Research on Trajectory Decision-Making Methods for Intelligent Robots Integrating LiDAR and Multimodal Perception such as Image

Hui Kou[1*]

[1] Foshan Rossum Robotics Co.,Ltd., Foshan, 528000, China

* 350415080@qq.com

https://doi.org/10.70695/IAAI202504A12

Abstract

In complex scenarios, single-sensor perception is unstable, trajectory planning lacks safety constraints, and there are problems with multi-source coordination. To address these issues, this paper proposes an intelligent robot trajectory decision-making method combining LiDAR and image processing, forming a multimodal perception system. This system requires a robot platform, proper extrinsic parameter calibration, and time synchronization. A LiDAR-Image feature acquisition and attention fusion network is designed from a unified BEV perspective to generate an environmental cost map that considers both geometry and semantics. Based on this cost map, an RL/MPC trajectory decision-making model is constructed, introducing chance constraints and dynamic boundaries to ensure safety margins. Simulations and real-world experiments included indoor corridors, office areas, and crowded places. Results show that the multimodal approach outperforms DWA and single-modal RL in terms of mAP, distance RMSE, minimum obstacle distance, and task completion rate. Furthermore, it can run continuously on embedded platforms, demonstrating the effectiveness of the proposed method and its value for engineering applications.

Keywords Multimodal Perception; LiDAR-image Fusion; Trajectory Decision; Reinforcement Learning; Model Predictive Control

1 Introduction

Multimodal perception and intelligent decision-making are key research directions for robot autonomous navigation. Zhao et al. reviewed the human-computer interaction decision-making driven by multimodal perception, showing that the fusion of multiple sources of information such as vision, lidar and voice is beneficial to enhance environmental awareness, but there are still difficulties in spatiotemporal synchronization and modal collaboration [1]. Tan proposed to apply deep reinforcement learning and multi-sensor fusion to path planning, thereby improving obstacle avoidance ability in complex scenarios, but the convergence and generalization of training are still limited [2]. Duan et al. conducted a systematic evaluation of the perception-fusion-control integrated framework in manufacturing scenarios, and felt that multimodal collaboration can improve operational accuracy, but is limited by immediacy and data redundancy [3]. Luo et al. created a multimodal perception-decision framework, showing a mechanism for efficient mapping of perception features to the decision layer [4]; Wu et al. applied the attention mechanism to multimodal path planning, improving the flexibility of trajectory under dynamic obstacle conditions [5]. Fan et al. completed multimodal perception and decision-making in complex road scenarios based on the basic model, and proved the feasibility of cross-domain transfer [6]. Wang et al. and Zhu improved the trajectory planning algorithm in the framework of reinforcement learning and adaptive control, thereby improving the smoothness and safety of the trajectory [7-8]. Shi et al. proposed an ultrasonic self-localization and multimodal perception system for soft robots, showing that traditional rigid body platforms have limitations in sensor layout and fusion methods [9]. Han combined 3D CNN, LSTM and visual SLAM and applied them to logistics robots, achieving a joint improvement in path and control. Current research has made some progress in perception fusion and intelligent decision-making, but there are still shortcomings in modal alignment, robustness of dynamic scenes and safety constraint modeling [10]. Based on this, this paper proposes an intelligent trajectory decision-making method, which integrates lidar and images, can achieve multi-source feature collaboration, and can also improve perception accuracy and unify safety control.

2 System Overall Design and Multimodal Sensing Modeling

2.1 Intelligent Robot Platform and Sensor Configuration

As shown in Figure 1, the intelligent robot platform adopts a differential chassis with a 3D LiDAR and RGB-D camera integrated on the top; the CPU and GPU collect data and send commands via the network to complete end-to - end control.

$$\mathbf{s}_k = \left[p_x(k), p_y(k), \theta(k), v(k), \omega(k) \right] \tag{1}$$

Where, are $p_x(k)$ the $p_y(k)$ position coordinates of k on the plane at discrete time, $\theta(k)$ is the heading angle of the aircraft, $v(k)$ is the linear velocity, $\omega(k)$ and is the angular velocity. The multimodal observation combination is written as:

$$\mathbf{z}_k = \left[\mathbf{z}_k^L, \mathbf{z}_k^C \right] \tag{2}$$

Among them, $\mathbf{z}_k^L$ represents LiDAR point cloud or BEV, and $\mathbf{z}_k^L$ represents image or depth map features. Formulas (1)-(2) construct a unified state-observation description, which is beneficial for the explicit use of multimodal information in subsequent decision-making strategies. The key hardware specifications of the multimodal robot platform are summarized in Table 1.

