Multi-scale Dynamic and Hierarchical Relationship Modeling for Facial Action Units Recognition

Zihan Wang[1,2,3], Siyang Song[4*], Cheng Luo[5], Songhe Deng[1,2,3], Weicheng Xie[1,2,3] and Linlin Shen[1,2,3*]
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University,
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society,
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,
[4]University of Leicester, [5]Monash University

1. Details of the employed TCN and SC

We provide the details of the temporal convolution Network (TCN) and the similarity calculating (SC) strategy employed in the proposed MDHR model. While the MFD module computes short-term facial dynamics between neighboring frames, long-term facial dynamics (T frames) are failed to be considered, which may provide additional cues for AU recognition. Consequently, we incorporate channel-wise temporal convolution network into our MDHR framework. For the n_{th} AU, the n_{th} graph nodes corresponding to the T input frames (produced by the GAT layer) are considered as an AU sequence $\hat{V}_n = \{\hat{v}_n^1, \cdots, \hat{v}_n^t, \cdots, \hat{v}_n^T\} \in \mathbb{R}^{T\times b}$. As a result, N AU sequences are obtained, where each is fed to a TCN to model long-term facial dynamics. This way, the AU node sequence $\hat{V}_n$ is updated as:

$$\bar{V}_n = \text{Conv1D}_n(\hat{V}_n) \tag{1}$$

where the $\bar{V}_n = \{\bar{v}_n^1, \cdots, \bar{v}_n^t, \cdots, \bar{v}_n^T\} \in \mathbb{R}^{T\times b}$ denotes the obtained long-term facial dynamic-aware representations of the n_{th} AU for the input T frames; and the kernel size of the temporal convolution operation is 5×1. Then, the similarity calculating (SC) strategy is employed to predict the probability of each AU's occurrence. For the n_{th} AU, a trainable vector s_n that has the same dimension as $\bar{v}_n^t$ is shared across all T frames, based on which the n_{th} AU prediction of the t_{th} frame is made as:

$$p_n^t = \frac{\sigma(\bar{v}_n^t)^T \sigma(s_n)}{||\sigma(\bar{v}_n^t)||_2 ||\sigma(s_n)||_2} \tag{2}$$

where σ is an activation function.

2. Target AUs of each facial region

Table 1 displays the target AUs located in each facial region defined by our approach, which are labelled by either BP4D or DISFA datasets.

Facial regions	Predicted AUs
upper	AU1,AU2,AU4,AU7
middle	AU6,AU9
lower	AU9,AU10,AU12,AU14,AU15 AU17,AU23,AU24,AU25,AU26

Table 1. AU-region mapping rule

3. Dataset label distribution

We provide the AU occurrence label distribution of the employed BP4D and DISFA datasets in Figure 1 and Figure 2. It is clear that for the majority of target AUs, the number of inactivated frames are much more than the frames where they are occurred. Meanwhile, data imbalance also exists between each AU class.

4. Training details

During the training, we only randomly select one image sequence of T frames from each video at each epoch, i.e., not all training examples are used for training at each training epoch. Thus, we train our model for with maximum 200 epochs. In the testing phase, all videos are split into image segments of length T. The number of input frame T, the number of adjacent frames k and the mini-batch size are set to 16, 5 and 8, respectively, for all our experiments. The detailed hyper-parameter settings of two backbone (ResNet and Swin-Transformer)-based best systems on BP4D and DISFA datasets are provided in Table 2. **We used this hyper-parameter setting for training two backbone-based systems on both datasets**, suggesting that our approach is robust. In other words, individually and specifically tuning hyper-parameters for each system on each dataset may lead our approach to achieve even more promising performances.

