2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Vision-Based Pedestrian Gesture Recognition System Using Spatiotemporal Features

MD Aminul Islam[1[0009-0001-5985-2166]] and Xiaohui Cui[1*[0000-0001-6079-009X]]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{aminul,xcui}@whu.edu.cn

Abstract. Pedestrians, classified as Vulnerable Road Users (VRUs) due to their lack of protective equipment, face high risks in traffic collisions. While crossing roads, VRUs frequently use communicative gestures such as raising a hand with the palm outward-to signal drivers to stop. Although human drivers can intuitively interpret these gestures, autonomous and driver-assistance systems still lack robust dynamic gesture interpretation, contributing to safety-critical failures. Despite progress in human-vehicle interaction for autonomous driving, prior research has largely emphasized on traffic police signal recognition, pedestrian trajectory prediction, or movement-based intent analysis, neglecting VRUs' explicit gesture-based interactions. To address this gap, we present a systematic taxonomy of pedestrian gesture behaviors, grounded in real-world observations along with a custom dataset. We propose a robust recognition framework that combines spatial feature extraction and deep learning to interpret these gestures. Our method leverages geometric relationships in body keypoints to model spatial patterns, while temporal dynamics are captured using a Long Short-Term Memory (LSTM) network. This architecture processes sequential geometric features to identify distinctive spatiotemporal characteristics of pedestrian gestures. Experiments on our proposed custom dataset and the public CTPG dataset demonstrate a recognition accuracy of 95.18% with near real-time inference speeds, surpassing existing vision-based approaches for VRU gesture recognition.

Keywords: Autonomous driving, Vulnerable Road User, Spatiotemporal features, long-short term memory (LSTM).

1 Introduction

The latest Global Road Safety Report 2023 reveals an annual toll of 1.19 million road traffic deaths, highlighting both the scale of the issue and advancements in safety measures [1]. According to global health statistics, road traffic injuries are the leading cause of death among individuals aged 5–29 and rank as the 12th leading cause of death across all age groups [1]. Among these incidents, vulnerable road users (VRUs) account for 21% of fatalities, while cyclists contribute 5% [1]. In urban environments, safe navigation relies on effective interaction between vehicles, pedestrians, and traffic officers [2]. In traditional driving scenarios, such interactions heavily depend on nonverbal cues [3]. For instance, a pedestrian (VRU) crossing the road may raise one hand-typically

Corresponding author *

with the palm facing outward-to signal vehicles to stop before proceeding. Similarly, a VRU may perform a stop gesture in near-collision situations or use a waving gesture to attract a driver's attention. While human drivers can usually interpret these cues, misunderstandings or failures in recognition can lead to serious accidents. Integrating a system that enables autonomous vehicles to accurately interpret pedestrian gestures could significantly reduce fatal road incidents. However, current autonomous driving systems still struggle to recognize dynamic human gestures, posing a major challenge to achieving seamless human–vehicle interaction. Hence, it is urgent to develop a reliable human-vehicle interaction system that can accurately recognize VRU gesture to reduce the casualties and injuries caused by road accidents.

Current research on human behavior recognition in the context of autonomous driving predominantly focuses on three key areas: traffic police gesture recognition, pedestrian trajectory prediction, and binary intent classification (e.g., "cross" vs."not cross") using body pose analysis. However, significant gaps remain in these approaches. Most studies on traffic police command gesture recognition [4–9] significantly differ from general pedestrian gestures in both context and purpose. On other hand, pedestrian trajectory prediction models [10, 11] often treat pedestrians as rigid objects, relying solely on motion history while overlooking crucial contextual factors such as body language and multi-agent interactions. Similarly, intent recognition frameworks [12–15] frequently neglect subtle non-verbal communication cues, including group dynamics and the use of hand gestures in scenarios like sudden road crossings [16].

Although some VRU gesture-based behavior recognition systems have been proposed in recent years, these systems primarily adopt human action recognition methods to predict pedestrian behavior from gestures [3, 17–21]. However, most of these methods overlook important real-world factors such as the presence of multiple individuals in the scene [17], body orientation [21], and the challenge of detecting specific gestures within cluttered environment [3]. Furthermore, some approaches have been evaluated on pre-segmented video clips, [7, 20], which lack timestamp information. This omission is critical, as real-time autonomous driving systems require continuous video streams with temporal context to make immediate and informed decisions. A major challenge in pedestrian gesture recognition is the lack of comprehensive datasets capturing real-world pedestrian-vehicle interactions. Ethical constraints and the rarity of natural gestures in traffic limit data collection, while existing datasets like JAAD [22] and TASI [23] focus on pedestrian intention prediction but lack active gestures. While staged datasets attempt to address this by simulating gestures in controlled environments, their generalizability to real-world scenarios remains unproven. Bridging this synthetic-to-real gap is essential for developing robust and reliable pedestrian gesture recognition systems.

