Supplementary Materials of Upper-body Hierarchical Graph for Skeleton Based Emotion Recognition in Assistive Driving

Anonymous ECCV 2024 Submission

Paper ID #3697

1 Basic Emotion Theories of Body Movement

Darwin was the first to scientifically explore emotions, with his research highlighting the significance of body language and posture in expressing emotions [5]. Despite his groundbreaking work, contemporary emotion recognition systems have largely focused on facial expressions, acoustic cues, and physiological signals, often overlooking body movement. This oversight fails to recognize body movement's pivotal role in non-verbal communication, especially in conveying emotional information during social interactions [1]. Body movement offer several distinct advantages in emotion recognition tasks, making them a valuable tool for affect detection. Firstly, unlike facial expressions or speech, which may require high-resolution cameras or microphones for data capture, body movement can be more readily observed and analyzed. This accessibility is crucial in situations where advanced recording equipment is unavailable or impractical. Secondly, with the recent success of deep learning on large-scale datasets, concerns about privacy protection and ethical issues have started to emerge [7]. Body movement, conveying less identifiable information compared to faces or voices, offers a more privacy-preserving approach to emotion detection. Lastly, research indicates that individuals attempting to conceal their emotions often focus on controlling their facial expressions, neglecting their body movement [6]. This discrepancy suggests that body movement could be a more reliable indicator of suppressed or hidden emotions, as they are less likely to be consciously controlled.

2 Additional Details of UbH-Graph

Universality of UbH-Graph. It is easier to construct our UbH-Graph than existing handcrafted graph [8] even if ours is composed of more edges than the existing one. [8]'s graph requires every physically adjacent edges for human joints as shown in Algorithm 1. On the other hand, our UbH-Graph requires only the hierarchy-wise node sets as shown in Algorithm 2. It verifies that our UbH-Graph is more universal than the existing graph in that the requirements of the UbH-Graph are fewer than those of the existing one.

Algorithm 1 Upper-body Physically Adjacent Graph

Input: Physically adjacent inward edge set $\mathcal{E} = \{e_1, e_2, ..., e_{N_\mathcal{E}}\}$

AIDE:

$\mathcal{E} = \{(14,13), (13,1), (1,12),$
$(1,2), (2,4), (1,3), (3,5),$
$(13,6), (6,8), (8,10),$
$(13,7), (7,9), (9,11)\}$

Output: $\mathbf{A}$;

```
Initialize Adjacency matrix A ∈ ℝ^{3×N×N} to 0
Assign value of 1 to all diagonal components of A^{id} to get identity nodes.
for e to ℰ do
    Centripetal edges: A^{cp}[e] ← 1
    Centrifugal edges: A^{cf}[reverse(e)] ← 1
end for
Initialize degree matrix Λ ∈ ℝ^{3×N×N} to 0
for n = 1 to N do
    Λ ← the number of non-zero elements in column n of A
end for
Normalize adj. matrix with degree matrix: A ← Λ^{-1/2} A Λ^{-1/2}
return A
```

Class Activation Maps. To show how our model works, the activation maps of some skeleton sequences are calculated by class activation map [9], as presented in Fig. 1, in which the activated joints in several sampled frames are displayed. From this figure, we can find that the UbH-GCN model successfully concentrates on the most informative joints.

3 Effectiveness of Four-way Ensemble

Ensemble Coefficients. For most recent models [2–4] have underscored the necessity of selecting optimal ensemble coefficients. These coefficients, which vary from one model to another, are integral in determining the contribution of different data streams—namely joint, bone, joint motion, and bone motion—to the overall model performance. For instance, [2] advocates for ensemble coefficients of [1.0, 1.0, 0.6, 0.6], signifying an equal emphasis on joint and bone streams while assigning a lesser weight to motion streams. Similarly, [3] recommend a different set of coefficients, [0.7, 0.7, 0.3, 0.3], and [4] suggests [0.6, 0.6, 0.4, 0.4], each proposing a unique distribution of emphasis across these streams. This variability in coefficient selection highlights a critical limitation: the lack of universality in these models, necessitating manual adjustment of coefficients to optimize performance. However, our UbH-GCN exclusively utilizes joint and bone streams. By applying an ensemble strategy that assigns equal importance to all four models without distinguishing between different types of data streams, we significantly streamline the model's operation. This approach not only simplifies the model's architecture but also enhances its applicability and efficiency.

