Video-Based Facial Expression Recognition using a Blend of 3D CNN and ConvLSTM

Sumeet Saurav
Academy of Scientific & Innovative Research
CSIR-Central Electronics Engineering Research Institute
Pilani, India
sumeet@ceeri.res.in

Tarun Kumar
Department of Computer Science
Birla-Institute of Technology and Science
Pilani, India
kumar1998.tarun@gmail.com

Ravi Saini
Integrated System Group
CSIR-Central Electronics Engineering Research Institute
Pilani, India
ravi@ceeri.res.in

Sanjay Singh
Cognitive Computing Group
CSIR-Central Electronics Engineering Research Institute
Pilani, India
sanjay@ceeri.res.in

Abstract—**The 3-Dimensional Convolutional Neural Network (3D CNN) and Long Short-Term Memory Network (LSTM) have consistently outperformed many approaches in video-based Facial Expression Recognition (VFER). The vanilla version of the fully-connected LSTM (FC-LSTM) unrolls the image to a one-dimensional vector, which results in the loss of vital spatial information. Convolutional LSTM (ConvLSTM) overcomes this limitation by performing LSTM operations in terms of convolutions without performing any unrolling, as in the case with FC-LSTM. Motivated by this, in this paper, we propose a neural network architecture that consists of a blend of 3D CNN and ConvLSTM. The proposed hybrid architecture captures spatial-temporal information to produce competitive accuracy on three publicly available FER databases, namely the CK+, SAVEE, and AFEW. The experimental results demonstrate excellent performance without using any external emotion data with an added advantage of having a simple model with a comparatively fewer number of parameters and model size. Our designed FER pipeline is a suitable candidate for automatic recognition of facial expressions in real-time on a resource-constrained embedded platform.**

Index Terms—**Video Facial expression recognition (VFFE); 3D convolutional neural networks (3D CNN); long short-term memory (LSTM); convolutional LSTM (ConvLSTM).**

I. Introduction

Human-computer interaction (HCI) has seen a lot of development in recent years. Video-based facial expression recognition (VFER) remains the most central component of HCI with a variety of methods proposed to date. However, VFER remains a challenging problem due to several limitations associated with the 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. Recently, 3-dimensional convolutional neural networks (3DCNN) have shown great ability in FER tasks by modeling spatial and temporal information simultaneously. These 3DCNN architectures have often been combined in different ways with recurrent neural network (RNN) models to capture long term dependencies and feature enhancement, as demonstrated in [1], [2].

Some of the above hybrid models of 3DCNN and RNN perform very well in the VFER task. Based on the performance of ConvLSTM in several computer vision applications, in this paper, we explore the combination of 3DCNN and Convolutional Long-Short Term Memory (ConvLSTM) for the classification of facial expressions in a video sequence. ConvLSTM as proposed in [3], shows its superiority over fully-connected LSTM (FC-LSTM) in number of spatial-temporal tasks like Moving MNIST Dataset, Radar Echo Dataset [4]. The main advantage of ConvLSTM is that the input-to-state and state-to-state transitions are convolutional, and thus, 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. We evaluate our approach on three well-known FER databases, namely CK+ [5], SAVEE [6], and AFEW [7]. Without using any additional videos with emotion labels in our training set, our approach achieved competitive performance.

A typical FER algorithmic pipeline consists of three main stages, namely the face detection, feature extraction, and feature classification. The face detection stage often accompanied by the alignment is more or less similar in all the algorithms. The techniques mainly used are the Viola-Jones and more recent deep learning-based face detectors (OpenCV). Traditionally, the feature extraction stage is done using Local Binary Patterns (LBP) [8], Histogram of Oriented Gradients (HOG) [9], Local Phase Quantization (LPQ) [10]. Since these methods are hand-crafted for their specific application, they do not generalize well in different imaging conditions such as lighting, occlusion, subjects' ethnicity, etc. More recently, works based on CNNs have been proposed for facial expres-

978-1-7281-6916-3/20/$31.00 ©2020 IEEE

sion recognition and have demonstrated to achieve superior performance.

The remainder of the paper is organized as follows: Section 2 provides an overview of the related works on video-based facial expression recognition. Section 3 explains our proposed hybrid of 3D CNN and ConvLSTM. Experimental results and analysis are presented in Section 4, followed by conclusive remarks in Section 5.

