Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777609

FACIAL EXPRESSION RECOGNITION IN VIDEO USING 3D-CNN DEEP FEATURES DISCRIMINATION

Rajesh Singh
Department of Electronic Science, Kurukshetra University, Kurukshetra, India
RKSD College, Kaithal, Haryana, India
E-mail: rsdeshwal@gmail.com

Naval Kishore Mehta
CSIR-Central Electronics Engineering Research Institute, Pilani, India E-mail: naval.mehta95@gmail.com

Shyam Sunder Prasad
CSIR-Central Electronics Engineering Research Institute, Pilani, India E-mail: shyam.ece56@gmail.com

Anil Vohra
Department of Electronic Science, Kurukshetra University, Kurukshetra, India E-mail: vohra64@gmail.com

Sanjay Singh
CSIR-Central Electronics Engineering Research Institute, Pilani, India E-mail: sanjay@ceeri.res.in

Abstract: The focus of research work presented in this paper on improving performance in video facial expression recognition using a computationally efficient 3D convolutional neural network. An end-to-end 3D CNN method is proposed, which employs R(2+1)D Resnet18 as the backbone encoder block, followed by a spatio-temporal attention block divided into two heads for the evaluation of sparse center-loss and cross-entropy loss. The inclusion of a spatio-temporal block in the network has adaptively refined features in both the spatial and temporal domains. The network's combination of sparse center-loss and cross-entropy loss has extracted significant feature elements for enhanced discrimination in the embedding space.. The proposed architecture is evaluated on three in-the-wild publicly available datasets, i.e., DFEW, AFEW and DAiSEE. On the DFEW, AFEW and DAiSEE datasets, our method required only 38.66 Giga MACs for one batch prediction and achieved
accuracy of 59.28%, 59.63%, and 58.62%, respectively. The results show that, when compared to earlier works in the DFEW, AFEW and DAiSEE datasets, our method performed comparably well with significantly low computational load.

Index Terms: Facial Emotion Recognition, Convolutional Neural Network, Engagement prediction, Spatio-temporal network, Attention network.

1. Introduction

Facial expressions are one of humans' most natural, effective signals to express their emotional state and intentions. Automatic Facial expression recognition (FER) has a wide range of applications in human-computer interaction [1,2], emotional artificial intelligence (EAI) [2], and affective computing [1,2], and negative emotion detection[3]. Vision-based FER systems have become a popular choice among researchers in the past few decades due to recent advances in computer vision and deep learning algorithms. According to the feature representations, FER systems can be divided into two major categories: 2D static image-based methods [4,5] and 3D dynamic sequence video methods [6,7]. The feature representation encoded with only spatial data from the current single image is applied via static image-based approaches. In the static frames, the FER methods can be both hand-crafted based methods like 2D Gabor filters, Local binary patterns (LBP), and Support Vector Machine (SVM) with Boosted-LBP, as well as learned-based methods such as deep convolutional neural networks (CNNs). Video-based methods, contrasted with, take into account both the spatial and temporal relationships among the contiguous frames in the video stream. The widely used Spatio-temporal methods FER are 3D convolution neural networks (3D-CNN) and Long Short-Term Memory (LSTM). Although sequence-based approaches have improved FER, they still have drawbacks in that they often require significant computational complexity to describe facial expression movements in the video. Furthermore, most of the work does not reflect the degree of the person's emotional state. DFEW (Dynamic Facial Expression in the Wild) [6], AFEW (Acted Facial Expressions In The Wild)[16] dataset, which includes seven basic emotions in the real-world. On the DFEW dataset several approaches have reported the unweighted average recall (UAR) and weighted average recall (WAR) performance measures. We have also evaluated our methodology using the Dataset for Affective States in E-Environments (DAiSEE), a real-world student engagement level dataset [7]. DAiSEE dataset consists of the student's engagement levels recorded while watching the MOOC videos. Various deep learning techniques have been reported in the literature to target the facial expression recognition problem, but the majority of these techniques paid attention only to the 'accuracy' as the performance evaluation metric, ignoring the models' memory footprint and computational needs.

