Supplementary Materials for:

Modeling the Evolution Dynamics to Enhance Micro-Expression Recognition

Authors: Yuhong HE, Guangyu WANG, Wenchao LIU, Lin MA, Haifeng LI

1 Dataset

In the micro-expression recognition experiments, two spontaneous micro-expression databases were utilized: DFME [1] and CAS(ME)^3 [2].

The DFME dataset is currently the largest dynamic spontaneous micro-expression dataset with the highest collection image rate, comprising 7,526 micro-expression video samples from 671 subjects (656 valid). It includes seven emotion categories: happiness, anger, contempt, disgust, fear, sadness, and surprise. The image rate ranges from 200 to 500 fps.

The CAS(ME)^3 dataset contains 1,109 labeled micro-expressions and 3,490 labeled macro-expressions. This dataset includes approximately 80 hours of footage with a resolution of 1280 $\times$ 720 pixels and a image rate of 30 fps. Samples in CAS(ME)^3 part A are categorized into happiness, anger, fear, disgust, surprise, sadness and others.

2 Evaluation Criteria

Given the imbalance across different emotion categories within the dataset, we assess the experimental results using unweighted F1-scores (UF1) and unweighted average recall (UAR). For each emotion category, we calculate the true positives (TP), false positives (FP), and false negatives (FN). As indicated in Eq. (11), the F1-score for each category are determined as follows:

$$UF1_i = \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{1}$$

As shown in equation (12), the UF1 are defined as the averaged values of the F1-score across all emotion classes, C is the total number of categories.

$$UF1 = \sum_{i}^{C} UF1_i / C \tag{2}$$

The UAR is calculated as follows:

$$UAR = \sum_{i}^{C} TP_i / N_i \tag{3}$$

3 Experiments on the CAS(ME)^3 Dataset

We conduct both three-class and seven-class experiments on the CAS(ME)^3 dataset [2]. The seven classes include happiness, anger, fear, disgust, surprise, others and sadness, consistent with the sample annotations provided in the database. In the three-class experiment, the categories were grouped into negative, positive, and surprise. The negative category includes samples from anger, fear, disgust and sadness classes, the positive category consists of "happiness" samples, and the "others" category was excluded from the dataset.

Table 1: Comparison with other methods on CAS(ME)^3 dataset

	Classes	UF1	UAR	AVG
FeatRef [3]	3	0.2875	0.3228	0.3052
STSTNet [4]	3	0.3795	0.3792	0.3794
RCN-A [5]	3	0.3928	0.3893	0.3911
HTnet [6]	3	0.5767	0.5415	0.5591
μ-Bert [7]	3	0.5604	0.6125	0.5865
HSTA [8]	3	0.5930	0.6180	0.6055
HDRCL [9]	3	**0.6423**	0.6065	0.6244
Our method	3	0.6182	**0.6643**	**0.6413**
Baseline [2]	7	0.1759	0.1801	0.1780
Baseline(+Depth) [2]	7	0.1773	0.1829	0.1801
μ-Bert [7]	7	0.3264	0.3254	0.3259
Our method	7	**0.4166**	**0.4172**	**0.4169**

3.1 Experimental Setting

During training, only the CAS(ME)^3 dataset was used, with cross-entropy loss employed to calculate the loss value. The optimizer for parameter updates was set to Adam, with an initial learning rate of 0.005, which gradually decreased as training progressed. We used the leave-one-subject-out (LOSO) cross-validation method, the most common dataset partitioning approach in the micro-expression recognition field, to evaluate ME recognition performance. LOSO cross-validation ensures the model's performance is not influenced by specific subject characteristics.

3.2 Results and Analysis

Our method achieves a UF1 score of 0.6182 and a UAR score of 0.6643 in the three-class recognition task on the CAS(ME)^3 dataset, as presented in Table 1. The UF1 score represents the current state-of-the-art performance in this task. For the seven-class recognition task, the UF1 score is 0.4166, and the UAR score is 0.4172, both of which achieve the best performance. Notably, the performance improvement in the seven-class experiment is more significant compared to the three-class experiment. We attribute this to the fact that in the three-class experiment, the facial movements corresponding to different classes are more distinct. In contrast, in the seven-class experiment, the ME movements within the negative emotion categories (anger, fear, disgust and sadness) are more similar, making them harder to distinguish. Since our method retains more temporal dynamic information, it is more effective for classifying micro-expressions with subtle differences, leading to a more pronounced improvement in the seven-class experiment.

