International Core Journal of Engineering Volume 11 Issue 6, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

Multimodal Perception-Based Target Tracking for Child Safety Monitoring in Home Environments

Huijie Liu, Hongyan Zhang*, Xiaotao Wang, Renjie Sun, Haitao Zhang

School of Mechanical and Electrical Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China

*Corresponding author: Hongyan Zhang

Abstract

With the rapid development of artificial intelligence and computer vision technology, this paper proposes a multi-modal perception method based on object tracking for continuous monitoring and security detection of children's activities in family environment. YOLOv8 and DeepSORT were used for multi-target detection and tracking, and RTAB-Map was introduced to complete real-time positioning and map construction of three-dimensional family environment through closed-loop detection and memory management, combined with vision and lidar sensors, so as to realize real-time detection and behavioral trajectory analysis of children at home and ensure the safety of children at home. The experimental results show that the accuracy of multi-target tracking (MOTA) is 87.5% compared with the baseline method. Reasoning speed maintained at 25FPS; The root mean square error (RMSE) of trajectory tracking is controlled within 0.15m, which provides a reliable solution for real-time safety detection of children at home.

Keywords

Target Tracking; Multimodal Sensing; RTAB-Map; SLAM Construction; Behavioral Trajectory Analysis.

1. Introduction

Currently, mainstream tracking and detection algorithms have their own advantages. YOLOv8 can separate classification and regression tasks to reduce feature conflicts and improve detection accuracy, which is suitable for real-time detection and tracking tasks[1]. SORT target tracking algorithm, implemented by the Hungarian algorithm and Kalman filter, predicts the target position and uses the prediction results to correct the tracked target [2]. However, it does not perform well in high occlusion conditions. TrackFormer[3] can build a spatio-temporal relationship between the self-attention mechanism and the modeling target, thereby improving the robustness and efficiency of tracking. However, the requirements for computing power are high. visual SLAM is widely used in driverless driving, augmented reality, three-dimensional reconstruction, indoor navigation and other fields, and is a popular research direction in recent years [4,5]. It can simultaneously realize self-positioning and environmental map construction in unknown environments. Most of the existing systems only operate in two-dimensional space, lack of spatial semantic understanding, and it is difficult to complete scene semantic recognition and behavior analysis. ORB-SLAM2[6] and ORB-SLAM3[7] support monocular, binocular, and RGB-D inputs, enabling complete 3D reconstruction and real-time tracking. However, the ORB-SLAM series has high hardware requirements for feature point extraction and matching, large computing requirements, and limited multi-modal data fusion capability. LSD-SLAM[8], proposed by Jakob Engel et al., preserves the details of the environment to a great extent and can

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

build semi-dense point cloud maps of the environment. However, the loopback detection of LSD-SLAM is not as good as the mechanism of ORB-SLAM and RTAB-Map based methods, which cannot generate 3D obstacle maps and are prone to collision with low objects. RTAB-Map can fuse multi-sensor and liDAR data, combine appearance and geometry information for closed-loop detection, effectively filter dynamic interference [9,10], and carry out consistent global mapping. However, most of these systems operate independently and lack tightly integrated behavioral analysis[11]. Therefore, this paper proposes a multi-modal target tracking system, which integrates visual image and LiDAR depth information, introduces YOLOv8 and DeepSORT for multi-target detection and tracking, ensures real-time performance, improves detection accuracy, anti-occlusion ability, and improves target tracking robustness and spatial perception ability. RTAB-Map and multi-sensor fusion are introduced to realize 3D reconstruction of home environment and camera pose estimation, so as to provide technical support for children's home safety monitoring by dynamically capturing children's behavior, analyzing movement intention and predicting potential risks.

2. Dataset

The hardware system proposed in this study consists of sensors such as depth camera, Light Detection and Ranging (LiDAR), IMU and is managed and controlled by ROS system. It mainly includes three parts: target detection and tracking module, SLAM mapping module and behavior trajectory analysis module. The multi-modal combination of YOLOv8+DeepSORT+RTAB-Map is the first application in the field of children's safety, and the real-time perception and behavior analysis of children at home is completed in collaboration. As shown in Figure 1.

