[1]Lian Li

[2]Vladimir Y. Mariano

Multi-layer LSTM Traffic Police Gesture Recognition Integrating Limb Angle Features and Attention Mechanism

Abstract: - According to international self-driving technology standards, if self-driving vehicles are to be driven on the road, they must have the function of recognizing traffic police gestures. At present, traffic police gesture recognition methods are mainly divided into three categories, namely recognition based on bioelectric signals, sensor-based recognition, and machine vision-based recognition. This thesis mainly focuses on the situation that traditional machine vision technology easily ignores key coordinates and temporal features when processing dynamic traffic police gestures. This thesis proposes a multi-layer LSTM model that integrates the continuous sub-limb angle and attention models of traffic police. Based on Mediapipe, after unifying key points, the model trained with fusion of angle information has a higher accuracy than the model trained without fusion of angle information, and the model trained with 33 key points and their angle information of Mediapipe is more accurate than 501 key points and their angle information. Finally, based on the model proposed in this thesis, good test results were achieved on the Chinese traffic police gesture data set.

Keywords: traffic police gestures, multi-layer LSTM, continuous sub-limb angle, Mediapipe.

I. Introduction

As an important application in the field of artificial intelligence, autonomous driving technology is leading the revolution in the automotive industry. By combining key technologies such as perception, decision-making and control, cars can achieve driverless driving, bringing higher safety, efficiency, and comfort.

In autonomous driving technology, safety and flexibility are the most important considerations. To ensure the safety of autonomous vehicles, multiple safety measures are required, such as backup sensors and emergency braking. Flexibility is mainly reflected in the understanding of special road conditions, especially the understanding of traffic police on-site command during morning and evening traffic peaks or temporary traffic control situations. Currently, self-driving cars must recognize traffic police gestures before they can be approved for driving on the road. Governments and institutions in various countries are also actively formulating laws and policies related to autonomous driving to ensure its safety.

The global level definitions of autonomous driving are mainly based on the six-level definitions promulgated by the United States Department of Transportation in 2022. They are Level 0-Momentary Driver Assistance, Level 1-Driver Assistance, Level 2-Additional Driver Assistance, and Level 3. -Conditional Automation, Level 4-High Automation, Level 5-Full Automation [1]. Among them, levels four and above require the system to have the ability to dynamically respond to various real-time road conditions. Being able to recognize traffic police gestures is a necessary condition for achieving Level 4 of autonomous driving. When China encounters situations such as natural disasters, severe weather conditions, or major traffic accidents that seriously affect traffic safety, and it is difficult to ensure traffic safety by taking other measures, the traffic management department of the public security organ can implement traffic control [2].

To sum up, the recognition of traffic police gestures has become a necessary production indicator for autonomous vehicles.

II. Related research

Currently, there are many methods for traffic police gesture recognition. Traffic police gesture recognition can be divided into three categories based on different signal sources from the traffic police. The first type is traffic police gesture recognition based on bioelectric signals. This method requires the traffic police to wear bioelectric signal (mostly myoelectric signals) collection equipment and based on the signal data templates of different types of gestures, the currently acquired bioelectric signals are Data identification, this method requires the traffic police to wear signal collection equipment that directly contacts the skin, and the traffic police experience is not good [3]. The second category is sensor-based traffic police gesture recognition. For example, many modern smart devices

[1] College of Computing & Information Technologies, National University, Philippines.
School of Big Data and Artificial Intelligence, Anhui Xinhua University, China.
[2] College of Computing & Information Technologies, National University, Philippines.
Corresponding author: Lian Li

are equipped with inertial sensors (IMU). Because inertial sensors can capture rich motion information, the application sensor can capture the continuous movements of traffic police gestures [4]. The third category is traffic police gesture recognition based on machine vision. This type can be divided into three situations, namely recognition based on image features of traffic police pictures, recognition based on key node information of traffic police limbs, and recognition based on combination of image features and body node information. identification. Common machine vision processing methods include DTW (Dynamic Time Warping), RCNN (Region-CNN), FastRCNN, SSD (Single Shot MultiBox Detector), MobileNet, SSDMobileNet, YOLO, etc. [5]. Technology based on machine vision can alleviate the traffic police only need to perform the command gestures normally according to the road conditions. They do not need to wear equipment, and they do not need to always worry about the equipment falling off or being damaged, which will cause psychological pressure. At the same time, the application cost is also greatly reduced.

The research idea of this thesis is to locate the traffic police position based on the target detection method, then use the limb node information component to obtain the traffic police's main limb node coordinates, extract the angles of each connected sub-limb when executing the command gesture, and finally input the features after normalization The multi-layer LSTM model is used to train the feature data, and finally simulation recognition is performed based on the trained model.

