Intelligent and Adaptive Learning Techniques for Human Emotion Recognition

Sakshi Indolia

Department of Computer Science, Banasthali Vidyapith, Rajasthan 304022, India

sakshiindolia95@gmail.com

Abstract— **Human emotion recognition has a very wide variety of applications in medical treatment, sociable robots, human computer interaction, and recommendation systems. However, existing ML and DL algorithms for human emotion recognition are inefficient in terms of performance, limiting its adoption in widespread applications. The main objective of this work is to present intelligent algorithms for implementation in human emotion recognition through macro-facial expressions, as well as micro-facial expressions and EEG signals. For macro-facial expression recognition, we have exploited each of a feature fusion framework, self-attention mechanism, and discrete wavelet transform for classification of emotions. Humans possess an intrinsic ability to hide their true emotions through basic facial expressions, however, micro-expressions are subtle changes in facial muscles that are involuntary by nature and difficult to hide. To address this issue, a vision transformer model is used. Furthermore, while humans can hide their true emotion, this can be easily detected through physiological signals; thus, a bidirectional-long short-term memory is used for emotion recognition task. These models have been trained and tested on standard benchmark datasets.**

Index terms— **Deep Learning, Emotion Recognition, Self-Attention, Vision Transformer, Bidirectional Long Short-Term Memory, Transfer Learning, Data Augmentation.**

I. INTRODUCTION

Humans have an intrinsic ability to both express themselves and comprehend others' sentiments through emotions. Emotion can be defined as the response generated by humans against some external event [1]. Generally, human emotions are expressed by physically visible features, such as facial expressions [2], textual data [3], body gestures [4], verbal communication [5], eye movements [6], or physically invisible physiological signals such as electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG) [7], respiration (RSP) [8] and galvanic skin response (GSR) [9].

Human emotion recognition is an active research area due to its wide range of applications in human-computer interaction. There exist two different approaches to modeling emotions. The first approach discretely categorizes emotions as happy, sad, fear, disgust, anger, contempt, and surprise, whereas the second approach uses multiple dimensions to label emotions.

This paper aims to provide intelligent and adaptive learning algorithms for human emotion recognition based on facial expressions, micro-expressions, and physiological signals. In particular, effective deep learning models for the classification of human emotions using posed and spontaneous facial expressions are presented. These models exploit the self-attention mechanism and discrete wavelet transform for accurate computational algorithms. Furthermore, literature shows that transformer models are widely accepted in language processing tasks due to their remarkable performance. Thus, in this work, a modified version of an existing vision transformer has been exploited to perform human emotion recognition through micro-expressions. Moreover, a deep learning model has been proposed for human emotion recognition through physiological signals, specifically the EEG. The objectives of this work are twofold: (i) to perform emotion recognition through facial expressions and physiological patterns; and (ii) to propose intelligent and adaptive deep learning frameworks that address the existing issues of human emotion recognition.

Facial expression is a widely adopted approach for expressing emotions [10]. This research introduces three facial expression-based emotion recognition frameworks. To utilize both shallow and deep aspects of face expressions, first, a dual-stream feature fusion-based model has been presented, which shows that the proposed deep learning method attains higher classification accuracy with feature fusion. However, the performance of facial recognition models is affected by intra-class variation and inter-class similarity. Thus, the second framework addresses this issue by employing a modified self-attention mechanism. We also provide a detailed comparison of a primitive self-attention mechanism with the proposed self-attention model. The third framework focuses on improving feature representation through the wavelet domain combined with a self-attention mechanism. These three frameworks are validated on posed and spontaneous benchmark datasets: CK+ [11], JAFFE [12], MUG [13], FER2013 [14], and RAF [15]

Humans can hide their true emotions through facial expressions. Therefore, micro-expressions are another important category of facial expressions. They are typically seen when someone tries to hide their true sentiments. Accurate recognition of emotions through micro-expressions is helpful in a wide variety of applications. Thus, we have proposed a novel vision transformer that exploits local as well as global features of the input facial expressions. This model has been validated on three benchmark datasets: CASME-I [16], CASME-II [17], and SAM [18]. Finally, emotion recognition based on EEG signals is also analyzed. EEG signals are usually contaminated by unconscious movements such as eye blinks. For this purpose, we have used fast Fourier transform as a preprocessing technique. Furthermore, spatio-temporal features of EEG signals have been exploited by utilizing a bi-directional long-short-term memory deep learning model. This proposed model is validated on two benchmark datasets: DEAP and SEED.