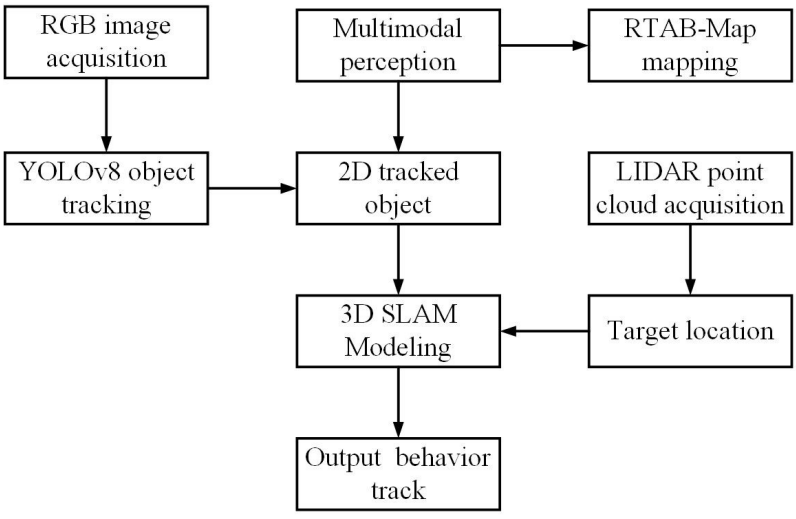

Figure 1. System flow chart

2.1 Multi-modal Fusion Target Detection and Tracking Module

The YOLOv8 object detection algorithm features a lightweight structure, high speed, and remarkable accuracy. Given the input image $I \in R^{H*W*3}$,a set of object detection boxes output by YOLOv8 are defined as:

$$\aleph = \{(x_i, y_i, w_i, h_i, s_i, c_i)\}_{i=1}^{N} \quad (1)$$

(x_i, y_i, w_i, h_i) indicates the center point, width and height of the detection frame, s_i indicates the confidence score, and c_i indicates the category label.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

Deep SORT uses Kalman filter for state prediction, Hungarian algorithm for data association, and CNN appearance feature embedding for re-recognition and matching. The appearance feature matching mechanism is introduced to effectively improve ID retention and occlusion resistance. It is particularly well-suited for multi-target scenarios and complex occlusion in home environments. At the same time, the lightweight features can be deployed to different devices. The state of the target is defined as:

$$X = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}, \dot{r}]^T \tag{2}$$

Where,(u,v) is the central position of the detection frame,s is the area of the detection frame,r is the aspect ratio, and($\dot{u}, \dot{v}, \dot{s}, \dot{r}$)is the velocity of each dimension. Deep SORT On the basis of the original SORT, the appearance similarity between the detection box d and the track t is calculated as follows:

$$\sin(d, t) = 1 - \cos(f_d, f_t) \tag{3}$$

Where f_d and f_t are the appearance feature vectors after L2 normalization. The system uses the Hungarian algorithm to match the target by considering the cosine similarity and the moving Mahalanobis distance. An improved similarity calculation method is proposed based on formula (3), where α is a learnable adaptive weight parameter:

$$\sin_{new}(d, t) = \alpha \cdot \sin(d, t) - (1 - \alpha) \cdot IoU(d, t) \tag{4}$$

The RGB image and depth map were concatenated at the channel level, the double-branch structure was inserted into the backbone of YOLOv8, and the features at the Neck were fused. The trunk handles RGB, and the branches handle depth maps and LiDAR projections. A multi-modal ReID feature extractor was constructed based on Deep SORT, and 3D reconstruction and spatial position tracking were carried out for the historical trajectory of the target, and the similarity between the spatial trajectory and the current observation point was calculated.In occlusion, illumination changes, pose complex, enhance the robustness of target detection and tracking.

2.2 SLAM Mapping Module

RTAB-Map is a closed-loop detection technology based on appearance real-time mapping. It supports multi-sensor input and can be used in large-scale closed-loop scenarios. Fusion of RGB image information and LiDAR point cloud data provides target spatial information. RTAB-Map uses a hierarchical memory management system to realize dynamic resource optimization through three modules: working memory (WM), long-term memory (LTM) and short-term memory (STM). In order to deal with children's behavior in a dynamic family environment, space-time information (location, speed, behavior) of target tracking is deeply embedded in SLAM's memory management process, and a closed-loop optimization route of detection → tracking → mapping → storage is established.

