IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2023, accepted 8 September 2023, date of publication 13 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314797

RESEARCH ARTICLE

Micro Expression Recognition Using Convolution Patch in Vision Transformer

SAKSHI INDOLIA[1,2], SWATI NIGAM[1,2], (Senior Member, IEEE), RAJIV SINGH[1,2], (Senior Member, IEEE), VIVEK KUMAR SINGH[3], (Senior Member, IEEE), AND MANOJ KUMAR SINGH[3]

[1]Department of Computer Science, Banasthali Vidyapith, Tonk, Rajasthan 304022, India
[2]Centre for Artificial Intelligence, Banasthali Vidyapith, Tonk, Rajasthan 304022, India
[3]Department of Computer Science, Banaras Hindu University, Varanasi, Uttar Pradesh 221005, India

Corresponding authors: Rajiv Singh (jkrajivsingh@gmail.com) and Vivek Kumar Singh (vivekks12@gmail.com)

This work was supported in part by the Ministry of Electronics and Information Technology (MeITY), Government of India, under Grant 3(9)/2021-EG-II; and in part by the Hewlett Packard Enterprise (HPE) Aruba Centre for Research in Information Systems, Banaras Hindu University (BHU), under Grant M-22-69.

ABSTRACT Humans possess an intrinsic ability to hide their true emotions. Micro-expressions are subtle changes in facial muscles that are involuntary by nature and easy to hide. To address these issues, several machine and deep learning models have been proposed in the past few years. Convolution neural network (CNN) is a deep learning method that has widely been adopted in vision-related tasks due to its remarkable performance. However, CNN suffers from overfitting due to a large number of trainable parameters. Additionally, CNN cannot capture global information with respect to an input image. Furthermore, the identification of important regions for the classification of micro-expressions is a challenging task. Self-attention mechanism addresses these issues by focusing on key areas. Furthermore, specific transformers, known as vision transformers are widely explored in vision-related applications. However, existing vision transformers divide an input image into a fixed number of patches due to which local correlation of image pixels is lost. Further, a vision transformer relies on self-attention mechanism which effectively captures global dependencies but does not exploit the local spatial relationships in an image. In this work, we propose a vision transformer based on convolution patches to overcome this problem. The proposed algorithm generates c number of feature maps from input images using c filters through convolution operation. These feature maps are then applied to a transformer model as fixed-size image patches to perform classification. Thus, the proposed architecture leverages advantages of both convolutional layers and transformer, and captures both spatial information and global dependencies respectively, leading to improved performance. The performance of the proposed model is evaluated on three benchmark datasets: CASME-I, CASME-II, and SAMM and compared with state-of-the-art machine and deep learning models, which generated classification accuracy of 95.97%, 98.59%, and 100%, respectively.

INDEX TERMS Facial expression recognition, deep learning, micro-expression recognition, self-attention, vision transformer.

I. INTRODUCTION

Micro-expressions (ME) are involuntary subtle facial muscle movements which represent true emotions of a person [1]. There are a variety of possible applications for micro-expression recognition (MER), including forensics, security, surveillance, education, entertainment, and health-care systems [2]. However, identification and classification of ME is a challenging task due to a variety of reasons. Typically, ME appear for a very short duration of time, i.e., 0:04 to 0:50 seconds [3]. Furthermore, ME show very subtle change in facial muscles, due to which identification and spotting of ME become difficult.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Traditional machine learning methods such as local binary patterns [4] and histogram of oriented gradients (HOG) [5], [6], [7], depend on handcrafted features for classification. This dependency has been avoided by the use of deep learning models. Convolutional neural network (CNN) is a deep learning method which has recently demonstrated remarkable performance in several vision based applications and outperformed both handcrafted features and shallow classifiers [8]. A deep fusion-based CNN model proposed by [9] has been implemented for facial expression recognition, which shows the impact of transfer learning and feature fusion on the performance of the model. Similarly, CNN and transfer learning are incorporated by [10] to determine the level of engagement of hearing impaired and hard-of-hearing students by analyzing their facial expressions and categorizing these expressions as highly engaged, nominally engaged or not engaged.

CNN requires large training dataset; however, most of the publicly available MER datasets are small in size, [11] used a data augmentation technique for CNN to increase the size of the facial expression datasets. Similarly, a CNN-based MER model proposed by [12], exploits optical flow information related to subtle muscle movements through apex and reference frame. Then, this information is passed to a CNN model for classification of an emotion. In the past few years, performance of CNN has been elevated by using it in stream or branch based networks.

However, implementation of CNN models in MER is limited due to variety of reasons: (i) CNN requires large number of trainable parameters (ii) CNN based models often suffer from overfitting (iii) convolution operation only captures local receptive field of a pixel and it is incapable of handling global receptive field, (iv) CNN does not effectively handle sparse spatio-temporal information, and, (v) ME consist of subtle movements of facial muscles which are difficult to handle.

As mentioned above, CNN is incapable of handling spatio-temporal information. Hence, 3D CNN has been explored by [13], [14], and [15] to address this issue for MER. A Siamese 3D CNN (MERSiamC3D) proposed by [13] is based on two-stage learning. The first stage applies an optical flow estimation technique to explain the spatio-temporal information, followed by a Siamese CNN model. The second stage adjusts the network parameters obtained from the first stage. Similarly, [15] also exploits a 3D CNN in combination with SqueezeNet. Another work proposed by [16], incorporates Squeeze-and-Excitation Networks with a 3D DenseNet to exploit spatio-temporal features.

The ability of attention mechanism to concentrate on certain locations makes it effective. Attention mechanism is either employed in conjunction with CNN or it replaces certain components of CNN. Accurate detection of ME plays a vital role in improving performance of the MER model. Attention mechanism can be used to effectively detect the presence of micro-expression in a video frame. A dual attention network known as LGAttNet, was proposed by [17] for automatic detection of micro-expression. Similarly, micro-expression analysis network (MEAN) proposed by [18], is used for simultaneous spotting and recognition of ME.

In this work, effective and accurate classification is performed 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. A simple vision transformer typically divides an image into fixed size patches. These non-overlapping patches form a linear embedding which is provided to the vision transformer. This architecture captures global dependencies but cannot capture spatial information. On the other hand, the proposed vision transformer architecture takes convolution feature maps as input patches; these feature maps contain spatial information. These feature maps are then provided as the input patches for the subsequent transformer layers. Thus, the proposed architecture leverages advantages of both convolutional layers and transformer and captures both spatial information and global dependencies for improved performance. Due to its remarkable performance, the proposed model can have a wide variety of real-life applications across different domains. For instance, the proposed model can be used for early detection and diagnosis of mental health issues such as anxiety and depression. It can also be used in security and law enforcement, where, security personnel can improve their ability to recognize possible threats by identifying ME associated with suspicious behavior. Further, MER can also play a very vital role for applications based on human-computer interaction and cross-cultural studies.

The major contributions of this paper are as follows:

1) We propose a deep learning framework for MER through a vision transformer with low computational cost.
2) Conventional vision transformers divide input image into fixed-sized patches; due to which it becomes difficult for the model to exploit the local correlation of pixels. The proposed model addresses this issue by maintaining the correlation of the target pixel with its neighbors through a local receptive field by using convolution patches.
3) We exploit global as well as local correlation in an image through a vision transformer and convolution patch respectively, which improves the overall classification performance of the model.
4) Extensive experiments have been performed on three benchmark datasets and comparison with existing state-of-the-art models validates the effectiveness of the proposed model.