* Corresponding author

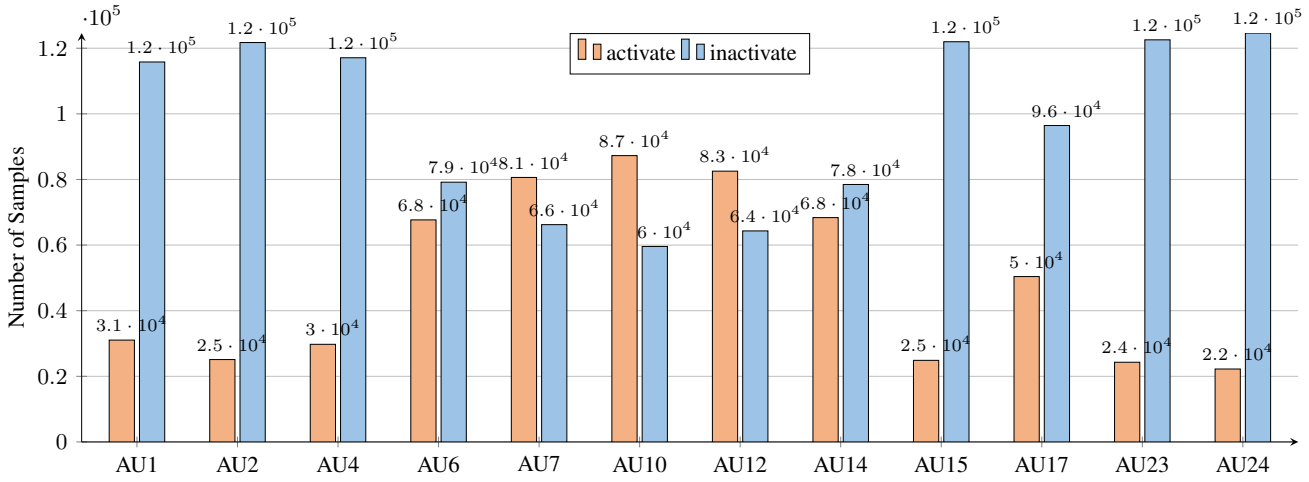

Figure 1. Label distribution on BP4D

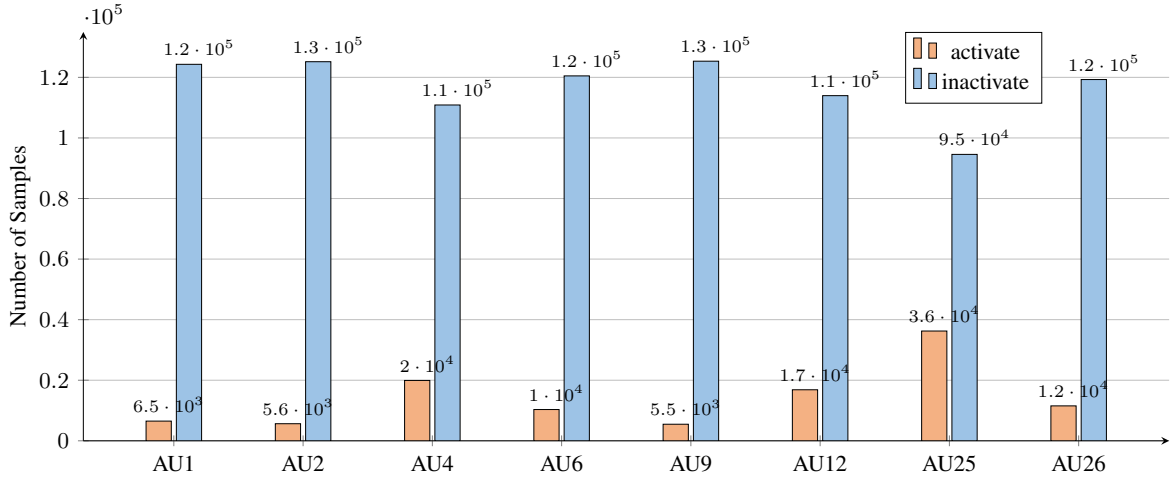

Figure 2. Label distribution on DISFA

parameters	values
Batch size	8
Learning rate	0.0001
Training epochs	200
Validation interval	25
Weight decay	0.0005
Crop size	224×224
T	16
k	5
λ	0.01
β_1	0.9
β_2	0.999