In this paper, we have addressed the aforementioned issues by presenting a novel hybrid R(2+1)D Resnet18 encoder-based architecture that captures both the spatio-temporal and deep discriminative feature elements for FER and achieves the right balance between computational cost and accuracy. By combining training with the sparse center loss and cross-entropy loss, the accuracy of the suggested model has increased compared to the solo cross-entropy loss. Even though the model was lightweight, we were still able to achieve competitive performance on two publicly available facial expression video datasets, DFEW (Dynamic Facial Expression in the Wild) [6], AFEW (Acted Facial Expressions in the Wild) and Dataset for Affective States in E-Environments (DAiSEE) [7].

The contributions of this paper can be summarized as follows:

- A novel R(2+1)D-based hybrid architecture is proposed for deep feature extraction and classification of facial expressions.
- We trained and evaluated our proposed model using four popular large-scale wild FER datasets (DFEW, AFEW, and DAiSEE) to illustrate the method's competitive performance.
- Our model requires only 47 GFLOPS computation for an inference (16 x 112 x 112 x 3 dimension image cube), making it suitable for real-time deployment in edge computing devices.

The rest of the paper is structured as follows. Related work is in Section 2. The specifics of our proposed architecture for facial expression recognition are explained in Section3. The datasets used in the experiments are summarized in Section 4. The findings are discussed in Section 5 and finally Section 6 gives the conclusion.

2. Related Work

Farzaneh et al. [8] proposed a method DACL to adjustably select a subset of notable feature elements for increased differentiation. This method combines attention mechanism to evaluate attention weights connected with feature importance with the use of halfway spatial feature maps in CNN as conditions. The approximate weights allow for the sparse formulation of centre loss to particularly attain intra-class compactness and inter class differentiation for the applicable information in the fixed space. In the similar way Fard et al. [9] suggested a fixing feature space which carry K feature vectors. Jiang X et al. [6] proposed a novel method Expression-Clustered Spatio-temporal Feature Learning (EC-STFL) that has superior performance because it reduces the inter-class distance and enhances the intra-class correlation. However, they have uncoupled the spatial and temporal information at different stages. The transformer proposed by Zhao et al. [8] also extracts spatial and temporal attention separately. Liu et al. [11] used local-global learning methods and a CNN to extract clip-level spatiotemporal data to build clip-aware dynamic facial expression. Recently, a spatio-temporal transformer was recently proposed by Ma F et al. [12], and the model's ability to distinguish between classes and to correlate within classes has been enhanced by the compact loss that was introduced. However despite the good performance in terms of the accuracy Liu et al.
[11] and Ma F et al. [12] are computationally expensive. Similarly there are several prior work that has focused research on facial expression recognition in traditional classroom and online learning platforms. Dukic et al. [16] built predictive networks that can assess facial expressions and emotions in the context of active teaching. Dataset for Affective States in E-Environments (DAiSEE) [7], using student recordings to evaluate students' engagement at four distinct intensity levels, from very low to very high, in a teaching classroom. Liao et al. [17] presented the Deep Facial Spatio-Temporal Network

(DFSTN) for measuring student engagement on DAiSEE. Zhang et al. [13] proposed a hybrid deep learning model for FER in video sequences. The suggested technique uses two distinct deep convolutional neural networks (CNNs), comprising a temporal CN network used to process optical flow images and a spatial CNN that processes static facial images, to learn high-level temporal and spatial features on the segmented video segments separately. To optimize

these two CNNs, target video facial expression datasets from a pre-trained CNN model are employed. Mehta et al. [14] proposed a three-dimensional DenseNet self-attention neural network (DenseAttNet) that may be used to track and assess student engagement in both modern and conventional educational programmes. Similarly, hierarchical deep network architecture to extract high-level spatial-temporal features was proposed by Cai et al [15].