To address these issues, we propose a real-time system for vulnerable road user (VRU) behavior recognition via gesture analysis. Our framework first identifies gesture performers in a scene, then preprocess the human skeletal data leveraging pose estimation algorithm. We construct spatial geometric feature on extracted skeletal data, and applies a custom Long Short-Term Memory (LSTM) network to model temporal dynamics. By excluding lower-body and facial features, we minimize noise and enhance discriminative spatial-temporal learning. A tracker maintains performer identity across frames,

ensuring continuity. We evaluate our approach on a custom multi-performer dataset (indoor/outdoor settings) and the CTPG dataset [24], which shares gesture classes with traffic police commands. Experiments demonstrate robust performance, highlighting the method's adaptability to pedestrian gesture recognition despite training on staged data. In short, the key contributions of this work are summarized as follows:

- To enhance road safety, research on pedestrian behavior using a gesture recognition system is carried out. This system classifies interactive gesture user on road and interpret their non-verbal gesture-based behaviors, bridging the gap in human-autonomous driving interaction.
- Our lightweight architecture resolves prior limitations by tracking multiple performers in dynamic scenes, achieving 95.18% recognition accuracy with minimal computational resources, even in outdoor settings.
- we introduce a behavior taxonomy derived from real-world observations and a custom dataset capturing diverse environments. Cross-validation on our dataset and public benchmarks CTPG, demonstrates superior generalizability.

The remainder of this paper is organized as follows: Section 2, reviews related work. Section 3, defines the problem. Section 4, details the proposed method. Section V presents the behavior taxonomy, dataset, experimental setup, evaluation metrics, results, limitations and future Finally, Section 6, concludes the paper.

2 Related Work

The autonomous driving industry has greatly benefited from advancements in deep learning and computer vision in recent years. Despite these advancements, pedestrian behavior recognition still faces challenges due to the dynamic nature of pedestrians and frequent occlusions, which result in lost information. While many robust studies have focused on pedestrian trajectory prediction [10, 11] and intention prediction [12–16] based on movement, the inability to account for early gestures at intersections and yield/non-yield signs in uncontrolled areas remains a critical gap in ensuring safety. This gap is particularly significant because gestural communication during road crossing represents a fundamental psychological and cultural practice [25, 26]. Recent experimental research [27, 28] also raises concerns about the effectiveness of systems that disregard pedestrian gestures at intersections or uncontrolled areas, suggesting that considering pedestrians' actions and gestures in studies is crucial for enhancing safety measures. In early efforts, Y. Zheng et al. [18] developed a k-nearest neighbor approach with a pyramid residual module, achieving 92% accuracy on the Udacity dataset. This outperformed conventional HOG methods but lacked temporal analysis. Yang et al. [21] expanded beyond binary pedestrian behavior classification to a more refined nine-category gesture classification using 2D pose estimation and Random Forest, improving performance by 5%. Q. Deng et al. [19] employed Part Affinity Fields for head and neck tracking, effectively recognizing behaviors like path clearing and hand waving in the TASI [23] and JAAD [22] datasets, though performance declined in the presence

of occlusion. Fang and López [29] combined monocular 2D pose estimation and CNNs to predict pedestrian and cyclist intentions through gestures. Xu et al. [20] used a dynamic adaptive graph convolutional network (DAGCN) on 3D pose data for multi-object interaction recognition but faced response time delays. Wiederer et al. [2] introduced a 3D skeleton-based dataset for European traffic control gestures, benchmarking eight deep neural networks across cross-subject (CS), cross-view (CV), and real-world (RW) validations. Bi-LSTM achieved the highest accuracy in CS (87.24%) and CV (87.37%), while LSTM performed best in RW (77.88%). However, their dataset lacks original video sequences. S. Wang et al. [5] proposed a two-stage framework combining upper-body geometric and keypoint co-occurrence features, achieving 89.67% accuracy. Wang et al. [6] later optimized processing time by removing redundant skeletal features, improving LSTM-based methods by 8%. J. He et al. [24] used convolutional pose machines (CPM) with handcrafted features and LSTMs, reaching 0.956% edit accuracy with a 1063 ms response time. Building on previous research demonstrating the effectiveness of Long Short-Term Memory (LSTM) networks in classifying temporal patterns from spatial skeleton keypoints for dynamic gesture recognition, we have adopted LSTM architectures in our work.

3 Problem Formulation

This study aims to enhance the safety of Vulnerable Road Users (VRUs) by enabling autonomous vehicles to recognize and interpret their intent-expressive gestures. Through real-world observations in Wuhan, China, we identified two primary gesture categories directed toward vehicles: yielding gestures (e.g., “Clear”: hand position near hip with palm directed to direction of vehicles, signaling clear to pass) and non-yielding gestures (e.g., “Stop”: vertical arm extension with palm outward, “Slow-down”: straight arm outward, and “Wave”: repetitive waving motion, hand position up shoulder level). later, we have adopted one type of yielding gesture “good” from virtual reality-based experiment [28]. While gestures observed in natural interactions, we note that VRUs consistently orient their body and face toward target vehicles during gesture execution, as illustrated in **Fig. 1**. A gesture $G(t)$ at time t is represented by a set of skeletal keypoints P_t. For each frame t, we extract geometric features, Spatial relationships between these keypoints are encoded through geometrically normalized features (length and angle). which collectively model spatiotemporal gesture dynamics. These features are concatenated into a unified vector and processed by LSTM network to capture temporal dependencies and classify gestures. It is important to note that; our gesture taxonomy reflects regional behavioral patterns observed in Wuhan and may not generalize universally. Cultural and contextual factors could influence gesture semantics across cities or regions. This formulation provides a foundational framework for context-aware gesture recognition in autonomous driving systems.

