Shallow convolution and attention-based models for micro-expression recognition

Prateek Upadhya and Tanmay Tulsidas Verlekar

Department of CSIS, BITS Pilani
K.K. Birla Goa Campus, Goa - India, 403726

Abstract. The use of deep learning models for micro-expression recognition is challenging because of the absence of large datasets. This paper proposes the construction of shallow models to address this problem. It explores block-wise 3D convolutions, 2D convolutions over frames and 2D convolutional long short-term memory (ConvLSTM) over the video and combines them with multi-headed self-attention to obtain three different models. The evaluation indicates that the proposed 2D ConvLSTM and attention-based model performs the best, beating the state-of-the-art by obtaining an accuracy of 74%. It also has a parameter size that is 10 times smaller than the state-of-the-art.

1 Introduction

Facial expressions voluntarily and involuntarily communicate human emotion. The voluntary part is called macro-expressions and lasts between 0.5 and 4 seconds. The involuntary reactions are called micro-expressions. They are subtle and occur within a fraction of a second [7]. Due to their involuntary nature, they can be considered genuine emotional responses of an individual. Thus, detecting and recognizing micro-expression is of great interest to the research community, with applications ranging from criminal investigation to psychological and medical assessment.

Research in video-based micro-expression recognition began with the use of handcrafted features such as local binary patterns (LBP) and classifiers, such as support vector machines (SVM) [8]. The LBP captured only the local context in a single image. The variations in the face across time were also observed to be unique for each micro-expression. Thus, global movement across time was captured using optical flow and was used as a feature to perform micro-expression recognition in [10]. Recently, deep learning models have also been explored for micro-expression recognition. In most cases, CNNs such as AlexNet and Resnet are pre-trained on macro-expressions and fine-tuned for micro-expression recognition [1]. This approach has become popular because of the small sample size of most publicly available datasets for micro-expression recognition. Different convolutional structures, such as 3D convolutions, have also been used to capture spatiotemporal information across frames for micro-expression recognition. They are used in a dual-stream setup where each stream focuses on a specific region of the face [4]. While most models focus on CNN to capture spatial information, CNNs can also combined with recurrent neural networks (RNN) to capture the spatiotemporal context [3]. The spatiotemporal context can also be

captured in a single step using a convolutional LSTM [6]. Recently, transformer structures have also been explored for micro-expression recognition [9]. These models combine all the publically available datasets to create a composite dataset for fine-tuning. The need for composite datasets arises from the fact that most micro-expression recognition datasets have a small sample size. Among the publically available video datasets, the CAS(ME)2, SAMM, and MMEW datasets contain 57, 159 and 300 video samples, respectively [1]. Acquiring large datasets can be challenging as even trained professionals equipped with high frame rate cameras find it difficult to annotate the ground truth [2].

Thus, this paper focuses on constructing shallow neural networks with a few million trainable parameters for micro-expression recognition. These models can be trained on datasets with small sample sizes without any pre-training. Different convolutional structures are explored to capture features corresponding to a micro-expression successfully. It is motivated by the idea that special convolutional structures, such as 3D convolution, 2D convolution aggregated across time and 2D ConvLSTM, can effectively capture spatiotemporal context for a micro-expression from an input video. It then improves the global spatiotemporal context using multi-headed self-attention. Evaluation of these models indicates that the proposed models are equivalent to or better than the state-of-the-art models that pre-train on macro-expressions. The best among the three, the 2DConvLSTM and attention-based model, achieves this while having a parameter size that is 10 times smaller than the state-of-the-art.

2 Proposal

The paper follows a pre-processing pipeline to similar [3], where optical flow maps are used to detect the onset, apex, and offset frames. The onset frame indicates the beginning of micro-expressions, the apex represents the peak manifestation, and the offset frame represents the end. It is followed by a uniform sampling of five frames between the onset and apex and five from apex to offset, respectively. The undersampling process results in 13 frames, which are resized to 90×90 pixels and arranged sequentially to form the input for the proposed micro-expression recognition models. The three models proposed in this paper are illustrated in Fig. 1. All three proposed models use multi-headed self-attention with four attention heads and a parameter size of 64. The basic structure of the attention layer remains the same across the three models. Given input feature vectors, each attention head first creates three distinct representations called the query q, key k, and value v through learnt weights. The attention vectors can then be computed following $softmax\left(\frac{qk^t}{\sqrt{d}}\right)v$, where $\sqrt{d}$ is the scaling term. The final two layers of the three models are also the same. The first fully connected layer among the two uses 128 parameters and the ReLU activation function. The second layer's parameters are set according to the number of micro-expressions considered for recognition, with a softMax activation function. The dropout of 0.2 is set between the two layers.

