ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747232

A NOVEL MICRO-EXPRESSION RECOGNITION APPROACH USING ATTENTION-BASED MAGNIFICATION-ADAPTIVE NETWORKS

Mengting Wei, Wenming Zheng, Yuan Zong, Xingxun Jiang, Cheng Lu, Jiateng Liu

Key Laboratory of Child Development and Learning Science of Ministry of Education,
School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China.

ABSTRACT

Micro-Expression recognition (MER) is a challenging task due to the short duration and low intensity of Micro-Expressions. A popular method to tackle this is magnifying MEs so as to enlarge the expression intensity to make recognition easier. However, the single fixed magnification strategy, widely used in existing works of MER, is not appropriate for different subjects, because each subject has specific expression intensity corresponding to different MEs. To cope with this issue, we propose a novel Attention-based Magnification-Adaptive Network (AMAN) to learn adaptive magnification levels for the ME representation. The network consists of two modules: magnification attention (MA module) to adaptively focus on appropriate magnification levels of different MEs, and frame attention (FA module) to focus on discriminative aggregated frames in a ME video. Extensive experiments on three widely used databases manifest that our method yields state-of-art results compared with other methods.

Index Terms— Facial expression magnification, Attention mechanism, Micro-expression recognition, Transfer learning

1. INTRODUCTION

Micro-Expression (ME) is a temporary facial expression that appears unconsciously when people try to hide their true feelings [1]. For this characteristic, it is more likely to reflect people's real emotions and thus contribute to potential applications in terms of clinical diagnosis and intelligence. Whereas, compared with concious expressions, MEs are characterized by lower intensity and shorter duration, which makes accurate Micro-Expression Recognition (MER) a challenging task.

Targeting at the low intensity of MEs, some works attempt to enhance the intensity of facial movements to improve recognition performance. Lei *et al.* [2] adopted learning-based video motion magnification network to magnify facial movements of the apex frame. In this way, relations of different facial area would be more apparent so as to promote the performance of the posterior recognition task. Li *et al.* [3] used Eulerian Motion Magnification (EMM) [4] to magnify the apex frame from original MEs which could enlarge the difference among different ME categories so that the network is easier to learn discriminative features. In these works, when choosing the level of motion magnification, they just set one fixed level for all subjects and MEs. Nevertheless, since different subjects have different facial anatomical structure, even for the same stimulus, different subjects would present facial muscle movements at different intensities. In this case, a fixed magnification level is apparently not appropriate for every subject, and such problem would potentially decrease the performance of subsequent classification tasks, as shown in Fig. 1.

To address the issue aforementioned, in this paper, we propose a novel Attention-based Magnification Adaptive Network (AMAN) to learn appropriate magnification levels for MEs in a more flexible way. To alleviate the influence of individual differences and different pattern of MEs, we try to design a magnification-adaptive method. Instead of setting a fixed magnification level for all ME sequences, we use a set of different amplification factors (AFs) for the same image. Our network consists of two attention modules:

1. One module is for weighting different magnification levels, namely magnification attention (MA module). In MA module, we employ attention mechanism to assign different weights among a set of AFs so the network is able to adaptively learn discriminative magnification levels.

2. The other module is designed to weight different frames in a video, namely frame attention (FA module). In FA module, we introduce attention mechanism to assign different weights among a set of selected frames so the network is able to focus on more discriminative aggregated frames.

2. PROPOSED METHOD

2.1. Basic Idea

Videos in different ME datasets have various temporal length [5–7]. Considerating this, for a given ME sequence, we firstly use a Time Interpolation Model (TIM) to reconstruct the video so that all the videos are consistent in length. Then we successively sample n number of frames from the beginning at a certain interval. Each selected frame is then magnified with a set of amplification factors (AFs). After that, for a

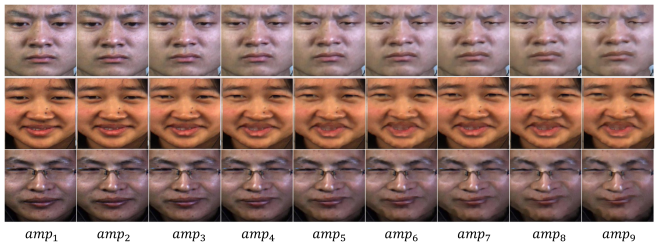

Fig. 1: A set of consistent magnification levels on Micro-Expressions from different subjects. The first, second and third row displays 'disgust', 'happy' and 'happy' respectively. Magnification levels most able to reflect real emotions vary among different subjects and Micro-Expressions.

set of AFs corresponding to a single frame, we use a pre-trained CNN to extract their feature representations. Next, we use our magnification attention module (MA module) to aggregate these representations into a frame representation. Subsequently, for these frame representations in a video, we employ our frame attention module (FA module) to aggregate them into a final representation of the video. This final representation is eventually sent into a fully-connected layer for classfication. The whole framework is shown in Fig. 2.