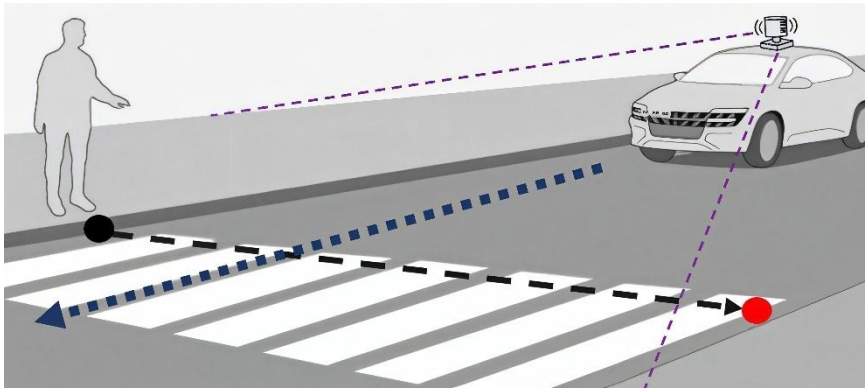


Fig. 1: VRUs gesture instance while crossing the road in uncontrolled area. the black and red point denote current and end position of VRU, black, blue and purple dash line denote the VRU motion, vehicles motion and camera coverage

We formalize the interaction between pedestrian gestures and vehicle control as follows. The pedestrian's position $x_p(t)$ at time t is given by:

$$x_p(t) = x_p(0) + v_p t\,, \tag{1}$$

Where $X_p(0)$ is the initial position and v_p is the pedestrian's constant walking velocity. The vehicle's position $X_v(t)$ is dynamically adjusted based on recognized pedestrian gestures:

$$x_v(t) = x_v(0) + \int_0^t v_v(\tau)d\tau\,, \tag{2}$$

Where $v_v(\tau)$ is the vehicle's time-dependent velocity. This velocity is modulated by gesture recognition outcomes:

$$v_v = \begin{cases} v_v^{\text{reduced}}, & \text{if a VRU perform crossing gesture,} \\ v_v^{\text{normal}}, & \text{otherwise}\,, \end{cases} \tag{3}$$

with $v_v^{\text{reduced}} < v_v^{\text{normal}}$. A pedestrian is deemed to have safely crossed the road when their trajectory satisfies:

$$x_p(t) \geq x_{\text{end}} \tag{4}$$

where x_{end} denotes the pedestrian's target position beyond the vehicle's path. The vehicle's velocity adjustment Δv_v is triggered by gesture recognition within a reaction time Δt :

$$\Delta v_v = -kI(G(t)), \tag{5}$$

Where k is a safety margin constant, and $I(\cdot)$ is an indicator function activated by non-yielding gestures. **Note:** We did not experiment with the decision-making pro-

cess in field tests or simulations integrating our framework with autonomous vehicles. The decision-making model is presented here solely to clarify the intention of our proposed framework.

4 Methodology

Pedestrian behavior recognition based on gestures in the real world is a complex and challenging task. The overall end-to-end pipeline of our proposed model is shown in **Fig. 2**.

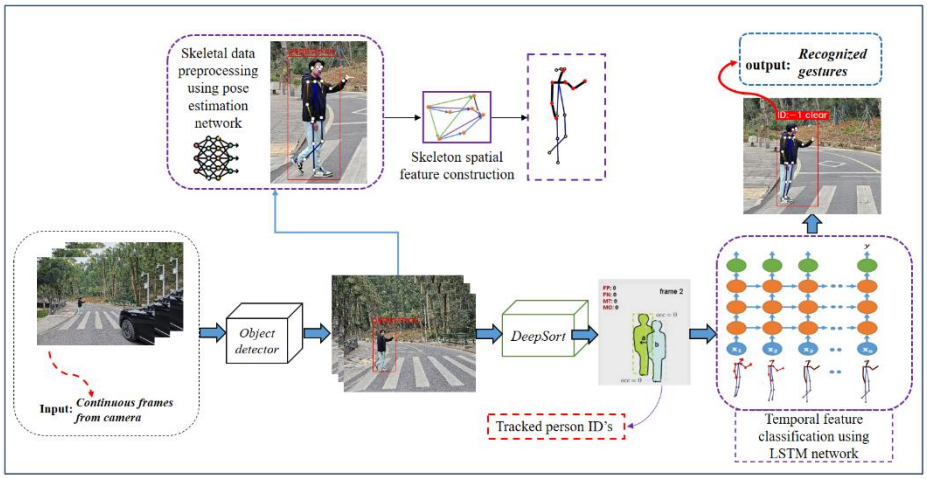


Fig. 2: The overall architecture of our proposed framework based on human skeleton pose estimation. Object detector detect the VRUs, pose estimation algorithm extract skeletal coordinates, skeletal feature construction is performed on each pedestrian in scene for input vector. After feature construction, it passes through the tracker to obtain the bounding box and unique tracker id, inside the bounding box pedestrian gesture is determined. Finally, our LSTM network recognize and classify the gesture based on trained data.

4.1 Pedestrian Detection, Classification and Tracking

Pedestrian Detection: Our framework begins by detecting individuals in the input monocular RGB frame sequences using YOLOv8, a state-of-the-art single stage detector. Known for its effectiveness in detecting small objects and robust performance in real-time applications, YOLOv8 is well-suited for autonomous driving scenarios, where both accuracy and efficiency are critical. Its ability to handle complex environments makes it the preferred choice for our object detection needs.