In this work, various deep learning approaches are used to explore and analyze the recognition of human emotions through facial expressions, micro-expressions, and EEG signals. The following are the major contributions of this work:

i. An exhaustive literature review that includes the detailed descriptions and representations of facial expressions, micro-expressions, and EEG signals, along with insightful discussions of datasets, ML and DL methods for human emotion recognition.

ii. Deep learning models for facial expression recognition with self-attention mechanism are proposed to address the issues of illumination, head pose, intra-class variations, and inter-class similarity. The proposed models exploit the use of data augmentation and discrete wavelet transform (DWT) to improve performance.

iii. For micro-expression recognition, a new vision transformer model is proposed using a convolution operation for generating feature maps to exploit global receptive field.

iv. A method for EEG based emotion recognition is proposed using bi-directional long short-term memory (BiLSTM) model in conjunction with a fast Fourier transform (FFT).

Overall structure of the paper is as follows. Section II presents the proposed models for macro-facial expression recognition. Section III illustrates the proposed vision transformer for micro-expression recognition. In section IV, the proposed bidirectional model is discussed. Finally, section V provides a brief conclusion of the paper.

II. Facial Expression Recognition

Facial expressions can be defined as a medium through which an individual can convey an emotional state to others [19]. Several applications based on facial expression recognition (FER) have been developed in recent years, including virtual reality [20], autonomous robotics [21], autonomous driving [22], entertainment, healthcare [23] and gaming [24]. Motivated by the wide variety of applications, we have proposed three deep learning models. The first model incorporates a dual stream feature fusion-based deep model that exploits shallow as well as deep features of facial expressions. Experimental results of this model show that feature fusion improves the performance of the model; thus, in the other two proposed models, feature fusion is exploited. Furthermore, studies show that the performance of FER model can be improved by considering only relevant facial regions. So, the second model incorporates self-attention mechanism to identify the relevant facial regions required for FER. The third proposed model is a fusion framework that exploits self-attention mechanism in wavelet domain for better feature representation.

A. Deep Feature Fusion for Posed FER

A dual-stream feature fusion-based deep model (as shown in Figure 1) is proposed to exploit shallow as well as deep features of facial expressions [25]. For preprocessing, the facial portion in an image is detected through the Voila Jones algorithm [26]. It returns coordinates for region of interest (facial region), which is used to crop the facial region. Then, the cropped facial

