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

Jiehui Wu[1], Jiansheng Chen[1†], Qifeng Luo[1], Siqi Liu[1], Youze Xue[2], and Huimin Ma[1]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing, China

Abstract. Emotion recognition plays a crucial role in enhancing the safety and enjoyment of assistive driving experiences. By enabling intelligent systems to perceive and understand human emotions, we can significantly improve human-machine interactions. Current research in emotion recognition emphasizes facial expressions, speech and physiological signals, often overlooking body movement's expressive potential. Existing most methods, reliant on full-body poses and graph convolutional networks with predetermined adjacency matrices, face challenges in driving scenarios, including limited visibility, restricted movement and imbalanced data distribution, which affect model generalization and accuracy. To overcome these limitations, we introduce an innovative emotion recognition method tailored for assistive driving. Our method leverages upper-body skeleton sequences, overcoming the constraints of full-body pose capture in driving scenario. Our architecture employs an upper-body hierarchical graph (UbH-Graph) to dynamically capture upper-body movement and emotion relationships. We uniquely incorporate class-specific variations during training, balancing feature distribution and enhancing emotion recognition. Our method outperforms existing multimodal approaches on the assistive driving dataset and demonstrates robustness and adaptability on the daily action dataset. Code is available at https://github.com/jerry-wjh/UbH-GCN.

Keywords: Emotion recognition · Assistive driving · Graph convolutional network · Upper-body movement · Imbalanced data distribution

1 Introduction

With the rapid advancement of artificial intelligence, autonomous driving has become a significant hotspot. Li et al. [21] demonstrated extreme driver emotions, such as anger, have long been one of the leading causes of road accidents worldwide. In the intelligent cockpit, accurately perceiving, understanding, and managing driver's anger through various interactive strategies can effectively minimize the risk of accidents caused by such emotions, thereby improving road

† Corresponding author: jschen@ustb.edu.cn

traffic safety [20]. This highlights the need for advanced emotion recognition methods that can identify and address driver's emotions, ultimately enhancing the safety and efficiency of autonomous driving systems.

Over recent decades, much of the work in emotion recognition has focused on vocal expressions, facial expressions, and electroencephalograms (EEG). However, body movement, one of the most fundamental and natural non-verbal channels in the process of emotional communication [24], has not received equal attention. Recognizing emotions through body movement offers several advantages. First, compared to verbal language and text, body movement exhibits greater universality and cross-cultural consistency [15, 32]. Second, compared to facial expressions, body movement is more likely to express a person's genuine emotions, as gestures are more difficult to suppress without training [28]. Third, compared to EEG, body movement provides a more direct and less interfering method of data collection, preserving the natural behaviour of the subjects [26].

Skeleton sequences, one of the input data types used in emotion recognition methods based on body movement, are considered an intuitive and effective way to represent body movement [23]. Additionally, skeleton sequences include rich spatial and temporal information about body movement [25], allowing us to explore the complex mapping relationship between emotions and full-body motions. With the development of advanced posture estimation algorithms [7, 8, 19, 36], we can easily and precisely extract joint coordinates and rotation angles, further promoting research in emotion recognition based on skeleton sequences.

Previous studies have primarily focused on analyzing full-body movement [6, 17, 33] for emotion recognition. However, visibility is limited to the driver's upper body in the semi-enclosed spaces of cockpits, presenting a unique challenge for movement analysis. Besides, current graph convolutional network (GCN) methodologies [28, 29], which leverage skeleton data for the analysis of body movement and gesture, are limited by their reliance on handcrafted graphs. These graphs primarily focus on the relationships between physically connected (PC) edges in the human skeleton, failing to account for the relationships between distant joint nodes. While graphs focusing on PC edges between joints have semantic significance, their exclusive reliance on these connections leads to a long-range dependency problem. Moreover, some datasets collected from realistic environment exhibit imbalanced data distribution, where a small number of categories dominate the majority of the samples.

Motivated by these limitations, we propose Upper-body Hierarchical Graph Convolutional Network (UbH-GCN) with Upper-body Hierarchical Graph (UbH-Graph). Our approach is designed to overcome the challenges inherent in previous GCN-based emotion recognition methods, particularly those arising from restricted visibility of upper-body movement within the cockpit environment. The framework of our proposed methods is shown in Fig. 1. UbH-Graph contains both meaningful adjacent and distant joint nodes by connecting all the nodes in neighboring hierarchy node sets and identifies the connectivity between those nodes for large receptive field. The architecture is crucial for accurately discerning the subtle dynamics of human emotion recognition, reliant on the analysis

of synchronized movement across different body parts. Notably, certain body parts, such as the hands and the head, although not physically linked, are interconnected through 'invisible edges'. These 'invisible edges' denote non-physical yet impactful relationships, pivotal for conveying emotions via coordinated body movement. The existing ensemble method uses two-stream data composed of the joint and bone streams [28], which are the original joint coordinates and spatial differential between joint coordinates, respectively. We have introduced a novel ensemble method which effectively employs two distinct UbH-Graphs, focusing on joint and bone stream. By leveraging UbH-Graphs in this manner, our method significantly enhances the accuracy of emotion recognition. Additionally, we address the challenge of data imbalance prevalent in naturalistic datasets through the adoption of a specialized loss function. This function, as outlined in [35], dynamically adjusts the weights of samples from various categories, which reduces the disparity between categories in the feature space. It ensures that our UbH-GCN can make precise recognitions across a broader spectrum of emotions, including those less frequently represented in the dataset.

