Int. j. inf. tecnol. (April 2023) 15(4):1819–1830
https://doi.org/10.1007/s41870-023-01183-0

ORIGINAL RESEARCH

# Facial expression recognition in videos using hybrid CNN & ConvLSTM

Rajesh Singh[1] · Sumeet Saurav[2] · Tarun Kumar[3] · Ravi Saini[2] · Anil Vohra[1] · Sanjay Singh[2]

Received: 2 September 2022 / Accepted: 15 February 2023 / Published online: 21 March 2023

**Abstract** The three-dimensional convolutional neural network (3D-CNN) and long short-term memory (LSTM) have consistently outperformed many approaches in video-based facial expression recognition (VFER). The image is unrolled to a one-dimensional vector by the vanilla version of the fully-connected LSTM (FC-LSTM), which leads to the loss of crucial spatial information. Convolutional LSTM (ConvLSTM) overcomes this limitation by performing LSTM operations in convolutions without unrolling, thus retaining useful spatial information. Motivated by this, in this paper, we propose a neural network architecture that consists of a blend of 3D-CNN and ConvLSTM for VFER. The proposed hybrid architecture captures spatiotemporal information from the video sequences of emotions and attains competitive accuracy on three FER datasets open to the public, namely the SAVEE, CK+, and AFEW. The experimental results demonstrate excellent performance without external emotional data with the added advantage of having a simple model with fewer parameters. Moreover, unlike the state-of-the-art deep learning models, our designed FER pipeline improves execution speed by many factors while achieving competitive recognition accuracy. Hence, the proposed FER pipeline is an appropriate candidate for recognizing facial expressions on resource-limited embedded platforms for real-time applications.

**Keywords** Video-based facial expression recognition (VFER) · 3D convolutional neural networks (3D-CNN) · Convolutional LSTM (ConvLSTM) · Long short-term memory (LSTM)

## 1 Introduction

The development of applications based on human–computer interaction (HCI) has been very rapid in recent years. VFER remains the most central component of HCI, with various methods proposed to date. However, automatic recognition of facial expressions in a video sequence is very challenging due to several limitations associated with physical factors such as age, gender, ethnic background, and lack of contextual information, as well as variations due to changes in lighting conditions, poses, and occlusion. 3D-CNN have recently demonstrated better performance in FER tasks. These networks have the inherent ability to model spatial and temporal information simultaneously. Further, the 3D-CNN architectures can be combined in different ways with the recurrent neural networks (RNNs) to capture long-term dependencies and feature enhancement [1, 2].

Based on the performance of ConvLSTM in several computer vision applications, in this paper, we explore the

✉ Rajesh Singh
rsdeshwal@gmail.com

Sumeet Saurav
sumeet@ceeri.res.in

Tarun Kumar
kumar1998.tarun@gmail.com

Ravi Saini
ravi@ceeri.res.in

Anil Vohra
vohra64@gmail.com

Sanjay Singh
sanjay@ceeri.res.in

[1] Department of Electronic Science, Kurukshetra University, Kurukshetra, India

[2] CSIR-Central Electronics Engineering Research Institute, Pilani 333031, India

[3] Department of Computer Science, Birla-Institute of Technology and Science, Pilani 333031, India

combination of 3D-CNN and ConvLSTM for the classification of facial expressions in a video sequence. ConvLSTM as stated in [3] have shown its superiority over fully-connected LSTM (FC-LSTM) in number of spatiotemporal tasks like Moving MNIST dataset, Radar Echo dataset, etc., [4]. The major advantage of ConvLSTMs over the FC-LSTMs in computer vision applications is due to their mode of operations, where the input-to-state and state-to-state transitions are convolutional. Therefore, these networks are inherently more suited to images, as there is no loss of spatial correlation as compared to the FC-LSTM, which operates on vectorized features.

A typical FER algorithmic pipeline consists of three main stages, namely face detection, feature extraction, and feature classification. Usually, the face detection stage is accompanied by a facial alignment stage, whose objective is to ensure that all the images are spatially symmetrical. For face detection, most of the FER framework mainly use the open-source implementation of the Viola-Jones and recently introduced deep learning-based face detectors use techniques offered by the OpenCV library. Also, for feature extraction, the traditional techniques for FER employs feature descriptors like the local phase quantization (LPQ) [5], histogram of oriented gradients (HOG) [6], local binary patterns (LBP) [7], local neighborhood difference binary patterns (LNDBP) [8]. Since these feature extractors are hand engineered, they do not generalize well in different imaging conditions such as lighting, occlusion, subjects' ethnicity, etc. Also, for classification, the traditional FER methods based on machine learning use popular classifiers such as naive Bayes, the support vector machine (SVM), etc. In contrast, the recently developed deep learning algorithms for FER are fully trainable and have shown to achieve superior performance.

The following are the major contributions of this work: (i) Proposed a novel lightweight deep learning model using a blend of 3D-CNN and ConvLSTM for VFER, (ii) Assess the model's performance on several benchmark facial expression dataset, namely the CK+, SAVEE, and AFEW, and (iii) Compare the performance of the proposed neural network model with several related state-of-the-art methods.

The rest of the sections are arranged as follows: The previous works on video-based facial expression recognition are summarized in Sect. 2. Section 3 explains the proposed hybrid 3D-CNN and ConvLSTM neural network for FER. Experimental results and discussion are presented in Sect. 4, followed by conclusive remarks in Sect. 5.

## 2 Related works

Facial expression recognition (FER) techniques available in the literature using deep learning are broadly classified into two major categories: video-based FER methods and the static image-based FER methods. Here, we briefly review some of the existing schemes for FER based on these two categories.

The static image-based techniques for FER use an end-to-end trainable convolutional neural network (CNN) for the classification of facial expressions using peak expression images. For instance, the work presented by Yu et al. [9] has used a collection of deep neural networks for FER in a single static image. Yang et al. [10] introduced a weighted fusion of features obtained from CNNs trained on gray-scale and local binary pattern (LBP) encoded facial images for FER. The authors in [11] presented a multi-channel deep neural network that can recognize face expressions in static images by learning and fusing spatial-temporal features. Here optical flow image is extracted between the peak expression face image and neutral face image and used as temporal information. Moreover, to extract spatial features, the framework has only used the peak expression image.