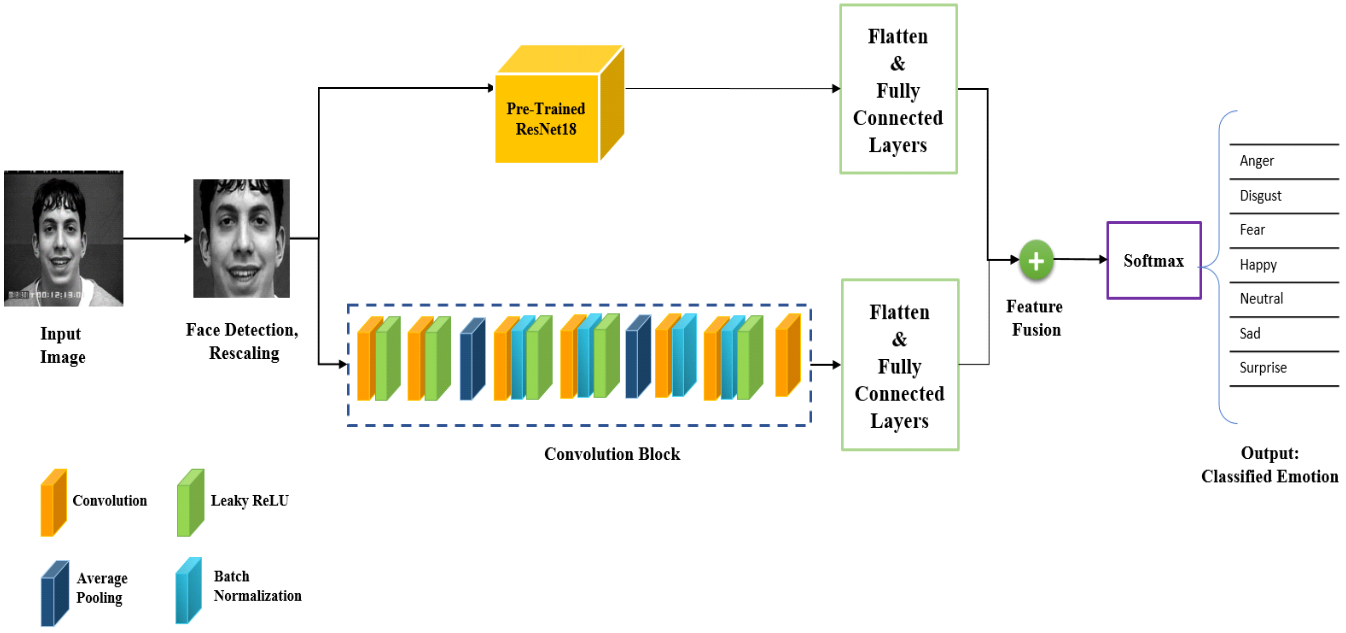

Fig. 1. FER through deep feature fusion for posed FER.

region is size normalized between 0 to 1. Due to the small size of the datasets, in-place data augmentation [27] is used to increase overall size of these datasets by the factor of 3 through application of geometric transformations such as scaling, width shift, height shift, rotation etc.

For the classification phase, we have divided this architecture into two branches. The preprocessed data is passed simultaneously to both the branches. The first branch incorporates a pre-trained ResNet18 model to extract features from the data. For experimental use, the last two layers of ResNet18 have been removed. Then, these features are passed to the fully connected block, which flattens the features and passes them to linear layers. The second branch consists of a sequence of seven convolutional layers with ReLU activation function. To reduce the size of input and the number of epochs, we exploited average pooling and batch normalization, respectively. Then, similar to the first branch, a fully connected block is used. Element-wise summation operation is performed on outputs received from both branches to perform feature fusion. Softmax activation function is applied to compute the probabilistic score required to classify the facial expressions.

Performance of the proposed network architecture has been evaluated in terms of classification accuracy on four benchmark datasets i.e., CK+, JAFFE, MUG, and YALE. The proposed model generated 90.62%, 98.36%, 88.36%, and 72.22% classification accuracy on JAFFE, CK+, MUG, and YALE datasets, respectively. Comparison with existing deep learning models shows that feature fusion helps the model achieve better classification accuracy.

B. *Self-Attention Mechanism and Transfer Learning*

CNN has played a very crucial role in FER. The convolution operation is performed by considering a local receptive field applied over input data. Due to this phenomenon, information associated with the entire data is lost. The attention mechanism addresses this problem by computing attention weights for a given input with respect to the entire input image.

Originally, an attention mechanism was introduced for natural language processing (NLP) applications. Conventional deep learning models in NLP are incapable of processing information in bidirectional sequence data and cannot handle long-term dependencies [28]. These issues can be addressed by using attention mechanism. Recent studies [29], [30] show that attention models are exploited in FER.

The performance of FER is degraded by intra-class variation and inter-class similarity. We resolve this issue by employing a specific variation of the attention models, i.e., self-attention mechanism, which identifies localized facial regions required for FER. Furthermore, the primitive self-attention mechanism has been improved by employing sigmoid activation function, and a detailed comparison of a primitive self-attention with the proposed self-attention model illustrates the efficacy of the proposed self-attention model.

The proposed dual stream ResNet18 attention (DSResNetAtt) based model (shown in Figure 2) is divided into two modules: preprocessing of faces and self-attention based classification [31]. The preprocessing phase is the same as that of the first model. For classification, the model takes the preprocessed and augmented images and passes them to the pre-trained ResNet18 network. For experimental purposes, the last two layers of ResNet18 have been removed. Therefore, ResNet18 generates feature maps of 512 × 7 × 7 pixel dimensions. These feature maps are further passed to convolution block 1 and convolution block 2, as shown in Figure 2. Convolution block 1 generates feature maps of Fc×h×wdimension that are passed to the self-attention block. Now, AttentionFc×h×wgenerated by attention block is passed through convolution block 3, and the generated result is added to the result produced by the convolution block