The remaining sections of the paper are arranged as follows. Section II discusses the related work. Section III presents the proposed vision transformer for MER. Section IV provides a description of the datasets, experimental setup, and hyper-parameters for training the model, results, and comparison

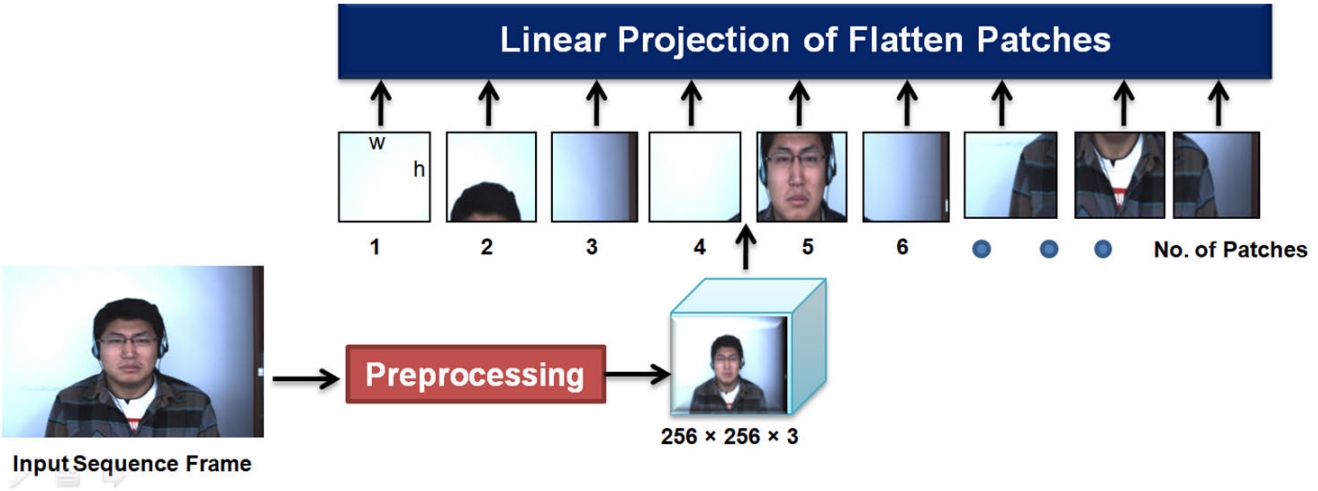


FIGURE 1. Flattening of image patches in conventional vision transformer.

of the proposed model with current state-of-the-art models. Conclusions are provided in Section V.

II. RELATED WORKS

A. MICRO-EXPRESSION RECOGNITION

Based on input data, MER models can be broadly categorized into single-image-based and sequence-image-based systems. Datasets such as AffectNet [19] and FER2013 [20] are single-image-based datasets, whereas, CASME-I [21], CASME II [22], SAMM [23], and SMIC [24] are sequence-image-based datasets. Sequence-image-based datasets are widely adopted for spotting and recognition of micro-expressions because they provide better insight into data. However, primitive sequence-image-based datasets such as USF-HD [25] and Polikovsky's [26] are not adopted at present because such datasets contain image sequences of posed expressions, and hence they cannot be used for practical implementations. Whereas, state-of-the-art ME datasets contain spontaneous image sequences captured in a laboratory-controlled environment. Due to the availability of these datasets, research in the MER domain has significantly accelerated.

Primitive approaches for MER rely on hand-crafted and low-level features such as local binary pattern (LBP) [4], gradient features and optical flow. Local binary pattern from three orthogonal planes (LBP-TOP) is a commonly used feature for MER which considers horizontal and vertical directions. However, LBP-TOP cannot capture muscle movements in oblique direction which is essential for MER. To address this problem, [27] proposed a new feature called LBP-FIP which could easily capture dynamic textures from images calculated through five intersecting planes. Similarly, [28] proposed an invisible emotion magnification algorithm (IEMA) which effectively magnifies the strength of facial muscle movement for better classification of micro-expression.

However, it is difficult to accurately interpret and represent ME through low-level features. Thus, a combination of several low-level features forming high-level features can be exploited for a better representation of ME. High-level feature representations can be obtained by deep learning models such as CNN. At an early stage, researchers exploited only spatial features [29], [30] through CNN, however, studies demonstrate that MER involves facial movement which can be captured through long image sequences. Thus, state-of-the-art MER models exploit both spatial as well as temporal information. A Deep 3DCNN-ANN model proposed by [31] performs micro-expression recognition by learning spatio-temporal features from the image sequences by combining deep 3DCNN and ANN through a feature called visual associations. However, it has been observed that CNN cannot capture the relationship of an entity with its parent as an image. To address this issue, [32] proposed CapsuleNet based on agreement routing mechanism for MNIST dataset. Inspired by its success, [33] experimented CapsuleNet for MER model on SMIC, CASME-II and SAMM datasets. It has been observed that training a model on a particular dataset may not necessarily perform well on other dataset. Thus to experiment with cross-dataset MER, [34] proposed a dual-inception network which exploits horizontal and vertical components extracted through optical flow.

B. TRANSFORMERS

Transformer model was originally designed for text based applications [35], where it has exhibited remarkable results. Inspired by its success, it has also been experimented in vision tasks [36]. Vision transformers (ViT) take image as an input and represent it as a series of fixed size image patches as shown in Figure 1. The obtained image patches are flattened and subjected to lower dimensional linear embedding. Due to flattening of patches, the correlation between adjacent patch might be lost. Therefore, positional embedding is added to

keep the correlation information intact. Furthermore, vision transformers rely on self-attention mechanism that provides global receptive field, unlike CNN, which yields local receptive field.

Considering the limitations of CNN models, vision transformer has been widely adopted for MER models. A late-fusion based vision transformer proposed by [37], exploits motion features through optical flow. Late-fusion and optical flow mechanisms allow the model to deal with small ME datasets. Similarly, a muscle motion-guided network proposed by [38], exploits the subtle muscle motion features for accurate classification of ME through a two branch model. The first branch comprises of a continuous attention block, which focuses on modeling muscle movement, whereas, the second branch comprises of a position calibration module which consists of a vision transformer.

Studies show that MER is difficult due to the fact that they are highly dynamic in nature and appear on localized facial regions. To solve this problem, [39] proposed a sparse transformer which exploits multi-head attention for sparse representation of emotions appearing in localized facial regions, whereas, temporal attentional fusion is employed to deal with dynamic nature of ME. Furthermore, studies [40] show that combination of local and global spatio-temporal pattern can improve classification accuracy of MER. To address the spatial patterns, a spatial encoder is employed, whereas, a temporal aggregator models the temporal patterns.

Another work proposed by [41], exploits two swin vision transformers F_transformer and S_transformer placed in two parallel streams. F_transformer exploits short term motion dynamics through optical flow sequences, whereas, long-term motion dynamics are utilized through S_transformer. Later, feature fusion is performed on features obtained from these two streams for classification of emotions.

However, the existing vision transformer models for MER divides the input image into n patches, due to which the local correlation of pixels with its neighboring pixels is lost. To address this issue, in this work, we exploit feature maps generated by convolution operation. Furthermore, convolution operation helps to capture local receptive field, and self-attention mechanism in vision transformer allows the model to capture global receptive field.

III. PROPOSED METHODOLOGY

Existing vision transformer models [36], [42] create fixed size patches from input image, which are flattened and provided to 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 patch deteriorates 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 generates c feature maps by applying c filters on an input image. These feature maps are considered as fixed size image patches and passed to transformer model for classification.

A. PRE-PROCESSING AND CONVOLUTION PATCH

Figure 2 presents detailed network architecture of the proposed model. First, the input sequence frames are provided to the network through a pre-processing stage. The input frames are subjected to pre-processing operations such as horizontal flip, normalization and resize to 256 × 256 pixels. After pre-processing, the images of 3 × 256 × 256 pixel dimension are generated. Next, to exploit local correlation, two subsequent convolution operations are applied. First convolution operation takes images of 16 × 3 × 256 × 256 pixel dimension, where, 16 is the batch size and applies 64 filters with stride equivalent to patch size i.e., 16. Then, Gaussian error linear unit (GELU) activation function proposed by [43] is applied, where GELU is computed by Equation 1.

$$GaussianErrorLinearUnit(z) = 0.5 \times z \times (1 + Tanh(\sqrt{\frac{2}{\pi}} \times (z + 0.44715 \times z^3))) \quad (1)$$

Thereafter, another convolution operation is applied which takes 64 feature maps and applies 3 filters with stride 1. Next, GELU activation function is applied to the obtained feature maps of dimension 16 × 3 × 256 × 256, which are reshaped to obtain 16 × 256 × 256 × 3 feature maps.