III. METHODOLOGY

The research ideas of this thesis are as follows. The first step is to obtain the data set, the second step is targeting detection, the third step is feature engineering, the fourth step is to perform traffic police gesture recognition based on multi-layer LSTM, and the fifth step is to apply the model.

A. Data

The data set selected in this thesis is the CTPGD(Chinese Traffic Police Gesture Dataset, https://github.com/zc402/ChineseTrafficPolicePose#police-gesture-dataset), which contains videos of Chinese traffic police command gestures and gesture label data for each video frame [6]. There are a total of eight gestures set in this data set. The eight Chinese traffic police command gestures are Stop(stop), Go straight(straight), Turn left(left), Turn right(right), Prepare for turning left(preleft), Slow down(slowdown), Change lanes(change), and pull over(aside). Fig 1 presents key frame sequences of eight traffic police command gestures.

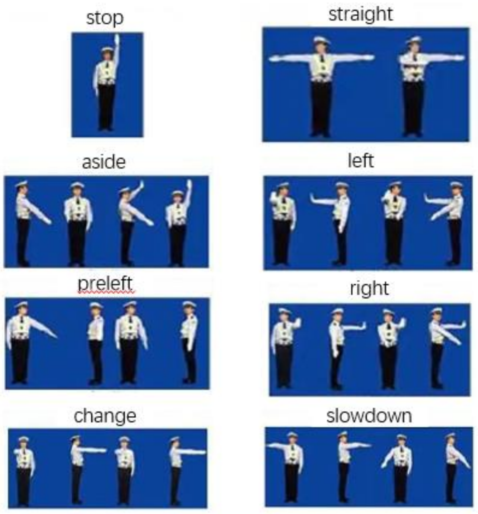


Fig 1 Eight types of traffic police gesture action key frame sequences

B. Target Detection

The recognition of traffic police gestures must have a prerequisite, which is to identify the specific location of the traffic police in a complex background. This process can be achieved through target detection methods. The purpose of target detection is to mark the area of the target to be recognized in a frame of image. At present, target detection methods are mainly divided into two categories, one is one-step detection, and the other is two-step detection. The difference between the two is RPN (Region Proposal Network). The two-step method uses RPN and has high detection accuracy. The one-step method does not use RPN and has fast detection speed. The representative technologies of the two-step detection method include RCNN/Fast RCNN/Faster RCNN. They need to generate a series of candidate frames as samples, and then perform classification and regression through the network. The representative technologies of the one-step detection method include YOLO/SSD/Retina-Net, one-step The method gives location information and target categories through the backbone network [7].

This thesis attempts to use three methods: Fast RCNN, YOLO5 and SSDMobileNetV2 to detect traffic police. In order to better meet the needs of identifying traffic police in different postures, scenes, and road conditions, this thesis also crawled 700 traffic police pictures from Baidu image search, and then created labels for the data through the ImageNet tool component. The defined categories include passers-by and traffic police. category. After data annotation, this thesis finally annotated 700 traffic police information and 700 passerby information. The training set and the test set are divided into a ratio of 7:3.

This thesis applies Faster RCNN, YOLO5 and SSDMobileNetV2 to the above-obtained data sets to locate traffic police. The evaluation indicators selected in this thesis include FPS and accuracy, The result of the comparison about the three target detection methods on this data set is shown in Table 1.

Table 1 Comparison of three target detection methods on this data set

Methods	FPS	Accuracy
Fast RCNN	17.58	0.83
YOLO5	98.74	0.81
SSDMobileNetV2	52.76	0.86

Faster RCNN is two-step method, which takes relatively longer than the one-step method. YOLOv5 and SSDMobileNetV2 are both classic algorithms in the field of target detection [8]. The main differences are shown in Table 2.

Table 2 Comparison of YOLOv5 and SSDMobileNetV2 algorithms on this data set

Metrics to compare	YOLOv5 and SSDMobileNetV2
Detection speed	is faster than SSDMobileNetV2 and can achieve real-time detection.
Accuracy	SSDMobileNetV2 performs better than YOLOv5 on the crawled traffic police data set.
Detection method	SSDMobileNetV2 uses multi-scale feature maps for detection, while YOLOv5 uses the FPN (Feature Pyramid Network) structure based on the backbone network to achieve multi-scale feature extraction and fusion.