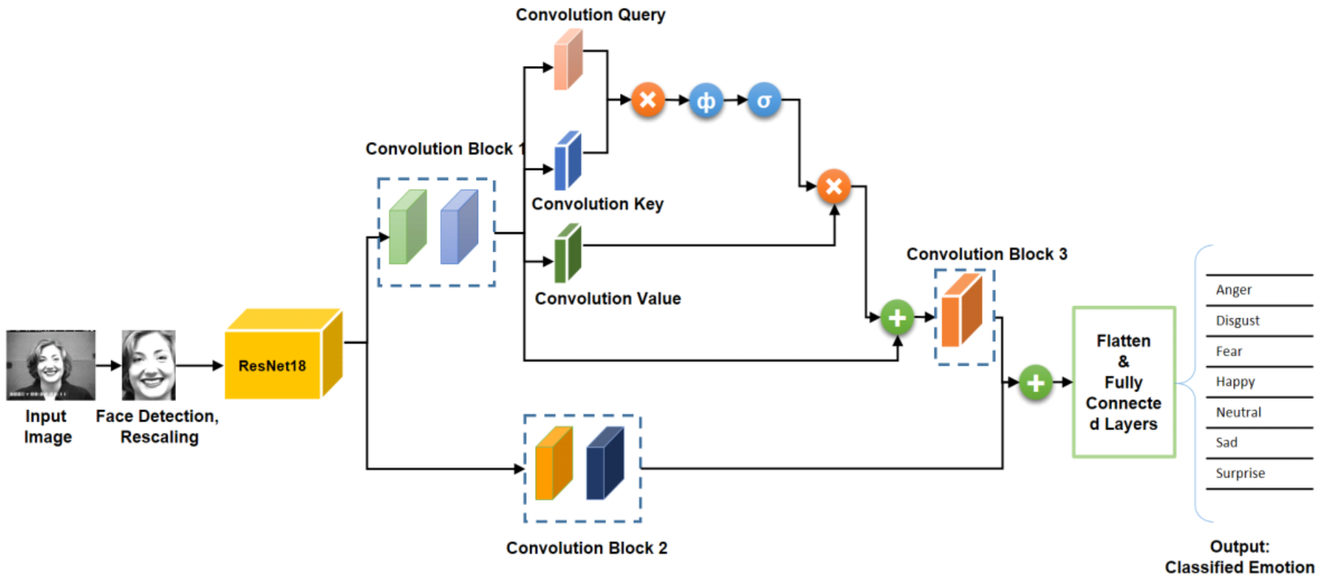

Fig. 2. FER through self-attention mechanism and transfer learning.

2, as shown in Figure 2. Thereafter, the result is flattened and passed to the block of fully connected layers. Then, we apply softmax activation σ(z) to compute the probabilistic score. Afterwards, the loss is computed through cross-entropy loss before the model is optimized by using the Adam optimizer.

Performance of the proposed model has been evaluated on JAFFE, CK+, RAF, MUG, YALE, and FER2013 dataset, shown in Table 1. It can be observed that the performance of the model improved when self-attention mechanism is incorporated.

TABLE I: RESULTS WITH SELF-ATTENTION BLOCK ON DIFFERENT DATASETS USING AUGMENTED SAMPLES

Dataset	Number of classes	Total number of samples	Accuracy without self-attention	Accuracy with self-attention
JAFFE	7	852	94.53%	97.00%
CK+	7	1218	98.91%	99.00%
RAF	7	15339	80.18%	81.06%
MUG	7	1203	91.68%	94.73 %
YALE	4	180	83.33%	87.03%
FER2013	7	32298	64.96%	64.89%

C. Self-Attention Based Fusion Framework in Wavelet Domain

This model is based on better feature representation through the wavelet domain combined with self-attention mechanism (as shown in Figure 3) [32]. The proposed model transforms the input image to wavelet domain through discrete wavelet transform. The framework employs two parallel branches for shallow and deep features, which are fused together for improved feature representation.

Initially, the preprocessing is employed to increase the overall size of small datasets such as JAFFE, MUG and YALE through data augmentation, and face detection is performed through Voila Jones algorithm. Then, the cropped image is passed through DWT to extract relevant features from an image. It decomposes the image into the following four sub-bands: LL, LH, HL, and HH, which provide approximate image, horizontal features, vertical features, and diagonal features extracted from the original image, respectively.

In this work, the LL sub-band is used as it has better feature representation capability and contains less noise. Thereafter, classification is performed by employing two parallel branches that take feature maps of 7 × 7 × 512 pixel size generated by the pre-trained ResNet18 model and follow the same steps as mentioned in Section II B. Performance of the proposed self-attention mechanism exploiting DWT domain is trained and validated on four benchmark datasets i.e., JAFFE, CK+, MUG, and YALE, which generates 96.87%, 99.18%, 94.18%, and 83.33% classification accuracy, respectively, which is comparable to the self-attention model without DWT.