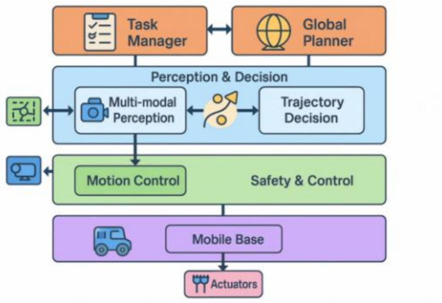

Fig. 1. System Overall Architecture Diagram

Table 1. Key Parameter Data Table for Sensors and Computing Units

Item	Spec 1	Spec 2	Spec 3
3D LiDAR	Range: 0.1–120 m	Angular resolution: 0.16°	Channels: 32
RGB-D camera	Resolution: 1280×720 pixels	Frame rate: 30 Hz	HFoV: 87°
CPU	Cores: 8	Base frequency: 2.4 GHz	TDP: 65 W
GPU	Compute: 4.0 TFLOPS (FP16)	Memory: 6 GB	Power: 60 W
System RAM	Capacity: 16 GB	Bandwidth: 50 GB/s	—
SSD storage	Capacity: 512 GB	Read speed: 550 MB/s	—

2.2 Coordinate System Establishment and Calibration, Time Synchronization

To achieve spatial and temporal alignment between the LiDAR and the image, the world coordinate system $\{W\}$, the body coordinate system $\{B\}$, the LiDAR coordinate system $\{L\}$, and the camera coordinate system must first $\{C\}$ be established. Then, the extrinsic parameters between these coordinate systems must be publicly solved $\{L\}$. $\{C\}$ In the LiDAR coordinate system, $\mathbf{x}_l$ mapping a 3D point to the camera coordinate system $\{C\}$ can be represented as:

$$\mathbf{x}_c = \mathbf{R}_{cl}\mathbf{x}_l + \mathbf{t}_{cl} \tag{3}$$

Here, $\mathbf{x}_c \in \text{ò}^{3}$ the point coordinates are below the Camera coordinate system, while $\mathbf{R}_{cl} \in SO(3)$ and $\mathbf{t}_{cl} \in \text{ò}^{3}$ are $\{L\}$ the rotation matrix and translation vector from to. By $\{C\}$ performing multi-pose observations in the two-sensor field of view on the chessboard calibration board and minimizing the reprojection error, can be estimated. Then $\mathbf{R}_{cl}$, by fusing the pose chains $\mathbf{t}_{cl}$ of $\{W\}$ − $\{B\}$ and, a complete coordinate transformation system is formed.

Regarding time synchronization, the LiDAR timestamp t_l and the Camera timestamp t_c need to be aligned via hardware triggering or software synchronization, with an upper bound constraint on the error.

$$|t_l - t_c| \le \Delta t_{\max} \tag{4}$$

The $\Delta t_{\max}$ maximum allowable time deviation is determined by the robot's maximum speed, control cycle, and distance to the nearest obstacle, and is used to control the impact of perception lag on trajectory safety margin. See Figure 2 for details.

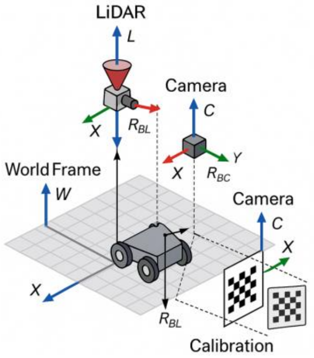

Fig. 2. Schematic diagram of multiple coordinate systems and calibration scenario

2.3 Multimodal Data Preprocessing and Mapping Modeling

When completing calibration and time synchronization, the LiDAR point cloud and image data need to be mapped onto a unified spatial representation so that subsequent fusion networks and trajectory decisions can be used directly. In the LiDAR channel, isolated noise is first removed using radius or statistical filtering, then ground points are segmented through plane fitting, and finally the obstacle point cloud is projected onto the Bird 's Eye View (BEV) grid to obtain a dense two-dimensional occupancy map.

