Investigating LSTM for Micro-Expression Recognition

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

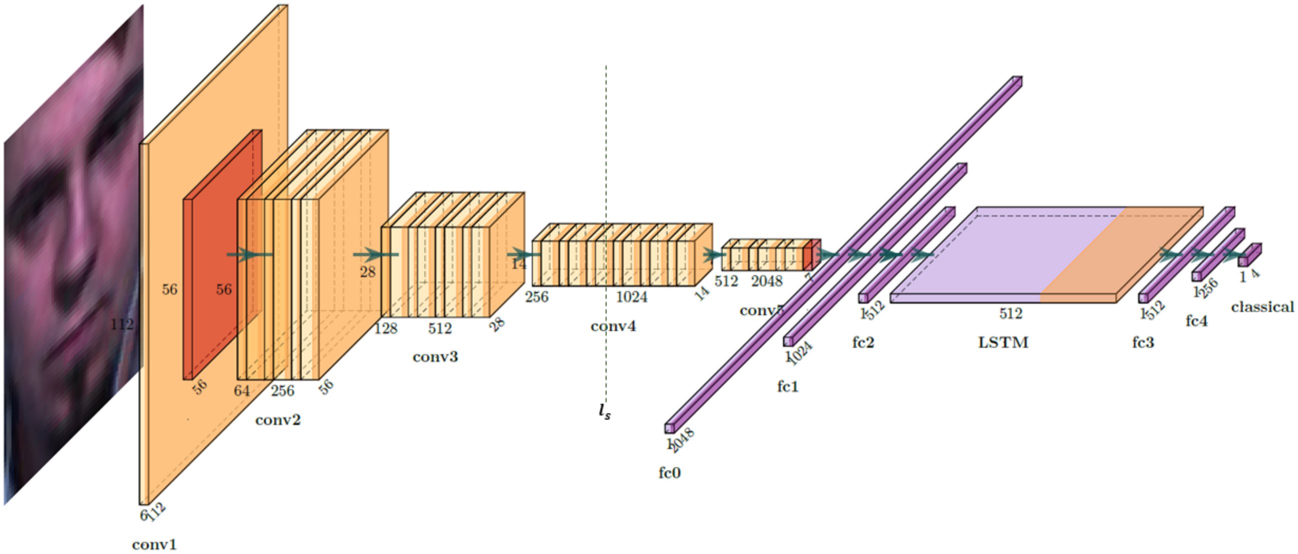


Figure 1: The proposed model structure: combining the VGGFace2 model with uni-directional and bi-diretional LSTM

ABSTRACT

This study investigates the utility of Long Short-Term Memory (LSTM) networks for modelling spatial-temporal patterns for micro-expression recognition (MER). Micro-expressions are involuntary, short facial expressions, often of low intensity. RNNs have attracted a lot of attention in recent years for modelling temporal sequences. The RNN-LSTM combination to be highly effective results in many application areas. The proposed method combines the recent VGFace2 model, basically a ResNet-50 CNN trained on the VGGFace2 dataset, with uni-directional and bi-directional LSTM to explore different ways modelling spatial-temporal facial patterns for MER. The Grad-CAM heat map visualisation is used in the training stages to determine the most appropriate layer of the VGGFace2 model for retraining. Experiments are conducted with pure VGGFace2, VGGFace2 + uni-directional LSTM, and VGGFace2 + Bi-directional LSTM on the SMIC database using 5-fold cross-validation.

CCS CONCEPTS

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

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425248

KEYWORDS

Micro-expression; Deep Learning; Long Short-Term Memory

ACM Reference Format:
Mengjiong Bai, Roland Goecke. 2020. Investigating LSTM for Micro-Expression Recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425248

1 INTRODUCTION

Based on Ekman *et al.*ś case study [4], depression patients show symptoms of self-deception and deception on other people, such as simulating the optimism expression and behavior. Micro-expressions can leak the clues about their emotional state to assist in psychological diagnosis with quantifiable evidence because they are involuntary.