The experimental evaluations conducted on relevant datasets have underscored the effectiveness of our proposed UbH-GCN method. The main contributions of this work are summarized below:

- We propose UbH-Graph to address the inherent limitations of previous GCN-based methods. Our approach enables the detection of relationships between both adjacent and distant joint nodes, a crucial advancement for accurate recognition of human emotions.
- We introduce a novel four-way ensemble method, employing two distinct UbH-Graphs. This method effectively overcomes the challenges faced by models relying solely on motion data, significantly enhancing the accuracy of emotion recognition.
- We use a special loss function added to UbH-GCN, tackling the issue of imbalanced data distribution, a common problem in naturalistic datasets.
- Our UbH-GCN outperforms the state-of-the-arts on the assistive driving dataset and demonstrates adaptability on the daily actions dataset.

2 Related Work

Emotion recognition has become a cornerstone in the advancement of human-computer interaction, drawing considerable interest from researchers aiming to improve user experiences and the develop intelligent systems. The ability of this technology to automatically recognise emotions is crucial to improving the quality of human-computer interaction.

Traditional Approaches and Their Limitations. Historically, the field has seen a reliance on traditional methodologies for emotion recognition through the analysis of body movement and posture. Studies such as those by De et al. [5], Li et al. [18], and Garber et al. [11] have laid the groundwork using techniques that,

while foundational, often depend on manually crafted features which may not capture the full spectrum of emotional expressions conveyed through body movements. These approaches, while pioneering, exhibit limitations in their ability to adapt and generalize across diverse datasets and scenarios.

Advancements through CNNs and RNNs. Recent research has pivoted towards leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Works by Avola et al. [1], Shen et al. [27], and Ilyas et al. [14] represent significant strides in this direction, employing deep learning models to automatically learn features directly from data. These methodologies have shown promise in enhancing the accuracy of emotion recognition. However, they often require extensive computational resources and large labeled datasets for training, posing challenges for scalability and efficiency.

The Emergence of GCN-based Approaches. A novel and less explored avenue within emotion recognition is the application of graph convolutional network (GCN)-based approaches. Shi et al. [28] have demonstrated the potential of GCNs in outperforming traditional and deep learning-based methods by leveraging the structural information of the human body. The approaches construct graphs that model the relationships among body joints, offering a more nuanced understanding of bodily expressions of emotion. Despite their advantages, existing GCN-based methods predominantly focus on handcrafted graphs that emphasize only adjacent, physically connected joints. This focus neglects the significant interactions between more distant joint nodes, limiting the comprehensiveness of emotion recognition.

3 Methodology

In Sec. 3.2, we detail UbH-Graph with fully connected (FC) edges to solve the problems of the conventional graph [3,37] with physically connected (PC) edges. In Sec. 3.4, we introduced how to mitigate the imbalance data distribution present in naturalistic datasets. In Sec. 3.3, we replace the widely used four-stream ensemble method [3, 30] with a new four-way ensemble without motion data streams. Finally, we introduce UbH-GCN, which uses these proposed methods.

3.1 PRELIMINARY TECHNIQUES

Notations. The spatio-temporal graph for human skeleton is represented by $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the joint and edge groups, respectively. $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is the set of N vertices. $\mathcal{E}$ is the edge set, which is formulated as an adjacency matrix $A \in \mathbb{R}^{N \times N}$ and its element a_{ij} reflects the correlation strength between v_i and v_j. Physically connected edges and fully-connected edges used in Sec. 3.2 are denoted as PC-edges and FC-edges, respectively.

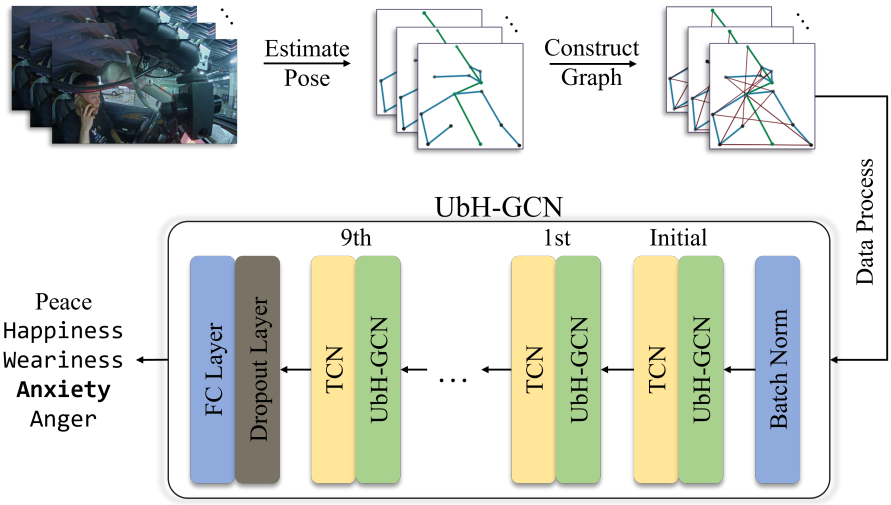


Fig. 1: The pipeline of UbH-GCN. Firstly, the AlphaPose algorithm is employed to extract human skeletal representations from in-vehicle images. Secondly, according to our method, all joint nodes between adjacent hierarchical node sets are connected to generate the UbH-Graph. Subsequently, these graph data undergo preprocessing methods before being inputted into our UbH-GCN approach. Ultimately, the model is capable of distinguishing and outputting five categories of emotions: peace, happiness, weariness, anxiety and anger.