2.3 Trajectory Analysis Module

Through the trajectory coordinate conversion technology, the image coordinate system and lidar coordinate system of the target are converted to the coordinate system of SLAM map, and the unity of the target trajectory and three-dimensional map is realized.The two-dimensional plane trajectory obtained by YOLOv8 + DeepSORT module was projected into the three-dimensional space constructed by RTAB-Map, and the target trajectory was generated based on time order.The projection transformation formula is as follows.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

$$P_{cam} = T_{LIDAR \overset{\Delta}{\to} cam} \cdot P_{LIDAR} \tag{5}$$

$$P_{global} = T_w^t \cdot P_{cam} \tag{6}$$

Assume that the coordinates of the center of the object detection box in the image plane are (u,v), and the camera internal reference matrix is:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

If the target depth Z is obtained from the depth chart D(u,v), the three-dimensional position in the camera coordinate system is:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = Z \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{(x - c_x)Z}{f_x} \\ \frac{(y - c_y)Z}{f_y} \\ Z \end{bmatrix} \tag{8}$$

For the transformation of the camera coordinate system to the world coordinate system, the RTAB-Map provides the camera pose transformation matrix T_t in each frame, and the 3D position P_t^map of the target in the SLAM map coordinate system. The image target coordinate is uniformly mapped to the 3D map by the rigid body transformation, and the closed loop of the image target and the physical 3D position and behavior trajectory is realized.

$$T_t = \begin{bmatrix} R_t & T_t \\ 0^T & 1 \end{bmatrix} \tag{9}$$

$$P_t^{map} = T_t \cdot \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{10}$$

3. Experiments and Analysis

3.1 Simulation Experiment

This study builds a virtual experimental environment based on a typical three-bedroom unit, as shown in Figure 2. The family environment includes functional areas such as living room, master bedroom and two secondary bedrooms, kitchen, balcony and toilet. The RGB camera, depth camera and LiDAR are installed in the corner of the room, and the visualization is synchronized with the Gazebo simulation platform and RViz SLAM, as shown in Figure 3. The 2D detection results are fused with the SLAM map to generate the behavior trajectory. Verify the ability of target tracking and trajectory analysis in home scenarios.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

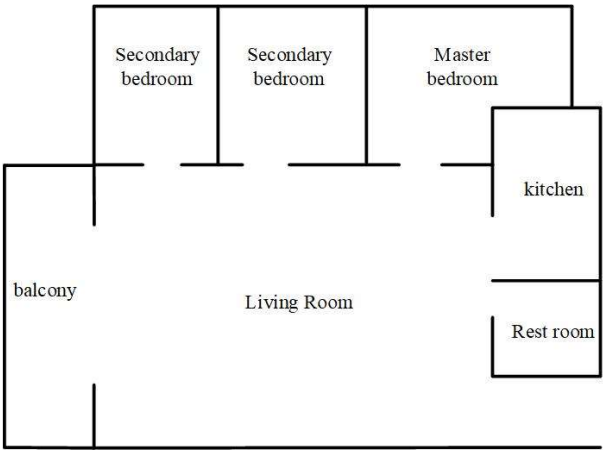

Figure 2. Three-bedroom flat plan

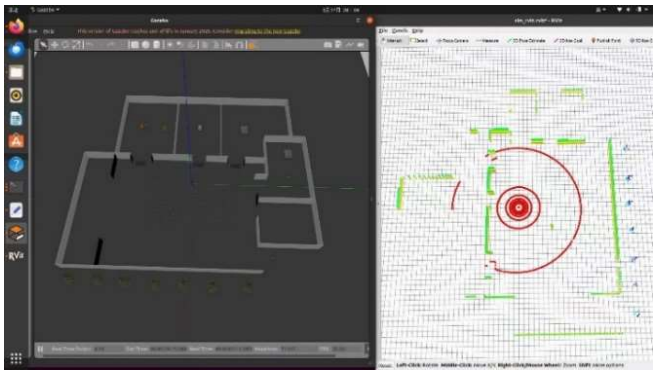

Figure 3. Home environment simulation and LiDAR scanning map

As shown in Table 1, the decision rules for visualization of children's behavior trajectory are designed, and the location of children is determined by the obtained real-time coordinates and time stamps of children. Set children's dangerous areas and dangerous goods, and realize trajectory visualization through space-time mapping and behavior label mapping. The semantic understanding of risky behavior is designed for the scene, as shown in Table 2.