Micro-expressions happen with a short period – at a length of about 1/5 to 1/25 of a second [30] and low intensity, they are difficultly captured by naked human eyes [5].

This paper proposes a novel MER algorithm using Long short-term memory(LSTM) [7] and Convolutional Neural Networks(CNN) deep learning models [23]. It focuses on the insight of the network's properties by getting rid of the manually pre-processing, visualizing the training phase, and quantifying the system's performance by visualization.

The proposed model, which was optimized in an end-to-end manner, developed based on VGGFace2 model [2], which yielded excellent face recognition performance. The proposed model, respectively, combined with uni-direction and bi-direction LSTM networks, trained and validated on the publicly available micro-expression benchmark SMIC datasets. Contributions can be summarized as:

(1) Exploited the visualization technique to quantify the network architectures.
(2) A novel technique to remove the manual face cropping and face alignment steps to forbid the potential error sources.
(3) Combined the VGGFace2 model and LSTM networks for the clear insight of the Spatio-temporal information process.

The rest of the article is organized as follows: Section 2 reviews the related works of MER. Section 3 introduces the proposed methods entirely. Section 4 explains the details of the experiments. We conducted three experiments in this paper and will refer to them by Exp-I, Exp-II, and EXP-III. Section 5 discusses the results. Section 6 concludes this research.

2 RELATED WORK

Previous research on the recognition of micro-expressions through artificial intelligence algorithms divided the task into two fundamental sub-tasks[32]: the facial feature extraction and micro-expression classification. The widely used methods include handcrafted work and deep learning methods.

2.1 Handcrafted Works

The Local Binary Patterns (LBP) algorithm is the main improvement direction to enhance the feature extraction capability of the system. Ruiz-Hernande *et al.* [21] used the Second Order Gaussian Jet to encode LBP for better discriminative properties. Huang *et al.* [8] argued that more comprehensive information could be beneficial to extract useful information, therefore raised methods of obtaining LBP projection on both horizontal and vertical. Wang *et al.* [27] proposed LBP-Mean Orthogonal Planes (MOP) method to improve the efficiency by only computing the LBP results on the three average planes instead of directly calculating each plane in LBP.

The video magnification methods were also considered for exaggerating motion by amplifying temporal information. In 2012, Wu *et al.* [28] proposed a way to reveal the tiny movement that is difficult to see with the naked eye by amplifying the temporal variations in videos from the Eulerian perspective. His research inspired Ngo *et al.* [17] to implement the Eulerian magnification method on the facial expressions motions. Furthermore, Peng *et al.* [18] proposed a consolidated Eulerian framework to expand the temporal duration and amplifies the muscle movements in micro-expressions simultaneously.

2.2 MER by CNN and RNN

Convolutional Neural Networks (CNN) [13] and Recurrent Neural Networks (RNN) brought about remarkable breakthroughs in the fields of image and sequence information processing, such as face recognition, expression detection, and affective computing. Based on these technologies boom, lots of outstanding works were done to explore the possibility of using CNN, RNN, or a combination of both networks for the MER task.

Takalkar *et al.* [24] used a specific data augmentation method to avoid overfitting in the training phase; meanwhile, Miao *et al.* [16] proposed a straightforward CNN model with only three layers with the assistance of the improved saliency map and a pipeline with the facial cropping function to address the overfitting issue. Li *et al.* [15] applied two deep convolutional networks for detecting the facial landmarks and estimating the optical flow features of the micro-expression, respectively. Besides, Reddy *et al.* [20] proposed a method that consists of two 3D-CNN models. Instead of extracting the features through the entire face, they attempted to capture the features from the eyes and mouth region(the micro-expression-aware areas) for more computational efficiency. Xia *et al.* [29] proposed a deep recurrent convolutional network in views of both facial appearance and geometry separately to connect the spatial information to the temporal domain.