Pedestrian Classification: Beyond pedestrians, other road actors (e.g., traffic police, cyclists) perform context-specific gestures with distinct semantic meanings. To prevent misinterpretation, we retrain YOLOv8 with three specialized classes:

- **Traffic Police**: Uniformed personnel directing traffic,
- **Cyclists**: Individuals on bicycles,
- **Pedestrians**: VRU in urban street scenes.

Tracking Process: To track VRUs in frame, we use DeepSort[30] algorithm in object detector to ensure precise identification and continuous monitoring. This algorithm leverages a Kalman filter for prediction and a Hungarian algorithm for data association, assigning each pedestrian a unique tracker ID.

4.2 Skeleton Spatial Feature Construction

Skeletal Data Preprocessing: To extract skeletal data for gesture analysis, this work employs OpenPose [31], an open-source pose estimation framework optimized for multi-person detection. Unlike conventional pose estimators, OpenPose demonstrates robust performance in challenging conditions such as partial occlusion, motion blur, and low-resolution imagery. The robustness of OpenPose is shown in **Fig. 5 (c)**. The pose estimation pipeline follows the MS COCO 2D 18-keypoint (0-17) format, which generates a sparse skeletal representation of body configuration.

Spatial Geometric Feature Encoding: Gestures are predominantly characterized by upper-body kinematics, specifically arm movements and torso orientation, with negligible influence from the lower limbs or facial expressions [5]. Therefore, inspired from previous work [6, 24] we select seven upper-body keypoints from the MS COCO [32] (0-17) keypoint model: nose, left and right shoulders, elbows, and wrists to encode the geometric features of the gesture region. Spatial relationships between these keypoints are encoded through geometrically normalized features (length and angle), ensuring translation and scale invariance for robust gesture recognition. Three connection types define the inter-keypoint relationships, as shown in **Fig. 3 (c)**. These connection types are:

1. Adjacency: Direct links between physiologically adjacent joints (e.g., shoulder → elbow → wrist)
2. Arm End: Connections between arm extremities (wrist ↔ ipsilateral shoulder)
3. Chain End: Cross-body links between contralateral keypoints (e.g., left wrist → right wrist)

For each connection *c*, a directional vector is computed:

$$\mathbf{V}_c = \overrightarrow{K_{c1}K_{c2}} = (x_{c2} - x_{c1}, y_{c2} - y_{c1}), \tag{6}$$

Where K_{c1} and K_{c2} denote connected keypoints.
Feature Derivation: Two normalized features are extracted per connection:

1. Length Feature: Scale-invariant limb length normalized by the reference distance d_0 (nose-to-shoulder midpoint):

$$f_l(\mathbf{V}_c) = \frac{\|\mathbf{V}_c\|_2}{d_0}, d_0 = \left\|\mathbf{K}_{\text{nose}} - \frac{\mathbf{K}_{\text{LShoulder}} + \mathbf{K}_{\text{RShoulder}}}{2}\right\|_2 \quad (7)$$

2. Orientation Features: Angular components relative to the vertical axis $\mathbf{n} = (0,1)$:

$$f_{\cos}(\mathbf{V}_c) = \frac{\mathbf{V}_c \cdot \mathbf{n}}{\|\mathbf{V}_c\|_2}, f_{\sin}(\mathbf{V}_c) = \frac{\det(\mathbf{V}_c, \mathbf{n})}{\|\mathbf{V}_c\|_2} \quad (8)$$

where $\det(\mathbf{a}, \mathbf{b}) = a_x b_y - a_y b_x$ computes the 2D pseudo-cross-product.

Feature Vector: Concatenating features across all connections yields a 30-dimensional vector ($10\ connections \times [1\ length + 2\ angle\ features]$). This encoding preserves limb configuration semantics while remaining invariant to absolute position and body scale, enabling reliable distinction between intent-driven gestures (e.g., "Stop" vs. "Wave").

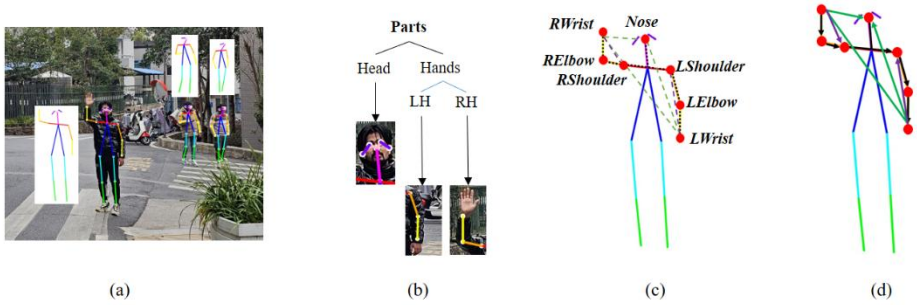


Fig. 4. (a) Generated 18 skeleton coordinates start from 0-17, where nose-0, left ear-17 in each frames, (b) Focused contour of human body where most gestures occur. (c) Connections. black, green, purple lines are three types of connection (d) Feature skeleton vectors are used to calculate the relative length and angle to the unit normal vector (0,1).

