A Neural Micro-Expression Recognizer

Yuchi Liu, Heming Du, Liang Zheng, Tom Gedeon
Research School of Computer Science, Australian National University, Australia

Abstract— **Recognizing micro-expressions underpins significant and critical research and significant application. We speculate that this problem requires the understanding of the subtle face movement, integration of face structures and a solution of limited training data. In this paper, we build an effective micro-expression recognition system that leverages techniques stemming from these speculations. First, we introduce an optical flow method based on the onset frame and the apex frame to encode the subtle face motion. This has already been validated by prior research. Second, to obtain discriminative representations from the rigid face structures, part-based average pooling is proposed to inject structure priors to the network. Finally, because the system suffers from small training sets, we propose to transfer domain knowledge from macro-expression recognition tasks to micro-expression recognition. Specifically, we adopt two domain adaptation techniques including adversarial training and expression magnification and reduction (EMR). Through experiment, we show that the proposed system achieves very competitive results on the 2nd Micro-Expression Grand Challenge (MEGC).**

I. INTRODUCTION

Micro-expressions (ME) are subtle muscle movements within 1/25 of a second. In addition to the expressions being short, micro-expressions are more likely to be those suppressed expressions in application settings. Compared with the expressions that are made consciously, micro-expressions are more likely to reflect true feelings and motivations. Research outcomes of micro-expression recognition (MER) can be applied to areas such as national security, clinical diagnosis, the judicial system, and political elections. MER is very challenging due to the short ME duration and the low intensity of facial muscle movements. These challenges mean that human performance on micro-expression recognition remain at a considerably low level. Therefore, it is important to design effective systems to recognize micro-expressions automatically.

Early work on MER mainly focused on extracting handcrafted features from micro-expression video clips. For example, Local Binary Pattern with Three Orthogonal Planes (LBP-TOP) [2] extracts discriminative features related to dynamic textures. It is applied as the feature descriptor in the micro-expression recognition task in [1] and is widely used as the baseline method in this area. Other variants of LBP-TOP like spatiotemporal LBP with integral Projection (STLBP-IP) [3] and discriminative spatiotemporal LBP (DSLBP) [4] have also been investigated.

Optical flow can extract representative motion features which are robust for the diversity of facial textures. Optical flow estimation can be used to enrich the input except for RGB channels [5]. Other work considers optical flow as a data prepossessing step for other handcrafted features based on optical flow. For example, The MDMO [15] computes the oriented optical flow vectors to form histograms from ROIs, which is discriminative for micro-expression recognition. By accumulating the derivatives of the optical flow, Bi-Weighted Oriented Optical Flow (Bi-WOOF) [6] utilizes the optical strain to generate the weighted histograms, which can be used to recognize micro-expressions.

Recently, motion magnification (MAG) was used to improve the accuracy of micro-expression recognition tasks significantly. As a data processing method, MAG magnifies the motion features of the original micro-expression video clips. A number of works [14], [7], [15] show the improved recognition accuracy by introducing MAG.

Deep neural networks have shown competitive learning ability on feature extracting and classification in many fields including micro-expression recognition. Dual Temporal Scale Convolutional Neural network (DSTCNN) [8] apply 3-D CNN models on CASME I and CASME II [17] datasets. Spatiotemporal recurrent convolutional networks (STRCN) model the spatiotemporal motion deformations and subtle changes by employing CNNs with recurrent connections [14]. However, deep approaches suffer from insufficient training samples. Even combining three micro-expression datasets, the total samples are no more than 500 in number. Therefore, it is worthwhile to design specific transfer learning techniques for micro-expression recognition tasks such that other facial datasets in the computer vision area could contribute to MER.

Furthermore, the motion information of eyes, eyebrow, nose, and mouth in the normalized and cropped face is highly structured and related to expressions. Previous research does not pay attention to these partial details of micro-expressions but feed the features extracted from the whole face into the classifier.

To solve the above issues, in this paper we proposed a part-based deep neural network with two domain adaptation techniques (adversarial domain adaptation and motion magnification and reduction). Our deep method can automatically learn to extract discriminative features related to facial parts. Additionally, the two domain adaptation techniques help to enrich the available training samples. Our code for this challenge is available on https://github.com/xiaobaishu0097/MEGC2019.