2.2. AMAN Model

2.2.1. MA module

Motion magnification techniques provide us convenience to capture small motions imperceptible to the naked eyes. In some cases, the motion is so small that the magnification results are inclined to noise, so MER methods using hand-designed filters to magnify micro-expressions may not be optimal. Considering this, we employ deep convolutional networks to learn the filters directly and use tranfer learning strategy to apply magnification characteristics learned in other databases [8] to micro-expression databases. As Wu *et al.* and Wadhwa *et al.* [4, 9] have defined on motion magnification, a single image in a consecutive video can be described as:

$$I(\mathbf{x}, t) = f(\mathbf{x} + \delta(\mathbf{x}, t)) \tag{1}$$

where $\delta(\mathbf{x}, t)$ is the motion field at position $\mathbf{x}$ and time t, our goal is to magnify the motion so the magnified image I_{mag} becomes

$$I_{mag}(\mathbf{x}, t) = f(\mathbf{x} + (1 + amp)\delta(\mathbf{x}, t)) \tag{2}$$

where amp is the magnification factor.

In our practice, for a single frame selected in a ME video, we firstly use the pre-trained network [8] to magnify it with a set of magnification factors ranging from 1 to k, denoted as amp_1 to amp_k. The network has three inputs: a magnification factor and two images between which there is a slight displacement (in different phases of the same video). This is formulated as:

$$I^i_{mag} = f(I_{on}, I_{sel}, amp_i) \tag{3}$$

where I_{on} is the onset frame. I_{sel} is another frame in the same video we select for magnification. I^i_{mag} is a generated image after magnification. amp_i represents amplification factor. $f(\cdot)$ is the network.

For a magnified image with a specific amplification factor amp_i, we use the backbone network to extract its feature representation, denoted as:

$$F^{\mathrm{T}}_i = g(I^i_{mag}) \tag{4}$$

where $g(\cdot)$ is the backbone network, and F^{T}_i denotes the feature representation of a single magnified image.

Since we expect that magnification levels contributing more to recognition are weighted more importance, we assign different attention scores on these feature representations of magnified MEs from the same frame. Our MA module is composed of a linear fully-connected (FC) layer and a sigmoid activation function. Attention weights among feature representations are calculated by:

$$\alpha_i = \sigma(F^{\mathrm{T}}_i q^0) \tag{5}$$

where q^0 is the parameter of FC layer, and $\sigma(\cdot)$ denotes sigmoid function. The new weighted feature representation of a single magnified frame can be formulated as:

$$F^{attn}_i = \alpha_i F^{\mathrm{T}}_i \tag{6}$$

With those weighted representations of all magnified frames, feature representation of the original frame is aggregated by:

$$F_m = \frac{\sum_{i=1}^{k} F^{attn}_i}{\sum_{i=1}^{k} \alpha_i} \tag{7}$$

2.2.2. FA module

Many methods have demonstrated that the apex frame (frame with largest movement intensity) contributes most to Micro-Expression recognition (MER) [10, 11]. Nevertheless, in a ME sequence, the intensity of the expression actually changes in a consecutive way, which means that more than one frames can be thought as 'apex'. In view of this, we use multiple frames instead of the single apex frame in the video. Each frame is managed with the MA module to get its feature representation. We assign a set of attention scores on these frame feature represention once again so the network could adaptively focus on more discriminative frames. The FA module has the same structure as MA module. In this module, attention weights among these feature representations are formulated as:

$$\beta_j = \sigma(F^{\mathrm{T}}_{m_j} p^0) \tag{8}$$

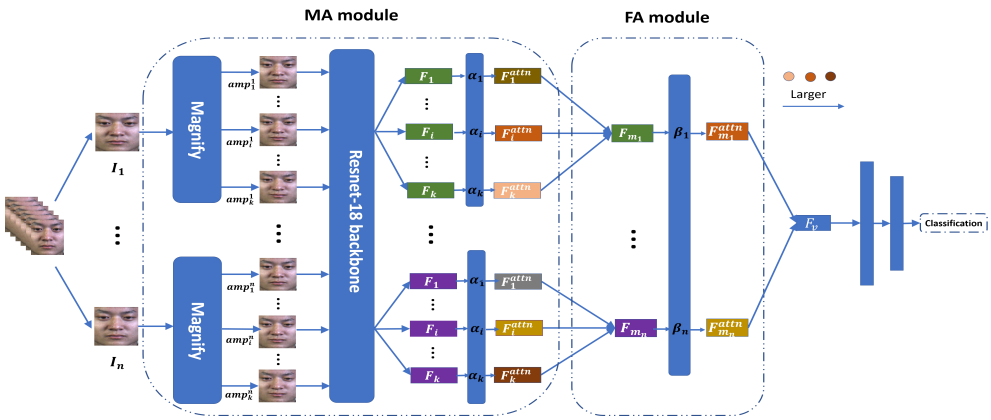

Fig. 2: Framework of Attention-based Magnification-Adaptive Network (AMAN).

where p^0 is the parameter of another FC layer, and F_{m_j} denotes the aggregated feature representation of the $j-th$ image in a ME sequence. Then, the new weighted feature representation of original representation is denoted as:

$$F_{m_j}^{attn} = \beta_j F_{m_j}^{\mathrm{T}} \tag{9}$$

We aggregate these new frame representations with their attached attention scores by:

$$F_v = \frac{\sum_{j=1}^{n} F_{m_j}^{attn}}{\sum_{j=1}^{n} \beta_j} \tag{10}$$

F_v is the global representation of a ME video. The network takes it as the final representation for the following classification.

3. EXPERIMENTS

3.1. Experimental Setup

3.1.1. Datasets Preprocessing and Experiments Detail.

Three ME databases are used to evaluate our method: CASME II [5], SAMM [6] and SMIC-HS [7]. On the CASME II and the SAMM, we select categories with more than 10 samples, following the rules most methods [12, 13] adopt for five classification. On the SMIC-HS, we use all samples in this database. In a ME sequence, we first calculate 68 landmarks of the whole face utilizing [13] in the onset frame and then align the face area in line with these landmarks. All the images are resized to 224×224.

We use Resnet-18 as our backbone to extract shallow features of expressions. The dataset employed to pre-train is FER2017. When fine-tuning on MEs, we freeze the former residual blocks to extract some shallow features and release the last residual block for fine-tuning. To avoid subject dependence in the process, we adopt the leave-one-subject-out (LOSO) cross validation.

3.1.2. Settings on Ceiling of Magnification Levels and Number of Frames.

Muscle movements in the facial area become more pronounced as the AFs increase during magnification. In order to ensure that most MEs are sufficiently enlarged, and meanwhile to avoid facial deformation caused by overlarge magnification levels, we conducted a series of comparsion experiments on the CASME II. Related results are shown in Tab. 1.

As Tab. 1 presents, recognition accuracy firstly grows as the maximum level increases and then begins to decline when the level ceiling is over nine, indicating that from this level there may appear excessive deformation in the facial area of most MEs. Therefore, we set $k = 9$ as the ceiling of magnification level.

Table 1: Evaluation on different ceilings of magnification level on the CASME II.

k	7	8	**9**	10	11
Acc(%)	69.85	72.34	**75.40**	69.23	66.41

To find appropriate number of frames selected in a video, in our experiments, we set n ranging from 1 to the maximum length(20) of the video on three datasets and plot the relation between recognization accuracy and number of frames, as shown in Fig. 3.

As can be seen from Fig. 3, recognition accuracy does vary with the change of n on three datasets, but not with large fluctuation, which further reflects the robustness of our method for selecting frames. We speculate that the robustness of this lies in our network's ability to voluntarily weight more on the aggregated frames near the apex. We set $n = 11$ uniformly on three databases.

Table 2: The accuracy (ACC(%)) and F1-score (%) of different methods under the LOSO protocal on three datasets.

MER method	CASME II		SAMM		SMIC-HS	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
LBP-SIP (2014) [14]	66.40	N/A	N/A	N/A	62.80	N/A
DSSN (2019) [16]	70.78	72.97	57.35	46.44	63.41	64.62
TSCNN-I (2019) [17]	74.05	73.27	63.53	60.65	72.74	72.36
LGCconD (2020) [3]	62.14	60.00	35.29	23.00	63.41	62.00
RNMA (2020) [18]	65.90	53.90	48.50	40.20	49.40	49.60
GEME (2021) [19]	75.20	**73.54**	55.88	45.38	64.63	61.58
AMAN(ours) (2021)	**75.40**	71.25	**68.85**	**66.82**	**79.87**	**77.08**

*N/A - no results reported.