To sum up, SSD and YOLOv5 have their own advantages and characteristics. After the target detection algorithm locates the traffic police, this thesis will also use the MediaPipe component to extract the body information of the marked traffic police for analysis. Therefore, this thesis selects SSDMobileNetV2 for traffic police detection and area marking. Fig 2 and Fig 3 show the results of identifying and labeling traffic police in single-type crowd background and multi-type crowd background respectively based on the SSDMobileNetV2 model. “tp” is the abbreviation of traffic police, and “cp” is the abbreviation of citizen person.

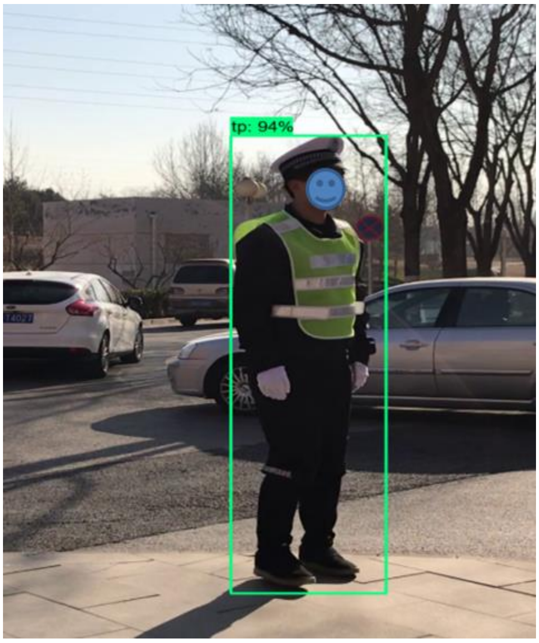


Fig 2 Traffic police detected.

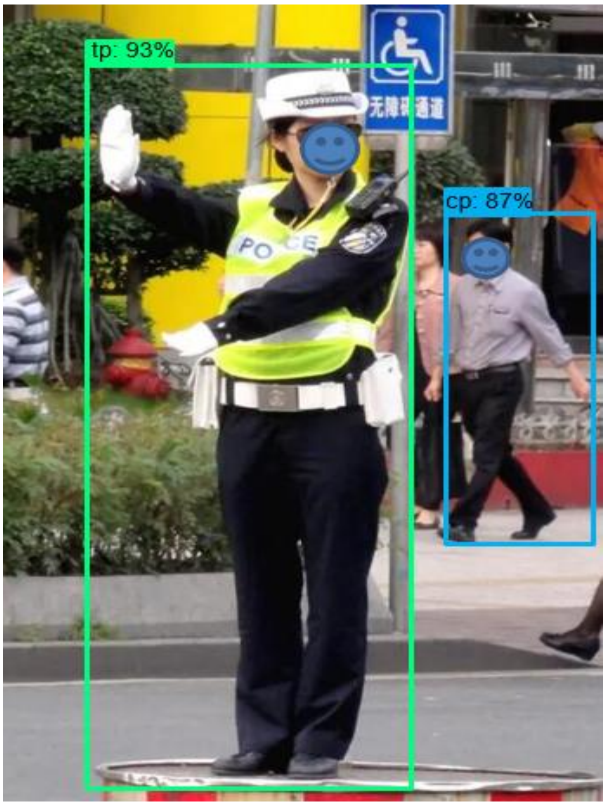


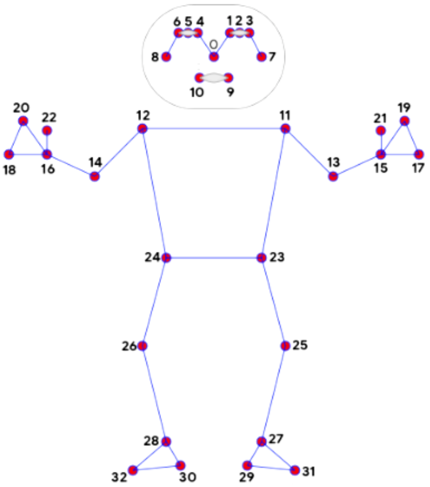
Fig 3 Results of detecting traffic police and pedestrians.

C. Traffic police gesture recognition based on multi-layer LSTM integrating limb angle features

1) Limb node information component.

Mediapipe is a framework for building machine learning pipelines that can help users process time series data such as video and audio. Mediapipe includes 16 technical solutions, including face detection, Face Mesh, iris, hand, posture, human body, character segmentation, hair segmentation, target detection, Box Tracking, instant Motion Tracking, 3D target detection, feature matching, AutoFlip, MediaSequence and YouTubeBe_8M [9].