B. VISION TRANSFORMER

Conventional vision transformer models divide an image of dimension h × w pixels into n × m number of fixed size patches (as shown in Figure 1), where each patch is of h/n × w/m pixel dimension. Thereafter, these patches are flattened and passed through linear projection.

However, in the proposed work, we exploit local correlation of images through convolution operation, shown in Figure 2. Here, feature maps of 16 × 256 × 256 × 3 dimension are flattened to form 16 × 256 × 768 feature vector. To maintain the order of sequence, we add positional embedding and perform reshape operation to generate feature vector of shape 257 × 16 × 256. It is further passed to six subsequent transformer encoders. The final feature vector is of shape 257 × 16 × 256, passed to multi-layer perceptron (MLP) head for classification of emotions. Figure 3 illustrates a single transformer encoder, which incorporates Multi-head attention, which is further based on self-attention mechanism.

Attention mechanism was introduced in encoder-decoder block of a neural sequence transduction model by [44]. It enable content-based summary of data from a variable length sentence. Attention mechanism is widely adopted because it has the ability to learn to focus on key areas. Self-attention mechanism also called intra-attention [35], allows the model to identify the inputs we should pay more attention to. It is used by [45] for facial expression recognition to deal with intra-class variation and inter-class similarity. It computes a weighted average of sequence elements where the weights are dynamically determined using the element

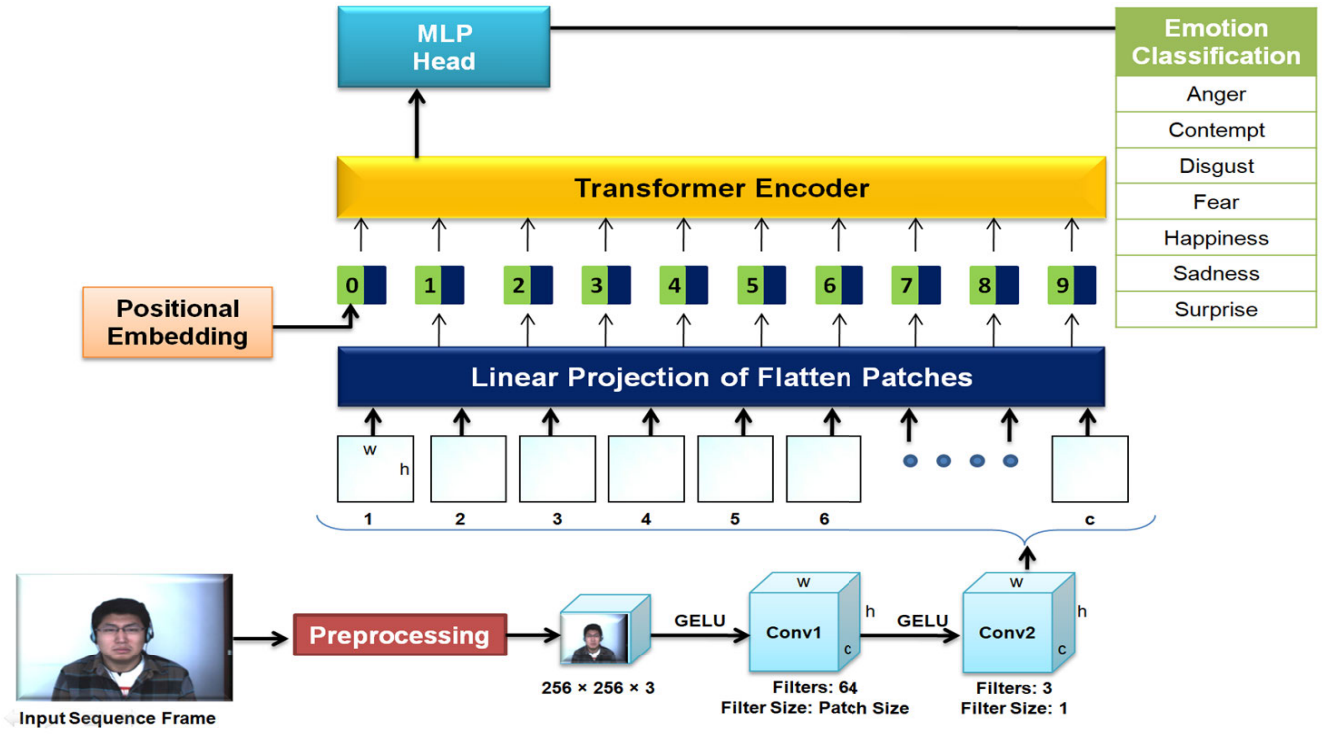


FIGURE 2. Detailed architecture of proposed model for MER using vision transformer.

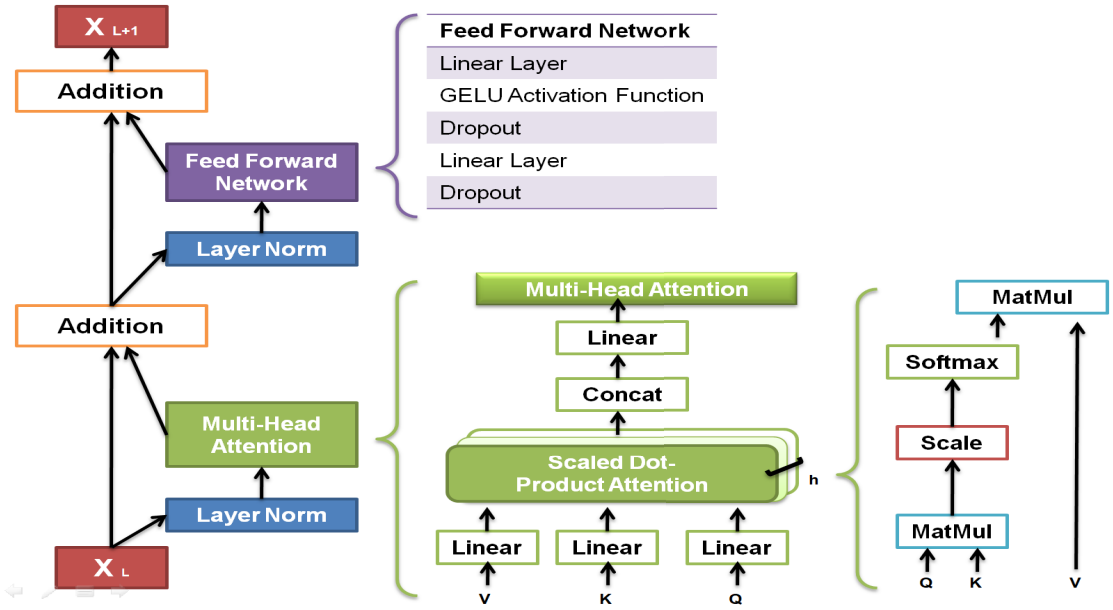


FIGURE 3. Transformer encoder.

keys and an input query. Attention mechanism rely on three feature vectors, key, query and value. In Figure 3, *Query* feature vector (represented by Q) attempts to identify the sequence-specific information the model is searching for. *Key* vector (represented by K), describes what the input element is offering. The *Value* vector (represented by V) is the one that we intend to average over. In this work, we exploit scaled dot product attention which takes $Query \in \mathbb{R}^{SL \times d_k}$, $Key \in \mathbb{R}^{SL \times d_k}$ and $Value \in \mathbb{R}^{SL \times d_v}$, where SL is sequence length, d_k and d_v are hidden dimensionalities. The scaled dot product