Conversely, the video-based methods for FER extract spatial and temporal knowledge from an expression video sequence using hybrid CNN & LSTM and 3D-CNNs. For instance, the VFER technique proposed by Hasani et al. [12] has used an amalgamation of deep neural network (DNN) and conditional random field (CRF). Here DNN has been used to understand the spatial relationship among the expression images, whereas a linear change of CRF captures the temporal relations. In another work discussed in [13], the authors introduced a novel technique for FER in videos. Their proposed VFER technique captures spatiotemporal information using a combination of CNN and nested LSTM. Using a nested LSTM with two-layered architecture, the framework captures the dynamics of the facial expressions. The first layer, LSTM, named T-LSTM, has been employed to model the temporal dynamics of the spatiotemporal features, and the second layer, LSTM, called the C-LSTM, has been employed to combine the outputs of all T-LSTMs. The nested LSTM network aids in encoding the multi-level information found in the model's intermediate layers. Also, the VFER technique introduced by Zhao et al. [14] has used a 3D-CNN to extract the spatial and temporal information from the RGB and optical flow sequences, respectively. Vielzeuf et al. [15], on the other hand, proposed several schemes for VFER using a hybrid CNN & LSTM and 3D-CNN & LSTM. The hybrid networks, initially trained independently, are fused to achieve enhanced performance. Another hybrid network for VFER proposed by Fan et al. [1] has used 3D-CNN and a recurrent neural network (RNN). Liu et al. [16] proposed a technique for VFER using a 3D-CNN model with deformable parts learning.

The FER technique proposed by Zhao et al. [17] has utilized local binary pattern (LBP) features extracted from three orthogonal planes (LBP-TOP). Their presented LBP-TOP operator derives expression features from the video sequences in the form of computed histograms from three orthogonal planes. Once extracted, the LBP-TOP features corresponding to expression video sequences are classified using a machine learning classifier. In [18], the authors presented a technique for VFER using a fuzzy ARTMAP neural network (FAMNN). Also, to determine the hyperparameters of the FAMNN network, their proposed framework has used particle swarm optimization (PSO). In [19], the authors proposed a technique for VFER using a Cross-channel Convolutional Neural network (CC-CNN). Zhang et al. [20] proposed a technique for VFER that used a part-based hierarchical bidirectional recurrent neural network (PHRNN). Also, their proposed framework has taken still frames from the expression sequence and extracted spatial information using a multi-signal convolutional neural network (MSCNN). Finally, both PHRNN and MSCNN are fused to extract the dynamic still, partial-whole, and geometry-appearance information that boosts the performance of the VFER system. A manifold learning-based framework has been explored for VFER by Liu et al. [21]. Ouyang et al. [22] demonstrated the use of transfer learning and spatiotemporal fusion for VFER. The work presented in [23] has proposed a fusion of audio and visual expression-related information for FER, and that introduced by Ebrahimi et al. [24] has suggested utilizing a hybrid of CNN & RNN.

In their paper, Kumar et al. [25] introduced a face and gait-based multimodal biometrics system that uses principal component analysis (PCA) and a simple deep neural network (S-DNN). Cross entropy, rather than the more traditional Euclidean distance, used in the S-DNN analysis. Similarly, Jayanthi et al. [26] proposed a unified framework that considers static facial images and speech modulation to identify an individual's emotional state. Furthermore, Pandey et al. [27] accentuated an algorithmic strategy that uses only computer vision techniques to detect tiredness in drivers at an early stage and with outstanding accuracy. Uddin et al. [28] presented an inexpensive intelligent mirror system, which would show different information to the user depending on what they wanted to see recommendations. The smart mirror utilizes IoT and AI to provide its users with information such as to-do list, the current weather, traffic updates, date, calendar, time, updated news headlines, the status of COVID-19 cases and much more. In addition, the authors integrated face detection into the smart mirror to make the whole system foolproof. Sabharwal and Gupta [29] presented a method that combines an existing face recognition system with statistics obtained from facial marks to improve the recognition rate between pre- and post-surgery photos of humans. Liu et al. [30] introduced a hybrid CNN and RNN model for video-based FER by adding a Graph Convolutional Network (GCN) layer. The hybrid framework uses CNN to extract features and the GCN layer to learn more significant facial expression features that focus on specific regions. Lee et al. [31] offered a scene-aware hybrid neural network (NN) with a novel blend of RNN, CNN and 3D-CNN. In the beginning, the network extracts global features from the video and local features from each frame. The framework uses a 3D-CNN with an auxiliary classifier to extract latent features representing the exact visual scenario of a video, while latent features comprising minute information are extracted from each frame using a fine-tuned CNN.

Pan et al. [32] proposed a video descriptor for FER to gather spatial and temporal convolutional attributes throughout an entire video. The developed system includes a trainable spatial-temporal feature aggregation layer and incorporates cutting-edge 30-stream CNN. Fan et al. [1] provided a system for identifying emotions using video that was submitted to the EmotiW 2016 Challenge. This system's fundamental component is a hybrid network that employs a late-fusion strategy to combine RNN with 3D-CNN(C3D). C3D and RNN have unique ways of encoding visual and kinematic data. Hasani et al. [2] also introduced a 3D-CNN for FER in videos. The 3D Inception-ResNet layers of the planned 3D-CNN architecture are followed by an LSTM unit, which conbinely extracts the spatial connections between facial images and the temporal relationships between various frames in the movie. Mehta et al. [33] introduced a 3D-DenseNet self-attention neural network (DenseAttNet)