Moreover, Kim *et al.* [12] improved the method mentioned above by using a Long short-term memory network to deal with the temporal information. They conducted their experiments on the CASME-II and utilized the labeled expression-state (such as on-set and off-set frames, which are not provided by SMIC dataset). Khor *et al.* [11] conducted the same CNN+LSTM architecture on SMIC and CASME-II dataset; both datasets provide pre-cropped samples. Moreover, they applied the Temporal Interpolation Model(TIM) to replace the data augmentation stage. By using VGG-16 and pre-trained VGGFace models, respectively, to enriched the spatial-temporal information capturing process. Verburg and Menkovski [26] conducted the experiments with the same architecture on the SAMM dataset, and also applied facial cropping, face align and ROI extracting for pre-precessing. By implementing optical flow in the encoding phase, the system indeed effectively improved the detection accuracy and lowered the dimension for the decoding phase.

The methods, as mentioned earlier that share the same CNN+LSTM architecture, all involve the manually pre-processing and mainly been conducted on the datasets with labeled expression-states, e.g., CASME II, SAMM. But our algorithms omitted the pre-processing, which in themselves are potential sources of errors, that is how the proposed method improves on previous works that have used a similar architecture.

3 PROPOSED METHODS

3.1 VGGface2 model

VGGface2 model is basically a ResNet-50 CNN trained on the VGGFace2 dataset, with the potential of facial features extracting, can undertake the MER task by transfer learning.

The ResNet CNN was proposed in 2016 by He *et al.* to solve the vanishing/exploding gradient problem in deep CNN by a residual learning framework. They reconstructed the architecture by adding identity mapping layers, and the other layers are copied from the learned shallower model [6]. The ResNet-50 CNN came from the ResNet family and was implemented on the VGGFace2 dataset with large-scale face images for more accurate face recognition results over pose and age. Its overwhelming performance inspired us to use the VGGFace2 model as the pre-trained network to speed up training while avoiding the risk of overfitting.

3.2 LSTM and Bi-directional LSTM

Long short-term memory (LSTM) networks are derived from the RNN family, contains robust computation and learning abilities for long-term dynamics, directly map the variable-length inputs. In LSTM, the conventional RNN hidden layer is replaced by a comprehensive unit containing an input-node, an input-gate, an internal state, a forget gate, and an output gate. This network converges by small weight could keep information in long time steps and activation ephemerally.

The experiments also used Bi-directional LSTM(Bi-LSTM). Bi-LSTM is an extension of conventional LSTM that trains the input sequences in both directional [10] to train the network on a reversed copy of the input sequence. This structure can provide additional context to the system and result in faster learning.

4 EXPERIMENTS

4.1 Database

In 2011, Li *et al.* published a spontaneous emotion corpus called spontaneous micro-expression (SMIC). In total, SMIC contains 210 minutes of video samples with 1,260,000 frames. Among them, the shortest recorded expression was about 0.11 seconds (11 frames at 100fps), and the average expression length was about 0.30 seconds (29 frames) [19].

The proposed experiments used HS data, samples in which were segmented and labeled as 'negative,' 'positive,' 'surprise,' and 'non-micro-expression' 4 categories. Every micro-expression clip in video sequence starts from a neutral frame, with the second frame as the starting point of expression-related facial muscle movement, and ends when the facial expression turns back to neutral. That means some frames do not contain the micro-expression.

4.2 Data augmentation

Data augmentation is an effective technique to avoid over-fitting [24]. The training phase used two methods.

For exp-I, due to training samples are on the frame level, the augmentation methods were mainly implemented by flipping and randomly rotating [24]. Because of the imbalanced problem in different micro-expression categories, the augmentation stage variable used multipliers were. (Table 1)

The augmentation stage in the exp-II and exp-III was on the sequence level and given as N frames, these frames are consecutive and are denoted as $f_1, ..., f_i, ..., f_N$. The augmentation task is extracting the frame clips F_N from completed video sequence F_e by a 'sliding window' with fixed size: N (Table 1).

