Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD26338

Real-Time Recognition of Dangerous Human Actions Using Lightweight Pose Estimation and Spatio-Temporal Graph Networks

Tiantian Miao

Liangjiang International College, Chongqing University of Technology, Chongqing, China
19112170654@163.com

Abstract. Identification of dangerous human actions is of vital importance for safety monitoring. In response to the limitations of traditional methods, this study has developed an efficient pedestrian dangerous behavior recognition system, aiming to enhance monitoring capabilities and having practical application value. The study adopts a modular design, consisting of four core modules: human detection, posture estimation, behavior recognition, and result output. It can detect and warn of five dangerous behaviors such as throwing objects and climbing obstacles in real time. Among them, the posture estimation module addresses the shortcomings of the classic HRNet by adopting the improved SCite-HRNet model. This model optimizes computational efficiency and feature expression ability, ensuring accuracy (COCOval2017 dataset AP value of 65.9) while significantly reducing computational load (parameter quantity is only 18% of MobileNetV2), significantly improving its applicability on mobile devices. Experimental verification has demonstrated the effectiveness of the model improvement. The system is developed using PyTorch on the Ubuntu system, utilizing CUDA acceleration and multi-threading processing. It achieves a real-time processing speed of at least 25 frames per second on the RTX 2080Ti graphics card. Tests based on 15,000 multi-scenario labeled images show that the system has good robustness in complex environments. The system provides an intuitive visual monitoring interface, and the final classification accuracy reaches 99%.

Keywords: Action recognition, Pose estimation, Graph neural networks, Lightweight models, Real-time systems

1. Introduction

In the early stage of human action recognition, traditional machine learning methods such as Support Vector Machine (SVM) and Random Forest were mainly relied upon. These methods required manual extraction of features like edge textures, and performed poorly in complex scenarios [1]. With the rise of deep learning, Convolutional Neural Network (CNN) achieved significant breakthroughs, enabling automatic learning of feature representations from raw images [2]. In the research of dangerous action recognition, scholars proposed various innovative solutions: He Qunying et al. used Graph Neural Network (GNN) to construct a video frame graph structure,

DOI: 10.54254/2755-2721/2025.LD26338

extracted spatial features through Graph Convolutional Network (GCN), and used temporal transformer to capture temporal information, effectively overcoming the defects of the skeletal method being susceptible to interference and having high computational cost [2]; Ma Zhiyou et al. adopted the efficient Blazepose algorithm for skeletal key point detection, combined with st-gru network to recognize dangerous driving actions, and demonstrated outstanding performance in model size and performance indicators [3]; Wang Di's team integrated the advantages of 2D and 3D CNN, detected the target through 2D CNN to generate temporal data, and modeled the temporal features through 3D CNN, achieving efficient recognition at 10 frames per second on edge devices [4]; Saif et al. combined Convolutional Long Short-Term Network (CLSTDN) of CNN and RNN, used spatial features to predict action intentions and verified it on multiple datasets [5]. Related technologies have been extended to industry applications, such as Bai Xingtao et al. optimized the network to improve the accuracy of power monitoring video recognition [6], Wang Tao integrated C3D and visual Transformer to achieve quantitative assessment of sports actions [7], Liu Haonan et al. improved the recognition effect of dual-stream CNN by embedding attention mechanisms [8].

Current research still faces challenges such as insufficient robustness in complex scenarios, high computational costs of resource-constrained devices, and insufficient in-depth research on multimodal fusion. Therefore, this study will deeply explore improved neural network methods for dangerous action recognition, focusing on enhancing recognition accuracy and robustness in complex environments, reducing computational complexity, and providing better solutions for practical applications.This study explores an integrated framework combining SCite-HRNet and ST-GCN to overcome these limitations.

2. Methodology

2.1. System architecture

The target detection model first performs human body detection on the input image to obtain the bounding box information of the human body. Then, the pose estimation model estimates the key points of the human body based on the detected bounding box. The human body is tracked between different frames to ensure the continuity of the time series data of the key points. Finally, the action recognition model performs action recognition based on the key point time series data obtained through tracking. Thus, through data transmission and collaborative work among the various models, accurate recognition of human dangerous actions is achieved.