Table 1. Visual decision rules of behavior trajectory

Form of expression	Decision basis	remark
Thermal map	Generate active hot spots based on historical track density	Red indicates the high frequency region and green indicates the low frequency
Path line	The color distinguishes the time period, and the line width determines the speed of movement	Blue for morning, orange for afternoon
Behavior tag	Different colors of the path, judge safety, warning, danger	The kitchen and balcony are dangerous areas
Scatter mark	Use ICONS to mark dangerous events	A red exclamation mark indicates a fall, and a yellow exclamation mark indicates proximity to a dangerous object

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

Table 2. Semantic understanding of risky behavior

Risk behavior type	Decision condition	Response mechanism
Climbing behavior	Z-axis coordinate change rate >0.5m/s for 3 consecutive frames	Alert and track marked red
Stay in the kitchen and balcony	Stay in the same area for >30 seconds	Push early warning information

3.2 Experimental Configuration

In this experiment, Ros system is installed on Raspberry PI, and a set of real-time perception and behavior analysis system for the safety monitoring of children at home is built through hardware and software collaborative optimization of multi-modal sensor (RGB-D+LiDAR+IMU) and lightweight algorithm (YOLOv8 + Deep SORT +RTAB-Map). The actual picture is shown in Figure 4. Table 3 lists the hardware configuration parameters.

Figure 4. Physical drawing

Table 3. hardware configuration parameters

name	Configuration parameter
Raspberry PI	Raspberry pi 4B
Laser radar (LiDAR)	YDLIDAR X3
Depth camera and RGB camera	Astra Pro Plus
Inertial Measurement Unit (IMU)	CMP10A

3.3 Ablation Experiment and Comparative Analysis

In order to verify the effect of each module and fusion, the following ablation experiment and comparison experiment were designed.

In this paper, ablation experiments were conducted to verify the effects of different module combinations on the performance of the system. The multi-modal sensors (LiDAR) and algorithms (RTAB-Map) were gradually introduced, and the effects on the key indicators such as target tracking accuracy (MOTA), 3D map quality (map@0.5) and trajectory reconstruction error (RMSE) were analyzed.The details of the experimental groups are shown in Table 4.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

Table 4. Analysis of ablation results

Experiment number	Module combination	map@0.5	MOTA	RMSE(m)
Exp-1	YOLOv8+DeepSORT	0.76	80.4%	-
Exp-2	YOLOv8 + DeePSORT + SLAM	0.85	83.7%	0.32
Exp-2	YOLOV8+DeepSORT+LIDAR	0.87	87.3%	0.25
This text	YOLOv8 + DeepSORT + LIDAR +RTAB-Map	0.92	87.5%	0.14

By comparing the three indexes of map@0.5, MOTA and RMSE, we can see that YOLOv8 + DeepSORT has good real-time performance and high accuracy. However, lack of spatial mapping and depth information support. After the introduction of visual SLAM, environment mapping can be realized without relying on depth sensor, and the occlusion robustness of target detection and tracking can be enhanced by adding LiDAR configuration. Finally, RTAB-Map is introduced to integrate visual and laser data to realize 3D reconstruction and semantic map construction. It has obvious advantages in target detection accuracy, spatial perception ability and tracking stability, especially suitable for children's behavior analysis in dynamic family environment.

The control variable method is used to verify the comprehensive advantages of the proposed method in multimodal perception, semantic graph construction and lightweight reasoning.The influence of YOLOv5, YOLOv7 to YOLOv8 on accuracy and speed is analyzed respectively, based on the differences between attention mechanism, shortest path association, traditional filtering and multi-modal fusion.The contribution of SLAM mapping to long-term tracking stability was highlighted. The details of the experimental group are shown in Table 5.