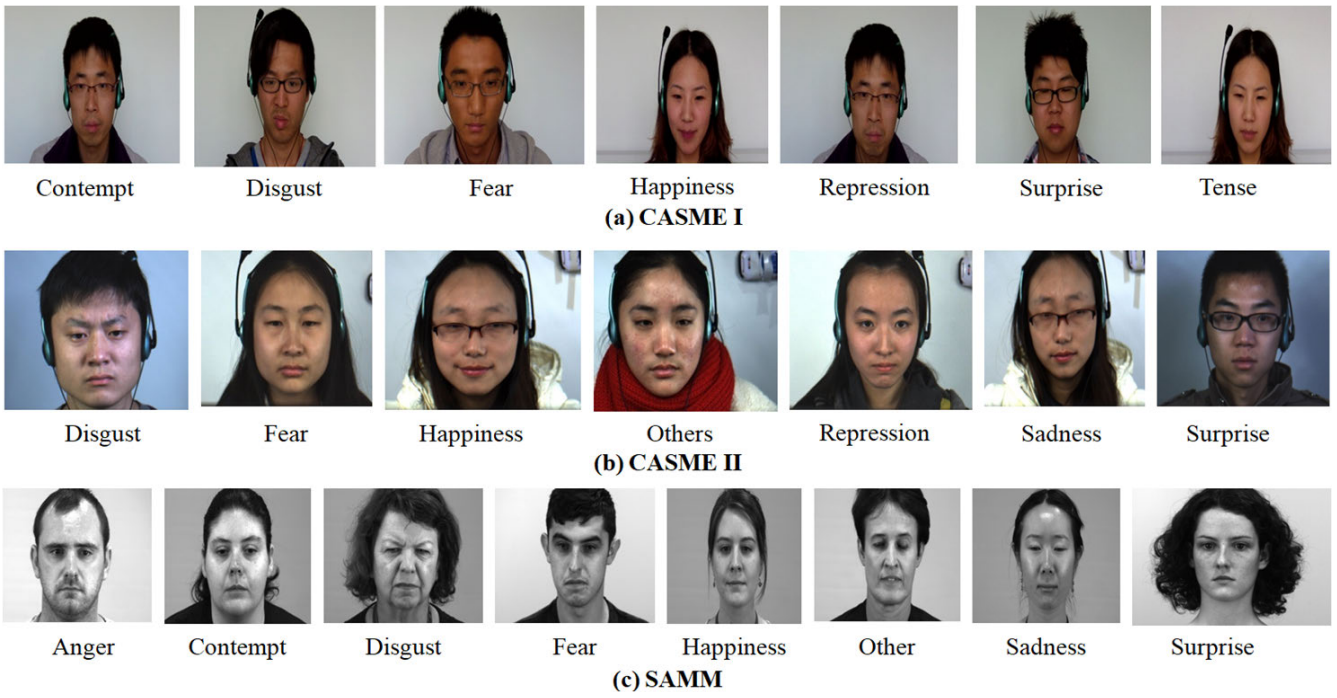


FIGURE 4. Sample images of (a) CASME-I (b) CASME-II (c) SAMM datasets.

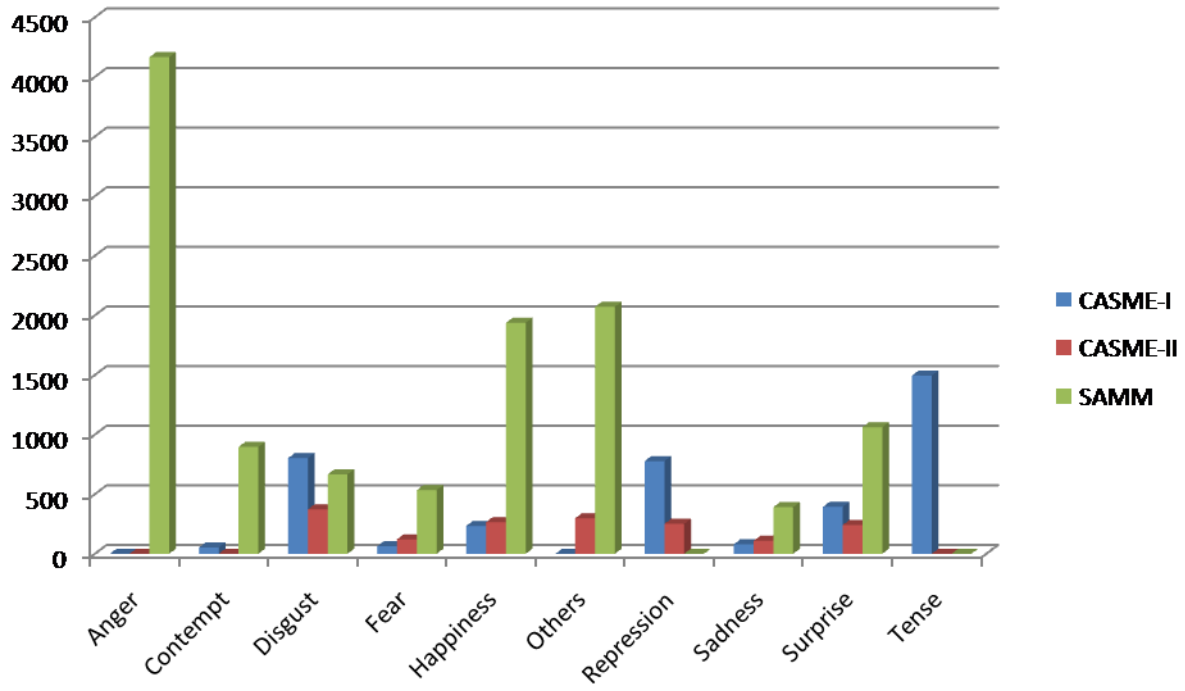


FIGURE 5. Unbalanced nature of emotion samples in datasets.

attention is computed by Equation 2.

$$Attention(Query, Key, Value) = Softmax(\frac{QueryKey^{SL}}{\sqrt{d_k}})Value \quad (2)$$

where, $\frac{1}{\sqrt{d_k}}$ is the scaling factor, used to monitor the variance of attention values. In Equation 2, *Query* and *Key* are two vectors with σ^2 variance, when a product operation is applied on *Query* and *Key*, it generates a scalar with d_k times higher variance. Thus, there is a need to scale down the variance back

TABLE 1. Description of datasets (a) CASME-I, (b) CASME-II, and (c) SAMM.

Characteristics	CASME-I	CASME-II	SAMM
Samples	96	255	159
Subjects	35	26	29
Ethnicity	1	1	13
FPS	60	200	200
Resolutions (in pixels)	1200 × 720	640 × 480	2040 × 1088
Class Labels	8	7	8

to σ^2, otherwise, softmax will make one random element saturate to 1 and other elements saturate to 0. Therefore, we use d_k for scaling, to maintain the optimal variance of attention values.

A network can pay attention to a particular sequence with scaled dot product attention. However, it does not allow sequence elements to attend to different features. This can be achieved through multi-head attention. Here, key, query, and value matrices are converted into h sub-keys, sub-queries, and sub-values respectively. Each of these sub-components is then independently passed through a h_i scaled dot product attention with weight matrices W_i^Q and W_i^K. Thereafter, these h heads are concatenated and it generates final weight matrix W^O.

$$Multi-head(Q, K, V) = Concat(h_1, h_2, \ldots, h_h)W^O \quad (3)$$

where, $h_i = Attention(QW_i^Q, KW_i^K, VW_i^K)$

$$W_{1\ldots h}^Q \in \mathbb{R}^{D\times d_k} \quad (4)$$

$$W_{1\ldots h}^K \in \mathbb{R}^{D\times d_k} \quad (5)$$

$$W_{1\ldots h}^V \in \mathbb{R}^{D\times d_v} \quad (6)$$

$$W^O \in \mathbb{R}^{h\cdot d_k \times d_o ut} \quad (7)$$

where, D is the input dimensionality.

IV. EXPERIMENT AND RESULTS

A. DATASETS

Performance of the proposed model has been tested on three benchmark datasets CASME-I [21], CASME-II [22], and SAMM [23]. Sample images of these datasets are shown Figure 4. Table 1 describes detail of these datasets on the basis of number of video samples, subjects, ethnicity, frames per second (FPS), resolutions (in pixels) and number of emotion labels. Figure 5 illustrates unbalanced nature of emotion samples in datasets. Furthermore, class-wise sample distribution is illustrated in Table 2. Video sequences containing the onset frame, progressing toward the apex emotion, and then ending with the offset frame are used to train the model.