However, despite having high computational demands, the method achieves low accuracy because the method is unable to capture distinguishing features in video clips due to harsh lighting, occlusions, and uneven pose changes. The primary challenge of a deep neural network on a video FER dataset is managing the model's demanding computational requirements while extracting the richer discriminatory features on the video.

3. Proposed Work

Performance of the facial expression recognition network suffers when there is partial occlusion of the face or uneven illumination because it is challenging for the network to detect the attention hotspots on the constantly changing areas. Attention-based networks can be used to overcome these problems [18]. In this section we have introduced an end-to-end pipeline for facial expression recognition in a dynamic scene as is illustrated in Fig. 1. We proposed an efficient hybrid R(2+1)D and spatio-temporal block that learns deep discriminative inter-class features while being supervised by sparse center loss [19] and cross-entropy loss. The proposed pipeline includes a pre-processing stage, a 3D-CNN feature extraction block, a 3D feature projection for sparse center loss calculation, and an emotion classification block.

3.1 3D-CNN Feature Extraction

Each video snippet is run through the DLib face detector. The extracted faces from each video clip are sampled at random into T = 16 frames, and the face frame is resized into H x W x C, where H,Height = 112, W,Width = 112, and C, Channel
= 3 of the frame. Finally, the input image cube of 16 x 112 x 112 x 3 is ready to be fed into the R(2+1)D with an 18 layers encoder model that pre-trained on the ImageNet dataset.

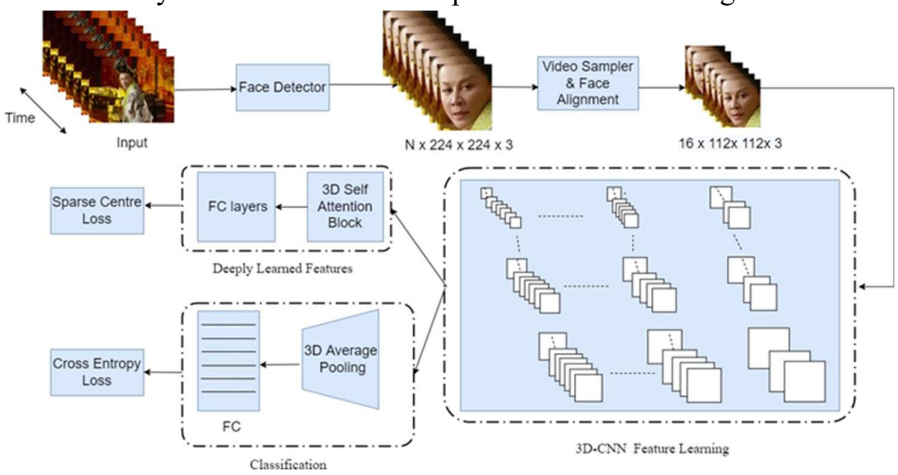


Fig. 1. The block-level representation of the proposed training pipeline for video-based FER

R(2+1)D achieves full 3D CNN by using the 2D ResNet blocks and 1D convolution that separate spatial and temporal components, making optimization much easier. Let f ∈ Rspatio-temporal output feature map of R(2+1)D convolutional layer. It is passed through the spatio-temporal attention block to link the local and global features in the spatial and temporal domain. The feature maps are then subdivided into Key (K), Query (Q), and Value (V) obtained by 1 x 1 x 1 convolution, as shown in Fig. 2. Equation 1 represents the spatial attention features across height and width, H' x W' dimension, and Equation 2 represents the temporal attention features across the temporal D, $Depth$ dimension. In the process of attention map generation, eight times compression of the features is employed for efficient computation and eventually $R^{D \times C \times H \times W}$ features are obtained across the spatial and temporal attention block.

$$s(f) = V_S\, Softmax\left(Q_S K_S^T\right) \quad (1)$$

$$d(f) = V_D\, Softmax\left(Q_D K_D^T\right) \quad (2)$$

$$p(f) = f + \gamma\,(s(f) + d(f)) \quad (3)$$

The resultant attention output given in Equation 3 is obtained by adding the spatial and attention features to the original input f by a scalar parameter γ.