II. Related Works

Techniques available in the literature for facial expression recognition (FER) could be broadly classified into two major categories viz image-based methods and video/sequence-based methods. 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 in [11] used an ensemble of multiple deep networks for FER in a single static image. In [12], the authors proposed a technique for FER using a weighted fusion of features extracted from the gray-scale and local binary pattern (LBP) encoded facial images. Expression related features from gray-scale facial images were extracted using ImageNet pre-trained VGG16 CNN and, that from the LBP encoded facial images using a shallow 2D-CNN trained from scratch. The authors in [13] presented a multi-channel deep neural network that learns and fuses the spatial-temporal features for recognizing facial expressions in static images. Here optical flow image is extracted between the peak expression face image and neutral face image and used as temporal information. Moreover, the emotional face image alone is used as the spatial information.

The video/sequence-based methods for facial expression recognition, on the other hand, extract both spatial and temporal information from an expression sequence, either using a hybrid of CNN & LSTM or a 3DCNN. In [14], the authors presented an integrated deep learning framework for facial expression recognition. Two deep learning models, one extracting the temporal appearance features from gray-scale images, and the other extracting the temporal geometrical features using landmark-based geometrical features, are first trained separately. Once trained, the models are combined using a novel integration method to boost the recognition accuracy. In [15], the authors proposed a technique for video-based facial expression recognition using a combination of deep neural network (DNN) and conditional random field (CRF). Here DNN has been used to capture the spatial relationship among the expression images, whereas a linear change of CRF is used to capture the temporal relations. In another work discussed in [16], a new technique has been proposed for facial expression recognition in videos. The proposed technique captures spatiotemporal information using a combination of CNN and nested LSTM. The spatiotemporal convolutional features from the image sequences are extracted using 3D CNN and the dynamics of the facial expressions are captured by nested LSTM with two-layered architecture. The first layer LSTM called T-LSTM is used to model the temporal dynamics of the spatiotemporal features in each convolutional layer, and the second layer LSTM called C-LSTM is used to facilitate the integration of the outputs of all T-LSTMs together. This helps in encoding the multi-level features present in the intermediate layers of the network. In yet another work presented in [17], the authors have used a 3DCNN architecture that learns to extract the static information using RGB sequences and dynamic information from optical flow sequences. The performance of the presented framework was tested on three publicly available FER databases namely the CK+ database, AFEW database, and SAVEE database. Two types of techniques were used for the extraction of optical flow. The first technique used for optical flow computation is the regular optical flow and uses the Gunner Farneback's algorithm. The second type contains the accumulative motion information of facial muscle movement, so it is called accumulative optical flow. In [18], the authors have proposed several schemes for facial expression recognition in an emotion clip using a hybrid of 2D-CNN & LSTM and C3D & LSTM. Moreover, a fusion of these schemes has also been explored to enhance the performance utilizing the effectiveness of the different schemes. Moreover, a hybrid of recurrent neural network (RNN) and 3D convolutional networks (C3D) has been explored for VFER in [1]. A technique for VFER using 3D CNN incorporating deformable parts learning has been demonstrated in [19]. Zhao et al. [20], proposed a method for facial expression recognition in a video sequence using Local Binary Pattern (LBP) features extracted from three orthogonal planes (LBP-TOP). Their presented LBP-TOP operator extracts expression features from a sequence of frames in the form of histograms computed from three orthogonal planes. In [21], the authors presented a technique for VFER using a fuzzy ARTMAP neural network (FAMNN). Hyper-parameters of the FAMNN has been determined using the particle swarm optimization (PSO). In [22], the authors proposed a technique for VFER using a Cross-channel Convolutional Neural network (CCCNN). In [23], the authors proposed a technique for VFER using a part-based hierarchical bidirectional recurrent neural network (PHRNN). Moreover, a multi-signal convolutional neural network (MSCNN) has also been proposed to extract spatial features from still frames of the expression sequence. Finally, both PHRNN and MSCNN fused extracts the partial-whole, geometry-appearance, and dynamic-still information, effectively boosting the performance of facial expression recognition. A manifold learning-based framework has been explored for VFER by Liu et al. [24]. Scheme for VFER using transfer learning and spatial-temporal fusion has been demonstrated by Ouyang et al. [25] whereas a hybrid of 2D CNN and RNN has been utilized in [26]. The work presented in [27], have proposed a fusion of audio and visual expression related information for FER. The also analyzed the impact of noise on the final decision of the system.