**Table 1** Summary of existing techniques for VFER in the literature (BE: Basic facial expressions)

| References | Year | Highlights | Dataset | 2D | Dynamic | E-xpressions | Accuracy % |
|---|---|---|---|---|---|---|---|
| Yu et al. [13] | 2018 | The proposed VFER technique captures Spatio- temporal information using a combination of CNN and nested LSTM. Their proposed method captures the spatiotemporal variations in the video sequences of emotions using a 3DCNN model | CK +<br>Oulu-CASIA MMI | N | Y | 6BE<br>6BE<br>6BE | 99.8(±0.2)<br>99.8(±0.2)<br>84.53(±0.67) |
| Zhao et al. [14] | 2018 | Used a 3DCNN architecture that learns to extract the static and information from the RGB sequences and optical flow sequences, respectively | CK + | N | Y | 7BE | 98.47 |
| Vielzeuf et al. [15] | 2017 | Proposed several schemes for VFER using a hybrid CNN & LSTM and 3DCNN & LSTM. The hybrid networks initially trained independently are fused to achieve enhanced performance | AFEW | N | Y | 7BE | 58.8 |
| Liu et al. [16] | 2014 | Proposed a technique for VFER using 3D CNN model with deformable parts learning | CK +<br>MMI<br>FERA | N | Y | 7BE<br>6BE<br>5BE | 92.4<br>63.4<br>56.1 |
| Barros et al. [19] | 2016 | Proposed a technique for VFER using a Cross-channel Convolutional Neural network (CC-CNN) | FABO | N |  | 10BE | 80.33 |
| Liu et al. [21] | 2014 | Explored manifold learning-based framework for VFER | CK +, Oulu-CASIA<br>MMI<br>AFEW | N | Y | 7BE<br>6BE<br>6BE<br>7BE | 94.19<br>74.59<br>75.12<br>31.73 |
| Ouyang et al. [22] | 2017 | Used transfer learning and spatiotemporal fusion to show how to use VFER | EmotiW2017 | N | Y | 7BE | 57.2 |
| Kumar et al. [25] | 2022 | Presented a multimodal biometrics system employing principal component analysis (PCA) and a simplified deep neural network (S-DNN) based on face and gait. Cross entropy rather than Euclidean distance is taken into account in the S-DNN analysis | CK +<br>Oulu-CASIA<br>MMI | N | Y | 6BE<br>6BE<br>6BE | 99.51 |
| Jayanthi et al. [26] | 2022 | Proposed a comprehensive framework that took into account both human static facial pictures and speech modulation to understand a person's mental state | Forkyknight | N | Y | 7BE | 94.26 |
| Liu et al. [30] | 2020 | Proposed a VFER model with Graph Convolutional Network (GCN) and the common CNN-RNN. The GCN layer is used to learn more significant facial expression features from specific facial regions after sharing information amongst retrieved CNN characteristics of nodes | CK +<br>Oulu-CASIA<br>MMI | N | Y | 7 BE<br>6BE<br>6BE | 99.54<br>91.04<br>85.89 |

**Table 1** (continued)

| References | Year | Highlights | Dataset | 2D | Dynamic | E-xpressions | Accuracy % |
|---|---|---|---|---|---|---|---|
| Lee et al. [31] | 2019 | A unique 3DCNN, CNN, and RNN-based scene-aware hybrid neural network (NN) has been proposed. The proposed system simultaneously extracts both frame-based local characteristics and global features based on the video. The latent features containing minor details from each video frame are extracted using a fine-tuned CNN & 3DCNN with an auxiliary classifier is u recovers the latent features of the entire scene of a given movie | AFEW<br>CK + | N | Y | 7BE<br>7BE | 49.9<br>98.2 |
| Pan et al. [32] | 2018 | Presented a video descriptor for FER that would allow it to combine spatial and temporal convolutional information over the whole length of a video. The created system features a trainable layer for aggregating spatial–temporal feature data and integrates a cutting-edge 30-stream CNN | BAUM-1 eNTERFACE05 | N | Y | 6 BE<br>6 BE | 46.51<br>43.72 |
| Fan et al. [1] | 2016 | Presented a system for recognizing emotions in videos and submitted for the EmotiW 2016 Challenge. This system's main element is a hybrid network that fuses RNN with 3DCNN (C3D) in a late-fusion fashion. RNN and C3D encode appearance and motion data in different ways | CK + | N | Y | 7BE | 59.08 |
| Hasani et al. [2] | 2017 | The proposed 3DCNN for videos is composed of 3D Inception-ResNet layers and an LSTM unit that work together to extract both the spatial and temporal correlations present in facial images and video sequences | AFEW | N | Y | 7BE | 62.34 |

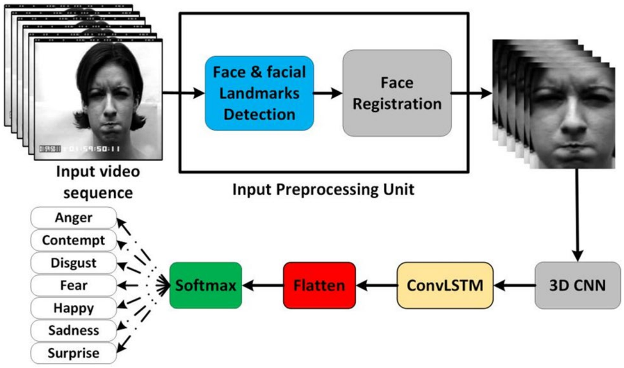


**Fig. 1** Proposed framework for video-based facial expression recognition

to evaluate student engagement in modern and conventional educational programs. The proposed DenseAttNet model outperformed all previous methods with the Dataset for the Affective States in E-Environments (DAiSEE). Table 1 highlights the existing techniques for VFER in the literature.

## 3 Proposed method

Figure 1 depicts the block diagram for our suggested algorithmic pipeline for VFER. The framework comprises a series of steps with dedicated functionality, as shown in Fig. 1. As given in Algorithm 1, the first step, namely the input pre-processing unit takes the video frames as input and returns the face and facial landmarks location. In the next step, based on the facial landmark information, the framework aligns the detected faces to attain spatial symmetry. In the subsequent step, the aligned facial images are scaled and cropped to a standard size.

Afterward, the hybrid 3D-CNN & ConvLSTM model takes the pre-processed facial images of fixed duration (# frames) as input and learns to classify them into one of the seven possible face expressions, namely disgust, anger, contempt, fear, happiness, surprise, and sadness. Below, we describe the various units used in our proposed VFER framework. In the subsections below, we provide the details of each stage of the proposed VFER pipeline.