DOI: 10.54254/2755-2721/2025.LD26338

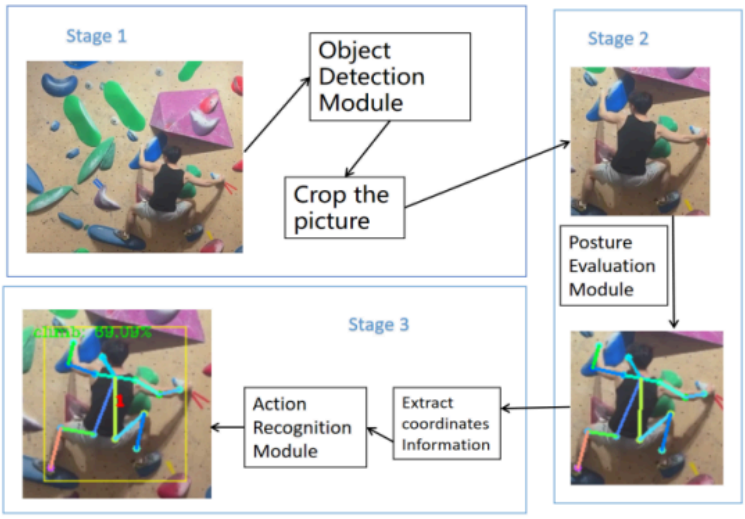


Figure 1. Algorithm flowchart

2.2. SCite-HRNet design

SCite - HRnet employs a four-stage progressive network architecture. Its key components are constructed by four parallel computing branches with successively decreasing spatial resolutions. As the network depth continuously increases, new computing branches are dynamically introduced at each stage. The spatial resolution of the new branches decreases progressively, with the specific scale precisely controlled at half of that of the previous branch. At the same time, the feature channel dimension grows exponentially to twice the original size. In the initial stage of processing the input data, the network cooperates with a 3×3 depthwise separable convolution layer and a channel recombination unit to complete the preliminary feature encoding of the input data. This ingenious design not only reduces the computational complexity but also effectively maintains the modeling ability for long-range spatial dependencies. Finally, the network outputs the feature map of the highest resolution branch to ensure that the detailed information is not lost.

2.3. Action recognition with ST-GCN

The behavior recognition module of this system adopts the ST-GCN (spatio-temporal graph network convolution model) as the action recognition module of the system. This network model mainly combines the graph convolution network (GCN) and the time convolution network (TCN) to expand into a spatio-temporal graph model, and designs a universal representation of the skeletal point sequence for behavior recognition. The ST-GCN model represents the human skeleton as a graph, where each node of the graph corresponds to each joint point of the human body.

3. Experimental evaluation

3.1. Dataset and implementation

Firstly, based on the principle of time-balanced sampling, we carefully extracted the original frame sequences from 70 monitoring video streams with an average duration of 60 seconds. These video streams covered various typical spatial distribution scenarios such as urban transportation hubs, commercial complexes, and enclosed venues, ensuring that the diversity of environmental variables

DOI: 10.54254/2755-2721/2025.LD26338

and lighting conditions could be fully represented. After a series of processing, we finally constructed an image set containing 15,000 different dangerous behaviors. Each image in the image set was equipped with precise bounding box annotations, behavior category labels, and 14 key point coordinates of human body postures, providing a solid data foundation for the training and validation of the model, which helps to further enhance the reliability and accuracy of the system in research.

2a: Behavior of climbing over obstacles

2b: Behavior of armed attack

Figure 2. Dataset examples

Table 1. Statistical table of dataset division

subset	sample size	proportion	main application	remark
training set	10,500	70%	Model parameter learning	Proportions of the five types of behaviors: Climbing 23% Quick approach 19% Throwing 21% Weapon attack 18% Gun aiming 19%
Validation set	2,250	15%	Hyperparameter tuning and early stopping mechanism	Maintain the same behavioral distribution as the training set
test set	2,250	15%	Final performance evaluation	Strictly isolate the source of the video

3.2. Results and analysis

3.2.1. Comparison of experimental results and analysis

Firstly, based on the principle of time-balanced sampling, we carefully extracted the original frame sequences from 70 monitoring video streams with an average duration of 60 seconds. These video

DOI: 10.54254/2755-2721/2025.LD26338

streams covered various typical spatial distribution scenarios such as urban transportation hubs, commercial complexes, and enclosed venues, ensuring that the diversity of environmental variables and lighting conditions could be fully represented. After a series of processing, we finally constructed an image set containing 15,000 different dangerous behaviors. Each image in the image set was equipped with precise bounding box annotations, behavior category labels, and 14 key point coordinates of human postures, providing a solid data foundation for the training and validation of the model, which helps to further enhance the reliability and accuracy of the system in research.