III. MICRO-EXPRESSION RECOGNITION

Facial expression is an important aspect of social interaction among human beings. Basic facial expressions include anger, contempt, disgust, fear, happy, sad, and surprise [33]. Facial expressions can be broadly categorized as macro and micro-facial expressions. Macro expressions are visible, prolonged in nature and can be easily identified by human beings. However, these expressions can be suppressed, posed, and disguised to hide the true emotion of a person. Micro-expressions on the other hand, are involuntary in nature, cannot be posed, are visible for a very short duration of time (0:04 sec to 0:50 sec) and reflect the true emotion of a person. Based on these properties, micro-expression recognition (MER) has a variety

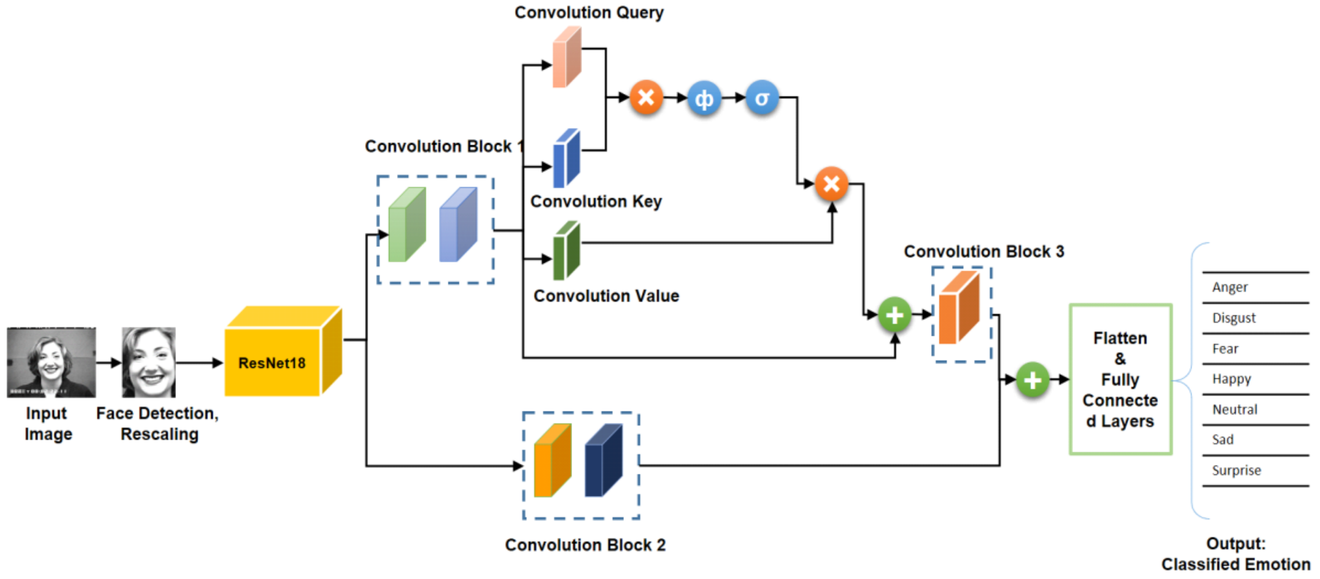

Fig. 3. FER through self-attention based fusion framework in wavelet domain.

of potential applications such as lie detection, security, surveillance systems, online learning, entertainment, healthcare systems (depression detection, clinical diagnosis), and forensics.

Attention mechanisms can be used to identify relevant facial region for classification. The ability of attention mechanisms to learn to concentrate on certain locations makes them effective. Attention mechanism is either employed in conjunction with CNN or it replaces certain components of CNN. In this work, we show that effective and accurate classification can be performed without deep CNN by exploiting vision transformer, which depends on self-attention mechanism. In the past few years, vision transformers have attained remarkable results on vision related classification tasks with substantially fewer computational resources.

Existing vision transformer models [34] create fixed-size patches from an input image, which are flattened and provided to the transformer for classification. However, this technique limits the performance of vision-based algorithms, because image pixels exhibit correlation with their neighboring pixels. Dividing images into fixed-size patches can deteriorate the correlation with neighboring pixels. Thus, a major limitation of this technique is that it cannot handle correlation among pixels in an image. To address this issue, the proposed algorithm (shown in Figure 4) generates c feature maps by applying c filters on an input image [35]. These feature maps are considered fixed-size image patches and passed to the transformer model that consists of transformer encoder (shown in Figure 5) for classification. The comparison of the number of heads in the transformer encoder in terms of classification accuracy is shown in Table II.

