Micro-expression recognition using a shallow ConvLSTM-based network

Saurav Shukla[1 [0000−0002−7973−073X]], Prabodh Kant Rai[1 [0000−0002−3637−3883]], and Tanmay T. Verlekar[1,2 [0000−0001−5534−4351]]

[1] Dept. of CSIS, BITS Pilani, K.K Birla Goa Campus, 403726 Goa, India
[2] APPCAIR, BITS Pilani, K.K Birla Goa Campus, 403726 Goa, India
{f20180653,f20180748,tanmayv}@goa.bits-pilani.ac.in

Abstract. Micro-expressions reflect people's genuine emotions, making their recognition of great interest to the research community. Most state-of-the-art methods focus on the use of spatial features to perform micro-expression recognition. Thus, they fail to capture the spatiotemporal information available in a video sequence.
This paper proposes a shallow convolutional long short-term memory (ConvLSTM) based network to perform micro-expression recognition. The convolutional and recurrent structures within the ConvLSTM allow the network to effectively capture the spatiotemporal information available in a video sequence. To highlight its effectiveness, the proposed ConvLSTM-based network is evaluated on the SAMM dataset. It is trained to perform micro-expression recognition across three (positive, negative, and surprise) and five (happiness, other, anger, contempt, and surprise) classes. When compared with the state-of-the-art, the results report a significant improvement in accuracy and the F1 score. The proposal is also robust against the unbalanced class sizes of the SAMM dataset.

Keywords: Micro-expression recognition · shallow neural networks · convolutional LSTM

1 Introduction

Micro-expressions are twitches on people's faces that reveal their genuine emotions. They are spontaneous reactions to an external stimulus that people consciously suppress, resulting in voluntary and involuntary emotional responses co-occurring and conflicting with one another [1]. The fleeting and subtle nature of the micro-expressions make their detection and recognition a challenging task.
Psychologists believe that 55% of people's emotional states can be conveyed through facial expressions [2] . However, most macro-expressions can be forced, making them unreliable. Micro-expressions being difficult to suppress, hide, disguise, or conceal, make them the preferred choice in understanding people's

emotional states. Thus, successful micro-expression recognition has several applications, ranging from criminal investigation to psychological and medical treatment.
Traditionally, professionals performed micro- expressions recognition through visual inspection. The technique was highly ineffective even after obtaining sufficient training, as most micro-expressions last less than 0.33 seconds [3]. With the development of high frame rate cameras, computer vision and machine learning, researchers have found some success detecting and recognising micro-expressions. This paper proposes a novel strategy for recognising micro-expressions using ConvLSTM-based network. The convolutional structures in both the input-to-state and state-to-state transition of ConvLSTM allow the network to capture complex information, which is not possible with just convolutional or recurrent neural networks (CNN, RNN). The structure, in turn, enables the construction of a shallow network for micro-expression recognition.

2 Literature review

The biggest challenge faced by the researchers in addressing the problem of micro-expression recognition is the availability of data. One of the first publically available datasets was collected in 2009, called the Polikovsky dataset [4]. The dataset contains 42 video sequences from 11 people captured at 200fps. Apart from the limited number of video sequences, another drawback of the dataset is that the participants posed the six different micro-expressions. The simulation of the micro-expressions was a common problem with most initial datasets. The spontaneous micro-expression corpus (SMIC) series dataset was the first to capture spontaneous micro-expression. The dataset contains 77 sequences of just six people recorded using a 100 fps camera [5]. The micro-expressions were induced by making people watch video clips with emotional stimulation. The captured micro-expressions were labelled as positive, negative and surprise. Finer micro-expressions such as happiness, fear, disgust, and sadness, among others, were made available in the Chinese academy of sciences micro-expressions (CASME) series datasets [6].
The most recent dataset called the spontaneous actions and micro-movements (SAMM) series dataset, has significant improvements over the other existing datasets [7]. It involves the participation of 32 people to capture 159 video sequences comprising of seven spontaneous micro-expression using a 200 fps camera. The participants are widely distributed in terms of age, gender and ethnicity, making it the preferred dataset for evaluating micro-expression recognition techniques. Hence, this paper reports its results using the SAMM dataset.
The initial work on micro-expressions involved detecting the time stamp indicating their beginning (onset), end (offset), and their peak manifestation (apex).They were detected by analysing the difference between consecutive frames using techniques such as local binary patterns (LBP) [8] and histogram of oriented gradients [9]. Recently, local temporal patterns have been used to detect facial move-

ment across frames. The specific signature associated with the micro-expressions identified in those facial movements has led to the most promising results [10].