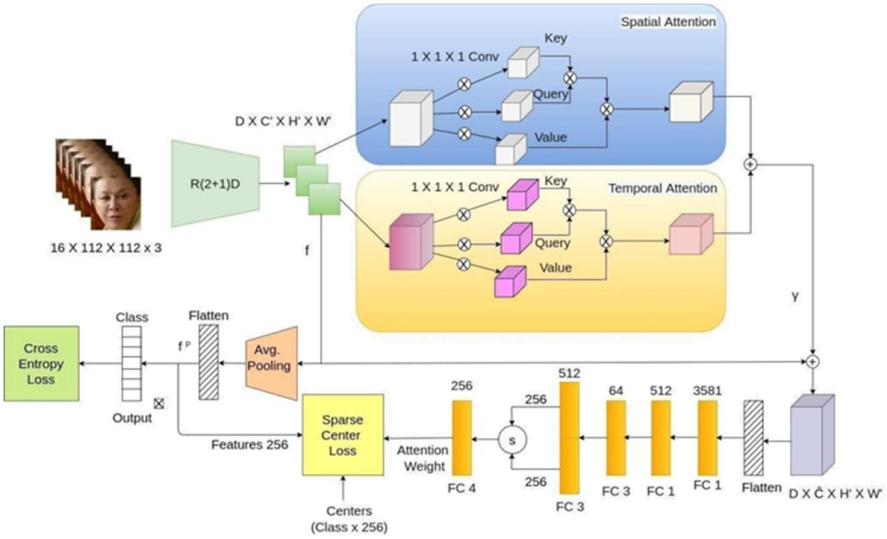


Fig. 2. The proposed video-level architecture for FER from video. Here ⊗ represents convolution operation, ⊕ represents feature addition, and 's' inside the circle represents Softmax operation.

3.2 Deeply Learned Features and Classification

The extracted features f from the R(2+1)D encoder are fed into two different layers, the 3D average pooling layer and flatten layer, which generates the feature maps fp. Pooled features fp are the deep features to the sparse center loss $l\,SCN$given in Equation 4. The 256 attention weights w for the sparse center loss is generated by passing features f through FC1-FC3 and converting it to 2 x 256 FC, which is then fed into the softmax layer shown in Fig. 2.

$$l_{SCL} = \sum_{n} \sum_{d} w * \left\| f^{\nu} - c \right\| \quad (4)$$

Where n is the batch size, p is the d dimensional features, c is the corresponding class center for the kth class, and attention weight w: $0 < w \leq 1$, and * represent element-wise multiplication.

The deep encoder features p
cross-entropy loss is evaluated.

are passed through the Flatten layer to the final FC layer for the classification, and the

$$l_{CE} = \sum^{K} Y \, log \, \hat{Y} + (1 - Y) \; log \, (1 - \hat{Y}) \quad (5)$$

Total loss of the FER detection is given as

$$l_{Total} = l_{SCL} + l_{CE} \quad (6)$$

In this section, we have discussed the datasets, followed by the experimentation to evaluate the proposed method on both datasets. The weighted and unweighted accuracy is used as an evaluation metric. A 5-fold cross-validation approach is used in this evaluation.

4. Dataset

This section gives an overview of the three publicly available datasets, i.e., Dynamic Facial Expression in the Wild (DFEW) [6], Acted Facial Expressions in the Wild (AFEW) [16] and Dataset for Affective States in E-Environments (DAiSEE) [7].

4.1 DFEW

DFEW is a collection of over 16,372 video snippets extracted from thousands of videos. These video snippets represent challenging interferences in real-world scenarios, such as harsh lighting, occlusions, and erratic position changes. The dataset is labelled with seven emotions: Happiness, Sadness, Disgust, Neutral, Surprise, Anger, and Fear.

4.2 DAISEE

The DAiSEE dataset contains 9062 10-second video snippets taken at 1920 × 1080 pixels at 30 frames per second (fps). The dataset was collected during an online learning scenario in which students were shown educational videos in various constraint environments. Each video is labelled with one of four affective states: engagement, boredom, confusion, or frustration. The affective state is classified into four levels: very low, low, high, and very high. We solely used the engaged affective state in our experiment.

