Detection of Micro-expression Recognition Based on Spatio-Temporal Modelling and Spatial Attention

Mengjiong Bai
Keira.Bai@canberra.edu.au
Human-Centred Technology, Faculty of Science and Technology, University of Canberra

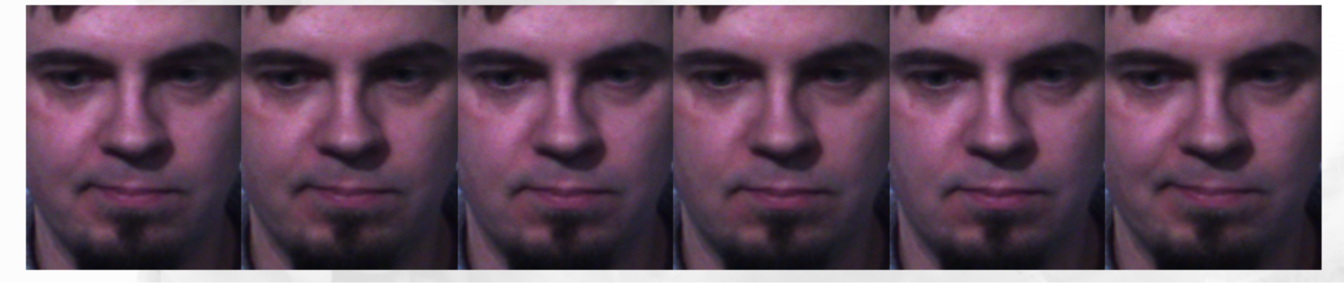

Figure 1: Micro-expression - samples from SMIC dataset

ABSTRACT

My PhD project aims to make contributions in the affective computing application to assist in the depression diagnosis by micro-expression recognition. My motivation is the similarities of the low-intensity facial expressions in micro-expressions and the low-intensity facial expressions ('frozen face') in people with psycho-motor retardation caused by depression. It will focus on, firstly, investigating spatio-temporal modelling and attention systems for micro-expression recognition (MER) and, secondly, exploring the role of micro-expressions in automated depression analysis by improving deep learning architectures to detect low-intensity facial expressions. This work will investigate different deep learning architectures (e.g. Temporal Convolutional Networks (TCNN) or Gate Recurrent Unit (GRU)) and validate the results on publicly available micro-expression benchmark datasets to quantitatively analyse the robustness and accuracy of MER's contribution to improving automatic depression analysis. Moreover, video magnification as a way to enhance small movements will be combined with the deep learning methods to address the low-intensity issues in MER.

CCS CONCEPTS

• **Human-centered computing → Gestural input**.

KEYWORDS

Affective computing; Micro-expression; Deep learning; Depression analysis

ICMI '20, October 25–29, 2020, Virtual event, Netherlands
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3421160

ACM Reference Format:
Mengjiong Bai. 2020. Detection of Micro-expression Recognition Based on Spatio-Temporal Modelling and Spatial Attention. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3382507.3421160

1 INTRODUCTION

Micro-expressions(shows in Figure 1) happen with a short time span – at a length of about 1/5 to 1/25 of a second [21]. The short duration and subtle movements of micro-expression cause the low intensity, they are difficultly captured by human naked eyes [4]. According with the development of artificial intelligence (AI), especially in deep learning techniques, AI techniques have been used to assist in improving the micro-expression detection and classification.

Based on Ekman *et al.ś* case study [3], depression patients show symptoms of self-deception and deception on other people, such as simulating the optimism expression and behavior. Yet micro-expressions can leak the clues about their emotional state to assist in psychological diagnosis with quantifiable evidence because they are involuntary.

Moreover, Psycho-motor retardation is an important symptom of a depressive condition. Psycho-motor retardation can cause a visible slowing of physical and emotional reactions, including speech and affect [15]. Therefore, the study of micro-expressions may assist in improving the techniques of detecting low-intensity facial expressions, which has great potential for making contributions to automated methods of depression analysis.

My research aims to explore how the research results of MER can improve the automated depression clinical diagnosis. With this as a goal, firstly I will carry out the MER task on different spatio-temporal modeling and investigate how the spatial and temporal information influences the recognition results respectively. The technical with strong recognition ability for low-intensity facial expression will be obtained during this stage. Followed by that, improve this technology to make it suitable for the depression diagnosis. Because the symptoms in the people with psycho-motor retardation have the same characteristics as micro-expression, it is feasible to improve the low-intensity facial expression recognition technology for clinical mental diagnosis on the basis of MER. However, due to the currently published spontaneous micro-expression databases all require participants to suppress their expressions during the collection process, the distinction between the micro- and macro- expression should be considered during the techniques transforming process.