II. METHODOLOGY

A. Reprocessing

Video clips recording the micro-expressions contain many variations in the natural scene that are not related to facial expressions, such as background and head posture. To minimize

978-1-7281-0089-0/19/$31.00

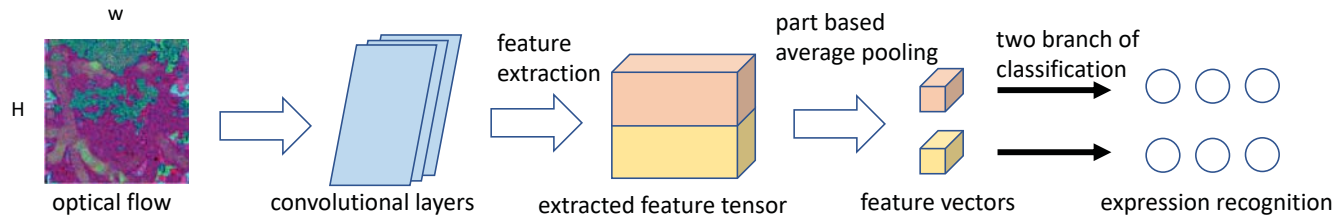

Fig. 1. Network structure. The convolutional layers extract the feature tensor (the feature maps) from the optical flow. Two average pooling are applied on the top and bottom part of the feature tensor separately to obtain the dimension-reduced feature vectors, following by a two layer fully connected classifier for each feature vector.

the negative impact of such irrelevant features, we introduced the following data preprocessing steps:

- The OpenCV pre-trained HOG and the Linear SVM object detectors are used to detect face regions.
- We utilize the method in [23] to identify facial landmarks in the above face regions, which has already been implemented in the dilb library.
- We obtain the normalized rotation, translation, and scale representation of the face based on the facial landmarks by using the OpenCV built-in facial alignment algorithm.

B. Motion Feature Extraction

The micro-expression features are highly correlated to subtle motion in the face area. We use optical flow methods to extract motion features. To reduce computation cost, the onset frame and the apex frame of the micro-expression clips are picked to compute the optical flow, which has the same image size as the original two video frames. The onset frame is the first frame and the apex frame has the maximum motion compared with other video frames.

Resnet [10] can extract discriminative image representations by supervised learning. We choose Resnet18 as the backbone of the optical flow encoder and the pre-trained weights based on the ImageNet2012 dataset as initialization.

C. Part-Based Classification

Local details in an input source contain discriminative information. Hence, intelligent systems need the ability to pay attention to local details in recognition tasks. The PCB method [16] splits feature map in the backbone of the convolutional neural network into several sub-tensors. They then used average pooling and 1*1 convolution to perform dimension reduction on each sub-tensor to obtain part-based feature vectors. Finally, multiple classifiers perform classification training based on the corresponding feature vector. The part-based mechanism has achieved competitive results in personal re-identification tasks. Inspired by this, we split the feature map extracted from the last convolutional layer of our proposed feature encoder into the top and bottom parts, which are more representative for the eyes area and the mouth area separately. Two individual branches of average pooling and classification are applied on the above two sub-tensors. At the same time, we concatenate the outputs of the first fully connected layer in two branches into a single vector to perform the expression recognition classification task by a third following fully connected layer. The details of the part-based mechanism are shown in Fig. 1.

D. Supervised Domain Adaptation

Because of the small samples size of the micro-expression recognition task, the domain adaptation is applied in our approach by introducing the macro-expression recognition task (CK+)[11]. We manually categorise the original macro-expression labels into the same label space with the target domain (MEGC2019 challenge) such that they could perform supervised trained in a shared model.

We propose a domain adaption technique called Expression Magnification and Reduction (EMR) for the micro-expression domain and the macro-expression domain. We assume that the apex of the micro-expression is an inevitable intermediate process for macro-expressions. The middle frame between the onset frame and the apex frame of the macro-expression video clip is picked. We call this step as macro-expression reduction. To maximize the similarity between the micro-expressions and macro-expression, we also perform micro-expression magnification. Motion Magnification (MAG) amplifies subtle motions and it is widely used in micro-expression recognition tasks [14], [15] to improve recognition accuracy. We utilize the open implementation of Eulerian Video Magnification (EVM) from Massachusetts Institute of Technology (MIT) to magnify micro-expressions. As illustrated in Fig. 2, the similarity of the intensity of the macro-expression and the micro-expression is increased as the result of EMR. Furthermore, to bridge the data distribution gap between the source domain and the target domain, adversarial based domain adaptation techniques are widely investigated in recent years to obtain domain invariant features [12], [13]. We use the input of the last fully connected layer of the classifier as the input of the discriminator, which consists of two fully connected layers. The model structure is shown in Fig. 3

E. Learning objective of the whole system

We consider L_t, L_b as the classification losses for the top and bottom part branches respectively in Fig 1. The classification loss and the adversarial loss for the concatenated feature vector in Fig 3 are called L_c and L_adv. In summary, the overall learning objective of the whole proposed system