4.3 AFEW

A dynamic temporal facial expression dataset is available in the Acted Facial Expressions in the Wild (AFEW) database that contains close to real-world emotions extracted from Movies and TV series [34,35]. The dataset contains 773 training clips, 383 validation clips, and 653 test clips. Since labels are not available for the test set, we perform our experiments only on the train and validation part. AFEW videos have no onset or offset of the emotion expressed. The dataset is labelled with seven emotions: Happiness, Sadness, Disgust, Neutral, Surprise, Anger, and Fear.

5. Experiments and Results

5.1 Experimental Setup

A 16 x 112 x 112 x 3 image cube with a batch size of 32 is used as an input to train the model. We used the Adam optimizer during training with an initial learning rate (LR) of 0.001. 5-fold cross-validation is implemented to obtain the result. All of the splits of the dataset individually were mutually exclusive. Cross-entropy loss is useful for learning separable features; however, it is insufficient for FER classification for videos with similar expressions. Our method employs a mix of sparse center loss and cross-entropy to get additional discriminative features. The experiments utilize two evaluation metrics: weighted average recall (WAR) and unweighted average recall (UAR), which are the most commonly used evacuation metrics for unbalanced problems. Equations 7 and 8 represent WAR and UAR, respectively. Finally, the computational requirements of our proposed neural network are evaluated in terms of Multiplication, addition and multiply-accumulate (MAC) units and compared to the existing literature. The experiments were carried out on the PyTorch platform with an NVIDIA Tesla V100 GPU.

$$Accuracy = WAR = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

$$UAR = \frac{TP}{TP+FN} * 0.5 + \frac{TN}{TN+FP} * 0.5 \quad (8)$$

5.2 Performance Evaluation

The performance of our model on the DFEW's seven emotions is summarized in Table 1. Our model showed the best performance on Happy, with an accuracy of 81.27%, and worst on Disgust, which contains the fewest facial emotions and is challenging to train and distinguish from the other facial emotions. Overall our method achieved a UAR value of 48.35% and a WAR value of 59.28%. Fig. 3(a) shows the model performance on the DFEW's one of the 5-fold splits. The model's confusion occurred highest in Disgust and Neutral emotions at around 48.3%. It was most likely due to a lack of training samples for the Disgust emotion or challenging instances, including extreme head positions, image lighting, and brightness [20].

Table 1. Performance of our proposed method on DFEW dataset.

	Emotions							**Metrics (%)**	
	Happy	**Sad**	**Neutral**	**Angry**	**Surprise**	**Disgust**	**Fear**	**UAR**	**WAR**
	81.27	55.07	58.90	60.66	51.53	3.45	22.88	48.35	59.28
No of Clips	2488	2008	2709	2229	1498	146	981	Total Clips = 12059	

In comparison to previous work on C3D with CE, C3D with EC-STFL[6], VGG11+LSTM with EC-STFL methods, our model outperformed them by 5.7 %, 3.8 %, and 2.8 %, respectively. Since the algorithm learns the deep feature center for each class, the deep features will aggregate towards their respective centers to generate deep features that will help the model perform even better. In terms of computational cost, our method required approximately 288 Giga less MACs, 244 Giga less MACs, and 32 Giga more MACs than C3D,

VGG11+LSTM, and 3D ResNet-18, respectively. In general, for deploying it on resource-constrained platforms, a lower MACs (computational complexity measure of the model) requirement is preferable. More recently, in Liu et al. [11] method each clip is split into 7 sub-clips that contain overlapping sequences across temporal dimensions. So, for the training, 7 sets of 5 x 224 x 224 x 3 image cubes were used. However, despite their strategy improving the WAR by 6.07 %, computational costs escalated by 25.14 Gigi MACs. Therefore, compared to CEFLNet [11], our method is appropriate for real-time edge computing platforms.