B. EXPERIMENTAL SETUP AND TRAINING HYPERPARAMETERS

The proposed model is trained using Nvidia A100 provided by Google Colab Pro+. Adam optimizer is used for optimization of model weights, learning rate is set to 0.0003 and batch size is 16. We initially tuned the number of heads for training the proposed vision transformer model; to ensure a fair comparison, same number of heads is used for ablation experiments. We have investigated the model based on 1, 2, 4, 8, and 16 heads. As shown in Table 3, it can be observed that the selection of 8 heads outperformed other variants. Thus, 8 heads are selected in multihead attention module of the transformer encoder for all the experiments. Other parameters used in the proposed transformer encoder are listed in Table 4. To avoid overfitting of our model, we have exploited dropout regularization technique and layer normalization. In our proposed model, we have chosen layer normalization technique over batch normalization. The reason is that, in batch normalization, each feature in the mini-batch is independently normalised, whereas, layer normalisation normalises each input in the batch across all features. Further, we compare our proposed model on the basis of number of trainable parameters and GFLOPS as shown in Table 5. It can be observed that the proposed model outperforms existing state-of-the-art transformer and CNN based models.

TABLE 2. Number of frames against each emotion for (a) CASME-I, (b) CASME-II, and (c) SAMM datasets, used for training the proposed model.

Emotions	CASME-I	CASME-II	SAMM
Anger	-	-	4165
Contempt	52	-	896
Disgust	802	373	666
Fear	63	121	534
Happiness	234	266	1937
Others	-	298	2071
Repression	777	251	-
Sadness	79	108	391
Surprise	393	241	1062
Tense	1495	-	-
Total	3895	1858	11723

TABLE 3. 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%

C. RESULTS AND DISCUSSION

1) PERFORMANCE ANALYSIS

The proposed model is trained and tested on three benchmarks datasets i.e., CASME-I, CASME-II and SAMM. The model is evaluated in terms of classification accuracy, precision, recall and F1-score. The training and validation accuracy of CASME-I, CASME-II and SAMM datasets are shown in Table 6. The validation accuracy of SAMM dataset is 100%, which raises concerns about potential overfitting.

In order to rule out this possibility, we have used layer normalization (LayerNorm) as a regularization technique, as shown in Figure 3, and also applied dropout technique with value 0.2 to mitigate overfitting in our proposed model. For further analysis, we have plotted the training and validation curves as mentioned in Figure 10 (where, x-axis represents epochs and y-axis represents accuracy), to closely monitor

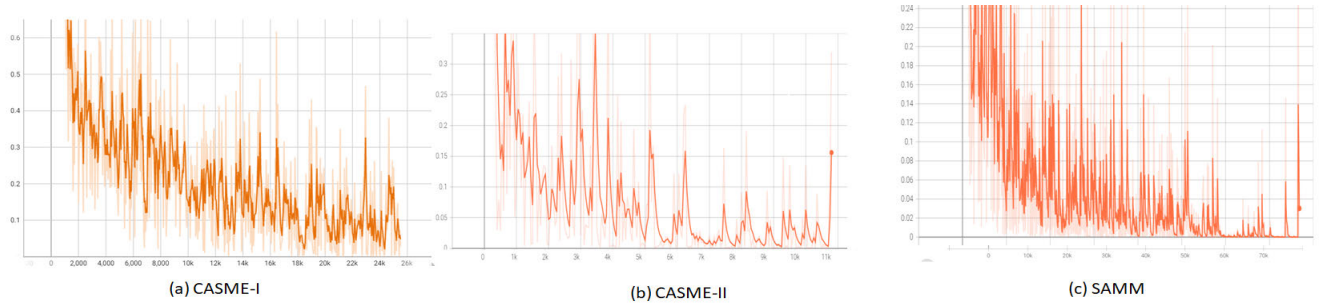


FIGURE 6. Training loss curve for proposed convolution patch based vision transformer.

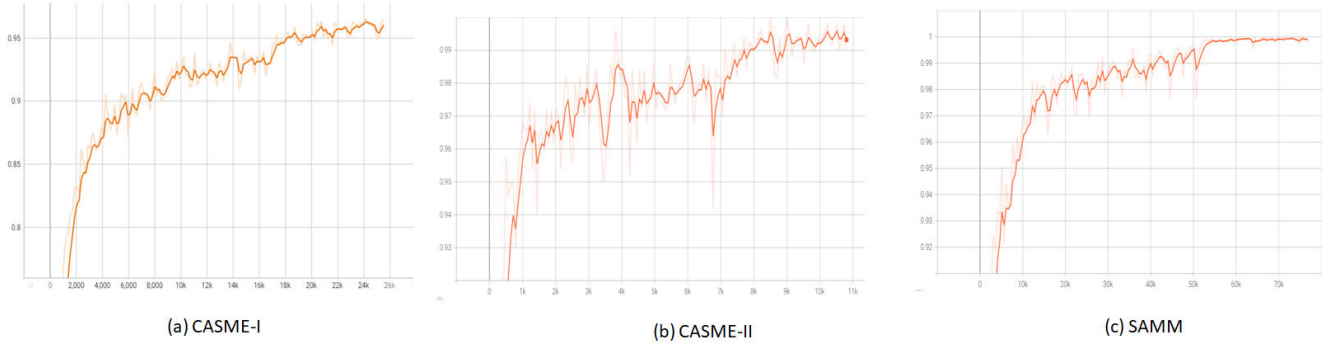


FIGURE 7. Validation accuracy curve for proposed convolution patch based vision transformer.

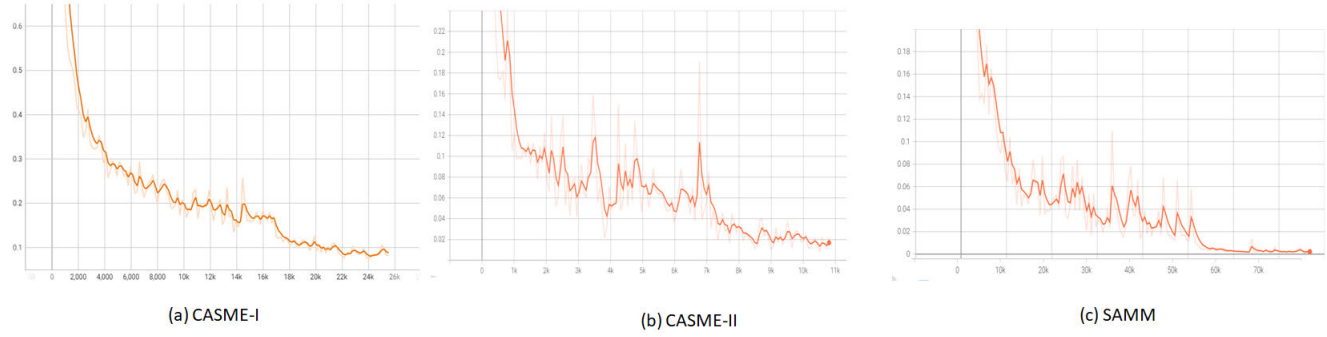


FIGURE 8. Validation loss curve for proposed convolution patch based vision transformer.

