ATTENTION BASED CONVOLUTIONAL NEURAL NETWORK FOR FACIAL EXPRESSION RECOGNITION

Jenisha A[1*], Aleesha Livingston L[2]

[1*]*ME/Communication Systems, Bethlahem Institute of Engineering*

[2]*Assistant Professor/Electronics & Communication Engineering, Bethlahem Institute of Engineering*

[*]***Corresponding author: Jenisha A***

Article DOI: https://doi.org/10.36713/epra12604
DOI No: 10.36713/epra12604

ABSTRACT

Many deep learning-based methods have been suggested in recent years to improve facial expression recognition performance, but they are unable to extract the small facial changes in skin textures, such as wrinkles and furrows, which indicate changes in expression. We put forward an attention mechanism-based Convolutional Neural Network (CNN) for face expression recognition to get around this issue. The attention module, classification module feature extraction module and the reconstruction module make up the architecture's four components. The feature extraction module, which consists of two distinct CNN processing streams—one for raw images and the other for Local Binary Pattern (LBP) feature maps. The LBP features extract image texture data before capturing the minute facial movements, which can enhance network efficiency. The neural network can pay more attention to useful characteristics by using an attention mechanism. To improve the attention model and produce improved results, we combine LBP features and convolutional features. We use the CK+ and Oulu-CASIA datasets to test the proposed methodology. The experimental findings show that the suggested method is workable and efficient.

KEYWORDS: *Convolution neural network, Local binary pattern, Facial expression recognition.*

1. INTRODUCTION

A major non-verbal method that humans use to communicate is facial expression [1]. Facial expressions can be categorized into six groups, including those that convey anger, disgust, fear, happiness, sadness, and surprise, in accordance with Ekman's theory of the six fundamental cross-cultural emotions [2]. Since its widespread use in areas like human-computer interface, virtual reality, intelligent course systems, and more over the past 20 years, facial expression recognition has seen significant advancements in the field of computer vision [3]. With the advancement of technology, facial expression analysis may be used in the field of education, allowing us to understand pupil commitment in a virtual classroom [4]. Additionally, facial expression recognition can be used in medicine when a doctor has automatically identified a patient's pain online [5]. Security applications of facial expression analysis include spotting suspicious people by watching their feelings. Facial expression recognition is used to identify the emotions in a particular image, even in animation. Because of this, the study of facial emotion recognition is fascinating and exciting in many ways [6]. However, because of the complexity, diversity, occlusion, and lighting, automatic face emotion recognition has become very difficult.

2. LITERATURE REVIEW

Some of the papers based on facial expression are reviewed below,

Frequency neural network (FreNet), a deep learning-based method developed by Tang *et al.* [7], is used to recognize facial expressions. To learn features in the frequency domain, the learnable multiplication kernel and build multiple multiplication layers are first introduced. In order to further produce high-level features, a summarization layer is then suggested after multiplication levels. Thirdly, based on the discrete cosine transform (DCT) property, they build the Basic-FreNet using multiplication layers and summarization layers, which can produce high-level features on the commonly used DCT feature. Finally, the Block-FreNet is introduced in order to further improve performance on Basic-FreNet, with a weight-shared multiplication kernel intended for learning features and a block subsampling designed for dimension reduction. This approach lowers the processing expense. However, it takes longer.

Deep residual network ResNet-50, which combines convolutional neural network for facial emotion recognition, was introduced by Li and Lima [8]. They make use of ResNet-50 as their network foundation. Convolutional neural networks are used to retrieve the features. Batch normalization and the activation function ReLU are then used to enhance the model's capacity for convergence. Although there are only a few images available for processing, this model has good accuracy and a good recognition impact in terms of average recognition accuracy.