4.3 Temporal Feature Classification

We employ Long Short-Term Memory (LSTM) network [33] to model temporal dependencies in gesture recognition using spatial geometric features. The LSTM processes sequential feature vectors while maintaining memory of long-range contextual patterns for continuous gesture interpretation.

Input Feature Structure Let $X = \{x_1, x_2, \ldots, x_T\}$ represent a sequence of T frames, where each frame's feature vector:

$$x_t = \left[f_l^{(1)}, f_{\cos}^{(1)}, f_{\sin}^{(1)}, \ldots, f_l^{(10)}, f_{\cos}^{(10)}, f_{\sin}^{(10)}\right] \in R^{30} \quad (9)$$

encodes normalized length and orientation features from all 10 kinematic connections (Fig. 3d).

Network Architecture The LSTM implementation consists of two layers with 128 hidden units each, governed by the following operations: Forget Gate: Regulates retention of prior cell state C_{t-1} :

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big), W_f \in R^{128\times(128+30)} \tag{10}$$

Input Gate & Candidate State: Control state updates:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{11}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{12}$$

Cell State Update:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{13}$$

Output Gate: Generates hidden state h_t :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{14}$$
$$h_t = o_t \odot \tanh(C_t) \tag{15}$$

Where: σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication, $W_{(\cdot)} \in R^{128\times(128+30)}$ are learnable weight matrices, and $b_{(\cdot)} \in R^{128}$ are bias terms.

Gesture Classification The final hidden state $h_T \in R^{128}$ encodes the complete spatio-temporal gesture context. This is processed through: Feature Projection:

$$z = \mathrm{ReLU}\big(W_{fc}h_T + b_{fc}\big), W_{fc} \in R^{c\times128} \tag{16}$$

Softmax Activation:

$$\hat{y} = \mathrm{softmax}(z) \tag{17}$$

yielding a probability distribution $\hat{y} \in R^c$ over c gesture classes.

5 Experiments

5.1 Pedestrian Interactive Gestures Taxonomy and Proposed Dataset

Pedestrian Interactive Gestures Based Behavior Taxonomy: Existing benchmark datasets like JAAD [22] and TASI [23] focus on pedestrian-vehicle interactions, particularly joint attention between pedestrians and drivers. However, they lack comprehensive representation of interactive gestures, making them insufficient for training or evaluating gesture-based behavior recognition systems. To address this gap, we propose a taxonomy of pedestrian interactive gestures, shown in **Table 1**, developed through real-world observations.

Table 1: Pedestrian Interactive Gesture Based Behavior Taxonomy.

Gesture Type	Gesture Class	Description
Good	Yield to Vehicle	Permission to pass
Clear	Yield to Vehicle	Path is clear to go
Stop	Non-yield to Vehicle	Requesting to stop
Slow-down	Non-yield to Vehicle	Requesting reduced speed
Wave	Non-yield to Vehicle	Gesture to move forward

Proposed Dataset: Our custom video dataset is built based on the proposed gesture taxonomy and consists of recordings made in both indoor and outdoor settings. The recordings were captured using a Nikon video camera and a Samsung cell phone. Seven actors participated in performing gestures, with one actor directing gestures towards the camera while other actors were present in the scene at varying distances. The outdoor environment included both crosswalk and non-crosswalk settings. In the indoor setting, two volunteers performed the gestures. The resulting dataset comprises a total of 1,400 interactive gesture incidents, with each video clip having an image resolution of 1920×1080 pixels and a frame rate of 25 fps. The total number of frames in the dataset amounts to 210,000. On average, each video clip containing a gesture incident lasts for 150 frames, equivalent to approximately 6 seconds. The dataset splitting for training and testing has presented in **Table. 2**.

Table 2. Pedestrian Interactive Gesture Based Behavior Taxonomy.

Description	Training	Testing	Total
Videos	5	5	10
Frames	180,000	105,000	285,000
Gesture Instances	1,200	700	1,900

5.2 Evaluation Metrics and Experimental Setup

Evaluation Metrics Following previous work on VRU gesture-based studies [2, 3, 18, 20], we selected three evaluation metrics: accuracy, F1-score, and the confusion matrix. Accuracy is defined as the ratio of the number of correctly classified samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{Total number of samples}}, \tag{18}$$

F1-score is defined as the harmonic mean of precision and recall:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

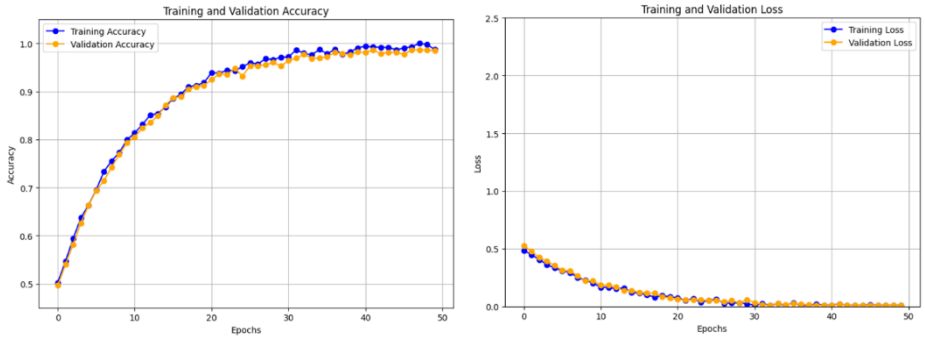
where precision is the ratio of true positives (TP) to the sum of true positives and false positives (FP):

