C_{j-\frac{k-1}{2}.j+\frac{k-1}{2}}$ and $C_{j-\frac{k}{2}+1.j+\frac{k}{2}}$ represent the context characteristics of C_j.

Proof: For a 1D convolution with a kernel height of k and a stride s of 1, the number of padding pieces required on the input, given an input height of H, is

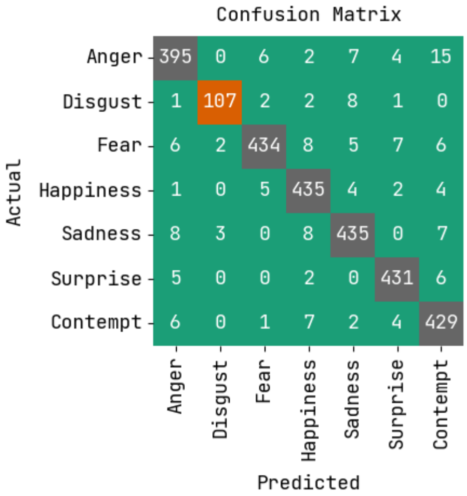
$$p_{need} = \left(\frac{H}{s} - 1\right) \times s + k - H$$

Following this, a max-over-time pooling operation is applied to each feature map to generate the output O∈P,that captures the most significant abstract characteristics associated with multi-grained attentions. These outputs are then concatenated together. The second-to-last layer consists of these characteristics and is passed via the dropout layer before being fed into the fully linked softmax layer. The "root mean square propagation" (RMSProp) technique is another term for an advanced AdaGrad modification that controls the learning rate at a rapid decline level. It is frequently compared to the Adadelta approach. Despite this, the Adadelta method unquestionably uses the RMSProp method of parameter changes carried out in the rule of numerator's updating.

RESULTS AND INTERPRETATIONS

The efficacy of the proposed approach for accurately identifying was evaluated by a classification test and k-fold cross-validation. The suggested model was trained and tested using all images in k-fold cross-validation, in which all data samples were arbitrarily split into k groups. For testing, one-fold was utilized, While k – 1 folds were utilized for training, this process should be repeated for the remaining k – 1 folds.This study employs the concepts of False Positive (FP), False Negative (FN), and 10-fold cross-validation. The classification performance was evaluated by measuring the area under the curves for accuracy, precision, sensitivity, specificity, f1-score, and MCC.

Fig 7 shows the confusion matrix for features used in ACGRNN-RMSProp testing, with rows representing the predicted class and columns representing the actual class of data relevant to micro facial expression identification. The crosswise colors reflect the testing networks that were appropriately and erroneously classified. The column on the right displays each anticipated class, while the row below depicts the execution of each actual class.



	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Contempt
Anger	395	0	6	2	7	4	15
Disgust	1	107	2	2	8	1	0
Fear	6	2	434	8	5	7	6
Happiness	1	0	5	435	4	2	4
Sadness	8	3	0	8	435	0	7
Surprise	5	0	0	2	0	431	6
Contempt	6	0	1	7	2	4	429

Fig 7: confusion matrix of ACGRNN- RMSProp for testing

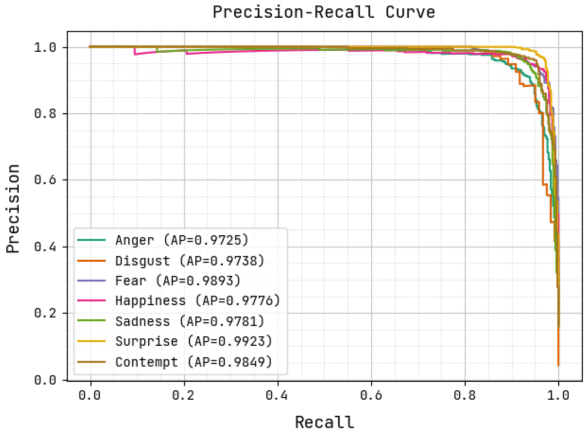


Fig 8: Performance of precision-recall curve of testing ACGRNN- RMSProp

Fig 8 depicts the precision-recall curve for testing ACGRNN-RMSProp, The horizontal and vertical directions reflect the false positive rate and the genuine positive rate, respectively. During the procedure, the AP achieves the maximum AP of 0.9893 for fear, 0.9923 for surprise, and 09849 for contempt.