Table 2. Pose estimation performance (COCOval2017)

model	backbone	Pretrain	input size	#params	GFLOPs	AP	AP^{50}	AP^{75}	AR
MobileNetV2	MobileNetV2	N	256×192	9.6M	1.4	63.2	85.2	70.8	69.4
ShuffleNetV2	ShuffleNetV2	N	256×192	7.6M	1.2	59.8	85.4	66.4	66.3
Lite-HRNet	Lite-HRNet-30	N	256×192	1.8M	0.3	65.1	87.3	72.8	71.5
SCite-HRNet	SCite-HRNet-30	N	256×192	1.8M	0.3	65.9	87.3	74.4	72.2

3.2.2. System experiment results and analysis

In order to evaluate the performance and effectiveness of the result output module, the system also conducted experiments in terms of real-time performance and visualization effects. In the evaluation of real-time performance, the system recorded the processing time of each frame, calculated the frame rate (FPS) and displayed it in the upper right corner of the video frame (as shown in Figure 2). The final experimental results showed that in different hardware environments and video resolutions, this module could maintain a high frame rate and meet the real-time requirements.

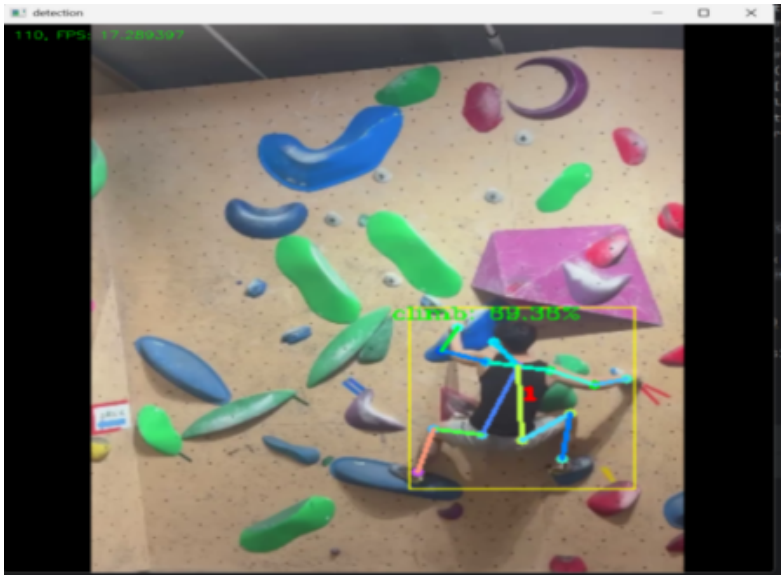
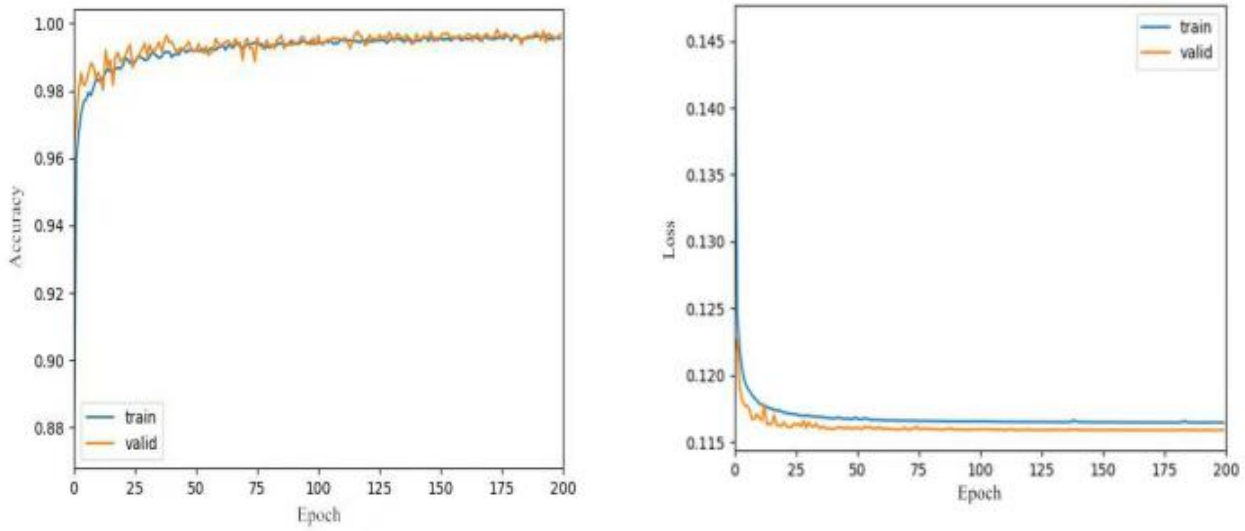


Figure 3. System output interface

As shown in Figure 3, the model training process was stable and efficient, and the final classification accuracy reached 99%. The experimental results indicate: The model is capable of

DOI: 10.54254/2755-2721/2025.LD26338

effectively learning the key features of the dataset. 2） The performance of the training set and the validation set is highly consistent (with a difference of less than 0.5%), indicating that the model has good generalization ability and no overfitting occurred. 3） This fully validates the effectiveness of the network design and the advantages of the underlying dataset: containing 15,000 multi-scenario images, it covers a wide range of features, effectively enhancing the model's adaptability and providing a reliable guarantee for high-precision predictions.