An appropriate and light-weight Facial Expression Recognition Network Auto-FERNet that is autonomously searched by a differentiable Neural Architecture Search model was presented by Li *et al.* [9] in their research. Additionally, they suggest a relabeling technique based on the similarity of facial expressions to reduce the uncertainty issue and avoid the model over fitting by altering these biased labels within similar expressions. This model only has a few features. The time commitment is greater. In order to recognize facial expressions, Wang *et al.* [10] created the oriented attention pseudo-siamese network (OAPSN), which uses both global and local facial information. The network has two branches: an attention branch with a UNet-like design to gather local highlight information, and a maintenance branch with several convolutional blocks to benefit from high-level semantic features. They specifically enter the face picture into the maintenance branch first. They determine the correlation coefficient among a face and its subareas for the attention branch. They then correlate the facial features and the correlation coefficients to create a weighted mask. The focus branch is then sent the weighted mask. The classification findings are then output after the two branches have been combined.

For the purpose of recognizing facial expressions, Kola & Samayamantula [11] introduced Local Binary Pattern with Adaptive Window (LBPAW). The feature length is shortened from 256 to 32 using this method. A new neighbourhood is created by averaging along radial direction for texture classification in order to have noise robustness. Based on the variations in intensity, a window that is adaptive is taken into account around each pixel in this study. Additionally, averaging of pixel intensities in each radial direction is taken into account to achieve robustness in the presence of noise; as a result, a 3 × 3 window is used to compute the LBP. This approach takes longer, but the computational complexity is reduced.

For the purpose of recognizing facial expressions, Gera *et al.* [12] proposed the spatio-channel attention net (SCAN). In order to make the FER model resilient to occlusions and pose variations, the local-global attention branch known as SCAN computes attention weight for each channel and each spatial position within each channel across all taken local patches. Additionally, it is demonstrated that all local fixes can share the local attention branch without noticeably degrading performance. The neighbourhood attention branch becomes light because of this. This technique provides more complementary information, but it is challenging to distinguish between the mouth and the noise region.

Swin transformer-based facial expression method (STFA) was suggested by Kim *et al.* [13] for the recognition of facial expressions. It is made up of three streams: an auditory stream, a temporal stream, and a visual stream. All streams are supported by the Swin transformer, and the visual stream-video shot that processes numerous frames uses the Shallow 3D CNN structure. Each visual stream-image, visual stream-video shot, and audio stream were separately trained before score fusion was used in the final inference. Efficiency is increased by this approach, but costs are higher. The summary of current facial emotion recognition is shown in Table.1.

Table: 1 Summary of existing facial expression recognition.

Author	Methods used	Datasets used	Advantages	Disadvantages
Tang *et al.* [7]	FreNet method	CK+,Oulu-CASIA, KDEF, FER2013	-This method reduces the computational cost.	-Time consumption is more.
Li & Lima [8]	ResNet-50 method	Self collected datasets	-This model has good accuracy and good recognition effect	limited number of images for processing.
Li *et al.*[9]	Auto-FERNet	FER 2013 datasets, CK+, & JAFFE	This model has limited number of parameters.	Time consumption is more
Wang *et al.* [10]	OAPSN method	Ck+	-solve the occlusion problem	-occurs error in classification
Kola & Samayamantula [11]	LBPAW method	Japanese Female Facial Expression database,Cohn-Kanade database	-reduces the computational complexity	- -Consumes more time

Gera *et al.* [12]	SCAN method	AffectNet,FERPlus, RAF-DB, SFEW,FED-RO, CK+,Oulu-CASIA	-gives richer complementary information	-recognizing mouth and noise region is difficult
Kim *et al.* [13]	STFA method	Aff-Wild2 dataset	-increases efficiency	-expense is more

3. PROBLEM DEFINITION

Numerous real-world uses for automatic facial expression recognition include improving human-computer contact and distance learning. Numerous techniques are developed for recognizing facial expressions. However, using automatic facial emotion detection is challenging due to irrelevant facial information (such as hair and hats) and complicated background clutter. Additionally, the adopted region-level technique is used to assess the significance of various regions, but this method is unable to infer skin textures from changes in facial features. As a result, face expression recognition and feature extraction become more accurate.