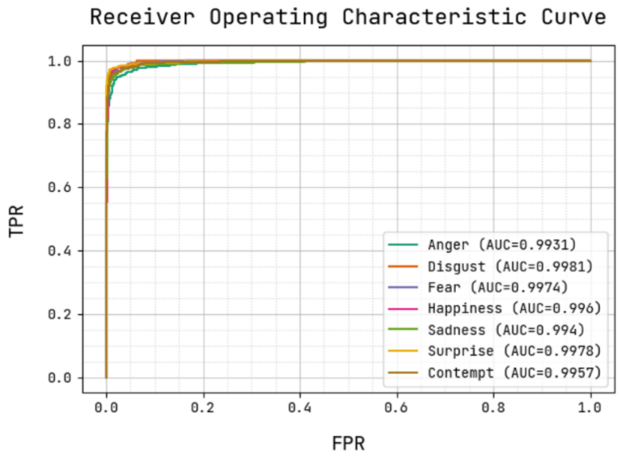
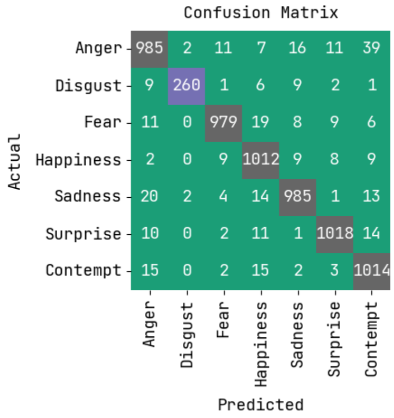


Fig 9: ROC curve of ACGRNN- RMSProp for testing

Fig 9 depicts a ROC curve for evaluating ACGRNN-RMSProp,the horizontal axis represents the false positive rate, while the vertical axis indicates the genuine positive rate.During the process the AUC achieves maximum range of 0.9978 for surprise, 0.9981 for disgust, and 0.9931 for anger.



	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Contempt
Anger	985	2	11	7	16	11	39
Disgust	9	260	1	6	9	2	1
Fear	11	0	979	19	8	9	6
Happiness	2	0	9	1012	9	8	9
Sadness	20	2	4	14	985	1	13
Surprise	10	0	2	11	1	1018	14
Contempt	15	0	2	15	2	3	1014

Fig 10: confusion matrix of ACGRNN- RMSProp for training

Fig 10 shows the confusion matrix for features trained for ACGRNN-RMSProp, with rows representing the predicted class and columns representing the actual class of data important to predicting microfacial expression detection. The crosswise colors reflect the testing networks that were appropriately and erroneously classified. The column on the right represents each projected class, but the row below illustrates the execution of each actual class.

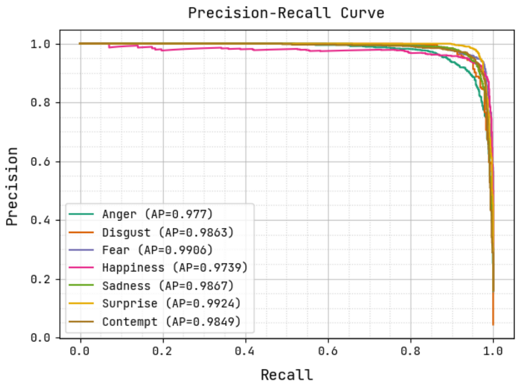


Fig 11: Analysis of precision-recall curve for training

Fig 11 depicts the precision-recall curve for training ACO_1DCNN,the horizontal axis represents the false positive rate, while the vertical axis indicates the genuine positive rate.During the procedure, the AP achieves highest value of 0.9924 for astonishment, 0.9849 for surprise, and 0.977 for anger.

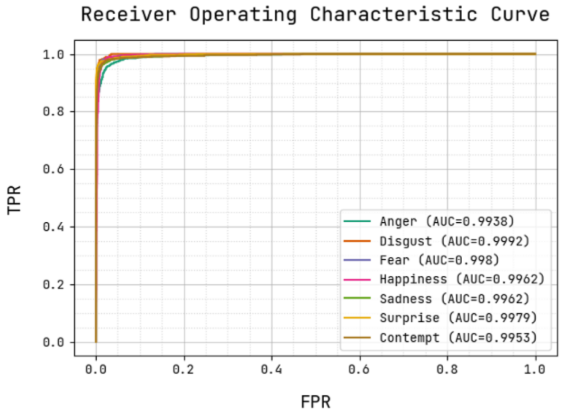


Fig 12: ROC curve of ACGRNN for training

Fig 12 depicts a ROC curve for training ACGRNN-RMSProp, The horizontal and vertical directions reflect the false positive rate and the genuine positive rate, respectively. During the process the AUC obtains the higher value of 0.9992 for disgust, 0.9962 for happiness and 0.9979 for surprise.