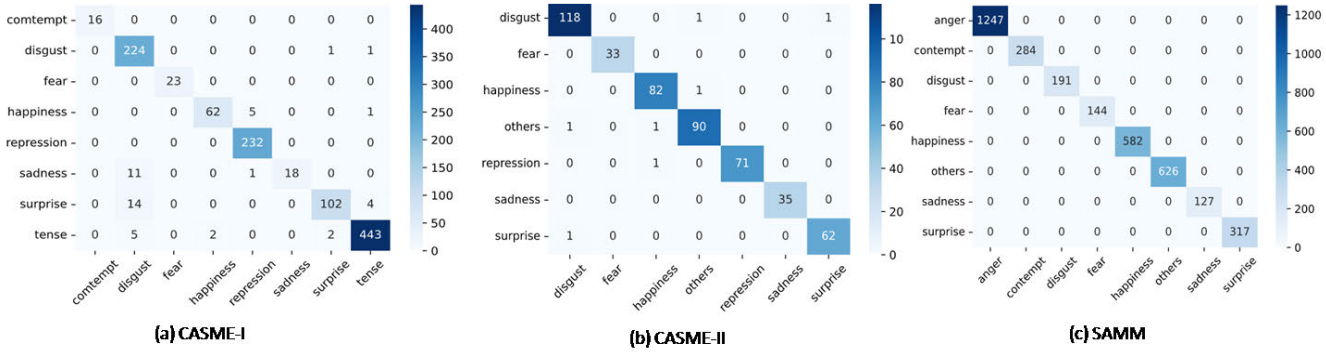


FIGURE 9. Confusion Matrices obtained using proposed vision transformer for CASME-I, CASME-II and SAMM datasets respectively.

the model's performance. Overfitting can be measured by observing a widening gap between the obtained training and validation curve. However, in our case, the training and validation curves exhibit a consistent alignment without a noticeable gap between them. Hence, it can be inferred that the model does not suffer from overfitting.

TABLE 4. Parameter values for Transformer encoder.

Parameters	Value
Embedding Dimension	256
Hidden Dimension	512
Heads	8
Transformer Encoder Layers	6
Patch Size	16
Input Channels	3
Number of Patches	257
Dropout	0.2

TABLE 5. Comparison of different models on the basis of number of parameters and GFLOPS.

Method	Image Size	Number of Parameters	GFLOPS
Swin-B [46]	384 × 384	88M	47.0G
CvT-21 [47]	384 × 384	32M	24.9G
DeiT-B [48]	384 × 384	86M	55.4G
Eff-B6 [49]	528× 528	43M	19.0G
Proposed Model	**640 × 480**	**13M**	**17.2G**

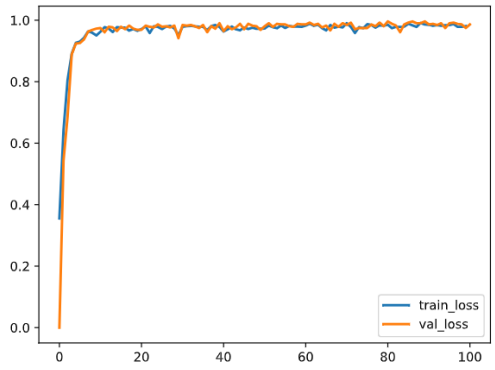


FIGURE 10. Training and Validation curve on SAMM dataset.

TABLE 6. Training and Validation accuracy for CASME-I, CASME-II, and SAMM datasets.

Datasets	Training Accuracy (%)	Validation Accuracy (%)
CASME-I	96.95	96.00
CASME-II	98.87	99.00
SAMM	99.10	100

The obtained evaluation metrics are shown in Table 7, 8, and 9 for CASME-I, CASME-II and SAMM datasets, respectively. Because of the severe class imbalances in CASME I and SAMM datasets, the F1-Score is more reliable while comparing performance of the proposed model. Figures 6 - 9 depict training loss curve, validation accuracy curve, validation loss curves and confusion matrix, respectively, where, number of iterations during training or validation are represented by x-axis, whereas, y-axis represents loss in Figures 6, 8 and accuracy in Figure 7. Validation accuracy and validation loss curves of CASME-II datasets in Figures 7 (ii) and 8 (ii) depict higher fluctuations as compared to other datasets. This might be due to lower number of training samples in CASME-II dataset. Figures 7 (iii) and 8 (iii) show less fluctuations for SAMM dataset as compared to CASME-II. However, despite of large number of samples, fluctuations in SAMM are higher than CASME-I dataset which is due to unbalanced training samples in SAMM dataset.

Table 7 shows evaluation metrics for CASME-I dataset. Based on evaluation of F1-Score, it can be inferred that the proposed model correctly classifies contempt and fear emotions, which contain least number of training samples i.e., 52 and 63 respectively as compared to other emotions (shown in Table 2). Thus, it can be concluded that the proposed model addresses the issue of smaller training samples required by state-of-the-art deep learning models. However, sadness emotion also contain fewer number of training samples i.e., 79, but the model could correctly classify only 75% samples. Figure 9 (a) shows confusion matrix obtained for the proposed model on CASME-I dataset. It can be observed that 11 samples of sadness emotion are wrongly classified as disgust. This is due to low inter-class variation among these two classes. Emotions such as disgust (802), happiness (234), repression (777), surprise (393) and tense (1495) generate F1-scores: 93%, 94%, 99%, 91%, and 98% respectively. The lower recognition rate might be because of overfitting of the model for emotions with higher number of training samples. The overall classification accuracy of the proposed model on CASME-I dataset is 95.97%.

Table 8 shows evaluation metrics for CASME-II dataset. It can be observed, that the proposed model generates 98.59% classification accuracy for CASME-II dataset. The model correctly classifies fear (121), and sadness (108) emotion. F1-score for repression (251), disgust (373), happiness (266), other (298), and surprise (241) are 99%, 98%, 98%, 98%, and 98% respectively. It can be observed that as the number of samples increases, the performance of the model drops for specific emotions. The reason behind this might be overfitting of the model.

Table 9 shows evaluation metrics for SAMM dataset. It can be observed, that the proposed model generates highest possible accuracy i.e., 100% for SAMM dataset. It is because of the availability of large number of training samples. Moreover, Table 2 shows that SAMM dataset is highly unbalanced, still the proposed model outperforms existing state-of-the-art models. Thus, it can be inferred that our model can easily handle unbalanced nature of the training datasets.

2) COMPARATIVE ANALYSIS

We contrast our proposed vision transformer model based on convolution patches with a number of state-of-the-art methods. We have compared the proposed transformer model with various machine and deep learning algorithms such as principal component analysis (PCA), CNN, CNN-LSTM, graph-CNN, and transformer models. From Tables 10-12, it can be observed that the proposed model outperforms several advance deep learning models and generates 95.97%, 98.59%, and 100% classification accuracy for CASME-I, CASME-II, and SAMM datasets respectively.

TABLE 7. Classification report over CASME-I dataset for 8 Classes.

Emotions	Precision	Recall	F1-Score
Contempt	1.00	1.00	1.00
Disgust	0.88	0.99	0.93
Fear	1.00	1.00	1.00
Happiness	0.97	0.91	0.94
Repression	0.97	1.00	0.99
Sadness	1.00	0.60	0.75
Surprise	0.97	0.85	0.91
Tense	0.99	0.98	0.98
Accuracy			0.96

TABLE 8. Classification report over CASME-II dataset for 7 Classes.

Emotions	Precision	Recall	F1-Score
Disgust	0.98	0.98	0.98
Fear	1.00	1.00	1.00
Happiness	0.98	0.99	0.98
Other	0.98	0.98	0.98
Repression	1.00	0.99	0.99
Sadness	1.00	1.00	1.00
Surprise	0.98	0.98	0.98
Accuracy			0.99

TABLE 9. Classification report over SAMM dataset for 8 Classes.

Emotions	Precision	Recall	F1-Score
Anger	1.00	1.00	1.00
Contempt	1.00	1.00	1.00
Disgust	1.00	1.00	1.00
Fear	1.00	1.00	1.00
Happiness	1.00	1.00	1.00
Others	1.00	1.00	1.00
Sadness	1.00	1.00	1.00
Surprise	1.00	1.00	1.00
Accuracy			1.00

TABLE 10. Comparison of the proposed method with existing models for CASME-I dataset in terms of classification accuracy.

