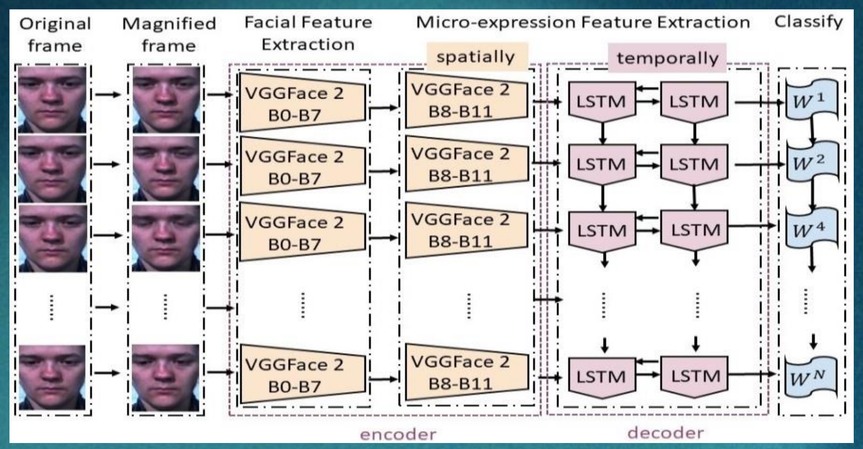
Poster number: 3078

MICRO-EXPRESSION RECOGNITION BASED ON VIDEO MOTION MAGNIFICATION AND PRE-TRAINED NEURAL NETWORK

Mengjiong Bai, Roland Goecke, Damith Herath
Human-Centred Technology, Faculty of Science and Technology, University of Canberra

Introduction

This paper investigates the effects of using **video motion magnification methods** based on amplitude and phase, respectively, to amplify small facial movements. We hypothesise that this approach will assist in the micro-expression recognition task. To this end, we apply the **pre-trained VGGFace2 model** with its excellent facial feature capturing ability to transfer learn the magnified micro-expression movement, then encode the spatial information and decode the spatial and temporal information **by Bi-LSTM model**.



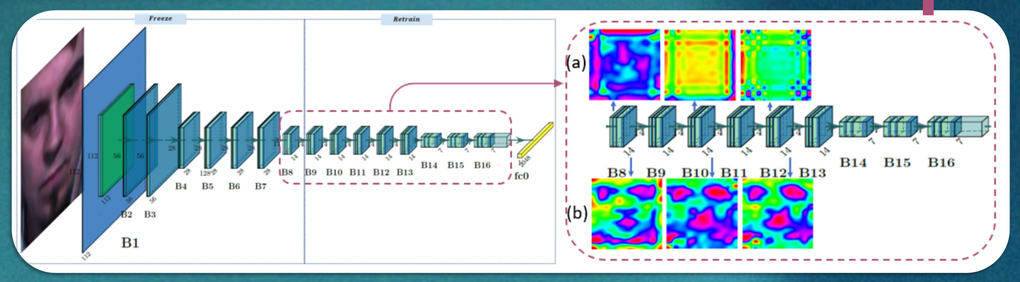
Methods

Motion Magnification Methods

$I(x, t)$ → an image profile at location x and time t
$\delta(t)$ → translation motion of the image.
$I(x, t) = f(x+\delta(t))$ while $I(x, t) = f(x)$
α → *amplification factor*
$\hat{I}(x, t) = f((x) + (1+\alpha)\,\delta(t))$ final goal is synthesising the signal based on the α[7]

√ Amplitude-based Motion Magnification(AMM) uses first-order Taylor expansion to approximate the image

√ Phase-based Motion Magnification(PMM) uses the Fourier series decomposition to approximate the motion field.



Encoding and Decoding Methods

√ The VGGFace2 model is a pre-trained network based on the ResNet-50 model and the VGGFace2 database, using the VGGFace2 model to extract the facial features. [13]
√ Employ a Bi-LSTM network to capture the micro-expression features from sequential samples. [16]

Experiments

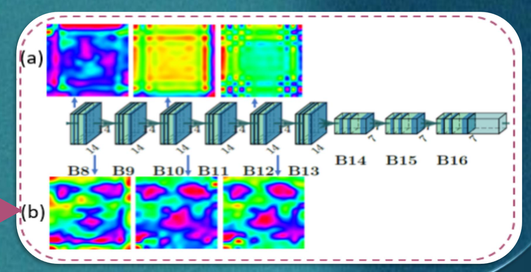
Dataset:
The spontaneous micro-expression corpus (SMIC) contains High Speed (HS) micro-expression samples recorded at 100fps.
√ shortest recorded expression was about 0.11s (= 11 frames) . √ Average expression length was about 0.30s (= 29 frames).
√ In total, 164 videos in the SMIC HS, 70 for negative, 51 for positive and 43 for surprise [17].