Of a point in the world coordinate system $\mathbf{x}_w$ onto the image pixel coordinates $(u,\ v)$ can be written as:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R}_{cw} & \mathbf{t}_{cw} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_w \\ 1 \end{bmatrix} \tag{5}$$

Where $\mathbf{x}_w \in \text{ò}^{3}$ is the coordinate of a point in the world coordinate system, $\mathbf{R}_{cw} \in SO(3)$ represents $\mathbf{t}_{cw} \in \text{ò}^{3}$ the pose transformation from {W} to {C}, $\mathbf{K}$ is the camera intrinsic matrix, $(u,\ v)$ is the pixel coordinate, and is the homogeneous scale λ factor. According to Equation (5), the LiDAR point cloud can be projected into a sparse depth map or confidence map, which can become geometric prior information in the image branch.

the spatial distance between points in the Camera coordinate system is defined as $\mathbf{x}_c = [X, Y, Z]$:

$$d = \sqrt{X^2 + Y^2 + Z^2} \tag{6}$$

Here, X, Y, Z represents the coordinates of a point on the three axes, d and represents the Euclidean distance from that point to the camera's optical center. d After distance quantization, it is written into the BEV raster and encoded together with the LiDAR occupancy probability, thus creating a multi-channel environment tensor containing geometric, texture, and distance information. The image undergoes distortion correction and normalization to ensure geometric accuracy while meeting real-time requirements.

3 Multimodal Fusion Sensing and Trajectory Decision-Making Method

3.1 LiDAR and Image Feature Extraction Network

After completing the multi-coordinate system calibration and BEV mapping in Chapter 2, a LiDAR-Image dual-branch feature acquisition network was created to provide unified raster features for multimodal trajectory decisions. The LiDAR branch takes the denoised and de-grounded point cloud P_k as input, obtains voxel features through voxelization and sparse 3D CNN, and then maps them onto the BEV plane using column pooling. The image branch takes the RGB image at the same time I_k as input, obtains local texture information using a shallow 2D CNN, and then feeds it into a lightweight Transformer to capture long-range dependencies. Its output resolution must match the BEV raster.

the two branches on the same raster domain Ω is represented as

$$F_k^L(x) = g_L(\cdot)\left(\mathrm{P}_k, x\right), \quad F_k^C(x) = g_C\left(I_k, x\right), \quad x \in \Omega \, (7)$$

In Equation (7), $\mathbf{F}_k^L(\mathbf{x})$ is k the LiDAR feature vector at $\mathbf{F}_k^C(\mathbf{x})$ the grid at time, $\mathbf{x}$ is the image feature vector at the same grid; $g_L(\cdot)$ is the voxel-BEV LiDAR encoding network, $g_C(\cdot)$ is the CNN-Transformer image encoding network; P_k is k the three-dimensional point cloud set at time, I_k is the two-dimensional image frame acquired simultaneously; $\mathbf{x} \in \Omega$ means $\mathbf{X}$ it belongs to the local BEV working region Ω . According to Equation (7), the geometric information of LiDAR and the image semantics are aligned in the same grid coordinate system, which provides the necessary input data for subsequent fusion and cost map formation.

3.2 Multimodal Data Preprocessing and Mapping Modeling

Based on grid- level feature alignment, and to fully explore the complementary aspects of laser and image, an attention-excited multimodal fusion module was employed. For each BEV grid $\mathbf{X}$, a lightweight attention network first predicts the modal weights based on local texture sharpness and point cloud density, $\alpha_L(\mathbf{x})$ and $\alpha_C(\mathbf{x})$ then generates the fused features.

$$\mathbf{F}_k^F(\mathbf{x}) = \alpha_L(\mathbf{x})\mathbf{F}_k^L(\mathbf{x}) + \alpha_C(\mathbf{x})\mathbf{F}_k^C(\mathbf{x}), \quad \alpha_L(\mathbf{x}) + \alpha_C(\mathbf{x}) = 1 \quad (8)$$

In equation (8), $\mathbf{F}_k^L(\mathbf{x})$ is the multimodal feature vector after fusion processing, where $\alpha_L(\mathbf{x})$ and $\alpha_C(\mathbf{x})$ refer to $\mathbf{X}$ the weight coefficients of LiDAR and the image at the grid, which are output by the attention network, and $\mathbf{F}_k^L(\mathbf{x})$ and $\mathbf{F}_k^C(\mathbf{x})$ represent single-modal features according to the definition of equation (7), and there is such constraint: $\alpha_L(\mathbf{x}) + \alpha_C(\mathbf{x}) = 1$, in order to ensure energy conservation and make training more stable.