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{20}$$

and Recall is the ratio of true positives to the sum of true positives and false negatives (FN):

$$\text{Recall} = \frac{TP}{TP+FN} \tag{21}$$

Experimental Setup: An overview of the dataset, including the training and testing distributions, is presented in the proposed dataset section and in Table 2. After investigating various combinations of LSTM layers, hidden unit layers, batch sizes, and learning rates, the hyperparameters used for model training are listed below. The input snippet length is 6 seconds, containing 150 frames, with a batch size of 16. The initial learning rate is set to 0.001, the ReLU activation function is used, and the Adam optimizer is applied. The number of hidden units in the LSTM layer is 128. The model is trained for 50 epochs, with an initial exponential decay rate for the learning rate of 0.02. Categorical cross-entropy loss with L2 regularization is employed to prevent overfitting. Additionally, a softmax activation function is used in a fully connected layer to enhance the accuracy of the model's predictions. Validation is performed after each epoch, and the model achieving the highest per-frame accuracy on the validation set is selected. Plots of training-validation accuracy and training-validation loss are presented in **Fig. 4**. The model, implemented using the TensorFlow 2.16.1 framework, has a size of 9.2 MB. In our experiments, an NVIDIA GeForce RTX 3080 GPU and an AMD Ryzen 7 5800H with Radeon graphics were used to train and test the model. With these parameter settings, the average response time of the system was approximately 1 second during online evaluation. Details of the results and time computation are provided in the Experimental Results section.



(a) Model training and validation accuracy (b) Model training and validation loss over-time

Fig. 4: Model training progress over time in custom pedestrian gesture dataset.

5.3 Experimental Results

Experimental Result of the Object Detector: The first stage of our framework focuses on pedestrian detection to prioritize gesture recognition and avoid ambiguity with other road users performing gestures (e.g., cyclists or traffic police). While the original YOLOv8 model detects all human roles under the generic “person” class in the COCO2017 dataset [32], this aggregation limits gesture specific analysis. To address this, we retrained YOLOv8 using a custom dataset suitable for traffic scenarios. The dataset, curated via the RoboFlow platform, comprises hundreds of urban street images featuring pedestrians, cyclists, and Chinese traffic police (wearing yellow reflective vests). Each category-pedestrian, cyclist, and traffic police was manually labeled as distinct classes (Fig. 5b). After retraining, the model achieves precise detection of these roles in traffic scenes (Fig. 5c), unlike the baseline YOLOv8 (Fig. 5a), which detects all human targets as "person."

(a) (b) (c)

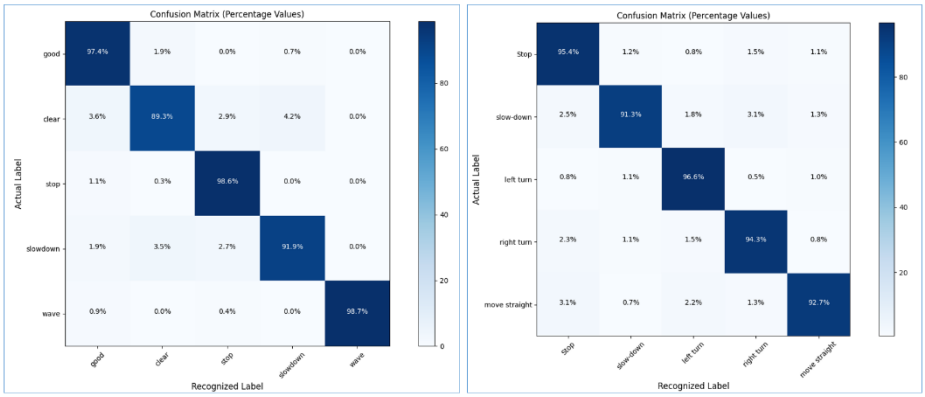
Fig. 5: Comparative analysis of detection performance between the official YOLOv8 baseline model (a), Data annotation process (b) and our customretrained model on proposed dataset video, where pedestrian, cyclist and traffic police are different class, featuring enhanced spatial feature extraction using Openpose (c).

Experimental Results of Pedestrian Gesture: The proposed method was evaluated on testing set from our custom dataset, where each gesture class contained an equal number of samples. As pedestrians typically face approaching vehicles frontally (with

head/body orientation aligned to the road), cross-view recognition is unnecessary and risks misinterpretation by unintended observers. Thus, we adopted a cross-subject evaluation protocol to assess generalization across performers. The details evaluation of our proposed method is presented in **Table 3**.