The different gestures of the traffic police are reflected in the data as the coordinate changes of the key points of the traffic police's body, so this thesis selects the posture key points of Mediapipe to support the recognition of the traffic police's gestures. Mediapipe's posture key points have a total of 33 coordinate information. It marks the human body model with a total of 33 key points from 0 to 32. For example, 6 corresponds to the outer corner of the right eye (right_eye_outer), 13 corresponds to the tip of the left elbow (left_elbow), 30 corresponds to the right heel (right_heel) and so on. The key points of the head are from 0 to 10, the key points of the left side of the body are odd numbers from 11, 13 to 31, and the key points of the right side are even numbers from 12, 14 to 32. The body parts corresponding to the two are symmetrically marked (such as Fig 4).



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

Fig 4 Mediapipe key points of human posture

2) Mediapipe key points of human posture.

When the traffic police perform gestures, the differences between different gestures are mainly concentrated in the following characteristic points: the angle between the left arm and the forearm, the angle between the right arm and the forearm, the angle between the arm and the horizontal or vertical direction, etc. The angular core characteristics of this gesture are described in Table 3 below.

Table 3 Angular macroscopic characteristics of each gesture

NO.	Traffic gestures	Macroscopic characteristics of angle
1	Stop	The angle between the left upper arm and the horizontal direction must be greater than 0 degrees
2	Straight	The angle between the right upper arm and forearm gradually increases from 0 degrees to 150 degrees
3	Aside	The angle between the left upper arm and the horizontal direction must be greater than 0 degrees, and the angle between the right lower arm and the vertical direction gradually increases from 0 degrees to 60 degrees.
4	Left	The angle between the right arm and the vertical direction must be greater than 80 degrees, and the angle between the left forearm and the vertical direction gradually increases from 0 degrees to 60 degrees
5	Preleft	The angle between the left upper arm and the forearm should be kept at 180 degrees, and the distance between the left palm and the ground should be
6	Right	The angle between the left arm and the vertical direction must be greater than 80 degrees, and the angle between the right forearm and the vertical direction gradually increases from 0 degrees to 60 degrees
7	Change	The angle between the right forearm and the vertical direction gradually increases from 0 degrees to 60 degrees
8	Slowdown	The angle between the right arm and the horizontal direction gradually increases from 0 degrees to 60 degrees

Based on the above Table 3, this thesis designs multi-angle fusion features. The specific contents are shown in Table 4.

Table 4 Multi-angle features

Serial number	Traffic police hand gestures	The angle between the left forearm and upper arm	The angle between the right forearm and upper arm	left upper arm and vertical direction	left forearm and vertical direction	right arm and vertical direction	Angle between right forearm and vertical direction
1	Stop	160-180	160-180	<10	<10	170-190	170-190
2	Straight	160-180	180-10	80-100	80-100	-	270-90
3	Aside	160-180	160-180	<10	<10	180-135	180-135
4	Left	160-180	160-180	180-225	180-225	260-280	260-280
5	Preleft	160-180	-	260-280	260-280	-	-
6	Right	160-180	160-180	80-100	80-100	180-135	180-135
7	Change	160-180	160-180	170-190	170-190	180-135	180-135
8	Slowdown	160-180	160-180	170-190	170-190	270-235	270-235

The coordinates of each key point can be obtained instantly according to the Mediapipe component. After the vector formula is obtained from the coordinates [10], various angles of the traffic police gesture can be obtained according to the angle formula between the two vectors, which can refer to Fig 5 and (1).

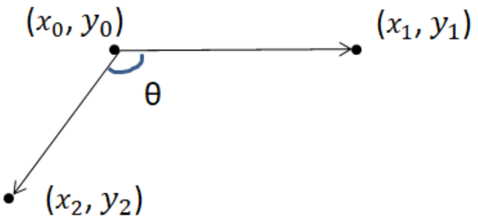


Fig 5 Three coordinates determine an included angle

$$\cos\theta = \frac{(x_1-x_0)(y_2-y_0)+(x_2-x_0)(y_1-y_0)}{\sqrt{(x_1-x_0)^2+(y_1-y_0)^2}\sqrt{(x_2-x_0)^2+(y_2-y_0)^2}} \tag{1}$$

3) *Traffic police gesture recognition based on optimized multi-layer LSTM.*

a) LSTM.