Expression	Original	Aug.-I	Aug.-II($\times f_N$)
Negative	2168	15000	1710
Positive	1856	14736	1685
Surprise	1408	15072	1661
Non-ME	5016	14664	1674

Table 1 Data augmentation results, the N was set as 13.

4.3 Exp-I: Pre-trained VGGFace2 Model

Retraining VGGFace2 model is an instance of learning spatial information from micro-expression described above: the scratch of layers was retrained using the transfer learning technique[31] then implemented experiments on data with frame level. Instead of only retraining the classifier layers, Grad-CAM technique [22]was used to interpret and quantify the transfer learning process. 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 produces a coarse localization map to highlight the overlap regions to determine the activation parts for the facial feature extracting stage.

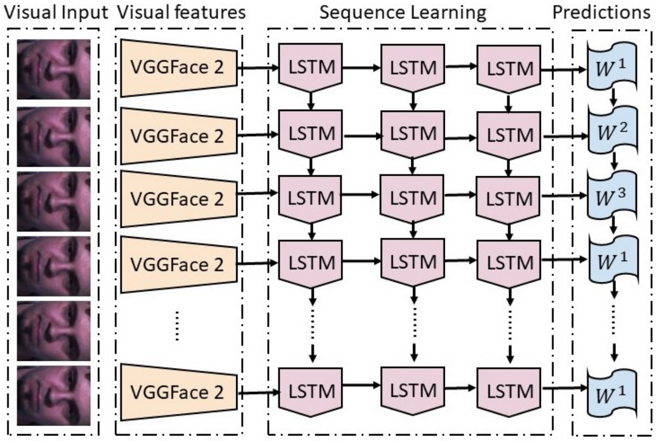


Figure 2: Exp-II was conducted on the sequence level to recognise and classify the micro-expression, which consists of encoding and decoding phases

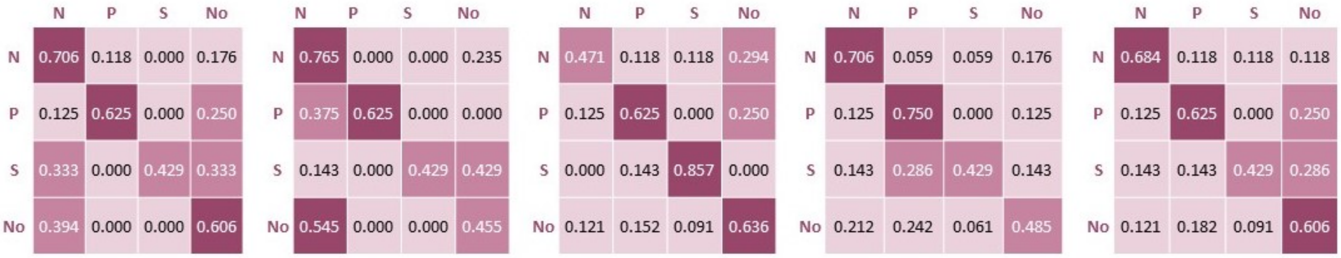
Grad-CAM could be applied to each convolutional layer to produce heat maps with different highlighted regions. The layer where the highlighted area was closest to the facial features is called specific layer l_s, layers after l_s $\{l_s, ..., l_{s+n}\}$ will be retrained under transfer learning policy, then layers $\{l_0, ..., l_{s-1}\}$ are frozen to keep the weights. Outputs from the learning layers would be fully connected to the classification layers $\{fc_1, fc_2, fc_3\}$ to decrease the feature vector size as Figure 1 shows. The result shows in Table 4.

4.4 Exp-II: VGGFace2 + LSTM

Inspired by [3], the proposed system abandon manual pre-processing (face cropping and face alignment), which in themselves are potential sources of errors. The exp-II was conducted on sequence level based on the network structure from experiment-I. In practical experiments, 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.' The best augmentation result based on different N sets shows in Table 3.