Table 2. hyper-parameter settings

5. Parameter sensitivity analysis

Time-window length k: Table 3 displays the parameter sensitive analysis results in terms of the 'k' on the BP4D dataset. It can be observed that this parameter moderately impacts the performance, with the best F1 results of 66.6% achieved at $k = 5$. Importantly, our model can consistently and effectively capture AU recognition-related temporal cues under most time-window sizes, indicating that our approach is robust and effective. Nevertheless, an appropriate context window sizes allows the model to more effectively capture useful facial dynamics and spatio-temporal cues, while excessive or insufficient adjacent frames provide less meaningful facial dynamic cues.

Loss weighting parameter λ: Table 4 presents the F1-scores achieved by our approach on the BP4D dataset with different λ values, which balances the importance between

k	1	2	3	4	5	6	7	8
F1-score	65.8	65.7	66.2	65.9	66.6	66.5	65.8	65.9

Table 3. Sensitive analysis of k in terms of average F1-scores (in %) achieved on BP4D dataset.

the main task loss and the auxiliary regional prediction loss during training. It can be seen that λ significantly influences model performance. Even using very small λ values (e.g., 0.0001 or 0.001) leads to improved performance compared to $\lambda = 0$, with F1-scores increasing from 65.9 % to 66.2% and 66.3% respectively. This indicates that a slight weighting on regional prediction task helps the model generalize better, where the optimal F1-score of 66.6% is obtained at $\lambda = 0.01$. However, when λ value is too large (i.e., more than 0.05), our model performance starts to degrades. This implies excessive emphasis on the regional prediction over the main task is harmful for the model training.

λ	0	0.0001	0.001	0.01	0.05	0.1	0.2	0.5
F1-score	65.9	66.2	66.3	66.6	66.0	65.6	65.3	64.8

Table 4. Sensitive analysis of λ in terms of average F1-scores (in %) achieved on BP4D dataset.

6. Model complexity analysis

Our model is very efficient as it has low FLOPs per frame, which makes our model very efficient for real applications. We compare our graph-based model with previous state-of-the-art ME-GraphAU model (based on their publicly available code) in Table 5, where our approach has clearly lower FLOPs per frame compared to ME-GraphAU but better AU recognition performance. This is because that the characteristics that ours takes a video clip as input and predicts for all frames together. In summary, our model has a powerful architecture but is still lightweight for video facial analysis. By predicting all frames together, it reduces computations compared to models that process each frame individually.

Method	Params	FLOPs
ME-GraphAU	93.3M	36.0G
MDHR(ResNet50)	91.09M	7.18G
MDHR(Swin-base)	105.94M	15.99G

Table 5. Parameters and FLOPs of our model.

7. Statistical significance analysis

We investigate the statistical significance differences between different variants of our approach by conducting paired t-tests on predictions achieved for the BP4D dataset, where the backbone is the ResNet50. As shown in Table 6, the returned P values of all paired comparisons are lower than the confidence threshold of 0.05. This indicates that: (i) the proposed MFD and HSR brought significant improvements in terms of AU recognition; and (ii) MFD and HSR can encode crucial complementary AU-related cues, leading MDHR to significantly better than Backbone+MFD and Backbone+HSR systems.

Method	Significant difference ?	P-value
Backbone+MFD vs Backbone	Yes	7.75×10^{-3}
Backbone+HSR vs Backbone	Yes	5.43×10^{-3}
MDHR vs Backbone	Yes	6.76×10^{-6}
MDHR vs Backbone+MFD	Yes	8.26×10^{-5}
MDHR vs Backbone+HSR	Yes	1.53×10^{-2}

Table 6. Statistical significance analysis results, where we set the confidence of 0.05.