Because the traffic police's gestures are a dynamic process with a fixed trajectory, the time series characteristics of different gestures can be used, so this thesis selected the LSTM model [11] as the basic model. LSTM has made great improvements to RNN, and to a certain extent overcomes RNN's problems with gradient disappearance and gradient explosion.

b) Multi-layer LSTM

To ensure the robustness of the model, the input sequence length of the LSTM network is set to be variable. This thesis sets 60 frames as the standard length of a segment as the input of the LSTM network. The first three layers of the multi-layer LSTM structure used in this thesis are LSTM layers, which are used to extract features of the input sequence. The next three layers are fully connected layers, which classify feature sequences. The first five layers of the activation function use ReLU, and the last layer uses Softmax. After the first three layers of LSTM networks extract temporal features, the feature data will flow into the three-layer fully connected neural network.

c) Multi-layer LSTM model integrating attention mechanism.

In fact, Mediapipe's 33 human limb node information is not average in its ability to represent traffic police gestures. For example, the representation of arm and face orientation has a greater impact on the expression of the physical meaning of the traffic police's gestures. Before neural network training, the attention mechanism can be used to increase the influence of key input information and improve the efficiency of the neural network. The influence here is mainly achieved by "weight".

This thesis adopts the key-value attention model [12], then N input information can be expressed as (KEY, VALUE) = [(KEY1, VALUE1), (KEY2, VALUE2), (KEY3, VALUE3)…(KEYN, VALUEN)], where KEY is used to calculate attention distribution, and VALUE is used to calculate aggregate information. The attention structure selected in this thesis is shown in Fig 6 below. It is mainly divided into three stages, namely similarity calculation, normalization, and weighted summation.

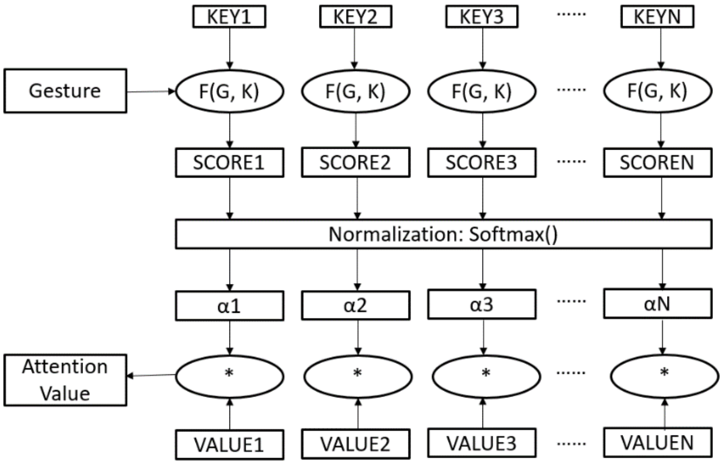


Fig 6 Apply key-value attention structure.

This thesis introduces an attention mechanism that can pay attention to the key frames of gesture changes. During the training process, more weights will be automatically assigned to the key frame sequences of gesture changes. This thesis attempts to integrate LSTM with the attention mechanism. The main process is as follows.

In the first stage, the output of LSTM is defined as the input of the Attention layer, and the similarity between the two is calculated based on Gesture (traffic police gesture type) and KEY (human body key point sequence) to obtain the original attention score. There are many ways to calculate similarity, and this thesis chooses the additive model to implement it. The specific calculation formula is as follows. The higher the score, the higher the attention received, and the greater the final weight assigned to it, where W is the weight coefficient and b is the offset.

$$SCORE_t = F(G, k_t) = \tanh\,(W_i H_t + b_i) \tag{2}$$

In stage two, normalization is used to process the attention scores obtained in stage one. This thesis uses the Softmax function to map the results to 0~1 to obtain the weight coefficient. The specific formula is as follows.

$$\alpha_t = softmax(SCORE_t) = \frac{\exp\,(SCORE_t)}{\sum_{j=1}^{N} \exp\,(SCORE_j)} \tag{3}$$

In stage three, the weight coefficients obtained in stage two are weighted and summed to VALUE (equal to KEY, which is the human body key point sequence). The specific formula is as follows.

$$Attention\big((KEY, VALUE), G\big) = \sum_{t=1}^{N} \alpha_t value_t \tag{4}$$

The LSTM-Attention structure designed in this thesis is shown in Fig 7 below.