III. Proposed Method

Figure 1 shows the block diagram representation of our proposed facial expression recognition framework. As could

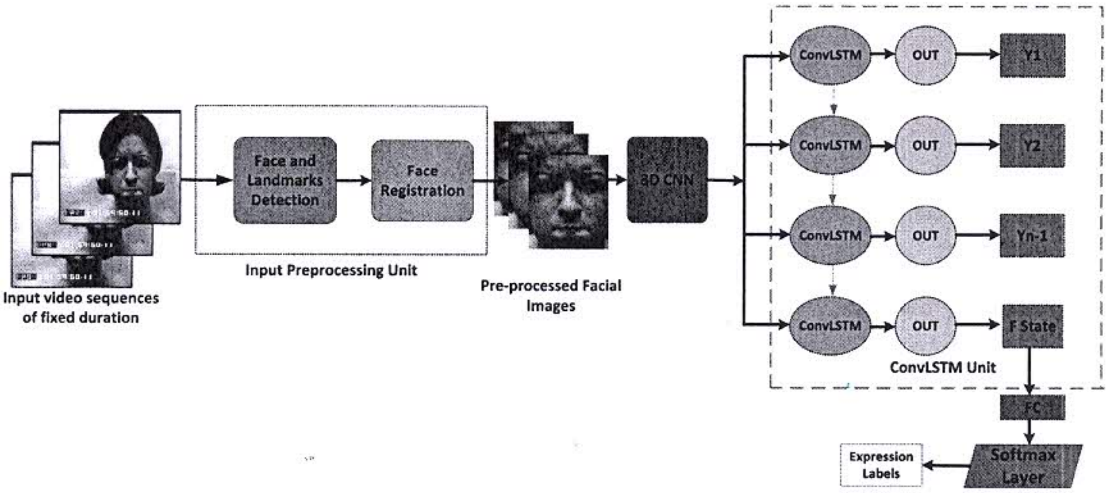

Fig. 1. Proposed framework for video-based facial expression recognition

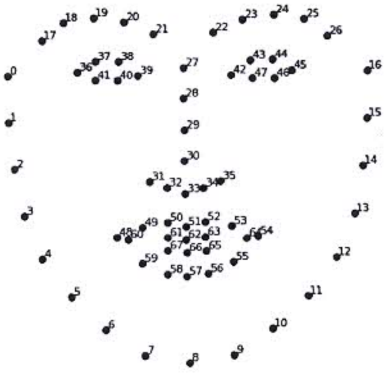

Fig. 2. Prototypical facial image with 68 landmarks

be seen from the figure, the framework consists of a sequence of steps with specific functionality. In the first step, the input pre-processing unit takes video frames as input and returns the location of the face and facial landmarks. Using the facial landmark information, the detected faces are aligned such that they share spatial symmetry. The aligned facial images are then cropped and scaled to a standard size. The second unit called the 3D CNN unit takes as input the pre-processed facial images of some fixed duration (# frames) and outputs a sequence of features maps learned using a set of convolutional and max-pooling layers. These feature maps are then fed to the ConvLSTM unit, which further extracts information from the feature maps and learns to model the evolution of expressions as contained in the input frames. The feature maps from the ConvLSTM unit are flattened and passed to a fully-connected and softmax layer. The softmax layer then classifies the extracted representation corresponding to an expression sequence into one of the expression labels. Below, we provide details of the various units used in our proposed FER framework.

A. Input pre-processing unit

As discussed above, the input processing unit takes as input the video frames and provides the registered facial image of 64 x 64 pixels resolution. For face detection, we used the open-source face detector and facial landmark detector available at [28]. The position of the 68-landmarks obtained from the landmark localizer is shown in Fig. 2. With the help of these landmarks, the facial image is aligned to share a similar spatial location of different facial components.

B. 3D CNN and ConvLSTM Unit

Better recognition accuracy is usually achieved by sequence-based facial expression techniques. These techniques using video sequences, extract both the spatial and temporal features, and learns to model the evolution of the facial expressions. Deployment of 3D CNN for such an application has a direct advantage as it could model both spatial and temporal changes associated with the change in facial expressions. Moreover, one could also find it more reasonable to learn short-term spatial-temporal features by 3D CNN and long-term spatial-temporal features by LSTM/RNN. In such a situation, the vanilla FC-LSTM is often used, however, it results in loss of spatial correlation information. Therefore, we used ConvLSTM on the output of the 3D CNN in our proposed approach as shown in Fig. 1.