Table 3: Performance Metrics for Different Classes

Class	Precision (%)	Recall (%)	F1-Score (%)
Good	92.85	97.40	95.07
Clear	94.00	89.30	91.59
Stop	94.26	98.60	96.38
Slow-down	94.94	91.90	93.39
Wave	100.00	98.70	99.35



(a) (b)

Fig. 6: Confusion matrix of both datasets. (a) custom pedestrian gestures, (b) CTPG dataset

The model achieved an overall accuracy of 95.18%, and averaged F1-score of 95.16%. "Wave and stop" (F1 score above 96%) gesture has best performances due to high precision and recall. Weakest Performance: "clear" (F1 = 91%), due to confusion with other classes as shown in **Fig. 6 (a)**. To verify the generalizability of our proposed method, we trained our model on publicly available Chinese traffic police dataset [24] with same experimental setup, and split dataset into five classes of gestures which originally has eight gesture class. The testing result in CTPG [24] dataset model achieved overall accuracy of 94.06% and average F1- score of 94.42%. the accuracy of model dropped in CTPG likely the changes of body direction during performing gestures, confusion matrix of evaluation result presented in **Fig. 6 (b)**. To validate the effectiveness of the proposed framework, we compare it with two state-of-the-art skeleton-based action recognition algorithms [34, 35] under the same experimental setup and input features. Additionally, we conduct a comparative analysis with existing VRU gesture-based behavior recognition methods [3, 18, 20, 21]. The results of these comparisons

are presented in **Table 4**. To the best of our knowledge, previous VRU gesture recognition approaches do not follow similar pipeline as ours. As a result, we did not directly experiment with the methods proposed in [3, 18, 20, 21].

Table 4: Comparison of Methods Based on Modality, Dataset, and Accuracy

Method	Modality	Dataset	Accuracy
Yolov8+OpenPose+LSTM (Ours)	Skeleton	Proposed	95.18%
YoloV8+Openpose+ST-GCN [34]	Skeleton	-	93.87%
YoloV8+Openpose+SGN [35]	Skeleton	-	94.42%
GAST-net+DA-GCN [20]	Skeleton	3D-HPT	95.47%
GAST-net+DA-GCN [20]	Skeleton	Custom	83.21%
CNN+PRM+KNN [18]	Skeleton	Udacity	92.00%
OpenPifpaf+Random Forest [21]	Skeleton	Proposed	82.90%
OpenPifpaf+SVM [3]	Skeleton	Proposed	94.00%

Limitations of the Proposed Framework and Future Work: In this paper, we proposed and demonstrated the feasibility of a pedestrian gesture-based behavior recognition framework for driving assistance systems in traffic environments. While our approach shows promise, its practical implementation still presents certain limitations, which are outlined below:

- During our experiments on a self-constructed dataset, the object detection and tracking processes required an average of 34.2 ms per frame. Additionally, the gesture recognition network took approximately 10 ms to process gestures, resulting in a total processing time of 44.2 milliseconds per frame. although, this processing time satisfy process data in most traffic scenario. The processing time might increase in highly cluttered environment. Higher computational setup will help normalize this limitation.
- Detection error: Our framework leverages the advanced YOLOv8 algorithm for object detection, known for its efficiency and high accuracy in most cases. However, traffic environments present significant challenges due to their complexity. In densely populated areas, the detector may occasionally miss objects or generate incorrect detections, which can impact the overall performance of action recognition.
- Changes of gestures: The types of gesture we included in our taxonomy is typically appeared in real traffic scenario. Changes of gesture by angle lead to misclassification.

In summary, our future work will focus on (1) Exploring more efficient algorithms to enhance computational speed while preserving accuracy. (2) Refining the object detection module to more accurately identify vulnerable road users (VRUs) both on and near the road. (3) Expanding the dataset to include a broader range of gestures with varying rotations in more complex scenarios, improving the framework's robustness and adaptability.

6 Conclusion

In this article, we prioritize pedestrian safety and propose a framework to enhance it. To this end, we developed a comprehensive taxonomy of pedestrian behavior and introduced a custom dataset. Additionally, we designed a robust data-driven pedestrian recognition framework tailored for traffic environments. Our framework improves each module, from object detection and spatial geometric feature extraction in human pose estimation to gesture recognition by incorporating optimizations specifically suited for traffic scenarios. A key element of our approach is the spatial feature construction module, which generates a robust feature vector. This refined representation enables our LSTM-based temporal model to achieve optimal performance in pedestrian behavior recognition. However, performance declines across different datasets have revealed certain limitations. Real-world traffic environments introduce additional challenges, including inaccuracies in human pose estimation in complex urban settings and difficulties in recognizing nonstandard pedestrian gestures. Variations in subject size, viewpoint, and diverse gesticulation styles further complicate gesture recognition. Our future research aims to address these challenges, enhancing the framework's adaptability to real-world traffic scenarios.