### 3.1 Input pre-processing unit

As discussed above, the input processing unit takes the video frames as input and provides the registered facial image having a resolution of 64 × 64 pixels. We deployed a Single Shot Multibox Detector (SSD) based face detector for face detection, with the MobileNetV1 CNN as the feature extractor [34, 35]. We used the transfer learning for the training face detector on the target wider face detection benchmark database [36]. The trained face detector operates at 60 frames per second and has great confidence in finding faces. Once detected, the preprocessing step transfers the x and y-coordinates of the faces to the facial landmark detector, which uses them to identify 68 facial landmarks, as shown in Fig. 2. With the help of these landmarks, the unit aligns the facial images such that they have a similar spatial location of different facial components. The landmark detector used in this work employs an ensemble of regression trees to localize the face landmarks in static images [37]. In our suggested FER pipeline, we utilized the landmark detector's open-source implementation provided by King [38].

**Algorithm 1**: Algorithmic pipeline of the proposed VFER

```
begin
Input: $I = [I_1, I_2, ......, I_N]$ Videos and $C = [C_1, C_1, ....., C_N]$ Labels
for i:=1 to  I do
    Obtain i^th video, D = Video Depth
    for j:=1 to D do
        Obtain j^th face features of i^th video, $U_j$ = Face Detector ($I_i(j)$)
    end for
Obtain the spatio-temporal features, $X_i$ = Hybrid 3D-CNN & ConvLSTM(U)
Facial Expressions Probability , Yi = Softmax (Xi)
Loss Evaluation, L = Categorical Cross-Entropy Loss ($Y_i$, $C_i$)
Neural network optimization based on the obtained L using Adam optimizer
end for
end
```

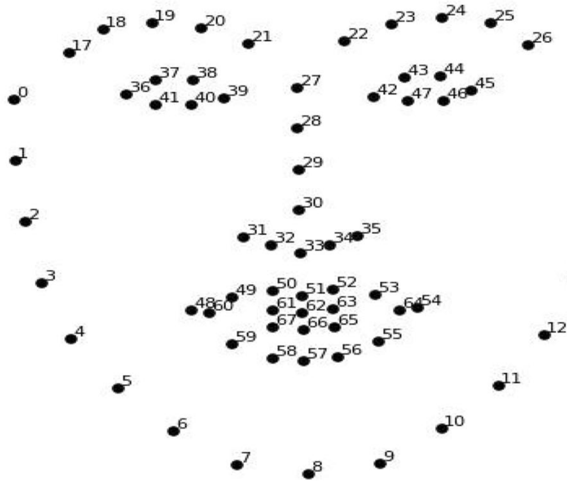


**Fig. 2** Prototypical facial image with 68 landmarks

### 3.2 Hybrid 3D-CNN and ConvLSTM

Since their advent, deep learning techniques, especially the convolutional neural networks (CNNs), have achieved tremendous success in several computer vision tasks related to image classification, object detection, human action

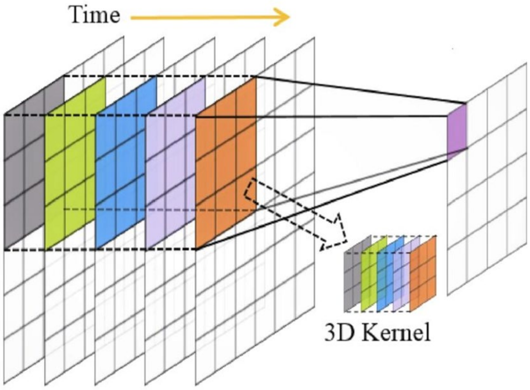


**Fig. 3** Illustration of 3D convolution operation

recognition, facial expression recognition, face recognition, etc. A typical CNN architecture consists of repeated convolution and pooling layers, along with a stack of fully-connected layers at the end. Convolutional layers at different stages of a CNN performs the role of local feature extractor, wherein the layers in the early stage extract primary details like edges, counters, and gradients. The convolutional layers at later stages of the networks extract high-level information necessary for the classification of the images. The max-pooling operations, on the other hand, reduce the resolution of the features. Finally, the FC layers are used for the non-linear transformation of the image features.

Depending on the input data dimensionality, the convolution operations and the CNNs are classified into 1D, 2D and 3D convolution operations and 1D-CNN, 2D-CNN and 3D-CNN models, respectively. Similarly, the max-pooling operations are also divided into 1D, 2D and 3D pooling operations. While 3D convolution is used to learn features from both the temporal and spatial dimensions, 1D and 2D convolution are both performed to compute information from the spatial dimensions. Since 3D convolution offers both spatial and motion-related information, it is appropriate for image sequence or video analysis issues. The 3D convolution output Z(x, y, t) for a given image cube I(x, y, t) can be obtained by convolving the image cube I(x, y, t) with the convolution kernel W(x, y, t) of size a×b×c as illustrated in Fig. 3 and expressed in Eq. (1).

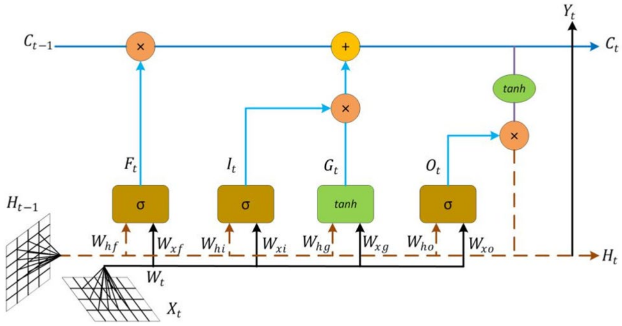


**Fig. 5** Internal elements of the ConvLSTM

$$W(x,y,t) = \sum\nolimits_{i=-a}^{a} \sum\nolimits_{j=-b}^{b} \sum\nolimits_{\tau=-c}^{c} I(i,j,\tau).W(x-i,y-j,t-\tau) \quad (1)$$

The architectural structure of the proposed hybrid 3D-CNN & ConvLSTM model is presented in Fig. 4. The proposed model consists of three 3D convolutional layers, one 3D max-pooling layer, one ConvLSTM block, one fully-connected layer, and a softmax classifier layer. To execute the non-linear transformation of features between the layers, the 3D convolution blocks are followed by a Rectified Linear Unit (ReLU) activation function. The output from the second max-pool layer is fed as input to the ConvLSTM layer which has 16 units. The subsequent fully connected layer takes as input the flatten output from the ConvLSTM layer. Lastly, the fully-connected layer is connected to the softmax layer, which categories the output from the facial sequences into one of the seven facial expression classes, namely anger, contempt, disgust, fear, happy, sadness, and surprise.