As shown in Fig. 3, the 3D CNN model has three 3D convolutional layers, one 3D max-pooling layer, one ConvLSTM block, one fully-connected layer, and a softmax classifier layer. Each 3D convolution block is followed by a ReLU activation function to perform the non-linear transformation of features between the layers. The output from the second max-pool layer is fed as input to the ConvLSTM layer which has 16 units. The subsequent layer is the fully connected layer, which takes as input the flatten output form the ConvLSTM layer. Finally, the fully-connected layer is connected to the softmax layer, which categories the output from the facial sequences into one of

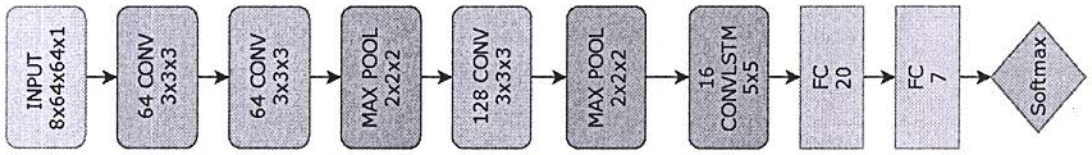

Fig. 3. Schematic representation of the proposed 3D-CNN

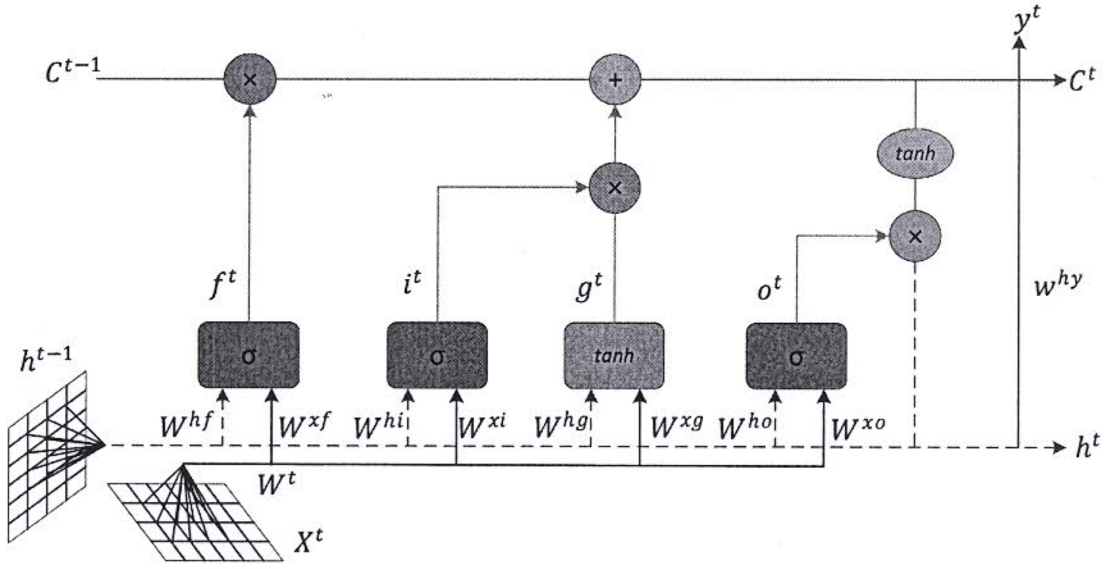

Fig. 4. Schematic representation of basic ConvLSTM architecture

the seven facial expression classes, namely neutral, surprise, happy, disgust, sad, fear, and anger.

Though LSTM is very successful in handling temporal informations, however, full connections in input-to-state and state-to-state transitions in LSTM does not take spatial information into account, and thus, leads to loss of spatial information, which is crucial modeling spatial-temporal information involved in fall and non-fall activities. To overcome this limitation of LSTM, Xingjian et al. [3] introduced convolutional LSTM (ConvLSTM), in which the state-to-state transition operations in the LSTM are replaced by convolution operations. Recently, several works have also been proposed exploiting the usefulness of ConvLSTM in human action related tasks [29], [30]. Figure 4 shows the internal details of the ConvLSTM layer, and necessary mathematical computation involved is depicted in Eqs. (1)-(6).