References

1. World Health Organization, Global Status Report on Road Safety 2023. Geneva: World Health Organization, 2023, licence: CC BY-NC-SA 3.0 IGO.
2. J. Wiederer, A. Bouazizi, U. Kressel, and V. Belagiannis, "Traffic control gesture recognition for autonomous vehicles," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10 676–10 683.
3. M. Zhang, W. Fei, and J. Yang, "Skeleton-based pedestrian gesture detection for autonomous driving," in 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2023, pp. 1140–1145.
4. S. Ono, A. Kida, T. WATANABE, M. KARG, and Y. SUDA, "Recognition system of hand signals of a police officer for automated driving," SEISAN KENKYU, vol. 72, no. 2, pp. 195–200, 2020.
5. S. Wang, K. Jiang, J. Chen, M. Yang, Z. Fu, T. Wen, and D. Yang, "Skeletonbased traffic command recognition at road intersections for intelligent vehicles," Neurocomputing, vol. 501, pp. 123–134, 2022.
6. S. Wang, K. Jiang, J. Chen, M. Yang, Z. Fu, and D. Yang, "Simple but effective: Upper-body geometric features for traffic command gesture recognition," IEEE Transactions on Human-Machine Systems, vol. 52, no. 3, pp. 423–434, 2021.
7. A. Mishra, J. Kim, J. Cha, D. Kim, and S. Kim, "Authorized traffic controller hand gesture recognition for situation-aware autonomous driving," Sensors, vol. 21, no. 23, p. 7914, 2021.
8. T. Baek and Y.-G. Lee, "Traffic control hand signal recognition using convolution and recurrent neural networks," Journal of Computational Design and Engineering, vol. 9, no. 2, pp. 296–309, 2022.
9. A. J. S. Ong, M. Cabatuan, J. L. L. Tiberio, and J. A. Jose, "Lstm-based traffic gesture recognition using mediapipe pose," in TENCON 2022-2022 IEEE Region 10 Conference (TENCON). IEEE, 2022, pp. 1–5.

10. H. Chen, Y. Liu, B. Zhao, C. Hu, and X. Zhang, “Vision-based real-time online vulnerable traffic participants trajectory prediction for autonomous vehicle,” IEEE Transactions on Intelligent Vehicles, vol. 8, no. 3, pp. 2110–2122, 2022.
11. H. Zhan, Y. Liu, Z. Cui, and H. Cheng, “Pedestrian detection and behavior recognition based on vision,” in 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 771–776.
12. F. Schneemann and P. Heinemann, “Context-based detection of pedestrian crossing intention for autonomous driving in urban environments,” in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016, pp. 2243–2248.
13. I. Kotseruba, A. Rasouli, and J. K. Tsotsos, “Do they want to cross? understanding pedestrian intention for behavior prediction,” in 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020, pp. 1688–1693.
14. A. D. Abadi, Y. Gu, I. Goncharenko, and S. Kamijo, “Detection of cyclists’ crossing intentions for autonomous vehicles,” in 2022 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2022, pp. 1–6.
15. D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, “Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention,” IEEE Transactions on Intelligent Vehicles, vol. 7, no. 2, pp. 221–230, 2022.
16. N. Sharma, C. Dhiman, and S. Indu, “Visual-motion-interaction guided pedestrian intention prediction framework,” IEEE Sensors Journal, 2023.
17. Z. Fu, J. Chen, K. Jiang, S. Wang, J. Wen, M. Yang, and D. Yang, “Traffic police 3d gesture recognition based on spatial–temporal fully adaptive graph convolutional network,” IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 9, pp. 9518 – 9531, 2023.
18. Y. Zheng, H. Bao, and C. Xu, “A method for improved pedestrian gesture recognition in self-driving cars,” Australian Journal of Mechanical Engineering, vol. 16, no. sup1, pp. 78–85, 2018.
19. Q. Deng, R. Tian, Y. Chen, and K. Li, “Skeleton model based behavior recognition for pedestrians and cyclists from vehicle scene camera,” in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1293–1298.
20. F. Xu, F. Xu, J. Xie, C.-M. Pun, H. Lu, and H. Gao, “Action recognition framework in traffic scene for autonomous driving system,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 22 301–22 311, 2021.
21. J. Yang, A. Gui, J. Wang, and J. Ma, “Pedestrian behavior interpretation from pose estimation,” in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 3110–3115.
22. A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 206–213.
23. R. Tian, L. Li, K. Yang, S. Chien, Y. Chen, and R. Sherony, “Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on tasi 110-car naturalistic driving data collection,” in 2014 ieee intelligent vehicles symposium proceedings. IEEE, 2014, pp. 623–629.
24. J. He, C. Zhang, X. He, and R. Dong, “Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features,” Neurocomputing, vol. 390, pp. 248–259, 2020.
25. Ö. Sert, İ. S. Mert, and Ö. Simsekoglu, “The power of hand gestures in traffic: Exploring non-verbal communication and yielding behavior among turkish drivers,” Available at SSRN 4990561.

26. E. Munzlinger, F. Batista Narcizo, D. Witzner Hansen, and T. Vucurevich, “Universal hand gesture interaction vocabulary for cross-cultural users: Challenges and approaches,” in International Conference on Human-Computer Interaction. Springer, 2024, pp. 56–69.
27. M. R. Epke, L. Kooijman, and J. C. De Winter, “I see your gesture: A vr-based study of bidirectional communication between pedestrians and automated vehicles,” Journal of advanced transportation, vol. 2021, no. 1, p. 5573560, 2021.
28. X. Chang, Z. Chen, X. Dong, Y. Cai, T. Yan, H. Cai, Z. Zhou, G. Zhou, and J. Gong, “" it must be gesturing towards me": Gesture-based interaction between autonomous vehicles and pedestrians,” in Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–25.
29. Z. Fang and A. M. López, “Intention recognition of pedestrians and cyclists by 2d pose estimation,” IEEE transactions on intelligent transportation systems, vol. 21, no. 11, pp. 4773–4783, 2019.
30. N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 3645–3649.
31. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 1, pp. 172–186, 2019.
32. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Computer vision– ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer, 2014, pp. 740–755.
33. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
34. S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.
35. P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, “Semantics-guided neural networks for efficient skeleton-based human action recognition,” in proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1112–1121.