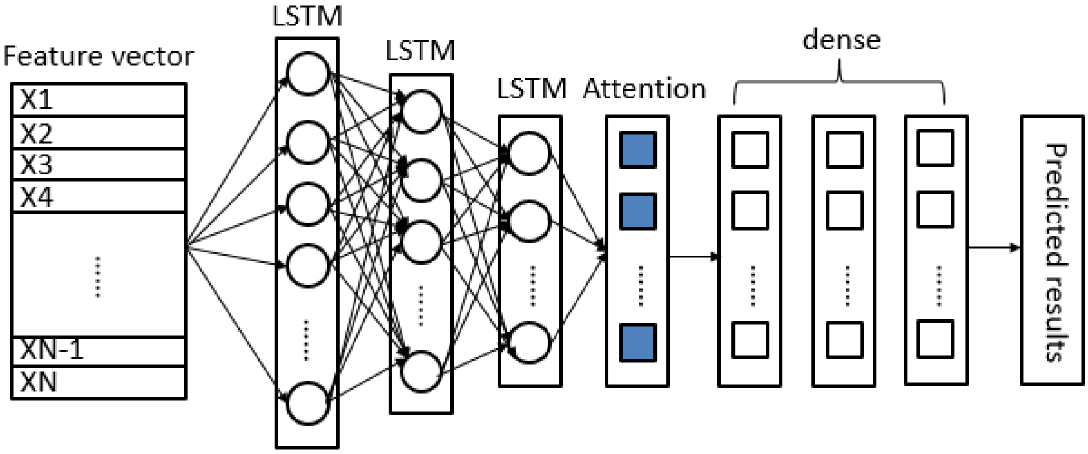


Fig 7 Traffic police gesture detection model based on Attention-multilayer LSTM.

4) Result analysis.

The traffic police gesture feature in this thesis requires the coordinate change features of 33 key points of the body. Comprehensive features are built based on these 33 key points, including the coordinate information of the 33 key points, the angle between the left forearm and the upper arm, the angle between the right forearm and the upper arm, the angle between the left upper arm and the vertical direction, the angle between the left forearm and the left forearm The angle between the vertical direction, the angle between the right arm and the vertical direction, and the angle between the right forearm and the vertical direction. The above comprehensive features are fed into the LSTM network after integrating the attention mechanism for training.

Because the change period and feature arrangement order of each gesture template are consistent, the coordinates and specific angle features need to be normalized when building the template, so that the amplitude of the sequence on the coordinate axis will be different. Regarding the normalization of coordinates, this thesis scales the data according to a certain ratio so that the range of the data is between 0 and 1. The normalization operation here mainly includes two steps, namely calculating the minimum value and range, and scaling each data point. The specific implementation is as follows.

Step 1: Calculate the minimum value and range, first you need to find the smallest x and y values in all data, and the range of x and y (i.e. the maximum value minus the minimum value). You can iterate through each data point to find the minimum value and range, or you can use the min, max, and ptp functions in the numpy library to calculate it.

Step 2: Scale each data point. For each data point (x, y), you can use the following formula to scale it to between 0 and 1:

$$\text{new\_x} = (\text{x} - \text{min\_x}) / \text{range\_x} \tag{5}$$

$$\text{new\_y} = (\text{y} - \text{min\_y}) / \text{range\_y} \tag{6}$$

Among (5) and (6), min_x and min_y are the minimum values of x and y calculated in step 1, and range_x and range_y are the ranges of x and y. Through such processing, the plane coordinates can be normalized to a value between 0 and 1 to facilitate subsequent calculation and analysis.

Regarding the normalization of each angle, because the angle of the core limbs when performing actions is between 0 and 360 degrees, the normalization of the angles in this thesis is to directly divide the angle by the maximum value, which is 360.

After the above results are classified, they will be imported into the multi-layer LSTM network fused with Attention. You can get traffic police gesture classifiers for different directions. The training process of the classifier is given, including the training accuracy change curve and the training loss change curve, as shown in Fig 8 and Fig 9 respectively.