Although LSTM is excellent at handling temporal data, full connections in input-to-state and state-to-state transitions do not take spatial data into account, which results in the loss of spatial data, which is essential for modeling the spatial–temporal data involved in facial expressions. Convolutional LSTM (ConvLSTM), which replaces the state-to-state transition operations in the LSTM with convolution operations, was developed by Xingjian et al. [3] to address this problem of LSTM. The effectiveness of ConvLSTM in tasks involving human activity has also recently been advocated in several

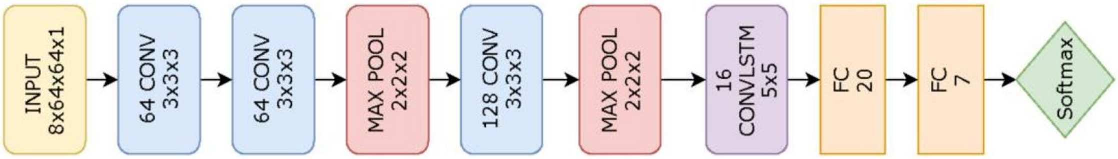


**Fig. 4** Schematic representation of the proposed hybrid 3D-CNN & ConvLSTM model

papers [39, 40]. The ConvLSTM layer's internal elements are illustrated in Fig. 5 and the essential mathematical computation is shown in Eqs. (2)–(7).

$$F_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b^f) \quad (2)$$

$$I_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \quad (3)$$

$$G^t = \sigma(Wg * X_t + Wg * H_{t-1} + b_g) \quad (4)$$

$$C_t = C_{t-1}\Theta F_t + G_t\Theta I_t \quad (5)$$

$$O_t = \sigma(Wo * X_t + W_{ho}H_{t-1} + bo) \quad (6)$$

$$H_t = \tanh(C_t)\Theta O_t \quad (7)$$

The input feature map $X_t$, the cell output $C_t$, the hidden state $H_t$, the forget gates $F_t$, the input gate $I_t$, and the output gate $O_t$ are all defined in Eqs. (1)–(6). The convolution operator and Hadamard product, respectively, are indicated by the symbols * and ⊙. The 2D convolutional kernels acting on the inputs and hidden states corresponding to the input gate, output gate, input modulation gate and forget gate are represented by $W_{xi}$ & $W_{hi}$, $W_{xo}$ & $W_{ho}$, $W_{xg}$ & $W_{hg}$ and $W_{xf}$ & $W_{hf}$, respectively.

## 4 Experimental results and discussion

In the present section, we describe the databases that will be used to evaluate our model. We also discuss the implementation details. Finally, we compare our findings with cutting-edge methods documented in the literature.

### 4.1 Dataset Details

We evaluated our method on extended Cohn Kanade (CK+), SAVEE, and AFEW datasets. All these datasets contain clips of various facial expressions. The CK+and SAVEE are lab-controlled, whereas the AFEW is captured in the wild. Brief details of these datasets have been discussed below.

The CK+database contains 593 sequences from 123 subjects. But only 327 sequences of them are labeled with emotions [41]. Sequences start from the neutral frame and reach up to the peak frame in one of the universal seven basic emotions—surprise, fear, disgust, contempt, sadness, happiness and anger. Figure 6 displays a sample sequence of happy facial expressions from the dataset.

The SAVEE database contains a total of 480 video clips from 4 subjects [42]. The video contains both audio and visual information. We discard the audio part and extract only the image frames for our experiment. Videos of SAVEE dataset express one emotion throughout each video clip, that is, there is no onset or offset of the emotion expressed.

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 [43, 44]. 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. Like SAVEE, AFEW videos too have no onset or offset of the emotion expressed.

### 4.2 Implementation details

First, we extract the faces from the frames of all the databases using Dlib frontal face detector and correlation tracker. The obtained sequences of faces are aligned and

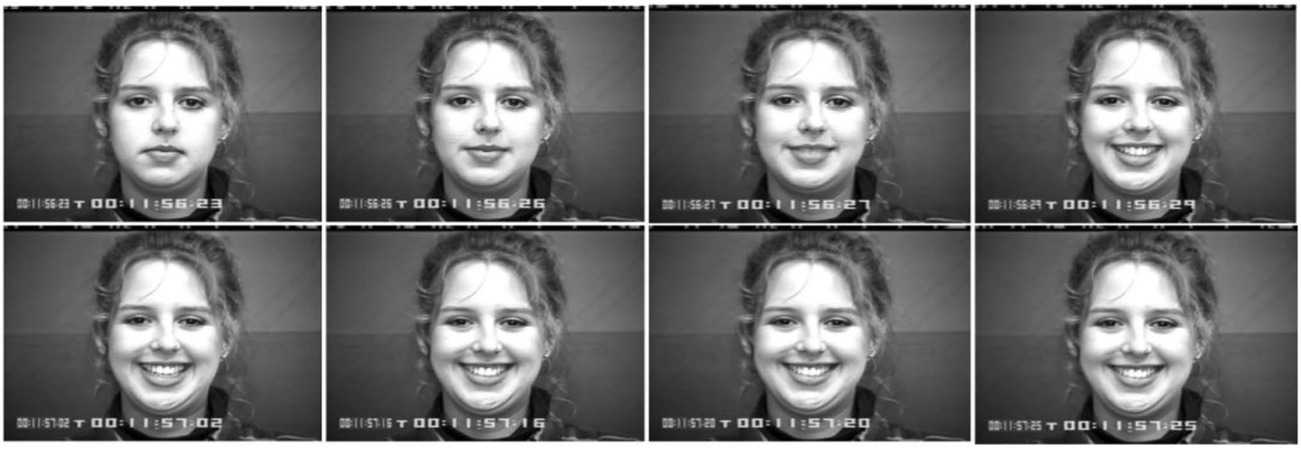
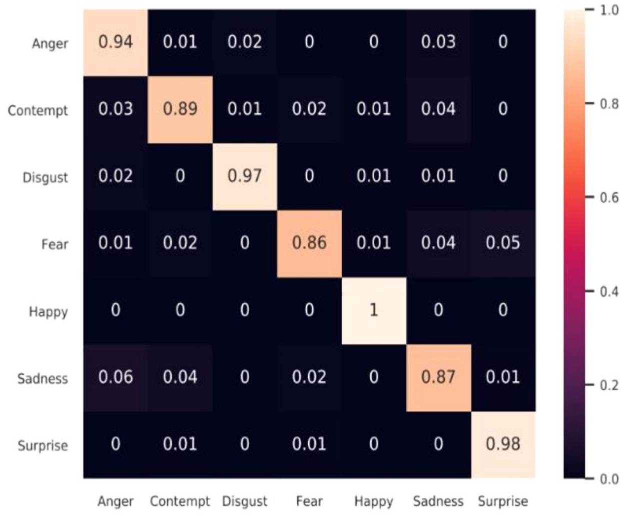