Table 2. Comparison to previous methods on DFEW dataset. The best results are in bold. Here CL represents the Center loss, and EC is Expression-Clustered Spatio-temporal Feature Learning (EC-STFL). The C3D, and VGG11+LSTM MACs calculations are extracted from Liu et al. [11] work.

Method	**Input Setting**	**MACs(Giga)**	**WAR(%)**
C3D[6]	sequence-based	>326.41	53.53
C3D,EC [6]	sequence-based	>326.41	55.50
VGG11+LSTM,EC [6]	sequence-based	>282.26	56.25
CEFLNet[11]	clip-based	63.80	65.35
Ours	sequence-based	38.66	59.28

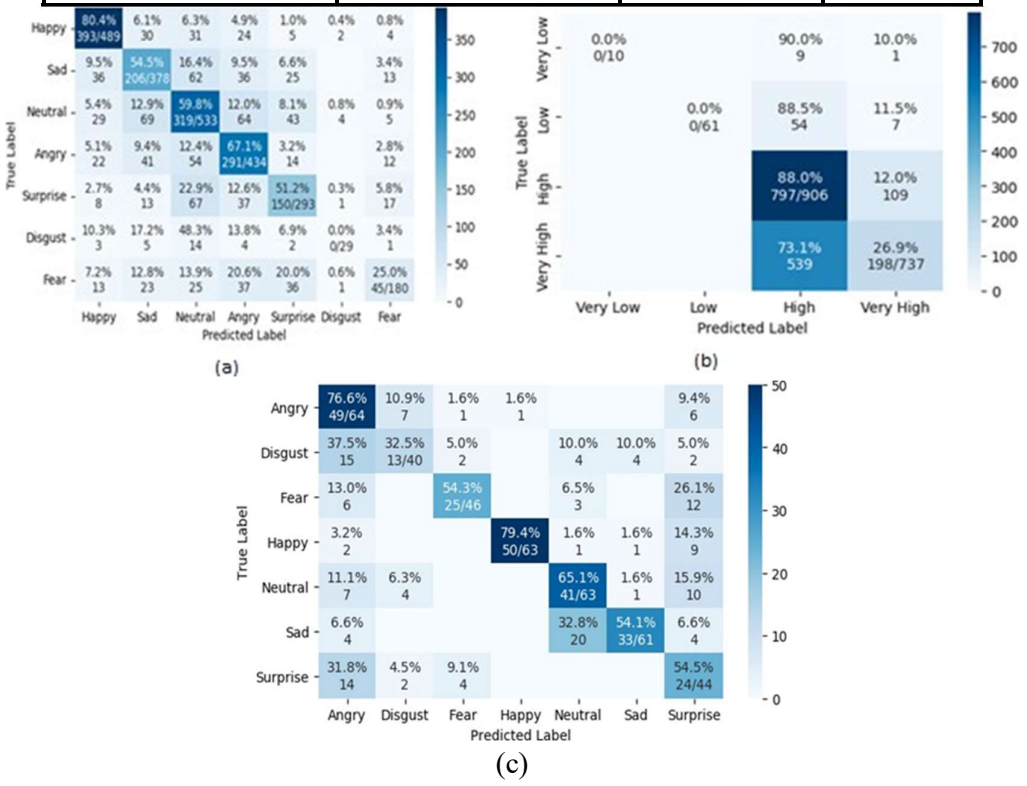


Fig. 3. Confusion matrices of the proposed model are being trained on (a) DFEW and (b) DAiSEE datasets (c) AFEW datasets. The result of one cross-validation experiment is shown.

Table 3. Performance of our model on the AFEW and comparison with previous approaches.