TABLE II COMPARISON OF NUMBER OF HEADS IN TRANSFORMER ENCODER

Number of Heads	Classification Accuracy
1	96.31%
2	95.62%
4	96.31%
8	97.08%
16	96.74%

The proposed model was evaluated on three benchmark datasets, CASME-I, CAMSE-II and SAMM, with classification accuracies of 95.97%, 98.59%, and 100% respectively.

IV. HUMAN EMOTION RECOGNITION BASED ON EEG SIGNALS

The human body generally exhibits some physical changes as a result of environmental events. Different emotional states are triggered in the human body as a response to these physical changes. Various physiological signals can be analyzed to monitor the emotional response to these physical changes. Although human beings can express their emotions through various visual factors, including body language and facial expressions depending on the situation, they may intentionally hide these visible emotions. Therefore, evaluating physiological data (which cannot be intentionally modified) gathered from various sensors can help determine a person's feelings in a variety of applications such as healthcare (depression, sleep disorders, epilepsy, Alzheimer, etc), human-computer interaction, surveillance systems, entertainment, and police interrogations.

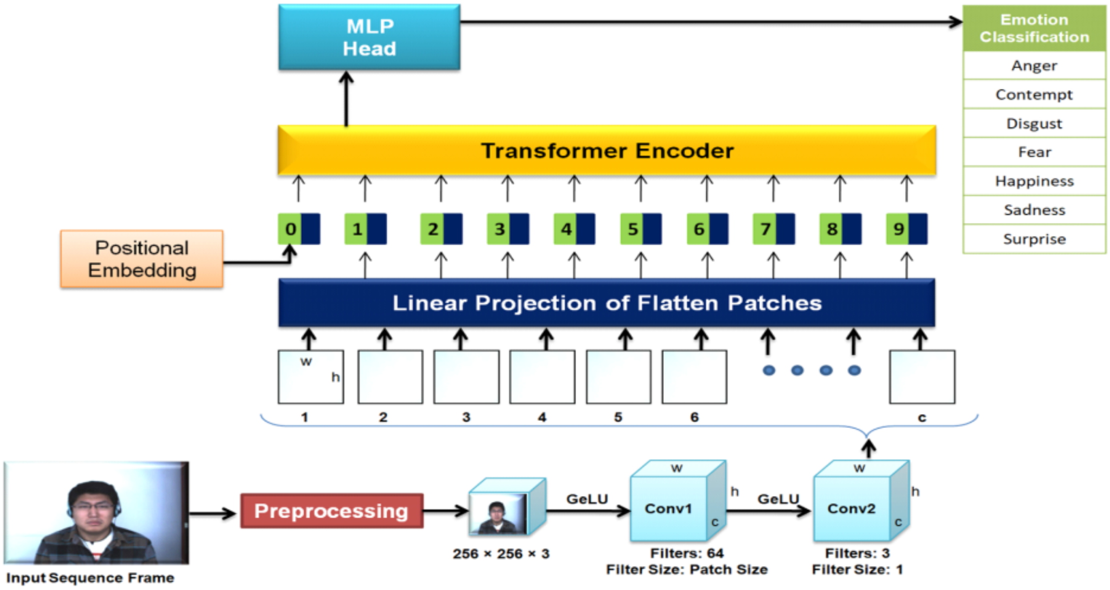

Fig. 4. Proposed vision transformer with convolution patches.

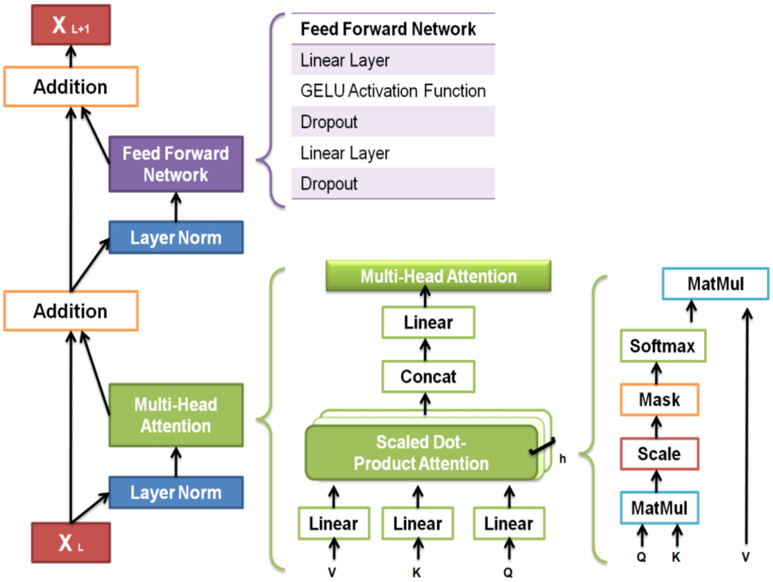

Fig. 5. Transformer encoder.