4. PROPOSED ATTENTION MECHANISM-BASED CNN MODEL

The proposed architecture of an attention-based CNN is shown in Fig. 1. To categorize various facial expressions, we suggest a novel attention method based convolutional neural network. The feature extraction module, the attention module, the reconstruction module, and the classification module make up the architecture's four components. Initial features are extracted from raw images in the feature extraction module and then sent to subsequent modules for processing. The initial features that are obtained are typically too coarse, but we need more extracted features to transmit to the later attention module in order to make it function properly right away. Therefore, we combine convolution features and LBP features in an attention module. The ability of the attention module can be enhanced in order to increase the recognition accuracy of the network with the aid of LBP features that offer texture information and can reflect small differences on the face.

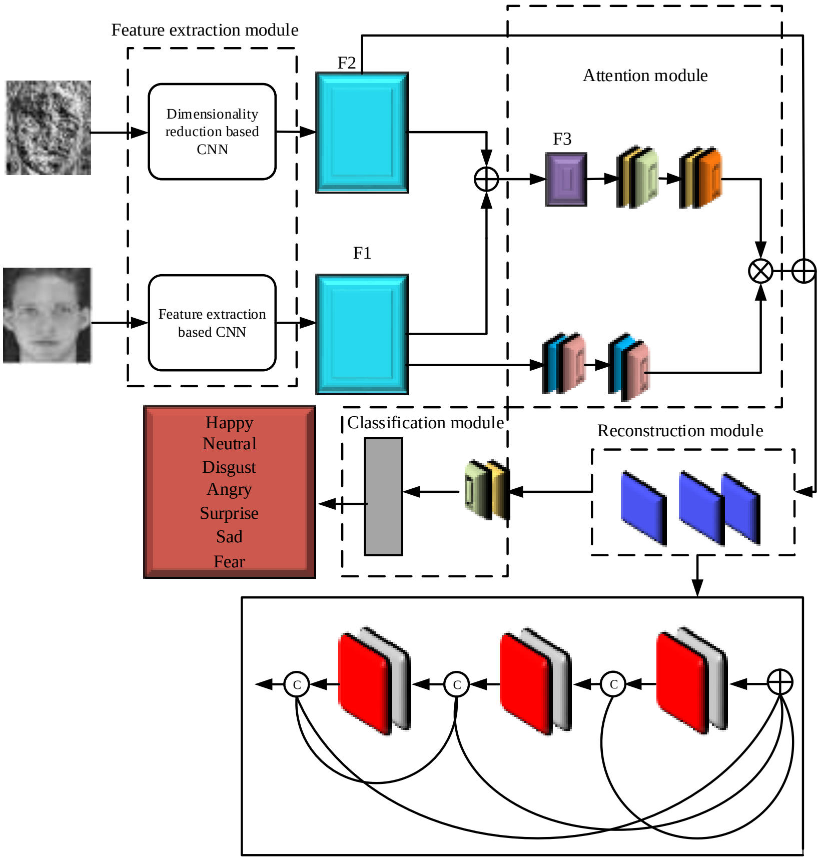
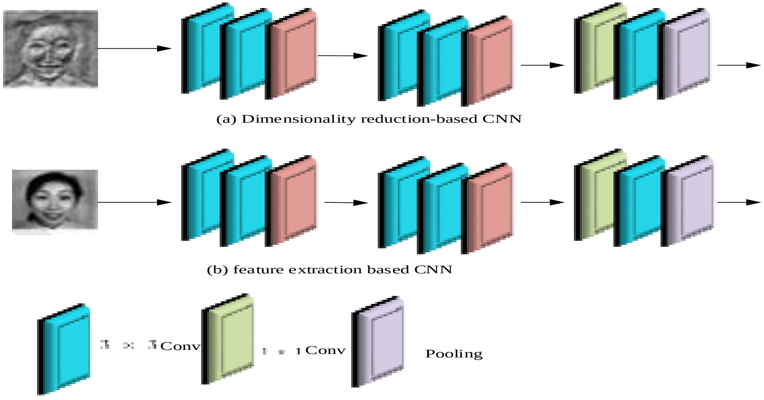


Figure 1: Architecture of proposed attention based CNN

The proposed Attention Mechanism-based CNN model involves four major steps and are explained below.