Table 5. Comparative analysis of experimental results

method	Detection model	Tracking strategy	map	MOTA	FPS
Trakformer	Trakforme	Self-attention based on Trakformer	-	83.8%	17
ByteTrack	YOLOv7	Shortest path association	-	84.2%	26
SORT+ORB+SLAM3	YOLOV5	Kalman+ORB SLAM	√	86.4%	19
This text	YOLOv8	DeepSORT+LiDAR +RTAB-Map	√	87.5%	25

In this paper, by comparing the traditional methods of multi-target tracking system Trackformer, ByteTrack and fusion SLAM, the effectiveness of the proposed method is further verified. Tracking strategy, mapping capability, MOTA accuracy, and frame rate (FPS) were compared. The self-attention mechanism of Transformer architecture achieves 83.8% MOTA without drawing, but its frame rate is only 17 FPS, and its real-time performance is weak. ByteTrack uses the lightweight YOLOv7 detector and the nearest path association strategy to obtain a high frame rate (26 FPS) and 84.2% MOTA, which is suitable for scenes with high speed requirements but lack of spatial understanding. The traditional SORT + OrB-SLAM method combines Kalman filtering and ORB SLAM to realize basic space mapping, and has high trajectory stability. However, the ability to detect small targets is limited. In contrast, the proposed YOLOv8 + DeepSORT + LiDAR + RTAB-Map method, based on the integration of YOLOv8, depth information and dense semantic mapping, achieves the highest MOTA (87.5%) while maintaining a processing speed of 25 FPS. Meet the real-time requirements.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0007

4. Conclusion

The experimental results show that the fully integrated method significantly improves the tracking stability and trajectory reconstruction accuracy, and demonstrates the effectiveness of multi-modal sensing and map location coordination. In addition, the method can improve the tracking accuracy and minimum trajectory error while maintaining the real-time performance, and is very suitable for the home environment with limited space and many obstructions. Future work will explore integrating audio sensing and risk behavior prediction via deep learning models to further enhance the robustness of child behavior monitoring.

Acknowledgments

Fund Project: Industrialization Cultivation Project of Education Department of Jilin Province (JJKH20240313CY).

References

[1] Lou H ,Duan X ,Guo J , et al.DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor[J].Electronics,2023,12(10):2323.

[2] Kaur H, Sahambi J S. Vehicle tracking in video using fractional feedback Kalman filter[J]. IEEE Transactions on Computational Imaging, 2016, 2(4): 550-561.

[3] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[4] Taketomi T, Uchiyama H, Ikeda S. Visual SLAM algorithms: A survey from 2010 to 2016[J]. IPSJ transactions on computer vision and applications, 2017, 9(1): 16.

[5] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M. Visual simultaneous localization and mapping: a survey[J]. Artificial intelligence review, 2015, 43: 55-81.

[6] Mur-Artal R, Tardós J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE transactions on robotics, 2017, 33(5): 1255-1262.

[7] Campos C, Elvira R, Rodríguez J J G, et al. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam[J]. IEEE transactions on robotics, 2021, 37(6): 1874-1890.

[8] Endo Y, Sato K, Yamashita A, et al. Indoor positioning and obstacle detection for visually impaired navigation system based on LSD-SLAM[C]//2017 International Conference on Biometrics and Kansei Engineering (ICBAKE). IEEE, 2017: 158-162.

[9] Wen Z. SLAM based vision self-navigation robot with RTAB-MAP algorithm[J]. Applied and Computational Engineering, 2023, 6: 1-5.

[10] Muharom S, Sardjono T A, Mardiyanto R. Real-Time 3D Modeling and Visualization Based on RGB-D Camera using RTAB-Map through Loop Closure[C]//2023 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE, 2023: 228-233.

[11] Zhou S, Li Z, Lv Z, et al. Research on positioning accuracy of mobile robot in indoor environment based on improved RTABMAP algorithm[J]. Sensors, 2023, 23(23): 9468.