00:11:56:23 T 00:11:56:23
00:11:56:26 T 00:11:56:26
00:11:56:27 T 00:11:56:27
00:11:56:29 T 00:11:56:29
00:11:57:02 T 00:11:57:02
00:11:57:16 T 00:11:57:16
00:11:57:20 T 00:11:57:20
00:11:57:25 T 00:11:57:25

**Fig. 6** A sample facial expression sequence from the CK+dataset (top to bottom and left to right): Displays evolution of the happiness expression

**Table 2** Performance comparison with state-of-the-art methods on the CK+database

| Method | Accuracy(%) |
|---|---|
| 3D-CNN-DAP [16] | 87.90 |
| 3D-CNN [14] | 98.77 |
| LBP-TOP [17] | 82.40 |
| DTAGN [34] | 97.25 |
| STM-ExpLet [21] | 94.19 |
| 3DIR [2] | 93.21 |
| PHRNN-MSCNN [20] | 98.50 |
| Hybrid 3D-CNN & ConvLSTM (ours) | 95.10 |



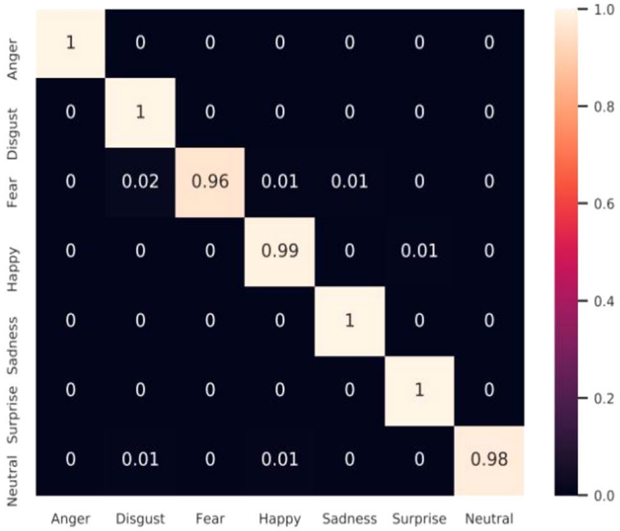
| | Anger | Contempt | Disgust | Fear | Happy | Sadness | Surprise |
|---|---|---|---|---|---|---|---|
| Anger | 0.94 | 0.01 | 0.02 | 0 | 0 | 0.03 | 0 |
| Contempt | 0.03 | 0.89 | 0.01 | 0.02 | 0.01 | 0.04 | 0 |
| Disgust | 0.02 | 0 | 0.97 | 0 | 0.01 | 0.01 | 0 |
| Fear | 0.01 | 0.02 | 0 | 0.86 | 0.01 | 0.04 | 0.05 |
| Happy | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Sadness | 0.06 | 0.04 | 0 | 0.02 | 0 | 0.87 | 0.01 |
| Surprise | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 0.98 |

**Fig. 7** Confusion matrix on the CK+database

then resized to 64×64-pixel resolution and converted to gray-scale. Additionally, for AFEW we perform gamma correction (γ = 1.5) for dark images. We also use a custom CNN trained on the FER2013 database to extract features to input into the ConvLSTM, as suggested in [1]. Thus, we have two features for AFEW, one from the hybrid 3D-CNN & ConvLSTM unit and other from the FER2013 pre-trained hybrid CNN & ConvLSTM. These two are added before passing onto the fully-connected layers. This processing is required because AFEW captures emotions in the wild and the small size of the database makes it difficult to learn the parameters of a deep model, considering the complexity of the emotions. The image sequences are normalized to a fixed length of 8 along the time axis. These eight frames are chosen in a uniform manner as follows: (1) the first and last frame of each sequence is taken to be the first and last frame of the 8-frame sequence (2) the rest 6 frames are selected in terms of an equal-interval scale.

**Table 3** Performance comparison with state-of-the-art methods on the SAVEE database

| Method | Accuracy (%) |
|---|---|
| CCCNN [19] | 93.90 |
| Banda and Robinson [23] | 98.00 |
| FAMNN [18] | 95.80 |
| 3DCNN [14] | 97.92 |
| 3DCNN-ConvLSTM (ours) | 98.83 |



| | Anger | Disgust | Fear | Happy | Sadness | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Anger | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Disgust | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Fear | 0 | 0.02 | 0.96 | 0.01 | 0.01 | 0 | 0 |
| Happy | 0 | 0 | 0 | 0.99 | 0 | 0.01 | 0 |
| Sadness | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Surprise | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Neutral | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 0.98 |

**Fig. 8** Confusion Matrix on the SAVEE database

To avoid over-fitting, we augment the datasets. For CK+, we rotate each image by different angle in range {−15°, −10°, −5°, 5°, 10°, 15°} and for SAVEE and AFEW the images are rotated by angle in range {−10°, −5°, 5°,10°}. Each of these images are flipped horizontally including the original image. The result is an augmented dataset having 14 times the size of the original for CK+ and 8 times the size of the original for SAVEE and AFEW dataset. The proposed network was implemented in Keras on the Tesla K80 GPU. While training the model, we employed Adam optimizer with a fixed learning rate of 0.0001. We used accuracy as our evaluation metric and categorical cross-entropy as our loss function. The network was trained for 80 epochs and the model with the highest validation accuracy was selected.