Data Preprocessing. Data preprocessing is essential for skeleton-based emotion recognition. In this work, the input features after various preprocessing are mainly divided into two classes: 1) joint positions, 2) bone features. Suppose that the original coordinate set of an skeleton sequence is $\mathcal{X} = \left\{x \in \mathbb{R}^{C_{in} \times T_{in} \times V_{in}}\right\}$, where C_{in}, T_{in}, V_{in} denote the input coordinates, frames, and joints, respectively. Then the relative position set is obtained as the normalized position features, i.e., $\mathcal{R} = \{r_i | i = 1, 2, ..., V_{in}\}$, where

$$r_i = x[:, :, i] - x[:, :, c], \tag{1}$$

and c represents the index of the center spine joint. Next, the input of joint positions is formed by the concatenation of $\mathcal{X}$ and $\mathcal{R}$. Moreover, the input of bone features consists of the bone lengths $\mathcal{L} = \{l_i | i = 1, 2, ..., V_{in}\}$. The lengths of each bone are calculated by

$$l_i = x[:, :, i] - x[:, :, i_{adj}] \tag{2}$$

where i_{adj} means the adjacent joint of the ith joint.

Graph Convolutional Networks. Skeleton sequences are represented by $X \in \mathbb{R}^{d \times T \times V}$, where T and V are the time step and the number of joints, respectively. GCN's operation with input feature map $F_{in} \in \mathbb{R}^{C \times T \times V}$ is as follows:

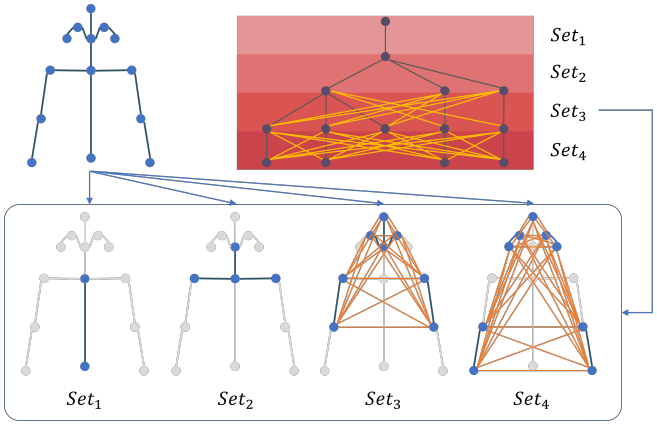


Fig. 2: UbH-Graph. The human skeleton graph is decomposed into a rooted tree, where PC edges are included in hierarchy sets. Edges between all nodes in the same semantic space are obtained by connecting all the nodes in adjacent hierarchy edge sets. Blue and orange lines stand for PC and FC edges, respectively.

$$F_{out} = \sum\nolimits_{s \in S} \hat{\mathbf{A}}_s F_{in} \mathbf{\Theta}_s, \tag{3}$$

where $S = \left\{ s_{id}, s_{cf}, s_{cp}, s_{cf^2}, s_{cp^2} \right\}$ denotes graph subsets, and $s_{id}, s_{cf}, s_{cp}, s_{cf^2}$, s_{cp^2} indicate identity, centrifugal, centripetal, 2-hop centrifugal and 2-hop centripetal joint subsets, respectively. $\mathbf{\Theta}_s$ denotes the pointwise convolution operation. The normalized adjacency matrix $\hat{\mathbf{A}}$ is initialized as $\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{A} \mathbf{\Lambda}^{-\frac{1}{2}} \in \mathbb{R}^{N_s \times V \times V}$, where $\mathbf{\Lambda}$ is a diagonal matrix for normalization.

3.2 Upper-body Hierarchical Graph

Transformation into an N-ary tree. To decompose the upper-body skeleton graph into an N-ary tree, we need to determine a root node that allows nodes from different body parts to coexist in the same set. For example, the left wrist joint and the right wrist joint, or the wrist joint and the ear joint, can exist in the same node set. Once the root node is established, we can convert the skeleton graph into an N-ary tree. This hierarchical structure of the tree allows nodes from different body parts to be categorized, thus extracting the hierarchical information of the graph. Finally, this defines the directed adjacency matrix $\overrightarrow{A} \in \mathbb{R}^{N_L \times V \times V}$.