2.1 Traditional micro-expression recognition techniques

Typically, once the onset and offset of a micro-expression are detected, the frames belonging to that time interval are used to recognise the micro-expression. Initial works on micro-expression recognition were evaluated on posed datasets. Because of the small size of the datasets, several hand-crafted techniques were developed. The most successful among them involved the use of LBP across the three orthogonal planes (LBP-TOP) of a stacked video sequence tensor [11]. The recognition was then performed using a support vector machine (SVM). The LBP-TOP was further improved by reducing the redundancy and sparsity in the feature descriptors. Redundancy was reduced by identifying six neighbourhood points associated with the three intersecting lines of the orthogonal plane, where the LBP could be applied [12]. The sparsity problem was addressed using components of each orthogonal plane, such as magnitude, sign and orientation, to generate compact feature descriptors [13]. Apart from LBP, optical flow was used to generate dynamic feature maps that captured the movements observed on the face during the manifestation of a micro-expression [14]. SVM then used these dynamic feature maps to perform micro-expression recognition. Other operations were also performed over the optical flow to obtain better feature descriptors, such as computing its derivative [15]. In contrast to global operations, local operations, such as assigning weights to specific regions of the dynamic feature maps, contributed to an increase in the micro-expression recognition accuracy [16].

2.2 Micro-expression recognition using neural networks

Recently, CNNs have been extensively used to address image/video-based problems. With larger datasets being captured in recent years, CNNs have also been explored to address the problem of micro-expression recognition. In works such as [17], cropped apex frames are used as inputs to deep CNNs to classify them across six different micro-expression classes. DeepCNNs have also been used to detect facial landmarks, which in turn are used to generate optical flow-based dynamic feature maps [18]. Popular deep CNNs such as ResNet trained on ImageNet, have also been repurposed to perform micro-expression recognition through fine-tuning [19]. These deep CNNs are altered by introducing attention units in residual blocks to capture subtle features associated with micro-expression.
Besides deep CNNs, shallow CNNs have also been explored to obtain lightweight networks that can effectively recognise micro-expressions. The shallow CNNs, in some cases, are truncated versions of deep CNNs such as AlexNet with optical flow-based dynamic feature maps as their input [20], [21]. They are used in a dual-stream configuration to capture discriminative features in salient facial areas. In other cases, novel architectures are proposed, such as shallow CNNs

with 3D convolutions to capture temporal information across frames [22]. Temporal information can be effectively captured using RNNs. The idea is explored in works such as [23], where spatial features captured using VGG-16 are fed to LSTM to perform micro-expression recognition.
The review suggests that within the state-of-the-art, CNNs are the most effective in performing micro-expression recognition. However, most CNNs cannot capture the spatiotemporal information available in video sequences. A second drawback of the state-of-the-art is the use of deep CNNs through fine-tuning. These networks are usually proposed for significantly more complex problems making them computationally expensive. Also, the features used for micro-expression recognition are usually more subtle than other vision problem that these networks address. Shallow CNNs have shown some promise, but their results have scope for improvement, especially on the SAMM dataset that containd unbalanced micro-expression classes.
This paper presents a shallow neural network that uses convolution in tandem with LSTM. The merging of convolution in LSTM allows the network to capture spatiotemporal information more effectively than other neural networks. Being shallow enables the network to be trained end-to-end on small datasets. Thus we hypothesise that using ConvLSTM in our proposed micro-expression recognition network allows it to perform better than other shallow CNNs, and deep CNN+LSTM networks.