## 5 Results

The result for the CK+7-expression database is reported as the tenfold subject-independent cross-validation accuracy. Subjects in any two folds are mutually exclusive.

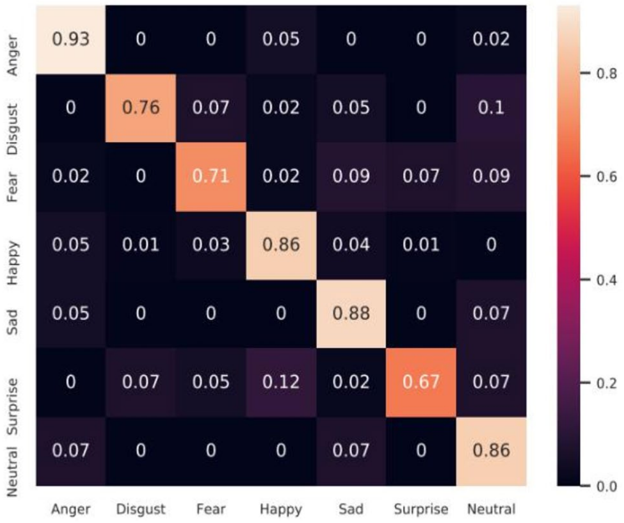


**Fig. 9** Confusion Matrix on the randomly selected 20% of AFEW samples

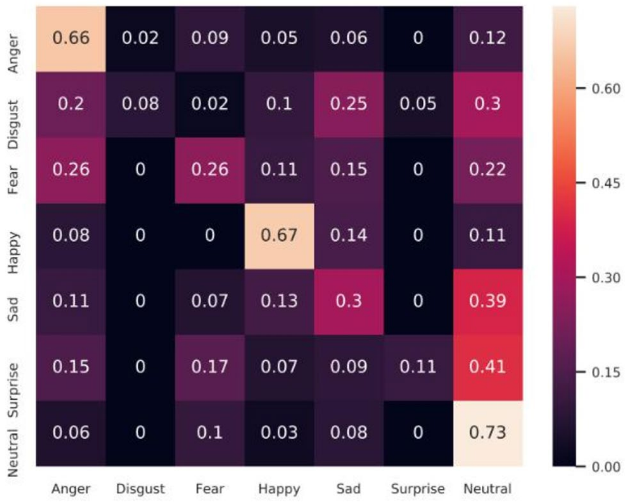


**Fig. 10** Confusion Matrix on the validation set of the AFEW database

As shown in Table 2, our proposed hybrid 3D-CNN & ConvLSTM model achieved over 95% accuracy. Although it does not beat the state-of-the-art, still it achieves competitive accuracy. Note that all the methods compared are sequence-based, that is, those which exploit spatial–temporal information. We achieved the accuracy without pretraining on any other emotion dataset, and we do not use facial landmark features as used in [2]. Figure 7 shows the confusion matrix of our result where fear and sadness achieve less recognition accuracy and are confused with each other. The emotions having less accuracy are the ones having the least amount of data available (18 sequences for contempt, 25 sequences for fear, and 28 sequences for sadness).

**Table 4** Performance comparison with state-of-the-art methods on the AFEW database

| Method | Accuracy (%) |
|---|---|
| 3D-CNN [14] | 38.12 |
| RNN [24] | 39.60 |
| ResNet-LSTM [22] | 46.70 |
| VGG-LSTM [15] | 48.60 |
| C3D [1] | 48.30 |
| Hybrid 3D-CNN & ConvLSTM (ours) | 43.86 |

The SAVEE dataset is partitioned into 80:20 ratio having 80% training data and 20% test data. Since there are only 4 subjects, the subject independent test is not viable. The accuracy reported on the 20% part is averaged over 5 runs. Table 3 compares our method with other state-of-the-art techniques. We achieve state-of-the-art accuracy on the SAVEE dataset, although with a small margin. It is important to note that we did not use audio features and that all the other methods employ a multi-modal fusion of audio and video features. Figure 8 shows the confusion matrix of our result, where fear reveals relatively low recognition accuracy, otherwise, the method performs very well in classifying the rest of the expression sequences in the test set.

We performed two types of experiments on the AFEW dataset. The first involves partitioning the dataset (train + validation) into 80:20 ratio with training on 80% and testing on 20%. The accuracy reported on the 20% part is averaged over five runs and attained an average accuracy of 83.12%. The second experiment involves training on train set and testing on the validation set. Here we get an accuracy of 43.86%. Figures 9 and 10 show the confusion matrix of the two experiments, respectively. The confusion matrix for the 2nd experiment indicates that anger, happiness, and neutral achieve good enough accuracy, whereas disgust and surprise perform poorly. Table 4 compares our method with other techniques. For a method, different accuracy corresponds to different models, with the last one being a fusion of multiple models. We note here that our method performs better than some of the individual models but worse than the fusion because we do not employ fusion techniques. Also, we do not use audio modality or any extra emotion videos to train our model. Note that the fusion accuracy reported except [15] does not incorporate audio features in the fusion process.

## 6 Conclusion

In this work, we presented a blend of 3D-CNN and ConvLSTM for facial expression recognition in videos (VFER).