$$\overrightarrow{\mathbf{A}} = \left[\mathcal{E}(H_1 \rightarrow H_2), ..., \mathcal{E}(H_{N_H - 1} \rightarrow H_{N_H}) \right], \tag{4}$$

where H_k denotes the k-th level node set, and $\mathcal{E}(H_k \rightarrow H_{k+1})$ represents the set of edges from H_k to H_{k+1}. N_L and N_H represent the number of levels in the hierarchical structure and the number of edges on the border of the hierarchical

structure, respectively. Specifically, $N_L = N_H - 1$. However, $\overrightarrow{\mathbf{A}}$ only includes directed centripetal edges. To ensure consistency with existing methods, all reverse edges from the leaf nodes in Fig. 2 to the root node should be included in the adjacency matrix to cover centripetal edges. In addition, the identity matrix for each set of level nodes should be considered to obtain the characteristics of the nodes themselves. Thus, the adjacency matrix $\overleftrightarrow{\mathbf{A}} \in \mathbb{R}^{N_L \times N_S \times V \times V}$ is defined as follows:

$$\overleftrightarrow{\mathbf{A}} = [\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_{N_L}], \tag{5}$$

$$\mathcal{E}_k = \mathcal{E}\left(H_k \bigcup H_{k+1}, H_k \to H_{k+1}, H_{k+1} \to H_k, H_k \xrightarrow{2} H_{k+1}, H_{k+1} \xrightarrow{2} H_k\right), \tag{6}$$

where $\mathcal{E}_k$ denotes the concatenation of the five edge subsets of $S = \{s_{id}, s_{cp}, s_{cf}, s_{cp^2}, s_{cf^2}\}$ and $s_{id}, s_{cp}, s_{cf}, s_{cp^2}, s_{cf^2}$ indicate the identity, centripetal, centrifugal, 2-hop centripetal and 2-hop centrifugal edge subsets, respectively. Through this construction policy, we create a skeleton graph with bidirectional and identity edges.

Fully Connected Edges. UbH-Graph has a different number of edges in its edge set compared to traditional graphs, but the edges themselves are the same. In order to identify the relationships between major long-distance joint nodes, especially those between nodes in different body parts, we connect all nodes between adjacent level node sets.

The graph in [37] only includes the connectivity of PC edges and doesn't contain long-distance relationships, which makes the receptive field of this sparse graph very small. By applying our FC edges to the rooted tree, the graph becomes denser, expanding the receptive field compared to before and making long-distance connections more meaningful. To enhance training adaptability and stability, we normalize the adjacency matrix with the degree matrix and treat all elements in the matrix as learnable parameters.

UbH-Graph Convolution. UbH-GCN architecture includes five parallel convolutional layers, which are specifically designed to extract correlations between human body joints. To reduce computational complexity, all five branch operations utilize linear transformations. For these branches, we employs the same way as Chen et al. [3], performing subset-wise GCN operations on hierarchical edge sets, each containing five subsets of edges. Additionally, we merge the results from the branch that operates with $\mathbf{A}_{s_{id}}$ into the results of the other four branches. Instead of simply summing the output values of these four branches as shown in Eq. (3), we concatenate these output values along the channel dimension in Eq. (7). To enhance the network's learning capability, the concatenated results pass through a batch normalization layer and are added to the input residual. Finally, non-linearity is introduced through a ReLU [12] activation layer.

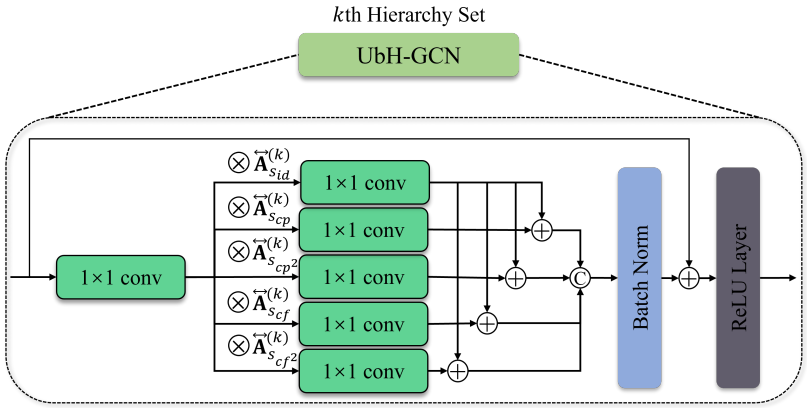


Fig. 3: UbH-Graph convolution operation block. The $\overleftrightarrow{\mathbf{A}}$ is an adjacency matrix, and $s_{id}, s_{cf}, s_{cp}, s_{cf^2}, s_{cp^2}$ indicate identity, centrifugal, centripetal, 2-hop centrifugal and 2-hop centripetal joint subsets, respectively. The $\times$, $+$ and C operations denote matrix multiplication, element-wise addition and concatenation.

$$F_{UbH}^{(k)} = \|_{s \in S} \left\{ \overleftrightarrow{\mathbf{A}}^{k}_{UbH;s} \Phi(F_{in}) \mathbf{\Theta}^{k}_{s} \right\}, \tag{7}$$

$$F_{UbH} \leftarrow \sum_{k=1}^{N_L} (F_{UbH}^{(k)} || F_{in}), \tag{8}$$

where $F_{UbH}^{(k)}$ denotes the output feature map of the UbH-Graph convolution and function Φ denotes a linear transformation with parameter $W \in \mathbb{R}^{C' \times C}$.Note that $\|$ is a concatenation operation.