2 PRIOR WORK

In the early stages, limited by the scale and availability of micro-expression datasets, MER researches were mainly focused on using handcrafted computer vision features and roughly fall into two categories: describing the skin texture of face by appearance-based methods such as Local Binary Patterns (LBP) algorithm and shaping the facial features by geometric-based algorithms such as optical flow or optical strain pattern. Huang *et al.* [5] proposed methods of obtaining LBP projections on images from both horizontal and vertical perspectives. Others proposed constructing features by concatenating LBP histograms on three orthogonal planes (LBP-TOP) – XY, XT, and YT, respectively – as a better method. For example, Wang *et al.* [17] proposed the LBP-Mean Orthogonal Planes (LBP-MOP) method to improve the efficiency by only computing the LBP results on the three average planes. Meanwhile, Yandan Wang *et al.* [22] proposed the Local Binary Patterns with Six Intersection Points (LBP-SIP) volumetric descriptor, this method used the unique distinct points that lie on the three intersecting lines that cross over the centre point of the three orthogonal planes to compute the spatio-temporal patterns. Shreve *et al.* [16] divided a subject's face into sub-regions to calculate the facial strain for distinguishing micro-expressions and macro-expressions. Meanwhile, Zhang *et al.* [10] proposed the Main Directional Mean Optical-flow (MDMO) feature for micro-expression recognition, which normalises the local statistic motion information and its spatial location based on the region of interest.

Currently, there are four publicly available spontaneous micro-expressions datasets: The spontaneous micro-expression corpus (SMIC) [12] by the University of Oulu, Finland. The Chinese Academy of Sciences Micro-Expression datasets CASME and CASME-II [20]. Samples in these two datasets are annotated with the onset, apex, and offset frames, with action units (AUs) marked and emotions labeled. The Spontaneous Micro-Facial Movement (SAMM) dataset developed by Davison *et al.* [2], which contains the largest amount of different ethnicities, resolutions, and age distribution when compared with any similar micro-expression dataset currently publicly available.

With the publication of more spontaneous micro-expression datasets, deep learning techniques, especially the convolutional neural network (CNN) and recurrent neural network (RNN), were increasingly being applied to MER. Takalkar *et al.* [14] used a method in their research for data augmentation to avoid overfitting caused by insufficient training data. To address the overfitting issue, Miao *et al.* [11] proposed a very simple CNN model with only three layers with the assistance of the improved saliency map, the function of this map is to cropping facial by visualizing the CNN model and saliency pixels from the different regions.

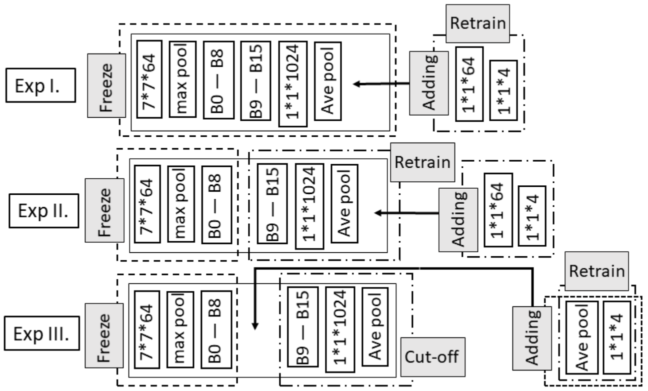

Figure 2: Experiment-I: only training the classifier layer. Experiment-II: frozen the layers before the specific layer. Experiment-III: removing the layers after the specific layer.

Based on these studies, Xia *et al.* [19] proposed a model consisting of several recurrent convolutional layers for capturing the spatial-temporal deformations of a micro-expression sequence. Meanwhile, Kim *et al.* [7] improved the aforementioned method, used the Long Short-Term Memory (LSTM) recurrent neural network to replace the RNN. Khor *et al.* [6] also used the same CNN+LSTM architecture for detecting the spatial and temporal information, but enriched the model on both spatio-temporal dimensions, respectively.

Handcrafted methods were also combined with a deep learning model. Li *et al.* [9] divided the face into 12 regions of interest (ROI) by facial landmarks and then extracted the optical flow features by utilising a deep multi-task convolutional network. Verburg and Menkovski [16] also used optical flow, facial cropping and face alignment techniques in their work.