In this work, we have employed bi-directional long short-term memory for emotion recognition through EEG signals (as shown in Figure 6). The EEG signals are contaminated by unconscious movements such as eye blink. Thus, fast-fourier transforms have been applied in the proposed model as a preprocessing and feature extraction technique for enhancing performance of the model. Literature shows that EEG signals contain temporal as well as spatial information, thus, in this work, bi-directional long short-term memory is incorporated in the proposed model to exploit EEG signals. Performance of the proposed model has been evaluated on two benchmark datasets i.e., DEAP [36] and SEED [37]. DEAP dataset generated 91.53%, 90.55%, 89.77%, and 90.24% for valence, arousal, dominance, and liking, whereas 98.68% classification accuracy was produced by the proposed model for the SEED dataset.

V. Conclusions and Future Works

The work described in this paper includes human emotion recognition through facial expressions, micro-expressions, and EEG signals. In this work, accurate facial expression recognition is demonstrated through the proposed models, which incorporate feature fusion, self-attention mechanisms, and wavelet domain techniques. These models generated state-of-the-art results for laboratory-controlled datasets, however, they did not perform as expected on in-the-wild datasets. Micro-expression recognition is a challenging task. Vision transformers have demonstrated their classification performance in various domains. Motivated by the success of

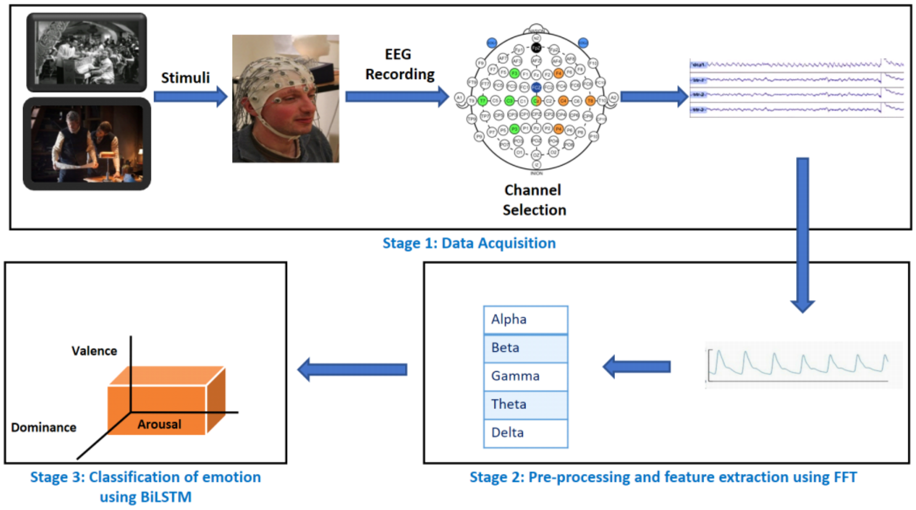
Stimuli
EEG Recording
Channel Selection
Stage 1: Data Acquisition
Valence
Dominance
Arousal
Stage 3: Classification of emotion using BiLSTM
Alpha
Beta
Gamma
Theta
Delta
Stage 2: Pre-processing and feature extraction using FFT

Fig. 6. Pipeline for human emotion recognition through EEG signals.

vision transformers, we proposed a convolutional patch-based vision transformer, which outperformed many state-of-the-art MER models. With the exceptional results generated by our proposed vision transformer on MER, we aim to test it on in-the-wild FER datasets and compare the results with self-attention-based models in the near future. Furthermore, we have performed human emotion recognition through EEG signals using the Bi-LSTM model. Spatial and temporal information from these signals can also be captured through the transformer model, which will be our future direction in this field.

References

[1] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," Sensors, vol. 18, no. 7, p. 2074, 2018.

[2] C. Darwin and P. Prodger, The expression of the emotions in man and animals. Oxford University Press, USA, 1998.

[3] Y.-S. Seol, D.-J. Kim, and H.-W. Kim, "Emotion recognition from text using knowledge-based ann," in ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications, 2008, pp. 1569-1572.