Our entire GCN process is illustrated in Fig. 3, and the computation is explained in Eq. (8). The branch outputs in the UbH-Graph are linked with the channel dimension and computed in the same way for all branch outputs of the hierarchical edge sets. As the number of joint nodes in each dataset varies, the number of hierarchical sets also differs. To address this, we use an additive strategy for the N_L level outputs and a union strategy for the N_S subset outputs. By using this method, we maintain dimensionality and follow a universal hierarchical set ensemble strategy for each skeletal dataset by adding all outputs with different numbers of hierarchical sets.

3.3 Four-Way Ensemble

Shi et al. [30] demonstrated that a four-stream ensemble method (joint stream, bone stream, joint motion stream, and bone motion stream) can be employed for efficient GCN. However, the motion stream exhibited relatively poorer performance compared to the joint and bone streams. To address this, we utilized

an new four-way ensemble method that only utilizes the joint and bone streams without any motion stream. we integrates four distinct models for two joint streams and two bone streams. Each model was trained with different UbH-Graphs rooted at the hip and nose nodes. By combining these four models, we leverage diverse perspectives, thereby significantly improving the performance of our UbH-GCN as shown in the supplementary. Many recent methods require the selection of optimal ensemble coefficients for specific datasets, leading to low model generalization ability. Our method eliminates the requirement of ensemble coefficient selection by letting each model contribute equally. As a result, our model can be more easily adapted to practical applications.

3.4 Loss Function

Suppose that the final FC layer in our model outputs a logits vector $\mathbf{z} \in \mathbb{R}^Q$, where Q is the number of emotion classes. The classification probabilities for the input sample can be computed by a Softmax function,

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{l=0}^{Q-1} e^{z_l}}, \tag{9}$$

where z_i denotes the ith element of $\mathbf{z}$. Then, the Cross-Entropy (CE) loss is calculated as the objective function for model optimization

$$L_{CE} = -\sum_{i=0}^{Q-1} y_i \log \hat{y}_i, \tag{10}$$

where $y \in \mathbb{R}^Q$ is the one-hot vector indicating the ground truth of emotion class.

The derivation from Eq. (9) to Eq. (10) provides insight into the basic relationship between optimizing the CE loss, which is denoted as L_{CE}, and the logit term, represented by $\mathbf{z}$. It is important to note that the dimension of logit $\mathbf{z}$ corresponds to the total number of categories and directly impacts the calculation of the loss, making it the most intuitive factor that influences the size of the classification feature space. Therefore, the key challenge is to effectively utilize the logit term to mitigate the problem of imbalanced data distribution. Building on the work of [35], this study introduces necessary modifications to the network predictions (i.e., logit $\mathbf{z}$ as referred to here) in Eq. (11).

$$\hat{z}_{i,j} = z_{i,j} + \frac{c_j}{\max_{i=0}^{Q-1} c_i} |\delta(\sigma)|, \; c_j = \log \frac{\sum_{j=0}^{Q-1} q_j}{q_j}, \tag{11}$$

where q_j is the number of the instances with category j and δ is a gaussian distribution with a mean of 0 and standard deviation of σ.

Eq. (11) is designed to adjust the variability within each class by differentially processing categories based on the number of instances. It assigns a smaller variation to the category which has more instances, while a larger variation to the categories with fewer instances. This approach expands the representation of

each instance in the feature space from a single point to a region with a certain range, which helps to balance the inter-class representation during the training process. Additionally, to ensure that predictions during the inference phase are reliable, the variations introduced during training are excluded.

3.5 Network Architecture

Our network architecture, as illustrated in Fig. 1, consists of 1 initial block and 9 GCN blocks that are stacked together. The output channels of each block are 64, 64, 64, 64, 128, 128, 128, 256, 256 and 256, respectively. Each block comes with a residual connection and is divided into two modules - the spatial module, where GCN operations are performed, and the temporal module, which comprises temporal convolutions. For the temporal module, our method utilizes the one from [3], which has four branch operations. Two of these are dilated temporal convolutions with kernel sizes of 5 and dilation rates of 1 and 2, while the remaining branch operations consist of point convolutions with a kernel size of 3 and max pooling. For the spatial module, we use an UbH-Graph convolution operation as detailed Sec. 3.2. To prevent overfitting, a dropout layer with a dropout rate of 0.25 is added after the Global Average Pooling (GAP) layer and before the final Fully Connected (FC) layer.

4 Experiments

4.1 Datasets and Experimental Settings

AIDE. AIDE [38] is an AssIstive Driving pErception dataset aimed at advancing research on vision-driven Driver Monitoring Systems (DMS). AIDE provides rich information from real driving scenarios, capturing both internal and external views of the vehicle through four cameras. This dataset is characterized by its multi-view setup, multi-modal data annotations related to driver features, and multi-task design for comprehensive driving assistance. With 2898 data samples and 521.64K frames, each sample includes 3-second video clips from four perspectives, each aligned with specific perception tasks. Bounding boxes and keypoints are estimated for the internal views. To meet the practical needs of the DMS, AIDE provides fine-grained (FG) criteria classifying coarse-grained (CG) emotions into positive, neutral, or negative categories.