The first model divides an input frame into a 3×3 grid. The frames are

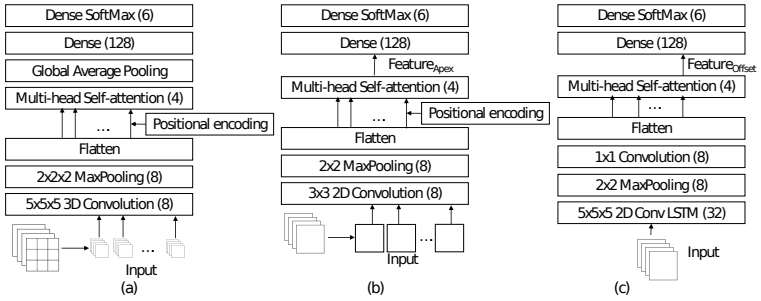

Fig. 1: Architecture of the proposed models: a) 3D convolution and attention, b) 2D convolution and attention and c) 2D ConvLSTM and attention models.

stacked together to generate nine $30 \times 30 \times 13$ blocks. Each block captures the change in a local region, such as around the eye or nose, across time. The blocks are then processed using 8 $5 \times 5 \times 5$ 3D convolution kernels. The size of convolution kernels is increased to $5 \times 5 \times 5$ from the typically used $3 \times 3 \times 3$ to effectively capture shape-related information available in the blocks, which is empirically verified. The 3D convolutions result in features that capture block-wise spatiotemporal information. The 3D Convolution is followed by a $2 \times 2 \times 2$ max-pooling. The output from max-pooling is flattened into feature vectors. Feature vectors corresponding to each block are used as inputs to the multi-headed self-attention. It combines each block's localised features to create feature vectors containing global spatiotemporal context with respect to each block. Since the multi-headed self-attention can't keep track of the sequence of the patches, positional embedding is added to the feature vectors. The output from the multi-headed self-attention is turned into a single feature vector using global average pooling and is passed on to the fully connected layers.

The second model applies 2D convolution to each frame. The layer considers eight convolution kernels of size 3×3. The convolutional layer captures the essential low-level spatial context for each frame. The convolution layer is followed by 2×2 max pooling. Similar to the first model, the output is flattened into a feature vector, and positional embedding is added to it. The resulting features are passed as input to the multi-headed self-attention with the same number of parameters as the first model. The multi-headed self-attention combines the spatial context of each frame to create feature vectors containing global spatiotemporal context across frames. Among the frames, the apex frame contains features when the expression is in its peak manifestation. Hence, the feature vector corresponding to the apex frame is passed onto the subsequent fully connected layers. The decision is empirically validated by observing the effects of onset, apex and offset feature vectors on accuracy.

The final model considers 2D ConvLSTM. It is a special type of recurrent neural network [5] that operates directly on 2D data. The overall structure of the

Table 1: Comparison between the proposed models and the state-of-the-art

Models	Accuracy (%)	Trainable Parameters
LBP-TOP [8]	39	-
MDMO [10]	66	-
ELRCN [3]	42	134 million
DTSCNN [4]	66	63 million
TLCNN [1]	69	62 million
Proposed 3D Convolution and attention	64	28 million
Proposed 2D Convolution and attention	70	55 million
Proposed 2D ConvLSTM and attention	74	6 million

2D ConvLSTM network is similar to that of the LSTM. However, 2D ConvLSTM processes its input using 2D convolution instead of vector multiplications. The proposed model sets 32 3×3 convolutions in the 2D ConvLSTM layer followed by max-pooling. To reduce the dimension of the output, it uses 8 1×1 convolution, which is followed by flattening the features. Since the 2D ConvLSTM processes the input sequentially, its output can be processed by the multi-headed self-attention without any positional embedding. Because of the sequential nature of 2D ConvLSTM, spatiotemporal context improves gradually in feature vectors from onset to the offset frame. The multi-headed self-attention further improves the spatiotemporal context of these feature vectors. The feature vector corresponding to the offset frame is then passed to the fully connected layers.

3 Evaluation

The paper evaluates the models using the MMEW dataset [1]. It contains videos of 36 participants registering seven different micro-expressions captured at 90 frames per second. They include Happiness (36), Anger (8), Surprise (89), Disgust (72), Fear (16), Sadness (13), and Others (66). Adhering to the evaluation protocol presented in [1], The proposed models are evaluated using happiness, surprise, disgust, fear, and sadness micro-expressions. The training and the validation set are kept mutually exclusive with respect to the participants, and a five-fold cross-validation is performed to obtain the results reported in Table 1. To address the issue of data imbalance, oversampling is performed using data augmentations. The samples are augmented with horizontal flip, change in brightness and salt and pepper noise. These augmentations are randomly introduced in the under-represented micro-expression. The three proposed models are trained using the ADAM optimiser with an adaptive learning rate initialised at 10^{-3} and a reduction factor of 0.5. The loss is set to categorical cross-entropy. The epochs are set to 100 with early stopping, and the batch size is 32.