3 MY WORK TO DATE

3.1 Retraining VGGFace2 Model for MER

Retraining VGGFace2 model is an instance of learning spatial information from micro-expression described above: the scratch of layers was retrained by using the transfer learning technique[**?**] then been implemented on data with frame level. Unlike the traditional transfer learning methods only focus on retraining the classifier layers, Grad-CAM technique [13] enables the process more quantitative with higher efficiency.

The mechanism of the CNN network is using individual units to respond to the stimuli only in a restricted region, then cover the entire visual field by partially overlap. Grad-CAM technique utilizes the nature of CNN to produce a coarse localization map to highlight the overlap regions, with which to determine the activation parts for the facial feature extracting process.

Grad-CAM could be applied on each convolutional layer, to produce heat maps with different highlighted regions. The layer where the highlighted area were closest to the facial features is called specific layer l_s, the transfer learning work would be carried out on layers after l_s, then layers $\{l_0, ..., l_{s-1}\}$ are frozen. Outputs from

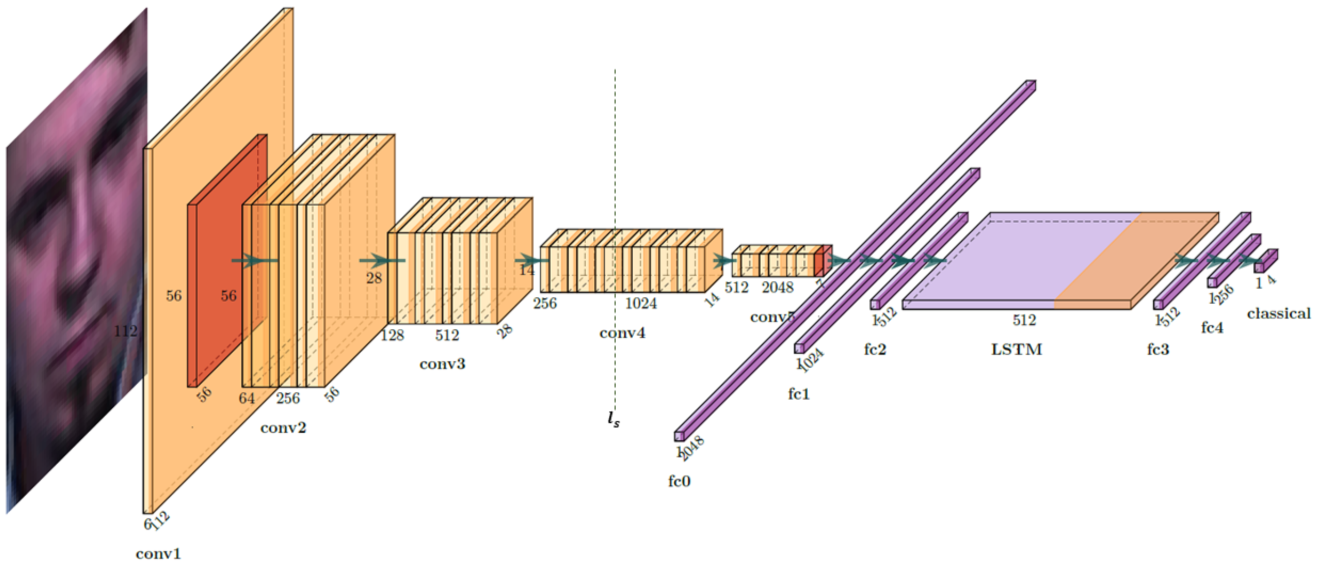

Figure 3: VGGFace2 model for encoding phase while LSTM model for the encoding phase, image sequences are directly fed into the model without complex pre-processing

the learning layers would be fully connected to the classification layers $\{fc_1, fc_2, fc_3\}$ to decrease the feature vector size.

The VGGFace2 model is basically a ResNet-50 CNN trained on the VGGFace2 dataset. In the experiment, part of the network was retrained on the SMIC database (shows in Figure 2). Learning result from this experiment reached 32.6%, which is not a outstanding number. But it provided a more efficient and controllable way to train a deep neural network to recognize micro-expression and uncover how much spatial information alone contributes to the accuracy in a MER task.

3.2 VGGFace2 Model + LSTM on MER

Based on the previous investigation, we expanded the work to an LSTM-RNN architecture (showes in the Figure 3). Here, the VGGFace2 model was combined with uni-directional and bi-directional LSTM to explore the different effects of spatial and temporal information in the MER process.