This work presented a blend of 3D-CNN and ConvLSTM for facial expression recognition in videos (VFER). The proposed pipeline consists of a face detector, a landmark detector for facial image alignment, and a deep convolutional neural network architecture. First, the proposed VFER framework has employed an input pre-processing step to obtain aligned facial images from the facial expression video sequences. Once obtained, the pipeline feeds the cropped face images corresponding to specified input frames as input to the lightweight 3D-CNN. The designed 3D-CNN incorporates the spatial–temporal correlations in its feature map. The resulting feature map is then fed to the ConvLSTM to extract long-term temporal information. We evaluated the proposed hybrid 3D-CNN & ConvLSTM model on the CK+, SAVEE, and AFEW databases. The proposed framework for VFER is compute-efficient and attained competitive accuracy on the benchmark datasets for video-based facial expression recognition. Our future work will include the facial landmark features, audio modality if available, and the use of pre-trained models to give a good initialization point and overcome the shortage of facial expression data. The future work also includes the implementation of the VFER algorithm in edge computing devices.

**Acknowledgements** The authors thank the Director, CSIR-CEERI, for his constant encouragement for research efforts at the institute.

**Availability of data and materials** The datasets will be made available on request.

## References

1. Fan Y, Lu X, Li D and Liu Y (2016) Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp 445–450
2. Hasani B and Mahoor MH (2017) Facial expression recognition using enhanced deep 3d convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 30–40
3. Xingjian S, Chen Z, Wang H, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in Advances in neural information processing systems, 2015, pp. 802–810.
4. Srivastava N, Mansimov E and Salakhudinov R (2015) Unsupervised learning of video representations using lstms. In: International conference on machine learning, pp 843–852
5. Wang Z and Ying Z (2012) Facial expression recognition based on local phase quantization and sparse representation. In: 2012 8th International Conference on Natural Computation. IEEE, pp 222–225
6. Carcagnì P, Del Coco M, Leo M, Distante C (2015) Facial expression recognition and histograms of oriented gradients: a comprehensive study. Springerplus 4(1):645
7. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816
8. Saurav S, Singh S, Yadav M and Saini R (2020) Image-based facial expression recognition using local neighborhood difference binary pattern. In: Proceedings of 3rd International Conference on Computer Vision and Image Processing. Springer, pp 457–470
9. Yu Z and Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 435–442
10. Yang B, Cao J, Ni R, Zhang Y (2017) Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access 6:4630–4640
11. Sun N, Li Q, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61
12. Hasani B and Mahoor MH (2017) Spatio-temporal facial expression recog-nition using convolutional neural networks and conditional random fields. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, pp 790–795
13. Yu Z, Liu G, Liu Q, Deng J (2018) Spatio-temporal convolutional features with nested lstm for facial expression recognition. Neurocomputing 317:50–57
14. Zhao J, Mao X, Zhang J (2018) Learning deep facial expression features from image and optical flow sequences using 3D-CNN. Vis Comput 34(10):1461–1475
15. Vielzeuf V, Pateux S and Jurie F (2017) Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp 569–576
16. Liu M, Li S, Shan S, Wang R and Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian conference on computer vision. Springer, pp 143–157
17. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
18. Gharavian D, Bejani M, Sheikhan M (2017) Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks. Multimedia Tools Appl 76(2):2331–2352
19. Barros P, Wermter S (2016) Developing crossmodal expression recogni-tion based on a deep neural model. Adapt Behav 24(5):373–396
20. Zhang K, Huang Y, Du Y, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203
21. Liu M, Shan S, Wang R and Chen X (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1749–1756
22. Ouyang X, Kawaai S, Goh EGH, Shen S, Ding W, Ming H and Huang D-Y (2017) Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp 577–582

23. Banda N and Robinson P (2011) Noise analysis in audio-visual emotion recognition. In: Proceedings of the 11th International Conference on Multimodal Interaction (ICMI). Citeseer
24. Ebrahimi Kahou S, Michalski V, Konda K, Memisevic R and Pal C (2015) Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp 467–474
25. Kumar A, Jain S, and Kumar M (2022) Face and gait biometrics authentication system based on simplified deep neural networks. Int J Inf Technol 15(2):1–10
26. Jayanthi K, Mohan S (2022) An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach. Int J Inf Technol. https://doi.org/10.1007/s41870-022-00900-5
27. Pandey NN, Muppalaneni NB (2021) A novel algorithmic approach of open eye analysis for drowsiness detection. Int J Inf Technol 13(6):2199–2208
28. Uddin KMM, Dey SK, Parvez GU, Mukta AS, Acharjee UK (2021) MirrorME: implementation of an IoT based smart mirror through facial recognition and personalized information recommendation algorithm. Int J Inf Technol 13(6):2313–2322
29. Sabharwal T, Gupta R (2021) Facial marks for enhancing facial recognition after plastic surgery. Int J Inf Technol 13(1):391–396
30. Liu D, Zhang H and Zhou P (2021) Video-based facial expression recognition using graph convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 607–614. IEEE
31. Lee MK, Choi DY, Kim DH and Song BC (2019) Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp 1–8. IEEE
32. Pan X, Guo W, Guo X, Li W, Junjie Xu, Jinzhao Wu (2019) Deep temporal–spatial aggregation for video-based facial expression recognition. Symmetry 11(1):52
33. Mehta NK, Prasad SS, Saurav S, Saini R and Singh S (2022) Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. Appl Intell 1–21
34. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y and Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
35. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S et al (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7310–7311
36. Yang S, Luo P, Loy C-C and Tang X (2016) Wider face: a face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5525–5533
37. Kazemi V and Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1867–1874
38. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758
39. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG (2018) Videolstm convolves, attends and flows for action recognition. Comput Vis Image Underst 166:41–50
40. Wang L, Xu Y, Cheng J, Xia H, Yin J, Wu J (2018) Human action recognition by learning spatio-temporal Features with deep neural networks. IEEE Access 6:17913–17922
41. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z and Matthews I (2010) 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, pp 94–101
42. Haq S, Jackson PJ and Edge J (2009) Speaker-dependent audio-visual emotion recognition. AVSP 2009:53–58
43. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia 3:34–41
44. Dhall A, Ramana Murthy O, Goecke R, Joshi J and Gedeon T (2015) Video and image-based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 423–426
45. Jung H, Lee S, Yim J, Park S and Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2983–2991