4 Experiments on the DFME Dataset

The experiments on the DFME dataset were conducted following the requirements of competition, named Automatic Micro-Expression Recognition, of the 4th Chinese Conference on Affective Computing (CCAC 2024) competition [10]. We will detail the task specifics, our experimental setting, and recognition results in the following section.

Table 2: Result of "test_A" of DFME dataset

	UF1	UAR	ACC
FeatRef (baseline) [3]	0.3410	0.3686	50.84%
jessica (rank 3) [10]	0.3462	0.3610	46.20%
Xiao el. (rank 2) [10]	0.4067	0.4074	46.41%
Our method (rank 1) [10]	0.4123	0.4210	48.73%

4.1 Task Introduction

The task of the competition [10] is seven-class facial micro-expression recognition using 2,629 micro-expression video samples from 259 subjects in the DFME dataset. The dataset is divided into three parts: a "train_data" set (containing 1,856 samples), a "test_A" set (with 474 samples), and a "test_B" set (with 299 samples). Subjects in these three parts are non-overlapping. Training set data, including the location of apex image and emotion category labels, is provided to all participants at the start of the competition. The test sets are not visible to participants before the evaluation phase. During the evaluation phase, participants are only given test set samples, sample names, and image rate information, but not emotion category labels.

4.2 Experimental Setting

During training, only the DFME dataset was used. The "train_data" is randomly divided into training and validation sets, with a ratio of 4:1 and non-overlapping subjects between the two sets. The model is trained using the training set, with cross-entropy loss used to compute the loss value. The optimizer for parameter updates is set to Adam, with an initial learning rate of 0.005, which decrease gradually during training. Model parameters from each training epoch are stored, and the model parameters with the best validation set performance are selected for testing set classification. The DFME dataset includes micro-expression videos at image rates of 500 fps, 300 fps, and 200 fps. To achieve temporal normalization, videos at different image rates are uniformly sampled at different frequencies, normalizing all video types to 100 fps.

4.3 Experimental Results

For seven-class emotion classification on the "test_A" dataset, the results show UF1 of 0.41 and UAR of 0.42. On the "test_B" dataset, the UF1 is 0.40, and the UAR is 0.40 [10]. As shown in Figure 1, according to the confusion matrix of the validation set test results, there is a significant confusion between micro-expressions of negative emotions such as disgust and anger. This may be due to the similarity in the facial movements associated with micro-expressions of negative emotions. As shown in Tables 2 and 3, our method exhibits significant improvement across all three metrics compared to other methods. We achieved the first-place in the competition (Automatic Micro-Expression Recognition) of the CCAC 2024.

The baseline method, FeatRef [3], recognizes micro-expressions by leveraging feature learning and fusion tailored to specific expressions, using the optical flow matrix between the onset and apex images as input. The third-place team of "test_B" adopts a transfer learning approach, incorporating macro and micro-expression datasets as pre-training samples to alleviate network overfitting. They also employ video motion magnification to

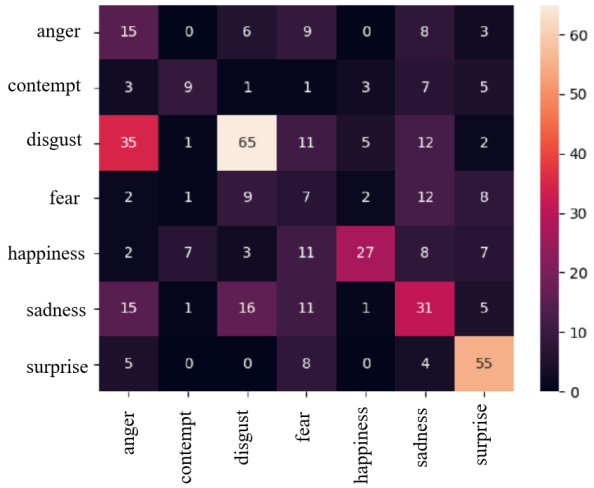

Figure 1: Confusion matrix of the validation set

Table 3: Result of "test_B" of DFME dataset

	UF1	UAR	ACC
FeatRef (baseline) [3]	0.2875	0.3228	36.45%
Zhang el.(rank 3) [10]	0.3356	0.3549	37.79%
Xiao el.(rank 2) [10]	0.3534	0.3661	38.13%
Our method (rank 1) [10]	0.4016	0.4008	41.47%