Table 1. Performance of Proposed ACGRNN- RMSProp for testing and training

Parameters	**Testing Data (%)**	**Training Data (%)**
Accuracy	0.9643	0.968
Precision	0.9457	0.9536
sensitivity	0.9381	0.9447
specificity	0.9906	0.9914
F-score	0.9415	0.9487
MCC	0.9324	0.9404

Table 2. Performance of existing and proposed methods

Metrics/methods	**GACN [12]**	**DS-3DCNN [14]**	**ACGRNN**
Accuracy (%)	95	91	96.43
Precision (%)	93	89	94.57
sensitivity (%)	92.3	87.2	93.81
specificity (%)	98	93.1	99.06
F-score (%)	92.4	93	94.15
MCC (%)	90	83.2	93.24

CONCLUSION

This research presents an innovative multi-task cascaded CNN framework designed for concurrent face identification and alignment, demonstrating superior performance compared to existing techniques on the FER2013 benchmark dataset for facial expression recognition.By combining these tasks, we remove the necessity for preliminary processes such as contrast enhancement and dimensionality reduction, resulting in improved accuracy and efficiency in facial expression recognition. Our model showcases enhanced recognition performance when compared to several leading models in the field.We propose to leverage the inherent relationship between face identification and other face analysis tasks to further enhance performance in future work.This study aims to create a unified framework for face identification and alignment through a cascaded CNN architecture. Experimental findings from the FER2013 dataset highlight the effectiveness of our method compared to current techniques.Our method eliminates the need for complex pre-processing, leading to improved efficiency and accuracy in face expression recognition.Future research will explore the synergistic relationship between face identification and other facial analysis tasks to improve overall performance.

REFERENCES

[1] Ekman, P., 2009. Telling Lies:Clues to Deceit in the Marketplace, Politics, and Marriage, W. W. Norton & Company.

[2] Polikovsky, S., Kameda, Y. and Ohta, Y., 2009., Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor", In 3rd International Conference of Crime Detection and Prevention, ICDP, London.

[3] Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003)., Facial expression recognition from video sequences: temporal and static modeling, Computer Vision and image understanding, 91(1-2), 160-187.

[4] Guo, Y., Tian, Y., Gao, X., & Zhang, X. (2014, July).,Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method, In 2014 international joint conference on neural networks (IJCNN) (pp. 3473-3479). IEEE.

[5] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu.,Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PloS one, 9(1):e86041, 2014.

[6] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap.,Samm: A spontaneous micro-facial movement dataset, IEEE transactions on affective computing, 9(1):116–129, 2016.

[7] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikainen. , Recognizing spontaneous facial micro-expressions, In 2011 international conference on computer vision, pages 1449–1456. IEEE, 2011.

[8] Yandan Wang, John See, Raphael C.-W. Phan, and Yee-Hui Oh.,Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, Computer Vision – ACCV 2014, pages 525–537, Cham, 2015. Springer International Publishing.

[9] Devangini Patel, X. Hong, and G. Zhao.,Selective deep features for micro-expression recognition, In 2016 23rd International Conference on Pattern Recognition (ICPR), pages 2258–2263, Dec 2016.

[10] Huai-Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao Lin.,Enriched long-term recurrent convolutional network for facial micro expression recognition, In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 667–674. IEEE, 2018.

[11] Ce Liu. ,Beyond Pixels: Exploring New Representations and Applications for Motion Analysis, Ph.D. Thesis, 2009.

[12] Liu, T., Li, J., Wu, J., Du, B., Wan, J., & Chang, J. (2024),Confusable facial expression recognition with geometry-aware conditional network, Pattern Recognition, 148, 110174.

[13] Hossain, S., Umer, S., Rout, R. K., & Tanveer, M. (2023), Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling, Applied Soft Computing, 134, 109997.

[14] Li, Z., Zhang, Y., Xing, H., & Chan, K. L. (2023), Facial Micro-Expression Recognition Using Double-Stream 3D Convolutional Neural Network with Domain Adaptation, Sensors, 23(7), 3577.

[15] Pan, H., Xie, L., & Wang, Z. (2023),C3DBed: Facial micro-expression recognition with three-dimensional convolutional neural network embedding in transformer model, Engineering Applications of Artificial Intelligence, 123, 106258.

[16] Hashmi, M. F., Ashish, B. K. K., Sharma, V., Keskar, A. G., Bokde, N. D., Yoon, J. H., & Geem, Z. W. (2021),LARNet: Real-time detection of facial micro expression using lossless attention residual network, Sensors, 21(4), 1098.

[17] Wu, C., & Guo, F. (2021),TSNN: Three-stream combining 2D and 3D convolutional neural network for micro-expression recognition, IEEJ Transactions on Electrical and Electronic Engineering, 16(1), 98-107.

[18] Dataset- I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D.-H. &. others Lee,Challenges in representation learning: A report on three machine learning contests, International Conference on Neural Information Processing, 2015.

[19] J. Wu, L. Wang, L. Wang, J. Guo, G. Wu, Learning actor relation graphs for group activity recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019, pp. 9964–9974.

[20] Sezin BARIN, Murat SARIBAŞ, Beyza Gülizar ÇİLTAŞ, Gür Emre GÜRAKSIN,Utku KÖSE, Hybrid Convolutional Neural Network-Based Diagnosis System for Intracranial Hemorrhage, BRAIN. Broad Research in Artificial Intelligence and Neuroscience,2021,12(4),pp. 01-27.

[21] S.Arun Kumar S., A.Geetha, Emotion Detection from Text using Natural Language Processing and Neural Networks, International Journal of Intelligent Systems and Applications in Engineering,2024,12(14s),pp. 609-615.