(a) Feature Extraction module

This module is made up of two distinct CNNs that extract two distinct features. In addition to raw images, LBP features also have texture information and can reflect small changes in skin texture on the face, making it possible to differentiate between minor expressions. After that, we combine the features F1 from the raw pictures with the features F2 from the LBP feature images before adding the combined features F3 to an attention module. The block layout for the feature extraction module is shown in Fig. 2.


(a) Dimensionality reduction-based CNN

(b) feature extraction based CNN

Figure 2: Block diagram of feature extraction module

(b) Attention Module

The attention module makes the network concentrate more on these features, which are essential for expression recognition, by increasing the weights of helpful features. The network can recognize various expressions more effectively in this manner. The block layout of the attention module is shown in Fig. 3.

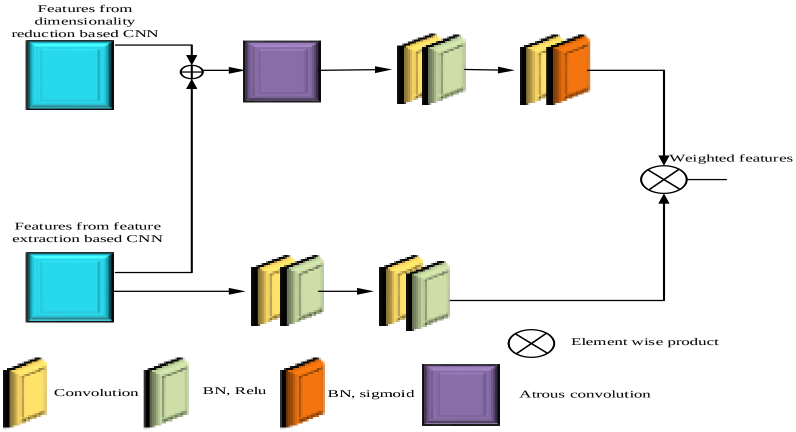

Features from dimensionality reduction based CNN
Weighted features
Features from feature extraction based CNN
Element wise product
Convolution
BN, Relu
BN, sigmoid
Atrous convolution

Figure 3: Block diagram of attention module

(c) Reconstruction Module

After the attention module, we modify the attention map using a dense atrous convolution as the reconstruction module to produce an improved feature map for the classification module. The feature-maps of all former layers as well as characteristics F1, are combined as inputs for each layer, and its own feature-maps are used as inputs into all layers that came after it. The result of the attention module is used to perform element-wise sum operations on the fused feature maps F3. This module can reuse helpful features and extract deeper features in addition to assisting in the resolution of the disappearing gradient issue. The reconstruction module's block layout is shown in Fig. 4.

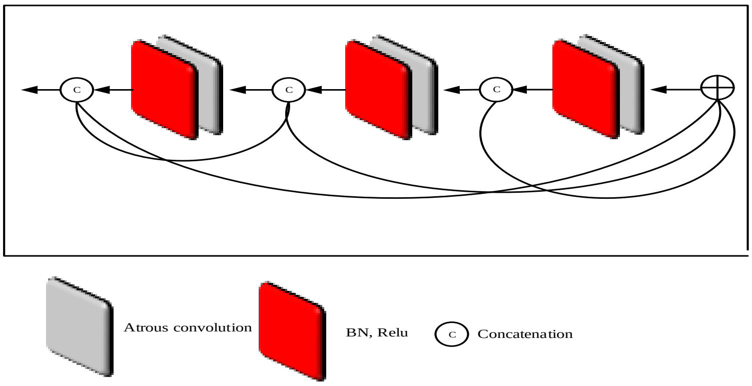


Figure 4: Block diagram of reconstruction module

(d) Classification module

At last, fully connected layers with softmax are used for classification. We use the batch normalization after each layer to speed up the convergence of the network and avoid over-fitting. Block diagram of classification module. Fig.5 shows the block diagram of classification module