amplify micro-expression movements and introduce a balanced optical flow feature extraction method to address the issue of low effective information content caused by interference in the optical flow map. The second-place team develops a novel micro-expression recognition framework that utilizes composite optical flow as input to enhance facial muscle movements while mitigating the impact of head pose variations. Additionally, they integrate an efficient attention mechanism within the model to focus on relevant regions and filter out noise, thereby improving recognition accuracy and efficiency. Both the second and third-place teams, along with the baseline method, rely on information from the onset and apex images for model construction and utilize the optical flow method to capture micro-expression movements. This demonstrates the effectiveness of optical flow in detecting subtle micro-expressions and further validates the strength of our approach in preserving more temporal dynamic information by modeling all micro-expression images. On the other hand, for "test_A" dataset and "test_B" dataset, we applied exactly the same model parameters. The experimental results reveal that the recognition performance across the two datasets is quite consistent, with only a slight UF1 decrease of 0.0107. In contrast, the methods by Xiao[10] and FeatRef[3] exhibit a significant drop in performance, with UF1 decreases exceeding 0.04. This demonstrates the strong generalization capability of our approach.

5 Ablation Experiment

To validate the effectiveness of retaining more micro-expression developmental states as model inputs, we conducted an ablation study on the DFME dataset. This study tested the results of seven-class recognition using N images from uniformly sampled micro-

Table 4: Ablation experiment of ME recognition method with fixed-number image sequence as input

Input images	UF1	UAR
4 images	0.3725±0.00035	0.3798±0.00035
8 images	0.3787±0.00043	0.3862±0.00040
12 images	0.3805±0.00058	0.3872±0.00049
16 images	0.3843±0.00049	0.3907±0.00046
All images	0.3877±0.00033	0.3931±0.00029

expressions as model inputs.

5.1 Experimental Setup

Unlike Section 3.4, in this ablation study, "train_data" was used as the training set, "test_A" as the validation set, and "test_B" as the test set. This setup minimizes the impact of random partitioning of the training and validation sets on the experimental results. During training, model parameters from each epoch are saved, and the parameters with the best performance on the validation set are chosen for testing. For each experiment with input length N, we use 30 random seeds and calculated the average experimental results to minimize the effects of randomness. Other experimental settings remained consistent with those in Section 3.4.

5.2 Experimental Results

Table 4 presents the results of experiments using different numbers of images for modeling. It can be observed that as the number of images used for modeling increases, the recognition results improve accordingly. The best results are obtained when using all available images for modeling. This demonstrates that retaining more temporal dynamic information is beneficial for micro-expression recognition. Additionally, it indicates that the proposed self-attention mechanism-based micro-expression temporal feature analysis network can reduce the influence of redundant information in the ME image sequence and effectively handle the long-range temporal dependencies.

References

[1] Zhao S, Tang H, Mao X, Liu S, Zhang Y, Wang H, Xu T, Chen E. Dfme: A new benchmark for dynamic facial micro-expression recognition. IEEE Transactions on Affective Computing, 2024, 15(3): 1371–1386

[2] Li J, Dong Z, Lu S, Wang S J, Yan W J, Ma Y, Liu Y, Huang C, Fu X. Cas(me)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 2782–2800

[3] Zhou L, Mao Q, Huang X, Zhang F, Zhang Z. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognition, 2022, 122: 108275

[4] Liong S T, Gan Y S, See J, Khor H Q, Huang Y C. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). 2019, 1–5

[5] Xia Z, Peng W, Khor H Q, Feng X, Zhao G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Transactions on Image Processing, 2020, 29: 8590–8605

[6] Wang Z, Zhang K, Luo W, Sankaranarayana R. Htnet for micro-expression recognition. Neurocomputing, 2024, 602: 128196

[7] Nguyen X B, Duong C N, Xin L, Susan G, Han-Seok S, Luu K. Micron-bert: Bert-based facial micro-expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023

[8] Hao H, Wang S, Ben H, Hao Y, Wang Y, W W. Hierarchical space-time attention for micro-expression recognition. arXiv, 2405.03202v1

[9] Zhu J, He W, Wang F, Chang H, Cheng L, Zong Y. Exploring holistic discriminative representation for micro-expression recognition via contrastive learning. Image and Vision Computing, 2024, 149: 105186

[10] Ccac2024 technical evaluation task 4: Automatic micro-expression recognition. In: proceedings of The Fourth Chinese Conference on Affective Computing