Inspired by [?], the proposed system abandon manual pre-processing (face cropping and face alignment), which in themselves are potential sources of errors. This experiment was conducted on sequence level and the size of frame sequence N was tested from 11 to 17 to keep the tradeoff between 'enable the network to learn enough critical information' and 'augment more samples to avoid overfitting'.

In this architecture, the pre-trained VGGFace2 was repackaged for the encoding phase, the D-dimension one-hot-vector outputs from this process acted as inputs for the following decoding task. The coding phase was expected to precisely capture the micro-expression-aware region and pass the state for the decoding phase. The decoder consists of LSTM to receive the stack of spatial state from the encoder to analyse the temporal information for the final classify.

The recognition accuracy has been greatly improved (60.06%) when compare with the state-of-art technique and our initially attempt.

3.3 Evaluation results

In addition to locating vggface2 model layers that needed to be retrained, the Grad-CAM technique can also be used to evaluate the final learning performance. Comparing the two heat maps in Figure 6 shows that the highlighted region has changed significantly before and after training (Figure 4 left and right). The highlighted regions gradually gather to the eyes, eyebrows, nose wings, corners of the mouth, etc demonstrate that the activation parts in the network are more focused on the micro-expression-aware areas [19].

The goal of this study was to better understand how much the spatial and temporal information would contribute to the accuracy in an MER task respectively, based on a quantified computing process. The aforementioned evaluation results demonstrate that learning sequential dynamics with an LSTM model will greatly improve the accuracy while recognizing the micro-expression, and this result shows why we need to care about temporal information when detecting and recognizing the micro-expression.

4 MY PLANNED WORK

In my future research, I first plan to investigate different spatio-temporal pattern modelling. Based on previous work, the experiment will be expanded to explore the Temporal Convolutional Networks (TCNN), Gated Recurrent Unit (GRU), and spatial attention mechanism.

4.1 Temporal Convolutional Neural Networks

Temporal Convolutional Neural Networks (TCNN) [8] is a unified approach that uses pooling and upsampling to efficiently capture long-range temporal patterns. It was initially used for action segmentation, but we could integrate the technique into a CNN model for investigating the possibility of effectively obtaining both spatial and temporal information in the MER process, which is hypothesised to lead improved accuracy.

4.2 Attention

In psychology, attention means the cognitive process of selectively concentrating on one or a few things while ignoring others. In machine learning, an attention mechanism is also an attempt to implement the same action of selectively concentrating on a few relevant things [1]. As shown in the Heatmap (Figure 4), the micro-expression frequently happens on the specific facial region(Highlight region) instead of the whole face, therefore, I plan to use the spatial attention model to extract the important parts of the face, while ignoring others in the model training process.

4.3 video Magnification

Eulerian Video Magnification (EVM) [18] is a video magnification method, which could reveal low-intensity temporal variations in videos that are difficult or impossible to see with the naked eye (hidden information) and display them in an enhanced manner. I plan to explore the Eulerian video magnification method and integrate it with the GRU and TCNN deep learning models.

5 TIMELINE

5.1 *Expanding the experiments area from existing point*

July 2020 – June 2021:

- Implementing the existing deep learning fine-tuning methods (e.g. Fast AI) on model training for better training results.
- Expanding the experimental validation to the China Academy of Science Micro-Expression 2 (CASME II) and spontaneous micro-facial movement dataset (SAMM)
- Exploring a useful video magnification technique into deep learning methods to address the low-intensity problems in MER.
- Investigating the utilisation of CNN and GRU for MER.

5.2 *Investigating other spatial-temporal deep learning model*

June 2021-June 2022:

- Investigating the utilisation of spatial attention mechanism for MER.
- Exploring the role of micro-expressions in the automated depression analysis by improving deep learning architectures for the detection of low-intensity facial expression.
- Quantitative analysis of the contribution MER can make to improve the accuracy in automated depression analysis.