In exp-II pre-trained VGGFace2 was repackaged for the encoding phase as exp-I, 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 a 3-layer LSTM to receive the stack of spatial state from the encoder to analyse the temporal information for the final classify. The exp-I was conducted with static input,



	N	P	S	No
N	0.706	0.118	0.000	0.176
P	0.125	0.625	0.000	0.250
S	0.333	0.000	0.429	0.333
No	0.394	0.000	0.000	0.606

	N	P	S	No
N	0.765	0.000	0.000	0.235
P	0.375	0.625	0.000	0.000
S	0.143	0.000	0.429	0.429
No	0.545	0.000	0.000	0.455

	N	P	S	No
N	0.471	0.118	0.118	0.294
P	0.125	0.625	0.000	0.250
S	0.000	0.143	0.857	0.000
No	0.121	0.152	0.091	0.636

	N	P	S	No
N	0.706	0.059	0.059	0.176
P	0.125	0.750	0.000	0.125
S	0.143	0.286	0.429	0.143
No	0.212	0.242	0.061	0.485

	N	P	S	No
N	0.684	0.118	0.118	0.118
P	0.125	0.625	0.000	0.250
S	0.143	0.143	0.429	0.286
No	0.121	0.182	0.091	0.606

Figure 3: Recognition results for each group validation, totally, 65-66 sequences are allocated in validation groups.N means negative, P means positive, S means surprise, No means non-micro-expression. By this confusion matrix we can see, the little numeric unbalance of expression lead to the 'negative' obtain higher accuracy

static output: $x \mapsto y$, while exp-II was conducted with sequential input, static output: $\langle x_1, x_2, ..., x_T \rangle \mapsto \langle Y \rangle$.

4.5 Exp-III: VGGFace2 + Bi-directional LSTM

Bi-directional LSTMs (also known as Bi-LSTM) is an extension of traditional LSTMs that can improve model performance on classification problems [25]. Exp-III applied one-layer Bi-LSTM in the decoding phase for temporal information. We originally assumed that Bi-LSTM would perform better when processing samples with more frames (larger parameter N) since its architecture provides additional context for learning. But judging from the result of exp-III, although the Bi-LSTM indeed obtained higher recognition accuracy, it happened on a relatively smaller training sequence size. The result of $N = 13$ and $N = 15$ show in Table 3.

Pattern	f_N	Data Number	Score
3-Layer-LSTM	13	11276	57.58%
	15	9114	56.07%
Bi-LSTM	13	11276	60.60%
	15	9114	46.07%

Table.3 Training results on different training sequence length(f_N) on 3-Layer LSTM and Bi-LSTM respectively.

4.6 Five-fold Cross-validation

Five-fold cross-validation was used on the sequence level to evaluate the learning performance [1]. In the procedure, the data shuffle and distribution steps were conducted on the subject level to ensure the data in each group for validation is unique and came from the different subjects (original video sequence). The data augmentation stage happened inner each group area. The average evaluation score from all groups shows in Figure 3.

5 RESULTS AND DISCUSSION

In addition to locating the retraining vggface2 model layers, the Grad-CAM technique can also evaluate the final learning performance. Comparing the two heat maps in Figure 4 shows that the highlighted region has changed significantly before and after training (Figure 4 left and right). The highlighted areas 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 [29].

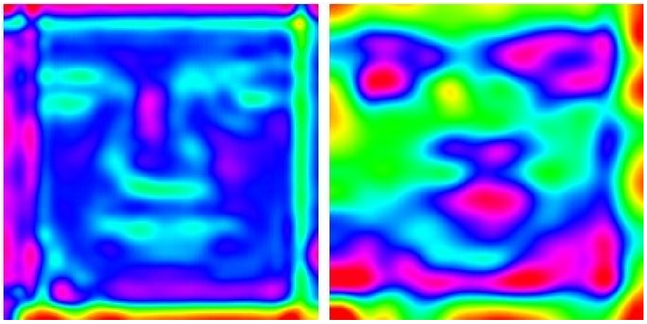


Figure 4: Left: heat map before training - less highlighted region and scattered around the edges of the image. Right: heat map after training - obvious highlighted region mainly focus on the micro-expression-aware areas

Table 4 also shows the comparison between the results of the experiment with the existing state-of-the-art methods.