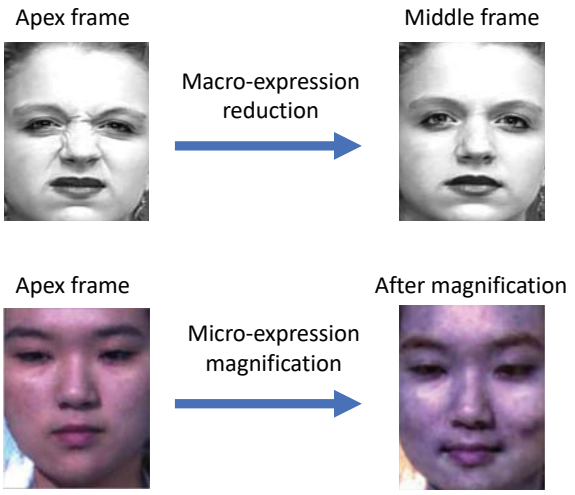

Fig. 2. Examples of EMR. At the top, the apex frame of one macro-expression sample is reduced by replacing it with a middle frame between the onset frame and the apex frame. At the bottom, the apex frame of one micro-expression sample is amplified by EVM.

is to minimize the following loss function:

$$L = \Sigma_{i=0}^{i=N} L_t(y_i, \hat{y_i}) + L_b(y_i, \hat{y_i}) + L_c(y_i, \hat{y_i}) - \lambda \Sigma_{i=0}^{i=N} L_a dv(y_i, \hat{y_i}),$$

where y_i is the ground truth and $\hat{y_i}$ is the prediction. The model parameters are updated through back propagation based on this objective function.

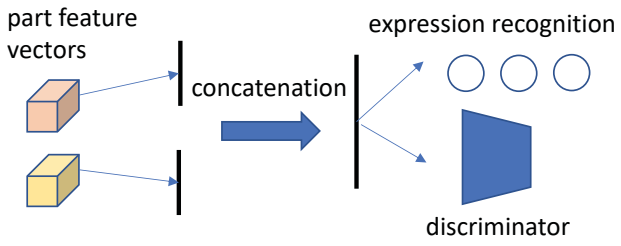

Fig. 3. The pipeline of classification and adversarial domain adaptation. The two feature vectors from the part-based average pooling are firstly reduced to the low dimension size by single separate fully connected layers. Then, the reduced feature vectors are concatenated into a single vector which is the input of the discriminator and the last classification layer.

III. EXPERIMENTS

A. Datasets

Three spontaneous facial micro-expression datasets are used in this challenge: CASME II [17] dataset, SAMM [18], [19] dataset, and SMIC [20] datasets. To compose the, into a single dataset and perform a unified evaluation metric, emotion classification labels in all three datasets are appropriately mapped into a common reduced subset including negative, positive, and surprise. This consolidation includes 442 samples (145 from CASME II, 133 from SAMM, and 164 form SMIC) from 68 subjects (24 from CASME II, 28 from SAMM, and 16 from SMIC). As described in the last section, the CK+ dataset is also introduced to implement domain adaptation. CK+ includes 327 video clips with expression labels which can also be relabeled into the above three labels.

B. Evaluation Metric

The Leave-One-Subject-Out (LOSO) cross-validation is applied to guarantee subject-independent evaluation. Therefore, 68 training and testing procedures are performed. Because of the imbalanced label distribution, Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) are considered as performance metric here to avoid the proposed method overfitting a certain class. Given True Positives (TP_c), False Positives (FP_c) and False Negatives (FN_c) for each class c (C classes in total) over 68 folds, UF1 can be calculated as:

$$UF1 = \Sigma_i^C UF1_i / C,$$

where:

$$UF1_c = \frac{2 * TP_c}{2 * TP_c + FP_c + FN_c},$$

and UAR can be formulated as:

$$UAR = \frac{1}{C} \Sigma_i^C Acc_c,$$

where:

$$Acc_c = \frac{TP_c}{n_c}.$$

C. Results and analysis

The LOSO experiment results based on the MEGC2019 official evaluation metric are shown in Table I. According to Table I, Part-based model plus Emotion Magnification and Recognition (EMR) outperform the baseline method LBP-Top significantly. The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) over three datasets achieved 0.7663 and 0.7531 respectively while they are only 0.5882 and 0.5785 for LBP-TOP. If we adopt the adversarial domain adaptation mechanism additionally, both UF1 and UAR can be improved over 20 percentage points on the composite dataset. Therefore, the domain adaptation techniques are vital for micro-expression recognition tasks with small datasets. The proposed methods also outperform the baseline method on each individual part of the composite dataset. More specifically, the system in the last line of Table I, which contains all techniques we proposed, achieves better results on both UF1 and UAR on SMIC dataset and SAMM dataset. However, the adversarial mechanism does not show the same effectiveness on the CASME II dataset. The inconsistency may be caused by insufficient training samples in the micro-expression dataset. Therefore, exploring better transfer learning techniques is essential for further works targeting micro-expression recognition tasks.