[4] S. Piana, A. Stagliano, F. Odone, and A. Camurri, "Adaptive body gesture ` representation for automatic emotion recognition," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 6, no. 1, pp. 1-31, 2016.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2203-2213, 2014.

[6] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eyetracking: taxonomy, review and current challenges," Sensors, vol. 20, no. 8, p. 2384, 2020.

[7] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, "Electrocardiogram-based emotion recognition systems and their applications in healthcare-a review," Sensors, vol. 21, no. 15, p. 5015, 2021.

[8] H. U. R. Siddiqui, H. F. Shahzad, A. A. Saleem, A. B. Khan Khakwani, F. Rustam, E. Lee, I. Ashraf, and S. Dudley, "Respiration based non-invasive approach for emotion recognition using impulse radio ultra wide band radar and machine learning," Sensors, vol. 21, no. 24, p. 8336, 2021.

[9] D. Ayata, Y. Yaslan, and M. Kamas¬∏ak, "Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods," IU-Journal of Electrical & Electronics Engineering, vol. 17, no. 1, pp. 3147-3156, 2017.

[10] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 2, pp. 97-115, 2001.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010, pp. 94-101.

[12] M. Kamachi, M. Lyons, and J. Gyoba, "The japanese female facial expression (jaffe) database," URL http://www. kasrl. org/jaffe. html, vol. 21, p. 32, 1998.

[13] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. IEEE, 2010, pp. 1-4.

[14] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learning: A report on three machine learning contests," in International conference on neural information processing. Springer, 2013, pp. 117-124.

[15] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2852-2861.

[16] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, 2013, pp. 1-7.

[17] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," PloS one, vol. 9, no. 1, p. e86041, 2014.

[18] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," IEEE transactions on affective computing, vol. 9, no. 1, pp. 116-129, 2016.

[19] C. Darwin, "Chapter viii. joy, high spirits, love, tender feelings, devotion," in The Expression of the Emotions in Man and Animals. University of Chicago Press, 2015, pp. 196-219.

[20] A. Nijholt, "Capturing obstructed nonverbal cues in augmented reality interactions: a short survey," in Proceedings of International Conference on Industrial Instrumentation and Control. Springer, 2022, pp. 1-9. 159

[21] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human-robot interaction," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 3, pp. 1473-1484, 2019.

[22] A. Marceddu, "Automatic recognition and classification of passengers' emotions in autonomous driving vehicles," Ph.D. dissertation, Diss. Politecnico di Torino Torino, Italy, 2019

[23] T. Altameem and A. Altameem, "Facial expression recognition using human machine interaction and multi-modal visualization analysis for healthcare applications," Image and Vision Computing, vol. 103, p. 104044, 2020

[24] P. M. Blom, S. Bakkes, C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers, "Towards personalised gaming via facial expression recognition," in Tenth Artificial Intelligence and Interactive Digital Entertainment Conference, 2014.

[25] S. Indolia, S. Nigam, and R. Singh, "Deep Feature Fusion for Facial Expression Recognition," in Second International Conference on Next Generation Intelligent Systems (ICNGIS) (pp. 1-6). IEEE, 2022.

[26] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1. Ieee, 2001, pp. I-I.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[28] N. Yu and D. Bai, "A visual self-attention network for facial expression recognition," in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1-8.

[29] Y. Fan, V. Li, and J. C. Lam, "Facial expression recognition with deeplysupervised attention network," IEEE transactions on affective computing, 2020.

[30] C. Gan, J. Xiao, Z. Wang, Z. Zhang, and Q. Zhu, "Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention," Image and Vision Computing, vol. 117, p. 104342, 2022.

[31] S. Indolia, S. Nigam, and R. Singh, "A framework for facial expression recognition using deep self-attention network," Journal of Ambient Intelligence and Humanized Computing, 14(7), 9543-9562, 2023.

[32] S. Indolia, S. Nigam, and R. Singh, "A self-attention-based fusion framework for facial expression recognition in wavelet domain", Visual Computing, 2023.

[33] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." Journal of Personality and Social Psychology, vol. 17, no. 2, p. 124, 1971.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[35] S. Indolia, S. Nigam, R. Singh, V. K. Singh., and M. K. Singh, "Micro Expression Recognition using Convolution Patch in Vision Transformer", IEEE Access, 2023.

[36] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," IEEE transactions on affective computing, vol. 3, no. 1, pp. 18-31, 2011.

[37] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 162-175, 2015.