Method	Accuracy (%)
LBP-TOP[21]	38.90
VGG-Face + BLSTM[22]	53.91
ResNet-18[23]	55.17
MobileNet-v1 [24]	56.10
Zhou et al. [25]	65.50
Ours	59.63

The model trained and evaluated on AFEW 8.0 dataset achieves 59.63% given in Table 3. Whereas Zhou et al. achieved the state-of-the art accuracy of 65.5% but their model were quite complex make use of audio-video fusion strategies. Fig. 3(c) shows the model performance on the AFEW's one of the 5-fold splits. Table 4 presents our results on DAiSEE dataset and compares them with several other known methods. On the DAiSEE dataset Gupta et al. [7] explored deep learning models like InceptionNet, C3D, and Long-term Recurrent Convolutional Network (LRCN) to categorize engagement states and reported that the LRCN model performed best, scoring 57.9% for engagement affective states whereas our method achieved the competitive accuracy of 58.62%. The Deep Facial Spatio-Temporal Network (DFSTN) was introduced by Liao et al. [17] for measuring student engagement. Their engagement detection model builds an attentional hidden state using an LSTM with global attention. In order to extract spatial characteristics from faces, a pre-trained SE-ResNet-50 is used. They put their method to the test on the DAiSEE dataset and discovered that it was 58.84 % accurate. When compared to the computation costs of the LRCN [7] and DFSTN [13] methods, our model is found to be more computationally efficient. Fig. 4 depicts the overlaid prediction result from both datasets.

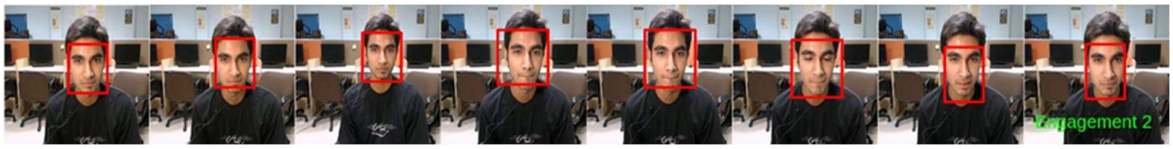
Table 4. Performance of engagement classification on the DAiSEE dataset and comparison with previous approaches.

Method	Accuracy (%)
I3D [7]	52.35
C3D (Fine Tuning) [7]	56.10
C3D (Focal Loss) [7]	56.20
LRCN [7]	57.90
DFSTN [17]	58.84
Ours	58.62

As a result of the DAiSEE dataset's extreme complexity and imbalance, the model's accuracy is below 60%. Fig. 3(b) displays the confusion matrix. Compared to the 'low' and 'very-low' engagement samples, the model performed exceptionally well in the 'high' and 'very-high'

engagement labels. The model's poor performance in 'low' and 'very-low' engagement can be attributed to three factors. First, because of the limited number of samples available for model training, and second, due to the data's imbalanced distribution ratio of 1:8:73:67 for four engagement labels, which allows for a minimal penalty on the misclassifying minority samples. Third, there are slight variations between videos in different levels of engagement [13]. The dataset reveals that there is little difference between any two levels of engagement that can be captured. As a result, a model that can distinguish micro expression changes from the video sequence is required.

(a) DFEW



(b) DAiSEE

Fig. 4. Eight consecutive sequential frames fed into the testing pipeline, with classifier results overlay in the last frame.

6. Conclusion

This study yields an effective 3D-CNN-based model for video-based FER. The proposed and implemented novel hybrid R(2+1)D attention block approach uses cross-entropy loss and sparse-center loss to introduce deep feature discrimination and improve final classification. The developed model is trained and tested on two publicly available in-the-wild datasets, namely DFEW and DAiSEE. It has achieved an accuracy of 59.28 % on DFEW, 59.63% on AFEW and 58.62 % on DAiSEE. Our model only requires 38.66 Giga MACs for the inference, and, therefore, is suitable for real-time facial expression classification applications. This study has empirically demonstrated that the proposed method has competitive performance on both the datasets with significantly low computational load. This approach may be extended in future to address the dataset's imbalances issue and to further improve the performance.

References

[1] Mohan K, Seal A, Krejcar O, Yazidi A. FER-net: facial expression recognition using deep neural net. Neural Computing and Applications. 2021 Aug;33(15):9125-36.