Emilya. Emilya [10] captured by the Xsens MVN system at a frame rate of 120 Hz. This dataset comprises 8,206 emotional posture segments depicting 8 distinct emotions: anxiety (Ax), pride (Pr), joy (Jy), sadness (Sd), panic fear (PF), shame (Sh), anger (Ag), and neutral (Nt). These emotional postures were enacted by 12 actors engaging in 8 common daily actions. Each posture segment includes the 3D positional coordinates and rotational information of 28 keypoints.

Experimental Settings. In our experiments, we adopt [30] as the backbone. The SGD optimizer is employed with a Nesterov momentum of 0.9 and a weight decay of 0.0004. The number of learning epochs is set to 90, with a warm-up strategy [13] applied to the first five epochs for more stable learning. We set the learning rate to decay with cosine annealing [22], with a maximum learning rate of 0.1 and a minimum learning rate of 0.0001. Our experiments are conducted using two distinct datasets: AIDE and Emilya. These datasets are selected for their relevance and the diversity of scenarios they present, which are critical for assessing the robustness of our model. All our experiments are conducted on a single RTX 2080Ti GPU. To evaluate the recognition performance of our model, we employ two metrics: classification accuracy (Acc.) and the weighted F1 score (F1) for the AIDE dataset. For the Emilya dataset, performance is evaluated using either a 3-fold and 10-fold cross-validation scheme.

Table 1: Comparisons of the Accuracy (%) and F1-score (%) against others methods on the AIDE dataset. The best one is in **bold** and the second one is underlined. †: 2-ensemble, ‡: 4-ensemble

Method	Modality	Body	Acc.	F1	CG-Acc.	CG-F1
AIDE [38]	Scene, Face, Body, Gesture, Posture	Full	74.87	72.56	76.52	74.92
CTR-GCN [3]	Posture	Upper	71.92	70.35	73.73	72.81
EfficientGCN-B0 [31]	Posture	Upper	67.65	66.68	71.10	70.49
EfficientGCN-B2 [31]	Posture	Upper	66.50	64.76	72.58	71.36
EfficientGCN-B4 [31]	Posture	Upper	67.49	65.87	70.61	69.98
HD-GCN † [16]	Posture	Upper	73.89	71.06	76.52	75.04
HD-GCN ‡ [16]	Posture	Upper	<u>76.68</u>	<u>74.32</u>	<u>77.50</u>	75.70
UbH-GCN †	Posture	Upper	75.37	73.13	77.34	<u>76.29</u>
UbH-GCN ‡	Posture	Upper	**77.50**	**75.70**	**78.33**	**77.19**

4.2 Comparisons with Other Methods

We uniquely employ upper-body data for emotion recognition, setting our method apart from others that rely on full-body data. This focused examination is presented through comparisons on the AIDE and Emilya datasets. Our method demonstrates superior performance against full-body data methods on the AIDE dataset, as seen in Tab. 1. This achievement is significant, showcasing that upper body data, even without motion streams, can effectively capture driver's emotions. The three action recognition methods [3, 16, 31] are similar to the emotion recognition task based on body movement. We obtain their original codes and evaluate their performance on the AIDE dataset using upper-body data. UbH-GCN outperforms the three action recognition methods mentioned above, which

reinforces the validity of our approach focusing on the upper body. On the Emilya dataset, UbH-GCN closely rivals methods that analyze full-body movement, as detailed in Tab. 2. This result underscores the potential of leveraging upper body information only for emotion recognition.

Table 2: Comparisons of the Accuracy(%) against other methods on Emilya dataset. The best one is in **bold** and the second one is underlined. †: 2-ensemble

Method	Body	Protocal	Acc.
RF-Motion Features [9]	Full	3-fold	75.00
MS-Shift [28]	Full	3-fold	92.00
ST-ITE [34]	Full	3-fold	**93.01 ± 1.81**
UbH-GCN †	Upper	3-fold	91.28 ± 0.36
SVM-$\mathcal{X}^2$ Kernel [4]	Full	10-fold	82.20
Multiscale CNN [2]	Full	10-fold	91.31
ST-ITE [34]	Full	10-fold	**94.42 ± 0.68**
UbH-GCN †	Upper	10-fold	93.98 ± 0.80

4.3 Ablation Study

In this section, we demonstrate the effectiveness of the proposed architecture. Performance is specified as fine-grained accuracy and fine-grained weighted F1 score on the AIDE [38] dataset.

UbH-Graph. To proceed with ablation study for UbH-Graph, we set Yan et al. [3]'s graph as conventional graph. The experiment results are shown in Tab. 3. We set the edges of UbH-Graph in different ways to show a gradual performance increase according to the type of graph. There are two main versions of UbH-Graph, the first of which is graph A containing only the PC edges and the second of which is graph B contains FC edges for N_H hierarchy joint nodes sets.