3 Proposal

Inspired by the shallow networks for micro-expression recognition discussed in section 2.2, we propose a ConvLSTM-based network to effectively capture the spatiotemporal information available in video sequences. The entire architecture of the proposal is presented in Fig. 1. Its details are discussed in the following three sub-sections:

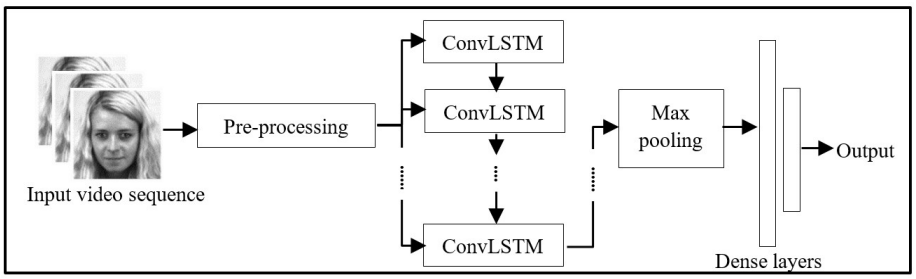

Fig. 1. The architecture of the proposed ConvLSTM based network for micro-expression recognition.

– Pre-processing: To prepare the video sequences for processing by the network;
– ConvLSTM: To capture spatiotemporal information available in the video sequence;
– Dense layers: To perform the final classification.

3.1 Pre-processing

Since each micro-expression lasts for a variable length, and the micro-expressions themselves are captured using a high fps camera, we first downsample the video sequence to t frames. These t frames include the onset, offset and the apex frames. The remaining frames are selected by uniformly sampling between onset and offset frames. Thus, given a video sequence, downsampling assures that the temporal redundancy across consecutive frames is reduced while still maintaining the facial movement information across frames. The frames are then resized to $m \times n$ and stacked together to create a $m \times n \times t$ dimensional input to the ConvLSTM-based network. The values m, n, and t are set according to the resolution and fps of the camera used. In the case of the SAMM dataset, the values m, n, and t are set to 90, 90, and 9, respectively.

3.2 ConvLSTM

The ConvLSTM proposed in [24] improves the capability of LSTM by capturing spatiotemporal correlations. LSTM, which forms the basis for the ConvLSTM, is an RNN structure capable of capturing long-term dependencies in sequential data. It achieves this by introducing a memory cell C_t within the RNN. The information within the memory cell is regulated using four special neural network layers called gates. They interact with each other to decide whether the memory cell will be accessed, written or cleared during the processing of an input X_t .The memory cell C_t maintains the long-term dependencies observed within the input X_t , thus addressing the vanishing gradient problem.
A limitation of the LSTM is that it creates too much redundancy when operating on video sequences. The ConvLSTM deals with this issue by arranging the input as 3D tensors and processing it through convolution. The steps involved are as follows:
Given an input X with dimensions $m \times n \times t$, the gate i_t, decides whether any new information will be added to the memory cell C_t according to equation (1). Where H_t is the hidden state, $*$ is the convolution operator and $\circ$ is the Hadamard product.

$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \tag{1}$$

The decision to clear the past content on the memory cell C_{t-1} is made by the gate f_t following equation (2). While, the actual update to the memory cell C_t is performed according to equation (3).

$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \tag{2}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right) \tag{3}$$

The gate o_t decides whether the content of C_t is accessible to the hidden state/output H_t according to equations (4) and (5).

$$o_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o\right) \tag{4}$$

$$H_t = o_t \circ \tanh\left(C_t\right) \tag{5}$$

Thus, in the context of micro-expression recognition, the ConvLSTM can capture facial movement across frames using the convolution kernels. Since the frames maintain their 2D structure, the network can capture better spatiotemporal correlation.
In this paper, to capture the spatiotemporal correlation across frames, we use a single layer of ConvLSTM. The number of kernels is set to 32, with a kernel size of 3×3 and strides of 1×1. The output from the ConvLSTM is downsampled using max-pooling with a kernal size of 2×2 and stride of 1×1.

3.3 Dense layers

The output of the max-pooling layer is flattened and passed on to the dense layers to perform recognition. In this proposal, we use two dense layers as the final two layers of the network. The first layer consists of 48 neurons with sigmoid activation function. Neurons in the second layer are set according to the number of micro-expressions to be recognised by the network. Being the final layer, it has a softmax activation function. We also introduce a dropout of 30% between the two layers.