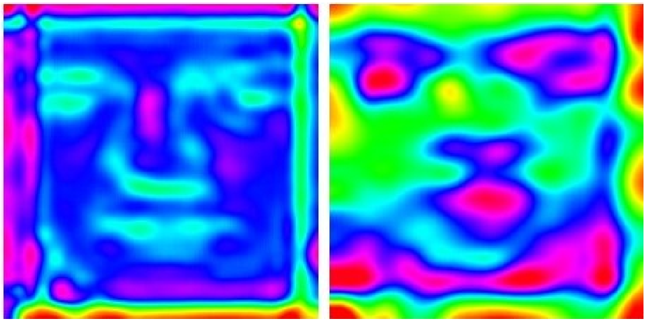

Figure 4: *Left:* heat map before training - less highlighted region and scattered around the edges of the image. *Right:* heat map after training - obvious highlighted region mainly focus on the micro-expression-aware areas

REFERENCES

[1] Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger. [n.d.]. Emotion Recognition with Spatial Attention and Temporal Softmax Pooling. ([n. d.]). https://doi.org/10.1007/978-3-030-27202-9_29 arXiv:http://arxiv.org/abs/1910.01254v2 [cs.LG]

[2] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* 9, 1 (2018), 116–129.

[3] Paul Ekman and Wallace V. Friesen. 1969. Nonverbal Leakage and Clues to Deception. *Psychiatry* 32, 1 (1969), 88–106. https://doi.org/10.1080/00332747.1969.11023575 PMID: 27785970.

[4] Malgorzata. Frank, Mark.Herbasz. 2009. I see how you feel: training laypeople and professionals to recognize fleeting emotions. In *the annual meeting of the International Communication Association*. Annual Meeting of the International Communication Association, New York City, NY.

[5] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. 2015. Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE. https://doi.org/10.1109/iccvw.2015.10

[6] Huai-Qian Khor, John See, Raphael C. W. Phan, and Weiyao Lin. [n.d.]. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition. ([n. d.]). arXiv:http://arxiv.org/abs/1805.08417v1 [cs.CV]

[7] Dae Hoe Kim, Wissam J. Baddar, and Yong Man Ro. 2016. Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations. In *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*. ACM Press. https://doi.org/10.1145/2964284.2967247

[8] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. [n.d.]. Temporal Convolutional Networks for Action Segmentation and Detection. ([n. d.]). arXiv:http://arxiv.org/abs/1611.05267v1 [cs.CV]

[9] Q. Li, S. Zhan, L. Xu, and C. Wu. 2018. Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. *Multimedia Tools and Applications* 78, 20 (Dec. 2018), 29307–29322. https://doi.org/10.1007/s11042-018-6857-9

[10] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu. 2016. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. *IEEE Transactions on Affective Computing* 7, 4 (2016), 299–310.

[11] Si Miao, Haoyu Xu, Zhenqi Han, and Yongxin Zhu. 2019. Recognizing Facial Expressions Using a Shallow Convolutional Neural Network. *IEEE Access* 7 (2019), 78000–78011. https://doi.org/10.1109/access.2019.2921220

[12] T. Pfister, X. Li, G. Zhao, and M. Pietikainen. 2011. Recognising Spontaneous Facial Micro-expressions. *IEEE International Conference on Computer Vision* (6 2011), 1449–1456. https://doi.org/10.1109/ICCV.2011.6126401

[13] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *IEEE Int. Conf. on Computer Vision (ICCV)*.

[14] Madhumita A. Takalkar and Min Xu. 2017. Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. https://doi.org/10.1109/dicta.2017.8227443

[15] Warren W. Tryon. 1991. *Activity Measurement in Psychology and Medicine*. Springer US. https://doi.org/10.1007/978-1-4757-9003-0

[16] Michiel Verburg and Vlado Menkovski. [n.d.]. Micro-expression detection in long videos using optical flow and recurrent neural networks. ([n. d.]).

arXiv:http://arxiv.org/abs/1903.10765v1 [cs.CV]

[17] Yandan Wang, John See, Raphael C.-W. Phan, and Yee-Hui Oh. 2015. Efficient Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition. *PLOS ONE* 10, 5 (may 2015), e0124674. https://doi.org/10.1371/journal.pone.0124674

[18] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics* 31, 4 (jul 2012), 1–8. https://doi.org/10.1145/2185520.2185561

[19] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. 2019. Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-expressions. *IEEE Transactions on Multimedia* (2019), 1–1. https://doi.org/10.1109/tmm.2019.2931351

[20] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. *PLoS ONE* 9, 1 (Jan. 2014), e86041. https://doi.org/10.1371/journal.pone.0086041

[21] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions. *Journal of Nonverbal Behavior* 37, 4 (July 2013), 217–230. https://doi.org/10.1007/s10919-013-0159-8

[22] Wang Y.and See J.and Phan R.CW.and Oh YH. 2015. LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition. *Computer Vision – ACCV 2014* 9003 (2015), 525–537. https://doi.org/10.1007/978-3-319-16865-4_34 Lecture Notes in Computer Science.