Table 3: Comparison of the different types of graph on the AIDE dataset. The best one is in **bold** and the second one is underlined. †: 2-ensemble

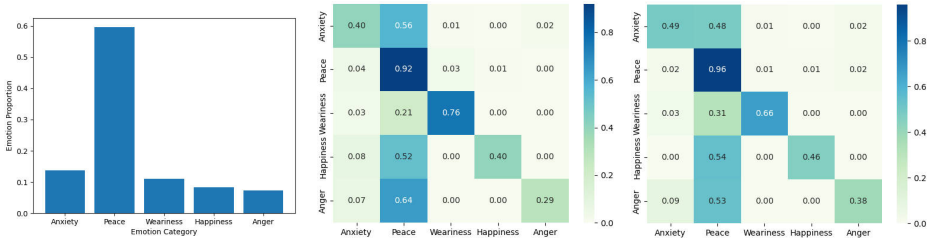
Method	Graph	Edges	Acc.	F1
CTR-GCN [3]	Conventional	PC	71.92	70.35
HD-GCN † [16]	A	PC	73.40	70.63
HD-GCN † [16]	B	FC	73.89	71.06
UbH-GCN †	A	PC	74.55	72.23
UbH-GCN †	B	FC	**75.37**	**73.13**

Four-Way Ensemble. We use the ensemble method by applying it to two graphs with different rooted nodes. Tab. 4 shows UbH-GCN with 4-way ensemble significantly outperforms the method proposed by [16]. This superior performance indicates that the features extracted from different rooted nodes are learned from distinct perspectives, enhancing the overall learning process.

Table 4: Comparisons of the Accuracy (%) and F1-score (%) against others on the AIDE dataset. ‡: 4-ensemble

Method	Loss Function	Acc.	F1
HD-GCN ‡ [16]	cross-entropy	76.03	73.40
HD-GCN ‡ [16]	ours	**76.68**	**74.32**
UbH-GCN ‡	cross-entropy	74.55	72.37
UbH-GCN ‡	ours	**77.50**	**75.70**

Loss Function. Fig. 4a vividly illustrates imbalanced data distribution for the emotion categories of the AIDE dataset, which is characteristic of naturalistic datasets. As shown in Tab. 4, we conduct ablation experiments on HD-GCN and UbH-GCN, applying both cross-entropy and special loss function. On the AIDE dataset, our method surpasses the performance metrics achieved through the conventional cross-entropy loss function. As evidenced by Fig. 4b and Fig. 4c, UbH-GCN demonstrates superior performance over the benchmark model in accurately recognizing four emotions including happiness and anger. The tendency to misclassify weariness as peace is likely attributable to the inherent limitations of emotion recognition from body movement. Despite this, it does not diminish the effectiveness of our method to address the class imbalance challenges.



(a) Percentage of samples in each category in the driver emotion recognition task on the AIDE dataset.

(b) Confusion matrices for the **benchmark** in the driver emotion recognition task on the AIDE dataset.

(c) Confusion matrices for **UbH-GCN** in the driver emotion recognition task on the AIDE dataset.

Fig. 4: Visualization of data distribution and experiments on the AIDE dataset.

Table 5: Comparison of complexity of the single-stream other methods on the AIDE dataset. The best one is in **bold** and the second one is underlined.

Method	Acc.	F1	GFLOPs	Param.
EfficientGCN-B4 [31]	67.49	65.87	0.24	2.02M
EfficientGCN-B0 [31]	67.65	66.68	**0.05**	**0.32M**
CTR-GCN [3]	71.92	70.35	0.14	1.64M
HD-GCN [16]	74.38	72.11	0.12	1.13M
UbH-GCN	**74.55**	**73.01**	0.11	1.04M

4.4 Comparison of Complexity with Other Models

To evaluate the efficiency of our model, we compare UbH-GCN against other methods based on accuracy, F1-score, and model complexity (FLOPs and number of parameters) on the AIDE dataset. By utilizing the original code supplied by these methods, we obtain results concerning model complexity trained on upper-body data where the time window size is fixed at 16. Despite our model employing multiple branched layers with multiple edge sets, its placement prior to the channel reduction layer ensures that it does not cause high complexity. As shown in Tab. 5, our approach demonstrates superior performance on the AIDE single-stream, with model complexity only second to EfficientGCN-B0.

5 Conclusion

In this work, we introduce a tailored emotion recognition approach for assistive driving, overcoming the limitations of traditional systems that rely on facial expressions, speech, and physiological signals. By focusing on the upper-body skeleton sequences, our method addresses the impracticality of full-body pose capture and leverages the emotional expressiveness of upper-body movements. Our novel UbH-Graph dynamically captures nuanced relationships between upper-body movement and emotion, while incorporating class-specific variations during training enhances model generalization and understanding of driver's emotion. Demonstrated improvements on the assistive driving emotion dataset and validation on a daily action dataset highlight our method's robustness and adaptability.

This work significantly contributes to emotion recognition within assistive driving systems and sets the stage for future research into real-time integration, adaptability to in-vehicle conditions, and the inclusion of other non-verbal communication forms. Addressing limitations such as imbalanced data distribution and exploring real-world applicability will be crucial. Ultimately, our research advances the journey towards autonomous vehicles, promising enhanced safety, user experience, and human-vehicle interaction harmony.

Acknowledgement.

This work was supported by the National Science and Technology Major Project (2022ZD0117902). This work was also supported by the National Natural Science Foundation of China (62376024, U20B2062).