Year	Method	Classification Accuracy (%)
2015	MDMO-SVM [52]	68.86
2015	LBP-TOP-ELM [53]	73.82
2017	CNN [54]	74.25
2018	Fusion motion boundary histograms [55]	61.33
2019	3D optical flow-based CNN [14]	54.44
2019	ResNet [56]	76.39
2019	Lateral Accretive Hybrid Network [56]	80.62
2022	Transfer learning with self-attention [57]	90.34
2023	Dual-stream incorporating optical flow and CNN [58]	61.20
2023	Two-stream 3D deep learning with iris biometric [59]	99.99
2023	Deep3DCANN [31]	87.00
	Proposed Transformer	**95.97**

A machine learning method proposed by [50], addresses two important characteristics of ME: low facial movement intensity and short duration of ME. The first issue is dealt by exploiting robust PCA and the sparse nature of ME in temporal domain is addressed by using local spatio-temporal directional features. This method generates 63.41% classification accuracy on CASME-II dataset. However, deep learning models such as CNN and LSTM generate

TABLE 11. Comparison of the proposed method with existing models for CASME-II dataset in terms of classification accuracy.

Year	Method	Classification Accuracy (%)
2014	LSTD [50]	63.41
2016	CNN-LSTM [29]	47.30
2019	3D optical flow-based CNN [14]	59.11
2019	STRCN [51]	80.30
2019	Lateral Accretive Hybrid [56]	76.57
2020	LFM [60]	73.98
2021	GEME [61]	75.20
2021	Transformer [37]	70.68
2022	MMNet [38]	88.35
2023	Deep3DCANN [31]	86.00
	Proposed Transformer	**98.59**

TABLE 12. Comparison of the proposed method with existing models for SAMM dataset in terms of classification accuracy.

Year	Method	Classification Accuracy (%)
2019	Dual-Stream Shallow Network [62]	63.41
2020	Graph-TCN [63]	75.00
2020	Knowledge distillation [64]	86.74
2021	MERSiamC3D [13]	68.75
2021	AU-GCN [65]	74.26
2022	SqueezeNet and 3D CNN [15]	81.33
2022	MMNet [38]	80.14
2023	Deep3DCANN [31]	93.00
	Proposed Transformer	**100.00**

remarkable performance as compared to machine learning models. Thus, to show a fair comparison we have compared our proposed model with state-of-the-art CNN models also. A 3D flow CNN proposed by [14], exploits a 3D convolution operation to extract spatio-temporal feature information along with optical flow. In this method, overfitting is avoided by using dropout mechanism and batch normalization technique. This method generates 59.11% classification accuracy on CASME-II dataset. To identify and analyse spatio-temporal deformations of ME, a recurrent CNN was proposed by [51] which generates 80.30% and 78.60% classification accuracy on CASME-II and SAMM datasets, respectively. Another category of recurrent neural network, known as long short term memory in conjunction with CNN was proposed by [29], generates 47.30% classification accuracy on CASME-II.

A vision transformer based model, muscle motion-guided network (MMNet), proposed by [38], exploits a two-branch network. The main branch of MMNet extracts motion-pattern related features through a continuous attention block, whereas a transformer encoder is exploited as a sub-branch of the model to generate positional embedding. Thereafter, the positionl embedding is added to motion-pattern features to generate 88.35% and 80.14% classification accuracy for CASME-II and SAMM datasets, respectively. Another vision transformer model based on optical flow and late fusion, proposed by [37], generates classification accuracy of 70.68% on CASME-II dataset.

V. CONCLUSION AND FUTURE WORK

When an existing vision transformer is exploited for micro-expression recognition, it divides the input image into small patches and a sequence of patch embedding is created by linearly embedding each patch. Due to this approach, the model may not exploit the local spatial relationships present in an image. To address this issue, in this work, a novel vision transformer based on convolution patches for micro-expression is proposed, which captures local receptive field through patches generated by convolution operation, and global receptive field is captured through a vision transformer based on self-attention mechanism.

While implementing the proposed network architecture, the following problems were handled: (i) Due to a large number of trainable parameters, self-attention-based operations, and long training time, high-performance computational resources are needed for the training of a vision transformer, thus, Nvidia A100 is utilized for training of the model which was provided by Google Colab Pro+, (ii) existing deep learning models are prone to over-fitting, thus we have employed layer normalization and dropout mechanism to avoid over-fitting which is usually caused by limited training data. The performance of the model is evaluated in terms of standard evaluation metrics such as precision, recall, F1-score, and classification accuracy. It has been demonstrated that the proposed model outperforms several state-of-the-art machine and deep learning models on three benchmark datasets i.e., CASME-I, CASME-II, and SAMM.

However, experiments show that the performance is still limited due to the following factors: (i) the existing micro-expression datasets are highly unbalanced in nature. It is evident from Table 2, CASME-II dataset is fairly balanced when compared to CASME-I dataset, thus, CASME-II generates better classification accuracy of 98.59% as compared to CASME-I i.e., 95.97%. Hence, it can be inferred that sample distribution plays a significant role in the performance of the model. It is to be noted that SAMM dataset is also not balanced (as shown in Table 2), but it contains large number of image samples for training, as compared to CASME-I and CASME-II, leading to the best possible classification accuracy i.e., 100%. Therefore, it is implied that large number of training samples can improve the performance of the model and help the model to overlook the unbalanced nature of a dataset. Thus, in future, we will address this issue by using data augmentation technique to generate a large number of samples for CASME-I and CASME-II datasets. Most of the existing MER datasets are laboratory controlled which limits the implementation of MER in real-life applications, thus, there is a need of in-the-wild datasets which contain a wide variety of images of individuals belonging to different age groups, gender, races, and cultural background. The existing deep learning models can only perform emotion classification based on pre-defined classes, to address this issue, deep continual learning can be explored which can identify an unknown emotion category [66].

REFERENCES

[1] J. Weiss and P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 2011.

[2] S. Zhao, H. Tang, S. Liu, Y. Zhang, H. Wang, T. Xu, E. Chen, and C. Guan, "ME-PLAN: A deep prototypical learning with local attention network for dynamic micro-expression recognition," *Neural Netw.*, vol. 153, pp. 427–443, Sep. 2022.

[3] P. Ekman and W. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[4] S. Nigam, R. Singh, and A. K. Misra, "Local binary patterns based facial expression recognition for efficient smart applications," in *Security in Smart Cities: Models, Applications, and Challenges*. Cham, Switzerland: Springer, 2019, pp. 297–322.

[5] S. Nigam, R. Singh, and A. K. Misra, "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28725–28747, Nov. 2018.

[6] Y. Guo, Y. Tian, X. Gao, and X. Zhang, "Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 3473–3479.

[7] X. Ben, P. Zhang, R. Yan, M. Yang, and G. Ge, "Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation," *Neural Comput. Appl.*, vol. 27, no. 8, pp. 2629–2646, Nov. 2016.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] S. Indolia, S. Nigam, and R. Singh, "Deep feature fusion for facial expression recognition," in *Proc. 2nd Int. Conf. Next Gener. Intell. Syst. (ICNGIS)*, Jul. 2022, pp. 1–6.

[10] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning," *Educ. Inf. Technol.*, vol. 28, no. 4, pp. 4069–4092, Apr. 2023.

[11] S. Indolia, S. Nigam, and R. Singh, "An optimized convolution neural network framework for facial expression recognition," in *Proc. 6th Int. Conf. Image Inf. Process. (ICIIP)*, vol. 6, Nov. 2021, pp. 93–98.

[12] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019.

[13] S. Zhao, H. Tao, Y. Zhang, T. Xu, K. Zhang, Z. Hao, and E. Chen, "A two-stage 3D CNN based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, Aug. 2021.

[14] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, Nov. 2019.

[15] S. Liu, Y. Ren, L. Li, X. Sun, Y. Song, and C.-C. Hung, "Micro-expression recognition based on squeezenet and C3D," *Multimedia Syst.*, vol. 28, no. 6, pp. 2227–2236, 2022.

[16] L. Cai, H. Li, W. Dong, and H. Fang, "Micro-expression recognition using 3D DenseNet fused squeeze-and-excitation networks," *Appl. Soft Comput.*, vol. 119, Apr. 2022, Art. no. 108594.

[17] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, and J. Yearwood, "LGAttNet: Automatic micro-expression detection using dual-stream local and global attentions," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106566.