4a: upward trend of the model's accuracy

4b: downward trend of the loss function

Figure 4. The trend during the training process

4. Discussion

This research has made progress, but there are still areas that can be improved. Currently, the system still struggles to run on small devices (such as cameras). In the future, the model needs to be further compressed to make it lighter and more efficient. To enhance the recognition ability in low-light conditions or when being obscured, other sensor data such as infrared can be considered. Currently, the system recognizes five specific dangerous actions. In the future, it should be expanded to warn of more potential dangerous behaviors (such as abnormal running, sudden acceleration). To promote the system, the operation interface and deployment process need to be improved to reduce the difficulty of use, and adjustments should be made according to the needs of different places such as schools, stations, and construction sites. In addition, when used in public places, personal privacy must be strictly protected. Effective methods (such as blurring sensitive information) need to be studied to minimize privacy infringement while ensuring the recognition effect.

5. Conclusion

The research focused on the actual needs for pedestrian safety monitoring in public places, and specifically designed and successfully implemented a pedestrian dangerous behavior recognition system based on deep learning. This system adopts a modular design concept, relying on the collaborative cooperation of three core modules: human body detection, posture estimation, and

DOI: 10.54254/2755-2721/2025.LD26338

behavior recognition, to achieve real-time detection and warning capabilities for five types of dangerous behaviors such as throwing objects and climbing obstacles.

Regarding the human body detection module, the system uses the lightweight YOLOv3 - Tiny algorithm and combines the single-class optimized TinyYOLOv3-onecls model. This combination, while ensuring the detection speed, effectively improves the positioning accuracy of human targets. Experimental results show that, based on the RTX 2080Ti platform, this module can achieve a processing speed of 25FPS, which truly meets the expectations for real-time monitoring.

The posture estimation module uses the SCite - HRnet algorithm model. Through a series of comparative experiments and smile experiments in this study, it can be proven that this algorithm model greatly contributes to the overall performance of the system. It also proves the advantages of the overall system in algorithm selection.

The behavior recognition module adopts the ST - GCN spatio-temporal graph convolution model. This model combines graph convolution networks and time convolution networks to construct a universal representation path for bone point sequences for behavior recognition. It represents human bones as a graph structure, where each node of the graph corresponds to a joint point of the human body. This achieves precise identification of human behaviors.

The final experimental results and analysis show that the system, built with the multi-level visual interface framework of OpenCV, significantly improves the interpretability and interactive analysis level of the monitoring system. The model exhibits obvious convergence properties during iterative training and ultimately achieves a classification accuracy of 99%. The performance indicators of the validation set are highly consistent with those of the training set, reflecting the excellent generalization characteristics of the model parameter space. This result not only verifies the rationality of the network architecture design but also highlights the advantages of the basic dataset in terms of data scale, annotation method, and the diverse distribution of the spatio-temporal dimensions, providing data support for high-precision prediction.

References

[1] YUSHAN LI. Research on Human Action Recognition Method Based on Machine Learning [J]. Highlights in Science, Engineering and Technology, 2024.

[2] QUNYING HE, LAN SHAN. Action Recognition Method Based on Graph Neural Network [C]. //2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), 2023, : 144-148.

[3] ZHENGYI MA, HAO ZHANG, YINGSHUO FENG, et al. Hazardous action recognition system based on blazepose and ST-recurrent neural network [C]. //Artificial Intelligence and Big Data Forum, 2023, : 1259311 - 1259311-6.

[4] DI WANG, ZHONGLIN YANG, GAOTIAN LIU. Research on Human Action Recognition Based on Edge Intelligence [C]. //2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), 2024, : 1058-1064.

[5] SAIFUDDIN SAIF, EBISA D WOLLEGA, SYLVESTER A KALEVELA. Spatio-Temporal Features based Human Action Recognition using Convolutional Long Short-Term Deep Neural Network [J]. International Journal of Advanced Computer Science and Applications, 2023.

[6] XINGTAO BAI, NINGGUO WANG, YONGLIANG LI, et al. Research on Power Safety Monitoring Action Recognition Algorithm Through Neural Network and Deep Learning [C]. //2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA), 2024, : 990-994.

[7] TAO WANG. Research on deep learning-based action recognition and quantitative assessment method for sports skills [J]. Applied Mathematics and Nonlinear Sciences, 2024.

[8] HAONAN LIU, WENZHEN KUANG. Research on deep learning-based instruction action recognition method for assistant duty officers [C]. //Other Conferences, 2023, : 127901P - 127901P-5.