$$F^t = \sigma\left(W^{xf} * X^t + W^{hf} * H^{t-1} + b^f\right) \tag{1}$$

$$I^t = \sigma\left(W^{xi} * X^t + W^{hi} * H^{t-1} + b^i\right) \tag{2}$$

$$G^t = tanh\left(W^{xg} * X^t + W^{hg} * H^{t-1} + b^g\right) \tag{3}$$

$$C^t = C^{t-1} \odot F^t + G^t \odot I^t \tag{4}$$

$$O^t = \sigma\left(W^{xo} * X^t + W^{ho} H^{t-1} + b^o\right) \tag{5}$$

$$H^t = tanh\left(C^t\right) \odot O^t \tag{6}$$

In Eqs. (1)-(6), X^t is the input feature map, C^t is the cell output, H^t represents hidden state, F^t is the forget gates, I^t is the input gate, and O^t is the output gate. Also, $*$ and $\odot$ denote the convolution operator and Hadamard product, respectively. Additionally, W^{xi} & W^{hi}, W^{xg} & W^{hg}, W^{xf} & W^{hf}, and W^{xo} & W^{ho} are the 2D convolutional kernels operating on the inputs and hidden states corresponding to the input gate, input modulation gate, forget gate, and output gate, respectively.

IV. Experimental results and discussions

In this section, we introduce the databases used for evaluating our model. We also discuss the implementation details. Finally, we compare our results with the state-of-the-art techniques available in the literature.

A. 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. Sequences start from the neutral frame and reach up to the peak frame in one of the seven basic emotions -

anger, contempt, disgust, fear, happiness, sadness, and surprise. The SAVEE database contains a total of 480 video clips from 4 subjects. 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. The acted facial expressions in the wild (AFEW) database is a dynamic temporal facial expression dataset that contains close to real-world emotions extracted from Movies and TV series. 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.

B. Implementation Details

First, we extract the faces from the frames of all the databases using Dlib's frontal face detector and correlation tracker. The obtained sequences of faces are aligned and then resized to 64x64 pixel resolution and converted to gray-scale. Additionally, for AFEW we perform gamma correction $\gamma = 1.5$ 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 3D-CNN-ConvLSTM unit and other from the (FER2013) 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. We normalize the image sequences along the time axis to a fixed length of 8. These 8 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.

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

C. Results

The result for the CK+ 7 database is reported as the 10-fold subject-independent cross-validation accuracy. Subjects in any 2 folds are mutually exclusive. As shown in Table 1,

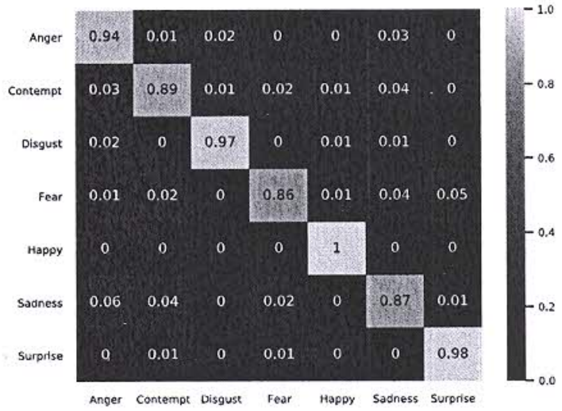

Fig. 5. Confusion matrix on the CK+ database

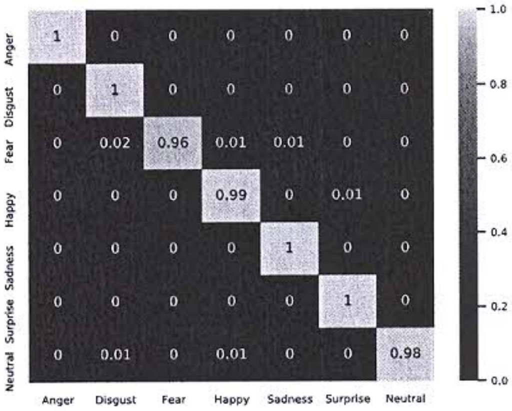

Fig. 6. Confusion matrix on the SAVEE database

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CK+ DATABASE

Method	Accuracy(%)
3DCNN-DAP [19]	87.90
3DCNN [17]	98.77
LBP-TOP [20]	82.40
DTAGN [14]	97.25
STM-ExpLet [24]	94.19
3DIR [2]	93.21
PHRNN-MSCNN [23]	98.50
3DCNN-ConvLSTM (ours)	95.10