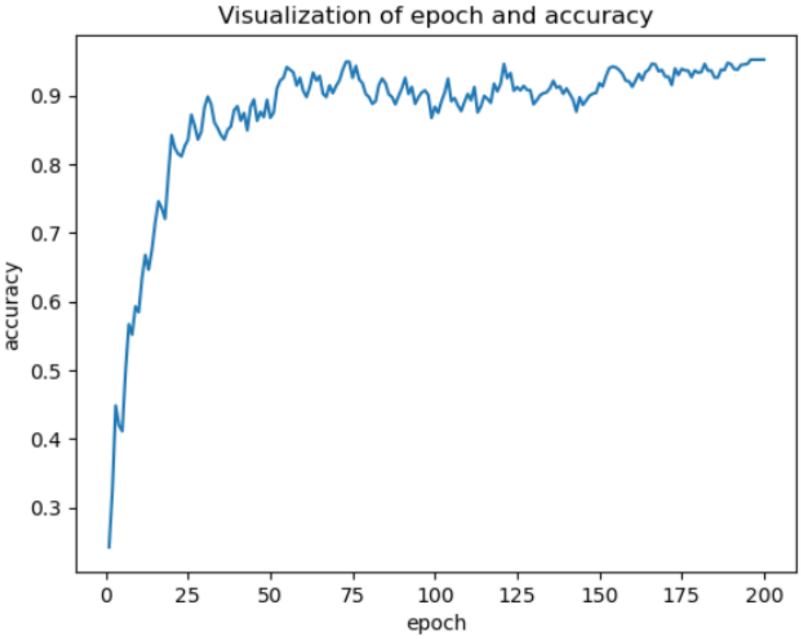


Fig 8 Model accuracy curve

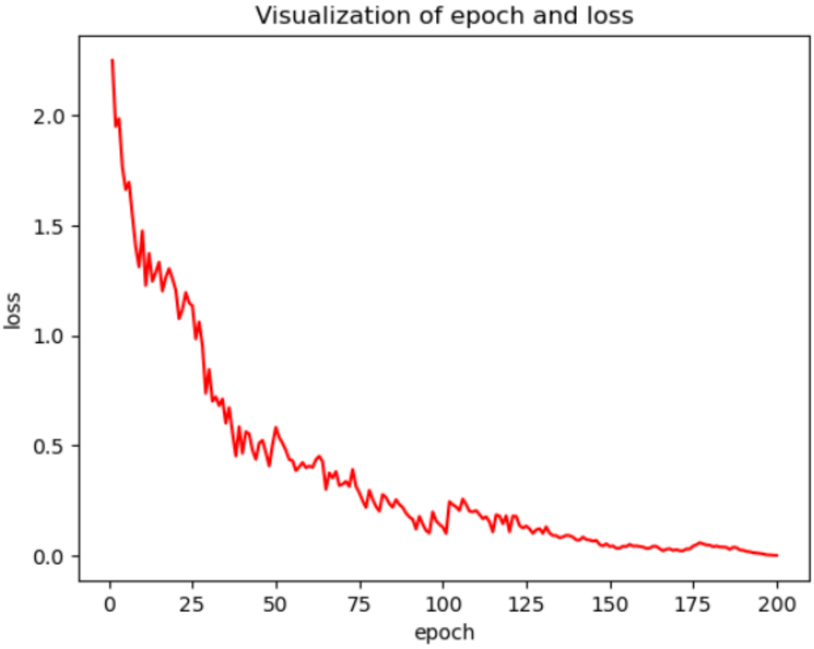


Fig 9 Model loss function curve

To better observe the experimental results, this thesis also gives the training results of the original LSTM. This thesis found that if all key point information (501) of Mediapipe is used, the accuracy will be low because the features are too complex and there are too many missing values. However, only 33 key point information of pose is used for feature engineering design, and the accuracy is It can be increased by a percentage point. Specific results information can be found in Table 5.

Table 5 Model results under different key point characteristics

method	index	501 Key Points	501 key points + angles	33 key points	33 key points + angles
LSTM	Accuracy	0.2962	0.3685	0.8257	0.8864
	learning rate	8.95E-3	8.23E-3	8.13E -4	7.05E-4
Attention -Multi LSTM	Accuracy	0.4831	0.5936	0.9018	0.9523
	learning rate	3.58E-3	3.04E-3	2.97E-4	2.46E-4

5) Result analysis.

Fig 10 shows some results of video testing on the test set in the Chinese Traffic Police Gesture Dataset based on the above model. The left subplot in Fig 10 is the result of recognizing lane change, the middle subplot is the result of recognizing pull over, and the right subplot is the result of recognizing slow down.