4 Evaluation Result

As discussed in section 2, the state-of-the-art dataset available for evaluating the proposal is the SAMM dataset. The dataset contains a total of 159 video sequences. Each of those sequences belongs to one of the seven micro-expressions or other class - see Fig. 2. However, the total number of sequences belonging to a micro-expression class is not equal. The disparity is quite significant, with the anger class having 57 sequences while the sad class has just 6. To address this issue, the works discussed in section 2 have introduced several evaluation protocols. We adopt a combination of protocols presented in [20], [21] and [22] to perform our evaluations.

4.1 Evaluation protocol

Following the protocol in [20] and [21], we evaluate our proposal using two subsets of the SAMM dataset containing 3 and 5 classes, respectively. The evaluation is carried out using stratified k-fold cross-validation, where the value of k is set to 5. This protocol is an upgrade over the 80-20 split employed in [22]. We ensure that

the five groups created by the k-fold cross-validation are mutually exclusive with respect to the participants. Also, the imbalanced class distribution is maintained in each group. The k-fold cross-validation is performed by iteratively selecting one group for validation while the remaining four groups are used for training. Once the validation is completed, the trained weights are discarded, and the process is repeated using the new groups. The performance is reported in terms of the accuracy and the F1 score. The F1 score is the harmonic mean of precision and recall, which is quite insightful when working with an unbalanced dataset.

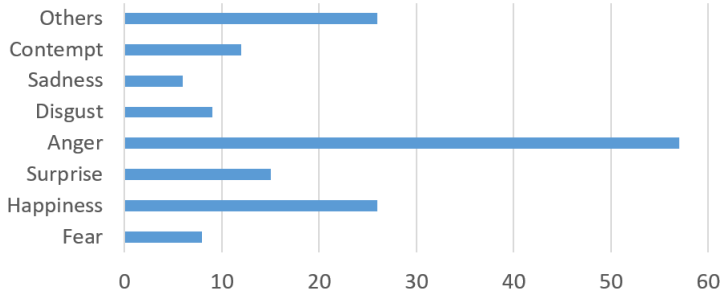

Fig. 2. Statistics of the SAMM dataset.

The problem of data imbalance is also partly addressed using data augmentation. We introduce a horizontal flip, change in brightness and salt and pepper noise in classes with fewer sequences, as illustrated in Fig. 3. The choice of introducing the augmentation is random, which prevents saturating a class with augmented data. We then train the model using the ADAM optimiser with a learning rate of 10^{-4}. The learning rate is set to reduce on plateau with a factor of 0.5. The loss is set to categorical cross-entropy loss, the epochs is set to 50 with early stopping, and the batch size is set to 16.

4.2 Result and discussion

Table 1 reports the result of the three class micro-expression recognition. The three classes are positive (happiness), negative (anger + contempt + disgust + fear + sadness), and surprise, containing 26, 92 and 15 sequences, respectively. Table 1 also compares the results of the proposal with the state-of-the-art. Similarly, Table 2 reports the results of micro-expression recognition across five classes. The five classes are happiness, other, anger, contempt, surprise, containing 26, 26, 57, 12 and 15 sequences, respectively – see Fig. 2.

The results in Table 1 report an improvement of 10.15% and 24.69% in accuracy and F1 score, respectively, over the state-of-the-art for the three class problem. The improvement in the five class problem is also significant, with a 10.29% and 15.25% increase in the accuracy and F1 score, respectively – see Table 2. The

Fig. 3. Example of augmentation introduced in the data.

Table 1. Accuracy and F1 score of three class micro-expression recognition on SAMM dataset.

Method	Accuracy	F1 Score
Weighted optical flow [16]	58.30	39.70
Shallow CNN [21]	68.18	54.23
Proposed ConvLSTM	**78.33**	**78.92**

Table 2. Accuracy and F1 score of five class micro-expression recognition on SAMM dataset.