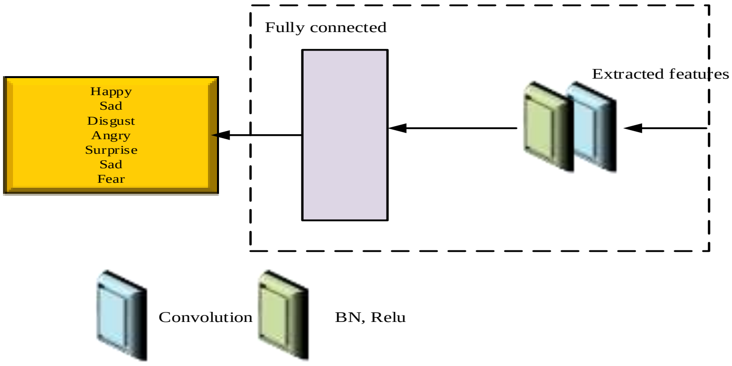


Figure 5: Block diagram of classification module

5. EXPERIMENTS AND RESULTS

We tested the proposed method on the dataset of CK+ and Oulu-CASIA facial expression datasets.

CK+ dataset: The training set and test set were built using the three peak expression frames from each expression sequence, yielding a total of 981 images.

Oulu-CASIA dataset: It includes videos that were shot under regulated lab circumstances. To create the training set and the test set, we choose the final three frames from each sequence that were recorded with visible light and strong illumination (consisting of 1,440 images in total).

5.1 Comparative Analysis

Table.2 shows the performance of proposed attention mechanism based CNN model with existing methods such as FreNet [7], OAPSN [10], and SCAN [12]. The proposed attention mechanism based CNN model achieved the better performance than other methods for facial expression recognition. Our proposed method achieves accuracy of 99.54 and 89.47 for CK+ and Oulu-CASIA dataset.

Table: 2 performance of proposed and existing methods in terms of accuracy for CK+ and Oulu-CASIA dataset

Methods	**Accuracy**	
	CK+ dataset	**Oulu-CASIA dataset**
FreNet	95.37	86.77
OAPSN	92.45	86.90
SCAN	95.78	87.01
Proposed	99.54	89.47