[18] G.-B. Liong, J. See, and C.-S. Chan, "Spot-then-recognize: A micro-expression analysis network for seamless evaluation of long videos," *Signal Process., Image Commun.*, vol. 110, Jan. 2023, Art. no. 116875.

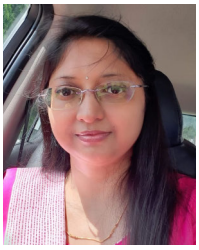
[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D.-H. Lee, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2013, pp. 117–124.

[21] 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 *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

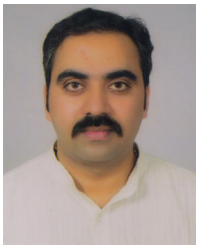
[22] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.
[23] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.
[24] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.
[25] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 51–56.
[26] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.
[27] J. Wei, G. Lu, J. Yan, and H. Liu, "Micro-expression recognition using local binary pattern from five intersecting planes," *Multimedia Tools Appl.*, vol. 81, pp. 20643–20668, Mar. 2022.
[28] A. M. Buhari, C.-P. Ooi, V. M. Baskaran, R. C. Phan, K. Wong, and W.-H. Tan, "Invisible emotion magnification algorithm (IEMA) for real-time micro-expression recognition with graph-based features," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 9151–9176, Mar. 2022.
[29] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2258–2263.
[30] V. Mayya, R. M. Pai, and M. M. M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 699–703.
[31] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep3DCANN: A deep 3DCNN-ANN framework for spontaneous micro-expression recognition," *Inf. Sci.*, vol. 630, pp. 341–355, Jun. 2023.
[32] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[33] N. V. Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.
[34] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[37] J. Hong, C. Lee, and H. Jung, "Late fusion-based video transformer for facial micro-expression recognition," *Appl. Sci.*, vol. 12, no. 3, p. 1169, Jan. 2022.
[38] H. Li, M. Sui, Z. Zhu, and F. Zhao, "MMNet: Muscle motion-guided network for micro-expression recognition," 2022, *arXiv:2201.05297*.
[39] J. Zhu, Y. Zong, H. Chang, Y. Xiao, and L. Zhao, "A sparse-based transformer network with associated spatiotemporal feature for micro-expression recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 2073–2077, 2022.
[40] L. Zhang, X. Hong, O. Arandjelovic, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," 2021, *arXiv:2112.05851*.
[41] X. Zhao, Y. Lv, and Z. Huang, "Multimodal fusion-based Swin transformer for facial recognition micro-expression recognition," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2022, pp. 780–785.
[42] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2978–2988.
[43] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[45] S. Indolia, S. Nigam, and R. Singh, "A framework for facial expression recognition using deep self-attention network," *J. Ambient Intell. Humanized Comput.*, vol. 14, pp. 9543–9562, May 2023.
[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[47] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[48] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[49] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[50] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, and C.-G. Zhou, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 325–338.
[51] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 626–640, Mar. 2020.
[52] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, Oct. 2016.
[53] Y. Guo, C. Xue, Y. Wang, and M. Yu, "Micro-expression recognition based on CBP-TOP feature with ELM," *Optik*, vol. 126, no. 23, pp. 4446–4451, Dec. 2015.
[54] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 1–7.
[55] H. Lu, K. Kpalma, and J. Ronsin, "Motion descriptors for micro-expression recognition," *Signal Process., Image Commun.*, vol. 67, pp. 108–117, Sep. 2018.
[56] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet: Dynamic imaging network for micro expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 1618–1627, 2020.
[57] S. Indolia, S. Nigam, and R. Singh, "Integration of transfer learning and self-attention for spontaneous micro-expression recognition," in *Proc. 7th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Nov. 2022, pp. 325–330.
[58] J. Tang, L. Li, M. Tang, and J. Xie, "A novel micro-expression recognition algorithm using dual-stream combining optical flow and dynamic image convolutional neural networks," *Signal, Image Video Process.*, vol. 17, no. 3, pp. 769–776, 2022.
[59] V. Esmaeili and M. M. Feghhi, "Real-time authentication for electronic service applicants using a method based on two-stream 3D deep learning," *Soft Comput. J.*, 2023.
[60] 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.
[61] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "GEME: Dual-stream multi-task GEnder-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, Feb. 2021.
[62] H.-Q. Khor, J. See, S.-T. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 36–40.
[63] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-TCN with a graph structured representation for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2237–2245.
[64] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1037–1043, Apr. 2022.
[65] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2871–2880.
[66] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep continual learning for emerging emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4367–4380, 2022.

SAKSHI INDOLIA received the M.Tech. degree in computer science and engineering from Banasthali Vidyapith, Rajasthan, India, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She has published and reviewed a number of research articles and successfully organized a few short-term programs on cognitive computing and artificial intelligence. Her research interests include computer vision, deep learning, emotion recognition, facial expression recognition, and machine learning. She was awarded the Gold Medal for the master's degree.

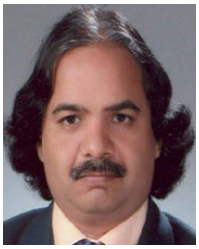


SWATI NIGAM (Senior Member, IEEE) received the Ph.D. degree in computer science from the Department of Electronics and Communication, University of Allahabad, India, in 2015. She has been a Postdoctoral Fellow under the National Postdoctoral Fellowship scheme of the Science and Engineering Research Board, Department of Science and Technology, Government of India. She is currently an Assistant Professor with the Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. She has authored more than 20 articles in peer-reviewed journals, book chapters, and conference proceedings. Her research interests include object detection, object tracking, and human activity recognition. She is a Professional Member of ACM. She was awarded a Senior Research Fellowship by the Council of Scientific and Industrial Research, Government of India. She has been the publication chair, the publicity chair, a TPC member, and a reviewer of various conferences. She is a Designated Reviewer of several SCI journals, such as IEEE Access, *Computer Vision and Image Understanding*, and *Journal of Electronic Imaging*.

RAJIV SINGH (Senior Member, IEEE) received the Ph.D. degree in computer science from the Department of Electronics and Communication, University of Allahabad, India. He is currently an Associate Professor with the Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. He has published more than 50 papers in refereed conferences and journals. His research interests include information fusion, computational cognitive science, medical image processing, computer vision, context-aware computing, and information security. He is a Senior Member of ACM and a Life Member of the Computer Society of India (CSI). He has served as a Reviewer for reputed journals, such as *Information Fusion*, IEEE Transactions on Biomedical Engineering, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, IEEE Transactions on Industrial Electronics, IEEE Transactions on Computational Imaging, and *The Visual Computer*, and many conferences.



VIVEK KUMAR SINGH (Senior Member, IEEE) received the bachelor's, master's, and Ph.D. degrees in computer science from the University of Allahabad, Allahabad, India. He is currently a Professor with the Department of Computer Science, Banaras Hindu University, Varanasi, India. His research is funded by the Department of Science and Technology (DST), Government of India; the Science and Engineering Research Board (SERB), India; and the Ministry of Electronics and Information Technology (MeitY), Government of India. His research interests include information systems, scientometrics, text analytics, and artificial intelligence. He is a member of ACM, ISSI, IETE, and CSI.



MANOJ KUMAR SINGH received the B.Sc. degree in physics and mathematics and the M.Sc. degree in computer science from the University of Allahabad, Allahabad, India, in 1994 and 1997, respectively, and the Ph.D. degree from the School of Information and Mechatronics, Gwangju Institute of Science and Technology (GIST), South Korea, in 2008. His Ph.D. dissertation concerned passive millimeter imaging systems and algorithms for forming high-quality images. From 2009 to 2010, he was a Postdoctoral Fellow with the Microwave Sensor System Laboratory, GIST. He is currently an Associate Professor with the Department of Computer Science, Banaras Hindu University, Varanasi, India. His research interests include image processing, compressive sensing, inverse problems in imaging, imaging systems, linear and nonlinear filtering, parameter estimation and application to signal processing, time-frequency analysis and application signal processing, and wavelet analysis. He is a member of the IEEE Signal Processing Society.