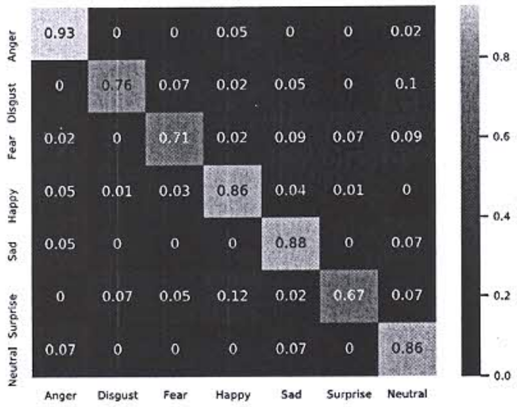

Fig. 7. Confusion matrix obtained on the randomly selected 20% of AFEW samples

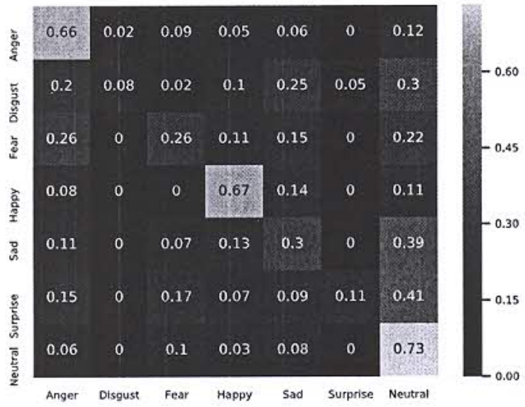

Fig. 8. Confusion matrix obtained on the validation set of the AFEW database

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SAVEE DATABASE

Method	Accuracy(%)
CCCNN [22]	93.90
Banda and Robinson [27]	98.00
FAMNN [21]	95.80
3DCNN [17]	97.92
3DCNN-ConvLSTM (ours)	98.83

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE AFEW DATABASE

Method	Accuracy(%)
3DCNN [17]	38.12
RNN [26]	39.60
Resnet-LSTM [25]	46.70
VGG-LSTM [18]	48.60
C3D [1]	48.30
3DCNN-ConvLSTM (ours)	43.86

our proposed hybrid of 3D CNN and ConvLSTM 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 to are sequence-based, that is, those which exploit spatial-temporal information. We achieved the accuracy without pre-training on any other emotion dataset, and we do not use facial landmark features as used in [14] and [2]. Fig. 4a 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).

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 2 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. Fig. 4b shows the confusion matrix of our result, where fear reveals a relatively low recognition accuracy, otherwise, the method performs very well in classifying the rest of the expression sequences in the test set.

We perform two types of experiments on AFEW. 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. We get an 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%. Fig. 5a and Fig. 5b shows the confusion matrix of the two experiments, respectively. The confusion matrix for the 2nd experiment tells that anger, happy, and neutral achieve good enough accuracy, whereas disgust and surprise perform poorly. Table 3 compares our method with other techniques. Our method performs fairly here. For a particular 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 [18] does not incorporate audio features in the fusion process.

V. CONCLUSION

In this work, we 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 input is fed to the 3D-CNN model which incorporates the spatial-temporal correlations in its feature map. The resulting feature map is then fed to the ConvLSTM to further extract temporal information. We evaluate our approach on CK+, SAVEE, and AFEW databases. Our experiments demonstrate competitive accuracy on all the databases. 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.

ACKNOWLEDGMENT

The authors would like to thank the Director, CSIR-CEERI, Pilani for supporting and encouraging research activities at CSIR-CEERI, Pilani. Constant motivation by the group head of the cognitive computing group at CSIR-CEERI is also acknowledged.

REFERENCES

[1] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 445–450.

[2] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30–40.

[3] S. Xingjian, Z. Chen, H. Wang, 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] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

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

[6] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, 2009, pp. 53–58.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, no. 3, pp. 34–41, 2012.

[8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[9] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, 2015.

[10] Z. Wang and Z. Ying, "Facial expression recognition based on local phase quantization and sparse representation," in *2012 8th International Conference on Natural Computation*. IEEE, 2012, pp. 222–225.

[11] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[12] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.

[13] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.

[14] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.

[15] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 790–795.

[16] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested lstm for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, 2018.

[17] J. Zhao, X. Mao, and J. Zhang, "Learning deep facial expression features from image and optical flow sequences using 3d cnn," *The Visual Computer*, vol. 34, no. 10, pp. 1461–1475, 2018.

[18] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 569–576.

[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[20] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[21] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2331–2352, 2017.

[22] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive behavior*, vol. 24, no. 5, pp. 373–396, 2016.

[23] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[24] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1749–1756.

[25] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 577–582.

[26] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467–474.

[27] N. Banda and P. Robinson, "Noise analysis in audio-visual emotion recognition," in *Proceedings of the 11th International Conference on Multimodal Interaction (ICMI)*. Citeseer, 2011.

[28] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[29] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[30] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17913–17922, 2018.