Method	Accuracy	F1 Score
LBP-TOP [11]	39.68	35.89
Truncated AlexNet [20]	57.35	46.44
CNN+LSTM [23]	50.00	52.44
Proposed ConvLSTM	**67.64**	**67.69**

difference in performance can be attributed to the fact that most state-of-the-art methods emphasise capturing spatial information. The features across frames are aggregated using simple concatenation techniques. The lack of good spatiotemporal features leads to poor micro-expression recognition in [16], [21], [11] and [20]. Some state-of-the-art methods that consider temporal features, such as [23], employ a two-step approach of using CNNs followed by the LSTM, which improves its F1 score. However, our proposal beats the use of CNN+LSTM with an improvement in accuracy and F1 score of 67.64% and 15.25%, respectively. These results support our hypothesis that capturing spatiotemporal information within a video sequence using ConvLSTM is significantly more effective than using shallow CNNs, deep CNN+LSTM or hand-crafted techniques.
Some concerns about the results can arise due to the data imbalance in the dataset. Hence, to obtain better insights, we report the confusion matrix for 5 class micro-expression recognition results in Table 3. The results suggests that all classes contribute approximately evenly to the proposed network's accuracy. In Table 3 surprise class has the highest accuracy of 75.93%, while the lowest accuracy is reported for the happiness class at 60.94%. This result is in sharp contrast to the Truncated AlexNet [20], where the anger class has an accuracy of 90%, while the happiness, contempt, surprise and other classes have an accuracy of 35%, 42%, 33% and 35%, respectively. Here, the accuracy of the state-of-the-art network largely depends on the correct recognition of the class with the highest number of sequences. The reasonably uniform accuracies reported in Table 3 suggest that the proposal handles the data imbalance significantly better than the state-of-the-art. The uniformity in the accuracy is also present in the three class micro-expression recognition problem, where the negative class containing 92 sequences and surprise class containing 15 sequences report an accuracy of 79.41% and 78.85%, respectively- see Table 4.

Table 3. Confusion matrix for the five class micro-expression recognition problem

		Prediction				
	Class	Happiness	Contempt	Anger	Other	Surprise
Ground Truth	Happiness	**60.94**	7.81	14.06	4.69	12.50
	Contempt	9.21	**64.48**	13.16	5.26	7.89
	Anger	8.70	2.17	**65.22**	17.39	6.52
	Other	3.34	3.34	10.00	**71.67**	11.67
	Surprise	9.26	5.56	9.26	0.00	**75.93**

Finally, we would like to comment on the other class. While anger and happiness are micro-expressions, other contain sequences that do not belong to any of the seven micro-expressions. Thus, a probable cause for the low accuracy in Table 3 can be the absence of similar features among the other class sequences. Therefore, we conducted another experiment using only the seven micro-expression

Table 4. Confusion matrix for the three class micro-expression recognition problem

		Prediction		
	Class	Positive	Negative	Surprise
Ground Truth	Positive	**76.74**	19.77	3.49
	Negative	13.24	**79.41**	7.36
	Surprise	2.88	18.27	**78.85**

available in the SAMM dataset. Following the protocol discussed in section 4.1, we achieved an accuracy of 72.64%. The increase in accuracy can be seen as support for our claim. However, larger datasets are needed to perform a more in-depth analysis.

5 Conclusion

In this paper, we address the problem of micro-expression recognition using a ConvLSTM-based network. The proposed shallow network can capture spatiotemporal information more effectively than other neural networks, such as CNNs and RNNs. We demonstrate the claim by comparing the proposal with the state-of-the-art comprising of hand-crafted, shallow CNN and fine-tuned deep CNN+LSTM networks. The proposed ConvLSTM-based network beats the state-of-the-art, thus emphasising its significance in capturing spatiotemporal information for micro-expression recognition. The proposed ConvLSTM based network also reports reasonably uniform accuracies across its classes in the face of the unbalanced SAMM dataset. The state-of-the-art performs poorly in such scenarios, as their performance is skewed by correct recognition of the class with the highest number of sequences.
The biggest challenge we still face is the absence of a large balanced dataset. Capturing a large dataset that allows us to explore deep ConvLSTM based networks can be a possible future direction. Since the downsampling of the input video sequences is a part of our pre-processing, the proposal performs micro-expression recognition with fewer frames than the state-of-the-art. Thus, its evaluation using low fps videos will also be a part of our future work.