References

1. Avola, D., Cinque, L., Fagioli, A., Foresti, G.L., Massaroni, C.: Deep temporal analysis for non-acted body affect recognition. IEEE Transactions on Affective Computing **13**(3), 1366–1377 (2020)
2. Beyan, C., Karumuri, S., Volpe, G., Camurri, A., Niewiadomski, R.: Modeling multiple temporal scales of full-body movements for emotion classification. IEEE Transactions on Affective Computing (2021)
3. 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)
4. Crenn, A., Meyer, A., Konik, H., Khan, R.A., Bouakaz, S.: Generic body expression recognition based on synthesis of realistic neutral motion. IEEE Access **8**, 207758–207767 (2020)
5. De Carolis, B., de Gemmis, M., Lops, P., Palestra, G.: Recognizing users feedback from non-verbal communicative acts in conversational recommender systems. Pattern Recognition Letters **99**, 87–95 (2017)
6. Ezzameli, K., Mahersia, H.: Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion p. 101847 (2023)
7. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
8. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)
9. Fourati, N.: Classification and Characterization of Emotional Body Expression in Daily Actions.(Classification et Caractérisation de l'Expression Corporelle des Emotions dans des Actions Quotidiennes). Ph.D. thesis, Télécom ParisTech, France (2015)
10. Fourati, N., Pelachaud, C.: Perception of emotions and body movement in the emilya database. IEEE Transactions on Affective Computing **9**(1), 90–101 (2016)
11. Garber-Barron, M., Si, M.: Using body movement and posture for emotion detection in non-acted scenarios. In: 2012 IEEE International Conference on Fuzzy Systems. pp. 1–8. IEEE (2012)
12. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
13. 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. pp. 770–778 (2016)
14. Ilyas, C.M.A., Nunes, R., Nasrollahi, K., Rehm, M., Moeslund, T.B.: Deep emotion recognition through upper body movements and facial expression. In: VISIGRAPP (5: VISAPP). pp. 669–679 (2021)

15. Kipp, M., Martin, J.C.: Gesture and emotion: Can basic gestural form features discriminate emotions? In: 2009 3rd international conference on affective computing and intelligent interaction and workshops. pp. 1–8. IEEE (2009)
16. Lee, J., Lee, M., Lee, D., Lee, S.: Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10444–10453 (2023)
17. Leong, S.C., Tang, Y.M., Lai, C.H., Lee, C.: Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. Computer Science Review **48**, 100545 (2023)
18. Li, B., Zhu, C., Li, S., Zhu, T.: Identifying emotions from non-contact gaits information based on microsoft kinects. IEEE Transactions on Affective Computing **9**(4), 585–591 (2016)
19. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10863–10872 (2019)
20. Li, W., Wu, L., Wang, C., Xue, J., Hu, W., Li, S., Guo, G., Cao, D.: Intelligent cockpit for intelligent vehicle in metaverse: A case study of empathetic auditory regulation of human emotion. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(4), 2173–2187 (2022)
21. Li, W., Zhang, B., Wang, P., Sun, C., Zeng, G., Tang, Q., Guo, G., Cao, D.: Visual-attribute-based emotion regulation of angry driving behaviors. IEEE Intelligent Transportation Systems Magazine **14**(3), 10–28 (2021)
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
23. Ma, Q., Shen, L., Chen, E., Tian, S., Wang, J., Cottrell, G.W.: Walking walking walking: Action recognition from action echoes. In: IJCAI. pp. 2457–2463 (2017)
24. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. IEEE transactions on affective computing **12**(2), 505–523 (2018)
25. Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3d skeleton-based action recognition using learning method. Cyborg and Bionic Systems (2020)
26. Saneiro, M., Santos, O.C., Salmeron-Majadas, S., Boticario, J.G., et al.: Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. The Scientific World Journal **2014** (2014)
27. Shen, Z., Cheng, J., Hu, X., Dong, Q.: Emotion recognition based on multi-view body gestures. In: 2019 ieee international conference on image processing (icip). pp. 3317–3321. IEEE (2019)
28. 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)
29. Shi, J., Liu, C., Ishi, C.T., Ishiguro, H.: Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network. Sensors **21**(1), 205 (2020)
30. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
31. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022). https://doi.org/10.1109/TPAMI.2022.3157033, https://doi.org/10.1109/TPAMI.2022.3157033

32. Tracy, J.L., Randles, D., Steckler, C.M.: The nonverbal communication of emotions. Current opinion in behavioral sciences **3**, 25–30 (2015)
33. Wang, J.Z., Zhao, S., Wu, C., Adams, R.B., Newman, M.G., Shafir, T., Tsachor, R.: Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion drawing insights from psychology, engineering, and the arts, this article provides a comprehensive overview of the field of emotion analysis in visual media and discusses the latest research, systems, challenges, ethical implications, and potential impact of artificial emotional intelligence on society. Proceedings of the IEEE (2023)
34. Wang, T., Liu, S., He, F., Dai, W., Du, M., Ke, Y., Ming, D.: Emotion recognition from full-body motion using multiscale spatio-temporal network. IEEE Transactions on Affective Computing (2023)
35. Wang, Y., Fei, J., Wang, H., Li, W., Bao, T., Wu, L., Zhao, R., Shen, Y.: Balancing logit variation for long-tailed semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19561–19573 (2023)
36. Xue, Y., Chen, J., Gu, X., Ma, H., Ma, H.: Boosting monocular 3d human pose estimation with part aware attention. IEEE Transactions on Image Processing **31**, 4278–4291 (2022)
37. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
38. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., et al.: Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20459–20470 (2023)