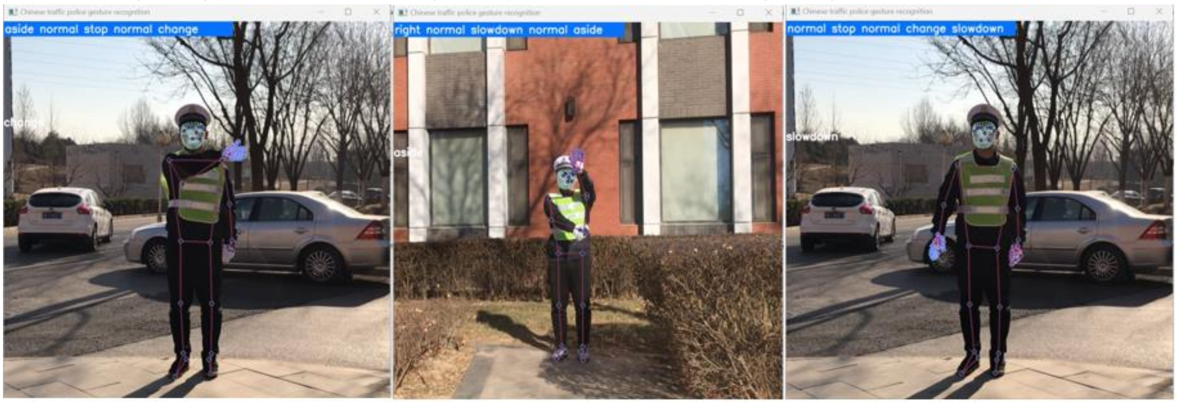

Chinese traffic police gesture recognition
aside normal stop normal change
change
Chinese traffic police gesture recognition
right normal slowdown normal aside
aside
Chinese traffic police gesture recognition
normal stop normal change slowdown
slowdown

Fig 10 Offline recognition of Chinese traffic police gestures based on the model proposed in this thesis

IV. CONCLUSION

This thesis provides an overview of relevant research on gesture recognition for traffic police. Based on the existing technology, a multi-layer LSTM optimization model that integrates limb key point coordinates and continuous limb angles is proposed. The research of this model is divided into the following stages: target recognition and positioning of traffic police, Mediapipe obtains the key point information of the traffic police limb (limb coordinates), The key continuous limb angles were calculated, and feature fused, and the multi-layer LSTM model of the attention mechanism was integrated. Finally, offline testing of the data set was conducted based on the model, and good recognition results were obtained. Although this thesis has achieved good results on the Chinese traffic police gesture data set, there are still some issues that require in-depth study. For example, the data set only has front-facing videos of traffic police and lacks comprehensive feature information from multiple angles. We will consider designing a multi-angle gesture data set in the future. During the experiment, it was discovered that the face orientation of the traffic policeman is also an important feature. Later, the face orientation feature of the traffic policeman was integrated.

ACKNOWLEDGMENT

This thesis thanks the following projects for their technical and financial support. Anhui Province provincial quality engineering projects: 2022jyxm660, AH202112216189, 2022jyxm659, 2022sx060, 2023cyts060. Anhui Xinhua University school-level projects: 2022jy006, 2020ylzy01.

REFERENCES

[1] https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-05/Level-of-Automation-052522-tag.pdf

[2] Li, C., & Yang, S. (2018, December). Traffic police gesture recognition for autonomous driving. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 1413-1418). IEEE.

[3] Ma, W., Song, G., Zeng, Q., Zhang, H., Zou, M., & Zhao, Z. (2024). FFCSLT: a deep learning model for traffic police hand gesture recognition using surface electromyographic signals. IEEE Sensors Journal.

[4] **ong, X., Wu, H., Min, W., Xu, J., Fu, Q., & Peng, C. (2021). Traffic police gesture recognition based on gesture skeleton extractor and multichannel dilated graph convolution network. Electronics, 10(5), 551.

[5] Li, Y. (2023, August). Traffic police command gesture recognition technology based on machine vision and two-stream spatio-temporal attention graph convolutional network. In Third International Conference on Computer Vision and Pattern Analysis (ICCPA 2023) (Vol. 12754, pp. 903-911). SPIE.

[6] He, J., Zhang, C., He, X., & Dong, R. (2020). Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features. Neurocomputing, 390, 248-259.

[7] Reddy, N. R., Priya, P., Aryan, S., & Ajay, P. (2023, July). Crime Detection System with Machine Learning using OpenCV, YOLO and CNN. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[8] Miao, Y., Shi, E., Lei, M., Sun, C., Shen, X., & Liu, Y. (2022, May). Vehicle control system based on dynamic traffic gesture recognition. In 2022 5th International Conference on Circuits, Systems and Simulation (ICCSS) (pp. 196-201). IEEE.

[9] Kim, J. W., Choi, J. Y., Ha, E. J., & Choi, J. H. (2023). Human pose estimation using mediapipe pose and optimization method based on a humanoid model. Applied sciences, 13(4), 2700.

[10] Thoutam, V. A., Srivastava, A., Badal, T., Mishra, V. K., Sinha, G. R., Sakalle, A., ... & Raj, M. (2022). Yoga pose estimation and feedback generation using deep learning. Computational Intelligence and Neuroscience, 2022.

[11] Zaheer, S., Anjum, N., Hussain, S., Algarni, A. D., Iqbal, J., Bourouis, S., & Ullah, S. S. (2023). A multi parameter forecasting for stock time series data using LSTM and deep learning model. Mathematics, 11(3), 590.

[12] **a, Z., Pan, X., Song, S., Li, L. E., & Huang, G. (2022). Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4794-4803).