References

1. Frank, S., Elena, G., M.: Empathy, emotion dysregulation, and enhanced microexpression recognition ability. Motivation and Emotion **40** (2016) 309–320
2. Pan, H., Lun, X., Zhiliang, W., Bin, L., Minghao, Y., Jianhua, T.: Review of micro-expression spotting and recognition in video sequences (2021)
3. Yan, W.J., Qi, W., Jing, L., Yu-Hsin, C., Xiaolan, F.: How fast are the leaked facial expressions: The duration of micro-expressions. Journal of Nonverbal Behavior **37** (2013) 217–230

4. Polikovsky, S., Yoshinari, K., Yuichi, O.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. International Conference on Imaging for Crime Detection and Prevention. IET (2009
5. Li, X., Tomas, P., Xiaohua, H., Guoying, Z., Matti, P.: A spontaneous micro-expression database: Inducement, collection and baseline. 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (2013)
6. Yan, W.J., Qi, W., Yong-Jin, L., Su-Jing, W., Xiaolan, F.: Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) (2013)
7. Yap, C.H., Connah, K., Moi, H.Y.: Samm long videos: A spontaneous facial micro- and macro-expressions database. 15th IEEE International Conference on Automatic Face and Gesture Recognition (2020)
8. Moilanen, A., Guoying, Z., Matti, P.: Spotting rapid facial movements from videos using appearance-based feature difference analysis. 22nd international conference on pattern recognition (2014)
9. Davison, A.K., Moi, H.Y., Cliff, L.: Micro-facial movement detection using individualised baselines and histogram-based descriptors. IEEE international conference on systems, man, and cybernetics (2015)
10. Li, J., Catherine, S., Renaud, S.: Local temporal pattern and data augmentation for micro-expression spotting. IEEE Transactions on Affective Computing (2020)
11. Zhao, G., Matti, P.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence **29** (2007) 915–928
12. Wang, Y., John, S., Raphael, C.W.P., Yee-Hui, O.: Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. Asian conference on computer vision (2014)
13. Huang, X., Guoying, Z., Xiaopeng, H., Wenming, Z., Matti, P.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing **175** (2016) 564–578
14. Xu, F., Junping, Z., James, Z.W.: Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing **8** (2017) 254–267
15. Liong, S.T., Raphael, C.W.P., John, S., Yee-Hui, O., , KokSheik, W.: Optical strain based recognition of subtle emotions. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) (2014)
16. Liong, S.T., John, S., KokSheik, W., Raphael, C.W.P.: Less is more: Micro- expression recognition from video using apex frame. Signal Processing: Image Communication **62** (2018) 82–92
17. Takalkar, M.A., Min, X.: Image based facial micro-expression recognition using deep learning on small datasets. International conference on digital image computing: techniques and applications (DICTA) (2017)
18. Li, Q., Jun, Y., Toru, K., Shu, Z.: Micro-expression analysis by fusing deep convolutional neural network and optical flow. 5th International Conference on Control, Decision and Information Technologies (CoDIT) (2018)
19. Wang, C., Min, P., Tao, B., Tong, C.: Micro-attention for micro-expression recognition. Neurocomputing **410** (2020) 354–362
20. Khor, H.Q., John, S., Sze-Teng, L., Raphael, C.P., Weiyao, L.: Dual-stream shallow networks for facial micro-expression recognition. IEEE international conference on image processing (ICIP) (2019)

21. Gan, Y.S., Sze-Teng, L., Wei-Chuen, Y., Yen-Chang, H., Lit-Ken, T.: Off-apexnet on micro-expression recognition system. Signal Processing: Image Communication **74** (2019) 129–139
22. Reddy, S.P.T., Surya, T.K., Shiv, R.D., Snehasis, M.: Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. International Joint Conference on Neural Networks (IJCNN) (2019)
23. Khor, H.Q., John, S., Raphael, C.W.P., Weiyao, L.: Enriched long-term recurrent convolutional network for facial micro-expression recognition. 13th IEEE International Conference on Automatic Face Gesture Recognition (2018)
24. Shi, X., Zhourong, C., Hao, W., Dit-Yan, Y., Wai-Kin, W., Wang-chun, W.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28 (2015)