Among the state-of-the-art models reported in Table 1, the best-performing

TLCNN [1] has an accuracy of 69%. It contains five 3×3 convolutional layers and 3 fully connected layers, resulting in around 62 million learnable parameters. Unable to be trained using only 300 micro-expression samples, the model is pre-trained on the macro-expressions, followed by fine-tuning on the micro-expressions. The work presented in [1] discusses the poor performance of this strategy on the underrepresented micro-expressions.Table 2 reports the confusion matrix where it can be seen that TLCNN recognises disgust and surprise with an accuracy of 100% but fails to recognise Sadness and Fear with an accuracy of 0%. ELRCN [3], a combination of VGG-16 and LSTM, is also pre-trained on macro-expressions. But even with LSTM to capture spatiotemporal context, it achieves an accuracy of 42%.

The proposed models perform equivalent to or better than the state-of-the-art without any pre-training. The best-performing 2D ConvLSTM and attention-based model achieve an accuracy of 74% while reducing the parameter size by a factor of 10 when compared to TLCNN [1]. The accuracy of the models can be attributed to the use of 2D convolution, 3D convolution, or 2D ConvLSTM layers, followed by multi-headed self-attention, along with oversampling of the training dataset through data augmentation. It enables the models to capture the local context of movement around landmarks such as lips, nose, and eyes, as well as the global context of association between landmarks and the changes in them across time.

The 3D convolution and attention-based model achieves an accuracy of 64%. It supports the idea discussed in [4] that creating localised blocks containing facial landmarks is an effective strategy for performing micro-expression recognition. However, without improving the precision of the blocks to capture a group of landmarks, the resulting discontinuity can affect the performance of the models. The second model captures spatial features using 2D CNN from a frame, a strategy adopted by the most state-of-the-art models. However, while the state-of-the-art models combine these features using RNN, the use of multi-headed self-attention allows the model to achieve an accuracy of 70%.

The final model uses 2D ConvLSTM to capture spatiotemporal features present in the video sequence. It allows the model to operate on images directly without splitting the process into spatial and temporal processing, as done in the ELRCN [3]. The multi-headed self-attention further improves the spatiotemporal features, resulting in an accuracy of 74%. The oversampling through data augmentation allows the models to perform better than the state-of-the-art on underrepresented micro-expressions, as reported in Table 2. The proposed 2D ConvLSTM and attention-based model recognises Surprise with an accuracy of 93% while performing the worst on Fear with an accuracy of 50%.

4 Conclusions and Future Work

This paper explores shallow models to perform micro-expression recognition. It focuses on 3D convolution, which captures spatiotemporal information within a localised block, 2D convolution, which captures spatial information within

Table 2: Confusion matrix comparison

Ground Truth		2D ConvLSTM and Attn model					TLCNN [1]				
		Happiness	Disgust	Fear	Surprise	Sadness	Happisess	Disgust	Fear	Surprise	Sadness
	Happiness	**87**	7	0	6	0	**55**	45	0	0	0
	Disgust	5	**65**	0	30	0	0	**100**	0	0	0
	Fear	11	20	**50**	19	0	0	60	**0**	40	0
	Surprise	0	7	0	**93**	0	0	0	0	**100**	0
	Saadness	6	20	0	0	**74**	0	100	0	0	**0**

frames, and 2D ConvLSTM, which captures spatiotemporal information across frames. The spatiotemporal context of these features is further improved using multi-headed self-attention. Among the models, the 2D ConvLSTM and attention model achieve the best accuracy of 74% while having a parameter size of just 6 million. The results appear promising but are reported using a single dataset. Future work will consider composite datasets, which will allow an increase in the depth of the proposed models.

References

[1] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5826–5846, 2021.

[2] S. Nag, A. K. Bhunia, A. Konwer, and P. P. Roy. Facial micro-expression spotting and recognition using time contrasted feature with visual memory. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2022–2026. IEEE, 2019.

[3] M. Peng, C. Wang, T. Bi, Y. Shi, X. Zhou, and T. Chen. A novel apex-time network for cross-dataset micro-expression recognition. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 1–6. IEEE, 2019.

[4] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[5] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[6] S. Shukla, P. K. Rai, and T. T. Verlekar. Micro-expression recognition using a shallow convlstm-based network. In *Proceedings of the Asian Conference on Computer Vision*, pages 17–28, 2022.

[7] E. Svetieva and M. G. Frank. Empathy, emotion dysregulation, and enhanced microexpression recognition ability. *Motivation and Emotion*, 40:309–320, 2016.

[8] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12*, pages 525–537. Springer, 2015.

[9] Z. Wang, K. Zhang, W. Luo, and R. Sankaranarayana. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024.

[10] F. Xu, J. Zhang, and J. Z. Wang. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*, 8(2):254–267, 2017.