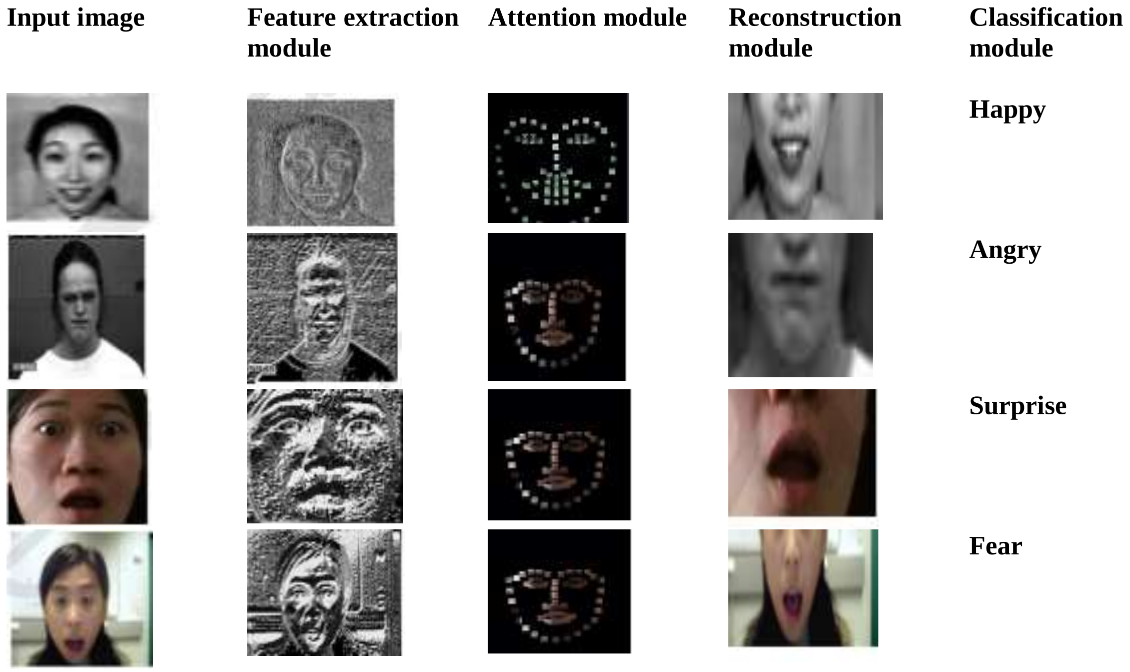


Figure 6: Results for Proposed attention based CNN model

Fig.6 illustrates the results of proposed attention based CNN model. Our method can achieve competitive result by using the facial expression data. In the feature extraction module, initial features are extracted from raw images and then sent to later modules for future processing. Generally, the obtained initial features are too coarse, but we need more refined features to send to the later attention module in order to make it work directly. Therefore, the feature extraction module is necessary in the network for extracting refined features. The attention module can make our network focus more on useful features and improve the accuracy. The reconstruction module can improve the accuracy by adjusting the attention map to create an enhanced feature map. We conclude that each module performs better for facial emotion recognition based on the experimental results.

6. CONCLUSION

In order to recognize facial expressions, this article introduces a novel convolutional neural network with an attention mechanism. The technique combines LBP features with convolution features before adding an attention mechanism to boost the network's performance. On the Oulu-CASIA and CK+ datasets, the proposed approach is assessed. The experimental findings demonstrate that, on these datasets, our technique outperforms a large number of other approaches. We will develop our architecture in the future to make it appropriate for video data, 3D face datasets, and our depth image data, and we will investigate improved machine learning techniques to strengthen the network.

REFERENCES

1. *Revina IM, Emmanuel WS (2021) A survey on human face expression recognition techniques. Journal of King Saud University-Computer and Information Sciences. 33(6):619-28.*
2. *Krumhuber EG, Küster D, Namba S, Skora L (2021) Human and machine validation of 14 databases of dynamic facial expressions. Behavior research methods. 53(2):686-701.*
3. *Liu Y, Sivaparthipan CB, Shankar A (2021) Human–computer interaction based visual feedback system for augmentative and alternative communication. International Journal of Speech Technology. 1-0.*
4. *Esra ME, Sevilen Ç (2021) Factors influencing EFL students' motivation in online learning: A qualitative case study. Journal of Educational Technology and Online Learning. 4(1):11-22.*
5. *Pikulkaew K, Boonchieng E, Boonchieng W, Chouvatut V (2021) Pain detection using deep learning with evaluation system. InProceedings of Fifth International Congress on Information and Communication Technology (pp. 426-435). Springer, Singapore.*
6. *Dzedzickis A, Kaklauskas A, Bucinskas V (2020) Human emotion recognition: Review of sensors and methods. Sensors. 20(3):592.*
7. *Tang Y, Zhang X, Hu X, Wang S, Wang H (2020) Facial expression recognition using frequency neural network. IEEE Transactions on Image Processing. 30:444-57.*
8. *Li B, Lima D (2021) Facial expression recognition via ResNet-50. International Journal of Cognitive Computing in Engineering. 2:57-64.*

9. *Li S, Li W, Wen S, Shi K, Yang Y, Zhou P, Huang T (2021) Auto-FERNet: A facial expression recognition network with architecture search. IEEE Transactions on Network Science and Engineering. 8(3):2213-22.*
10. *Wang Z, Zeng F, Liu S, Zeng B (2021) OAENet: Oriented attention ensemble for accurate facial expression recognition. Pattern Recognition. 112:107694.*
11. *Kola DG, Samayamantula SK (2021) A novel approach for facial expression recognition using local binary pattern with adaptive window. Multimedia Tools and Applications. 80(2):2243-62.*
12. *Gera D, Balasubramanian S (2021) Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recognition Letters. 145:58-66.*
13. *Kim JH, Kim N, Won CS (2022) Facial expression recognition with swin transformer. arXiv preprint arXiv:2203.13472.*