[2] Anil Audumbar Pise, Mejdal A. Alqahtani, Priti Verma, Purushothama K, Dimitrios A. Karras, Prathibha S, Awal Halifa, "Methods for Facial Expression Recognition with Applications in Challenging Situations", Computational Intelligence and Neuroscience, vol. 2022, Article ID 9261438, 17 pages, 2022.

[3] Negative Emotions Sensitive Humanoid Robot with Attention-Enhanced Facial Expression Recognition Network. Rongrong Ni, Xiaofeng Liu, Yizhou Chen, Xu Zhou, Huili Cai and Loo Chu Kiong. Intelligent Automation & Soft Computing.

[4] Bodavarapu PNR, Srinivas PVVS. (2021) Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques. Indian Journal of Science and Technology. 14(12): 971-983.
[5] Jeong, D.; Kim, B.-G.; Dong, S.-Y. Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition. Sensors 2020, 20, 1936.
[6] Jiang X, Zong Y, Zheng W, Tang C, Xia W, Lu C, Liu J. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. InProceedings of the 28th ACM International Conference on Multimedia 2020 Oct 12 (pp. 2881-2889).
[7] Gupta A, D'Cunha A, Awasthi K, Balasubramanian V. Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885. 2016 Sep 7.
[8] Farzana AH, Qi X. Facial expression recognition in the wild via deep attentive centre loss. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2021 (pp. 2402-2411)
[9] Fard AP, Mahoor MH. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access. 2022 Mar 3;10:26756-68.
[10] Zhao Z, Liu Q. Former-dfer: Dynamic facial expression recognition transformer. InProceedings of the 29th ACM International Conference on Multimedia 2021 Oct 17 (pp. 1553-1561).
[11] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, Zhongwen Luo, Clip-aware expressive feature learning for video-based facial expression recognition, Information Sciences,Volume 598,2022,Pages 182-195.
[12] Ma F, Sun B, Li S. Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild. arXiv preprint arXiv:2205.04749. 2022 May 10.
[13] Zhang S, Pan X, Cui Y, Zhao X, Liu L. Learning affective video features for facial expression recognition via hybrid deep learning. IEEE Access. 2019 Mar 4;7:32297-304.
[14] Mehta NK, Prasad SS, Saurav S, Saini R, Singh S. Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. Applied Intelligence. 2022 Mar 18:1-21.
[15] Youyi Cai, Wenming Zheng, Tong Zhang, Qiang Li, Zhen Cui, and Jiayin Ye. 2016. Video based emotion recognition using CNN and BRNN. In Chinese Conference on Pattern Recognition. Springer, 679–691.
[16] Duki´c, D.; Sovic Krzic, A. Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment. Electronics 2022, 11, 1240.
[17] Liao, J., Liang, Y., and Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence, 51(10):1–13.
[18] Zhu X, He Z, Zhao L, Dai Z, Yang Q. A Cascade Attention Based Facial Expression Recognition Network by Fusing Multi-Scale Spatio-Temporal Features. Sensors (Basel). 2022;22(4):1350.
[19] Farzana AH, Qi X. Facial expression recognition in the wild via deep attentive centre loss. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2021 (pp. 2402-2411).
[20] Fard AP, Mahoor MH. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access. 2022 Mar 3;10:26756-68.

[21] Dhall, A. (2019, October). Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In 2019 International Conference on Multimodal Interaction (pp. 546-550).

[22] Lu, C., Zheng, W., Li, C., Tang, C., Liu, S., Yan, S., & Zong, Y. (2018, October). Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (pp. 646-652).

[23] Kumar, V., Rao, S., & Yu, L. (2020, August). Noisy student training using body language dataset improves facial expression recognition. In European Conference on Computer Vision (pp. 756-773). Springer, Cham.

[24] Savchenko, A. V. (2021, September). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 119-124). IEEE. Available from: https://arxiv.org/pdf/2103.17107.pdf

[25] Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., & Qiao, Y. (2019, October). Exploring emotion features and fusion strategies for audio-video emotion recognition. In 2019 International conference on multimodal interaction (pp. 562-566).