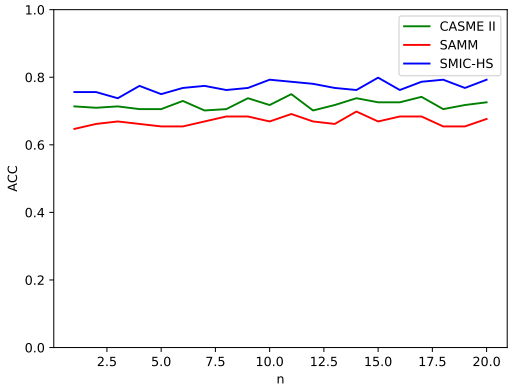

Fig. 3: Evaluation on the number of frames n.

3.2. Ablation Study

In order to testify the effectiveness of our two modules, we conducted a series of ablation experiments on three databases. We kept only one module each time, and compared the results with that of the whole network.

As shown in Tab. 3, the performance decreases greatly when we remove MA module, which proves that our multi-magnification method yields better performance than the single-magnification method. Besides, the performance is also imperfect when there is without FA module, which demonstrates that attention mechanism in this module to focus on discriminative frames is also indispensable.

Table 3: Accuracy(%) of ablation experimental results on the two modules.

	CASME II	SAMM	SMIC-HS
MA module	59.19	59.57	64.02
FA module	62.71	56.25	62.80
MA+FA module	75.40	68.85	79.87

3.3. Experimental Results

Experimental results of our experiments and comparison to the state-of-the-art approaches are shown in Tab. 2.

From Tab. 2, we can see that our method exceeds most methods using handcraft features or high-level learned features. Specifically, AMAN yields comparable accuracy on the CASME II but gets further improvement on the other two datasets compared with GEME [19]. On the SAMM, our accuracy is much higher than TSCNN-I [17] by 5.32%. On the SMIC-HS, compared to TSCNN-I, our method improves the accuracy by 7.13%. AMAN achieves further improvements compared to those approaches using single-magnification techniques [2, 3] to magnify MEs, indicating that our method can automatically find the appropriate degree of magnification levels.

4. CONCLUSION

Our paper presents a novel attention-based network that can be adaptive to different magnification levels for Micro-Expression recognition (MER). The network consists of two modules: magnification attention module (MA module) for weighting magnification levels and frame attention module (FA module) for weighting aggregated frames in a Micro-Expression video. Extensive experimental results prove that our method yield superior performance compared to other state-of-the-art methods.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program under Grant 2018YFB1305203, and in part by the NSFC under grants U2003207 and 61902064.

6. REFERENCES

[1] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of

micro-expressions," *JOURNAL OF NONVERBAL BEHAVIOR*, vol. 37, no. 4, pp. 217–230, 2013.

[2] Ling Lei, Jianfeng Li, Tong Chen, and Shigang Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2237–2245.

[3] Yante Li, Xiaohua Huang, and Guoying Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 249–263, 2020.

[4] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[5] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, pp. e86041, 2014.

[6] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE transactions on affective computing*, vol. 9, no. 1, pp. 116–129, 2016.

[7] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 2013, pp. 1–6.

[8] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik, "Learning-based video motion magnification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 633–648.

[9] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman, "Phase-based video motion processing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–10, 2013.

[10] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[11] Sze-Teng Liong, John See, KokSheik Wong, and Raphael Chung-Wei Phan, "Automatic micro-expression recognition from long video using a single spotted apex," in *Asian conference on computer vision*. Springer, 2016, pp. 345–360.

[12] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129–139, 2019.

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[14] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Asian conference on computer vision*. Springer, 2014, pp. 525–537.

[15] Yante Li, Xiaohua Huang, and Guoying Zhao, "Can micro-expression be recognized based on single apex frame?," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3094–3098.

[16] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 36–40.

[17] Baolin Song, Ke Li, Yuan Zong, Jie Zhu, Wenming Zheng, Jingang Shi, and Li Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.

[18] Chongyang Wang, Min Peng, Tao Bi, and Tong Chen, "Micro-attention for micro-expression recognition," *Neurocomputing*, vol. 410, pp. 354–362, 2020.

[19] Xuan Nie, Madhumita A Takalkar, Mengyang Duan, Haimin Zhang, and Min Xu, "Geme: Dual-stream multi-task gender-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, 2021.

[20] Radhouane Guermazi, Taoufik Ben Abdallah, and Mohamed Hammami, "Facial micro-expression recognition based on accordion spatio-temporal representation and random forests," *Journal of Visual Communication and Image Representation*, p. 103183, 2021.