Data pre-process:
Extracting the fixed-length video segments {So,S1, …, Sn} from original video V={fo, f1, …, fn} by overlapping sliding window on both continuous frames and extracts frames at even intervals; the size of intervals is based on the length of the original video, the number of the frames in each training sequence (training sample size) $N = Fn$ (the size of the based on the sliding window).

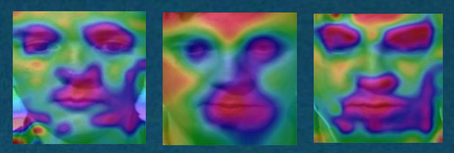
	N	7	8	9	10	11	12	13	14
AMM	**Average**	**0.67**	**0.67**	**0.74**	**0.68**	**0.68**	**0.64**	**0.62**	**0.73**
	Negative	0.54	0.70	0.61	0.57	0.73	0.63	0.64	0.72
	Positive	0.70	0.70	0.78	0.62	0.75	0.71	0.50	0.80
	Surprise	0.80	0.42	0.56	0.50	0.60	0.60	0.63	0.70
PMM	**Average**	**0.70**	**0.70**	**0.76**	**0.71**	**0.73**	**0.70**	**0.67**	**0.79**
	Negative	0.67	0.71	0.65	0.75	0.68	0.63	0.65	0.74
	Positive	0.75	0.71	0.75	0.67	0.86	0.60	0.80	0.75
	Surprise	0.73	0.67	0.67	0.71	0.71	0.67	0.64	1.00

As training sample size *N* leads to a loss of long-term dependence information, while large *N* reduces the number of sequences, we evaluated several combinations to find a suitable value for N.
As shown in the Table, the best accuracy by both amplitude and phase-based methods was achieved for N=14. We hypothesise that as 14 is around half of the average video duration, it is easier to capture the process of micro-expressions.

In order to locate the retraining parts , Grad-CAM technique couldvisualise the activation maps of each residual block before training,. We observed that the block 8 with activation regions is closest to the facial features. Hence, we froze the blocks before block 9–B9, while retraining blocks {B9 -B16} on the micro-expression database.



Conclusion



To visually explain the operating mechanism of the spatio-temporal network, we applied Grad-CAM to map the model. It highlighted activation regions from different face regions correspond to different micro-expressions. For *negative*, it focuses more on the eyes; for *positive*, on the mouth corners; and for *surprise*, more on the eyebrows.

To test individual contributions of each component within the framework:

√ Only the VGGFace2 pre-trained model with a fully-connected layer for classification, it achieved an accuracy of **41.9%**.

√ Replacing the VGGFace2 model with a ResNet-50 model pre-trained on ImageNet, it achieved an average accuracy of **37.9%**.

√ Comparing proposed experiments results to baseline experiments results, a performance improved from 60.60% to 75.76%. It infer that the Eulerian motion magnification exaggerated the micro-expression deformation at spatial and temporal level indeed increased the recognition rate.

Source paper	Method	Year	Accuracy
[19]	3D-FCNN	2018	55.49%
[20]	Hierarchical scheme	2018	55.49%
[21]	MicroExpSTCNN	2019	68.75%
[10]	EVM+TIM	2019	68.90%
[18](baseline)	VGGFace2+Bi-LSTM	2020	60.60%
[22]	SETFNet	2020	70.25%
Proposed	EVM+VGGFace2+Bi-LSTM	2021	75.76%

In comparison to the state-of-the-art, the proposed framework achieved a significantly improved average accuracy of 75.76%.

[7] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," ACM Trans. Graphics, vol. 31, no. 4, pp. 1–8, Jul 2012.
[10] W. Peng, X. Hong, Y. Xu, and G. Zhao, "A Boost in Revealing Subtle Facial Expressions: A Consolidated Eulerian Framework," in 2019 14th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2019
[17]T. Pfister, X. Li, G. Zhao, and M. Pietik¨ainen, "Recognising spontaneous facial micro-expressions," IEEE International Conference on Computer Vision, p.1449–1456, 6 2011
[16] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in 2013 IEEE Int Conference on Acoustics, Speech and Signal Processing, May 2013, pp. 6645–6649.
[13] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in 2018 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018). May 2018, IEEE.).
[22] L. Yao, X. Xiao, R. Cao, F. Chen, and T. Chen, "Three stream 3d cnn with se block for micro-expression recognition," in 2020 International Conference on Computer Engineering and Application (ICCEA), 2020, pp. 439–443.
[20] S. P. Teja Reddy, S. Teja 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), 2019, pp. 1–8.
[19] Y. Zong and X. Huang, "Learning from Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition," IEEE Trans. Multimedia, pp. 1520–9210, 2018.
[18] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," Pattern Analysis and Applications, vol. 22, no. 4, pp. 1331–1339, Nov 2018.
[21] M. Bai and R. Goecke, "Investigating LSTM for Micro-Expression Recognition," in 2020 Int. Conf. Multimodal Interaction (ICMI). Oct 2020, ACM.