IV. CONCLUSION

In this paper, we proposed a neural micro-expression recognizer to solve micro-expressions recognition tasks with small datasets. The part-based model and two domain adaptation techniques are our main contributions. The part-based

TABLE I

THE UNWEIGHTED F1-SCORE (UF1) AND UNWEIGHTED AVERAGE RECALL (UAR) OF THE BASELINE METHOD AND OUR METHODS

Method	Full		SMIC Part		CASME II Part		SAMM Part	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Part + EMR	0.7663	0.7531	0.7001	0.6859	0.8615	0.8398	0.7180	0.6994
Part + Adversarial + EMR	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152

model forces the encoder to learn representations focusing on local motions on the face, which is discriminative for expression reconnecting. Adversarial domain adaptation helps to extract cross domain invariant features between micro-expressions datasets and macro-expression datasets. Motion magnification and reduction reduces the distribution gap between the two types of expressions. The LOSO experiment results show that our proposed methods can achieve much higher UF1 on each dataset in the 2nd Micro-Expression Grand Challenge (MEGC).

REFERENCES

[1] Pfister T, Li X, Zhao G, Pietikinen M. Recognising spontaneous facial micro-expressions. In Computer Vision (ICCV), 2011 IEEE International Conference on 2011 Nov 6 (pp. 1449-1456). IEEE.

[2] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence. 2007 Jun;29(6):915-28.

[3] Huang X, Wang SJ, Zhao G, Piteikainen M. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In Proceedings of the IEEE international conference on computer vision workshops 2015 (pp. 1-9).

[4] Xiaohua H, Wang SJ, Liu X, Zhao G, Feng X, Pietikainen M. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Transactions on Affective Computing. 2017 Jun 7.

[5] Khor HQ, See J, Phan RC, Lin W. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition. In Automatic Face Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on 2018 May 15 (pp. 667-674). IEEE.

[6] Liong ST, See J, Wong K, Phan RC. Less is more: Micro-expression recognition from video using apex frame. Signal Processing: Image Communication. 2018 Mar 1;62:82-92.

[7] Peng W, Hong X, Xu Y, Zhao G. A Boost in Revealing Subtle Facial Expressions: A Consolidated Eulerian Framework. arXiv preprint arXiv:1901.07765. 2019 Jan 23.

[8] Peng M, Wang C, Chen T, Liu G, Fu X. Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology. 2017 Oct 13;8:1745.

[9] Yan WJ, Wu Q, Liu YJ, Wang SJ, Fu X. CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on 2013 Apr 22 (pp. 1-7). IEEE.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[11] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on 2010 Jun 13 (pp. 94-101). IEEE.

[12] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495. 2014 Sep 26.

[13] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. Domain-adversarial training of neural networks. The Journal of Machine Learning Research. 2016 Jan 1;17(1):2096-30.

[14] Xia Z, Hong X, Gao X, Feng X, Zhao G. Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-expressions. arXiv preprint arXiv:1901.04656. 2019 Jan 15.

[15] Liu YJ, Zhang JK, Yan WJ, Wang SJ, Zhao G, Fu X. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing. 2016 Oct 1;7(4):299-310.

[16] Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 480-496).

[17] Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H., Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PloS one, 9(1), e86041.

[18] Davison, A. K., Lansley, C., Costen, N., Tan, K., Yap, M. H. (2016). SAMM: A spontaneous microfacial movement dataset. IEEE Transactions on Affective Computing, 9(1), 116-129.

[19] Davison A, Merghani W, Yap M. Objective classes for micro-facial expression recognition. Journal of Imaging. 2018 Oct;4(10):119.

[20] Li, X., Pfister, T., Huang, X., Zhao, G., Pietikainen, M. (2013). A Spontaneous Micro-expression Database: Inducement, collection and baseline. In IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG) (pp. 22-26)

[21] Peng M, Wu Z, Zhang Z, Chen T. From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning. In Automatic Face Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on 2018 May 15 (pp. 657-661). IEEE.

[22] Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 480-496).

[23] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014 (pp. 1867-1874).