Algorithm 2 Upper-body Hierarchical Adjacent Graph

Input: Hierarchical node sets $\mathbf{H} = \{H_1, H_2, ..., H_{N_L}\}$;
AIDE:
$H_1 = \{14\}$,
$H_2 = \{13\}$,
$H_3 = \{1, 6, 7\}$,
$H_4 = \{2, 3, 12, 8, 9\}$,
$H_5 = \{4, 5, 10, 11\}$
Output: $\mathbf{A}$;
Initialize Adjacency matrix $\mathbf{A} \in \mathbb{R}^{(L-1)\times 3\times N\times N}$ to $\mathbf{0}$
for $l = 1$ to $L - 1$ **do**
  H_l and H_{l+1}, include all nodes of those subsets in the diagonal components of the adjacency matrix to get identity nodes: $A_l^{id}[H_l, H_l] \leftarrow 1, A_l^{id}[H_{l+1}, H_{l+1}] \leftarrow 1$
  for $i = 1$ to $length(H_l)$ **do**
    for $j = 1$ to $length(H_{l+1})$ **do**
      Centripetal edges: $A_l^{cp}[H_{l+1}(j), H_l(i)] \leftarrow 1$
      Centrifugal edges: $A_l^{cf}[H_{l+1}(i), H_l(j)] \leftarrow 1$
      Two-hop Centripetal edges: $A_l^{cp^2}[H_{l+1}(j), H_l(i)] \leftarrow A_l^{cp}[H_{l+1}(j), H_l(i)]^2 - A_l^{cp}[H_{l+1}(j), H_l(i)]$
      Two-hop Centrifugal edges: $A_l^{cf^2}[H_{l+1}(i), H_l(j)] \leftarrow A_l^{cf}[H_{l+1}(i), H_l(j)]^2 - A_l^{cf}[H_{l+1}(i), H_l(j)]$
    end for
  end for
  Initialize degree matrix $\mathbf{\Lambda}_l \in \mathbb{R}^{3\times N\times N}$ to $\mathbf{0}$
  for $n = 1$ to N **do**
    $\mathbf{\Lambda}_l[n, n] \leftarrow$ the number of non-zero elements in column n of $\mathbf{A}_l$
  end for
  Normalize adj. matrix with degree matrix: $\mathbf{A}_l \leftarrow \mathbf{\Lambda}_l^{-\frac{1}{2}} \mathbf{A}_l \mathbf{\Lambda}_l^{-\frac{1}{2}}$
end for
return $\mathbf{A}$

Additional Experimental Results. As we mentioned in our main paper, we propose the ensemble method with joint and bone streams without motion streams. Model with each stream is trained with two different UbH-Graphs, which have different rooted nodes; nose and hip. In other words, training ways for our ensemble methods are as follows: (1) joint stream with rooted of nose node, (2) bone stream with rooted of nose node, (3) joint stream with rooted of hip node, (4) bone stream with rooted of hip node. Tab. 1 shows every single experimental result for our four-way ensemble method.

4 Additional Details of Loss Function

The challenge of imbalanced data distribution across various categories is a significant hurdle in machine learning, particularly in scenarios where 'tail' categories—those with fewer instances—suffer due to their features being compressed

Fig. 1: Activated joints in 3 sampled frames of UbH-GCN for the sample emotions of different actions in AIDE Dataset. The red points denote the activated joints, while blue points represent non-activated joints.

Table 1: Classification accuracy and F1-score with different UbH-Graphs in AIDE dataset. ‡: 4-ensemble

Rooted	Stream	Acc.	F1	CG-Acc.	CG-F1
Nose	Joint	72.74	71.13	74.55	73.62
	Bone	74.55	73.01	76.03	74.99
Hip	Joint	74.06	72.78	74.71	74.04
	Bone	73.56	72.16	76.19	75.29
Ensemble ‡		**77.50**	**75.70**	**78.33**	**77.19**

into a constrained region of the feature space. This compression not only diminishes the representational capacity of these categories but also biases the model towards 'head' categories with abundant samples.

To mitigate this issue and promote a more equitable feature distribution, an innovative approach involving the introduction of class variation during the training phase has been used. This method diverges from traditional techniques by not projecting an instance onto a singular feature point. Instead, it maps each instance into a small, designated region within the feature space. This strategic perturbation is meticulously designed to be proportional to the category scale, resulting in smaller variations being allocated to head categories and larger variations to tail categories. By expanding the feature space for tail categories, it facilitates a more nuanced and comprehensive learning of their characteristics, thereby reducing the dominance of head categories.

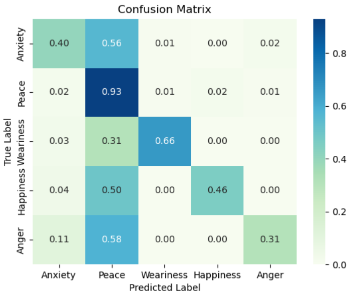

Fig. 2: Confusion matrix for the Ubh-GCN with Cross Entropy Loss Function in the AIDE dataset.

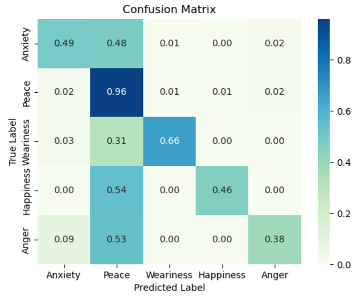

Fig. 3: Confusion matrix for the Ubh-GCN with the new loss function in the AIDE dataset.

We conduct ablation experiments on the loss function, applying both the cross-entropy loss function and our novel loss function. The confusion matrices presented in Fig. 2 and Fig. 3 demonstrate that the method utilizing our new loss function outperforms the cross-entropy loss function in terms of recognition accuracy for emotions such as anxiety, peace, and anger. This, to a certain extent, validates the effectiveness of our approach in addressing the issue of imbalanced data distribution.

References

1. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science **338**(6111), 1225–1229 (2012) 1
2. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13359–13368 (2021) 2
3. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 536–553. Springer (2020) 2
4. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 183–192 (2020) 2
5. Darwin, C., Ekman, P., Prodger, P.: The expression of the emotions in man and animals: Oxford university press. USA,(1872 reprinted 2002) (2002) 1
6. Ekman, P.: Darwin, deception, and facial expression. Annals of the new York Academy of sciences **1000**(1), 205–221 (2003) 1
7. Oh, S.J., Benenson, R., Fritz, M., Schiele, B.: Faceless person recognition: Privacy implications in social media. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 19–35. Springer (2016) 1
8. Shi, H., Peng, W., Chen, H., Liu, X., Zhao, G.: Multiscale 3d-shift graph convolution network for emotion recognition from human actions. IEEE Intelligent Systems **37**(4), 103–110 (2022) 1
9. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016) 2