Technology	Proposed year	Accuracy
STLBP-IP [8]	2015	59.51%
STCLQP [9]	2016	56.10%
3D-FCNN [14]	2018	55.49%
VGGFace2 (Proposed)	2020	**30.2%**
VGGFace2+LSTM (Proposed)	2020	**59.09%**
VGGFace2+Bi-LSTM (Proposed)	2020	**60.06%**

Table 4 Comparison between the results of the experiments with existing state-of-the-art methods.

6 CONCLUSION

The goal of this study was to gain insight into spatial and temporal information for the MER task. Visualizing and interpret how the training procedure process the Spatio-temporal data and how the information contributes to the final classification. This proposed method enables the network training phase to be more quantitative and controllable.

Moreover, the learning result shows that even remove the manual methods(face cropping and face alignment), the proposed methods can still obtain good results.

REFERENCES

[1] Michael W Browne. 2000. Cross-Validation Methods. *Journal of Mathematical Psychology* 44, 1 (mar 2000), 108–132. https://doi.org/10.1006/jmps.1999.1279

[2] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. https://doi.org/10.1109/fg.2018.00020

[3] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. [n.d.]. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. ([n. d.]). arXiv:http://arxiv.org/abs/1411.4389v4 [cs.CV]

[4] P. Ekman and W.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.

[5] 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.

[6] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] 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

[9] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. 2016. Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* 175 (jan 2016), 564–578. https://doi.org/10.1016/j.neucom.2015.10.096

[10] Zhiheng Huang, Wei Xu, and Kai Yu. [n.d.]. Bidirectional LSTM-CRF Models for Sequence Tagging. ([n. d.]). arXiv:http://arxiv.org/abs/1508.01991v1 [cs.CL]

[11] 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]

[12] 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

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[14] Jing Li, Yandan Wang, John See, and Wenbin Liu. 2018. Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Analysis and Applications* 22, 4 (nov 2018), 1331–1339. https://doi.org/10.1007/s10044-018-0757-5

[15] 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

[16] 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

[17] Anh Cat Le Ngo, Yee-Hui Oh, Raphael C.-W. Phan, and John See. 2016. Eulerian emotion magnification for subtle expression recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. https://doi.org/10.1109/icassp.2016.7471875

[18] Wei Peng, Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. 2019. A Boost in Revealing Subtle Facial Expressions: A Consolidated Eulerian Framework. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE. https://doi.org/10.1109/fg.2019.8756541

[19] 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

[20] Sai Prasanna Teja Reddy, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. 2019. Spontaneous Facial Micro-Expression Recognition using 3D Spatiotemporal Convolutional Neural Networks. arXiv:1904.01390 [cs.CV]

[21] John A. Ruiz-Hernandez and Matti Pietikainen. 2013. Encoding Local Binary Patterns using the re-parametrization of the second order Gaussian jet. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE. https://doi.org/10.1109/fg.2013.6553709

[22] 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)*.

[23] Leslie N. Smith. [n.d.]. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. ([n. d.]). arXiv:http://arxiv.org/abs/1803.09820v2 [cs.LG]

[24] M.A. Takalkar and M. 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

[25] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. 2018. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* 6 (2018), 1155–1166. https://doi.org/10.1109/access.2017.2778011

[26] 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]

[27] 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

[28] 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

[29] Z. Xia, X. Hong, X. Gao, X. Feng, and G. 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

[30] 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

[31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3320–3328.

[32] Yuan Zong, Wenming Zheng, Xiaopeng Hong, Chuangao Tang, Zhen Cui, and Guoying Zhao. 2019. Cross-Database Micro-Expression Recognition. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval - ICMR '19*. ACM Press. https://doi.org/10.1145/3323873.3326590