The fused features are further mapped to an environmental cost map, and $\mathbf{X}$ a comprehensive cost is defined for each grid location:

$$c(\mathbf{x}) = \lambda_{\text{obs}} c_{\text{obs}}(\mathbf{x}) + \lambda_{\text{smooth}} c_{\text{smooth}}(\mathbf{x}) + \lambda_{\text{comfort}} c_{\text{comfort}}(\mathbf{x}) \qquad (9)$$

In Equation (9), $c(\mathbf{X})$ represents the comprehensive grid cost, $c_{\text{obs}}(\mathbf{x})$ which depends on the occupancy probability and obstacle distance. is the collision risk cost, $c_{\text{smooth}}(\mathbf{x})$ which relates to the trajectory curvature and the smoothness of acceleration. $c_{\text{comfort}}(\mathbf{x})$ is affected by the speed limit fluctuation and lateral acceleration. The comfort cost is determined by it. λ_{obs} , λ_{smooth} , λ_{comfort} are the weights of these cost terms, which are obtained through offline parameter tuning or strategy learning (see Figure 3).

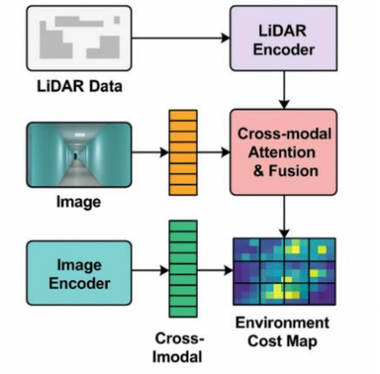

Fig. 3. Multimodal feature fusion network structure diagram (LiDAR branch, Image branch, fusion module, and finally output environment cost map).

3.3 Trajectory Decision Model Construction

A finite-time optimization model for local trajectory planning based on an environmental cost graph is constructed. The extended state vector includes the robot's motion state $\mathbf{s}_k$ and the local environment embedding $\phi\left(\mathbf{F}_k^F\right)$. The control input is a linear velocity-angular velocity pair $\mathbf{a}_k = \left[v_k, \omega_k\right]$. Discretized nonholonomic dynamics approximation of the differential chassis is also provided.

$$\mathbf{s}_{k+1} = f\left(\mathbf{s}_k, \mathbf{a}_k\right) \qquad (10)$$

In equation (10), $\mathbf{s}_k$ represents the robot's state vector at time k, which includes pose and velocity information. $\mathbf{s}_{k+1}$ is the state after one prediction step, $\mathbf{a}_k$ representing the current control input, v_k formed by linear velocity $f(\cdot)$ and angular velocity, ω_k and is a discrete motion model obtained according to wheel nonholonomic constraints. Under the RL or MPC framework, according to the prediction chain determined by equation (10), by $k = 0, \ldots, N-1$ minimizing the trajectory cost accumulated by equation (9) over the time interval and adding the terminal deviation term, a locally optimal trajectory that meets the dynamic constraints can be obtained.

3.4 Safety Constraints and Robust Design

The explicit optimization problem of safety distance and control boundary under multimodal perception uncertainty, given the state $\mathbf{s}_k$, the shortest distance from the robot to the nearest obstacle is given $d_{\min}\left(\mathbf{s}_k\right)$, and chance constraints are used to limit the collision risk:

$$\text{ə̶}(d_{\min}\left(\mathbf{s}_k\right) < d_{\text{safe}}) \le \varepsilon \qquad (11)$$

In Equation (11), $\text{ə̶}(\cdot)$ is a probability operator that takes into account the uncertainty of perception and prediction, $d_{\min}\left(\mathbf{s}_k\right)$ is $\mathbf{s}_k$ the minimum obstacle distance corresponding to the state, d_{safe} is the pre-set safety distance threshold, ε and is the upper limit of the maximum allowable collision probability. Equation (11) uses the occupancy probability and the upper bound of motion error to perform a conservative deterministic decision, thus forming the feasible region of RL/MPC together with velocity, acceleration, and curvature limits. Table 2 compares the performance of DWA, single-modal RL, and multi-modal RL/MPC in terms of success rate, minimum obstacle distance, trajectory length, total cost, and number of iterations.

Table 2. Offline performance comparison of different decision models

Method	Success rate [%]	Min. distance [m]	Avg. path length [m]	Normalized cost [-]	Avg. iterations [-]
DWA	82.3	0.24	18.7	1.00	15.2
Single-modal RL	89.5	0.27	17.9	0.83	30.6
Multi-modal RL/MPC	96.1	0.32	17.1	0.68	18.4

4 Experimental Design, Results and Analysis

4.1 Experimental Dataset and Scenario Configuration

To validate the multimodal perception and trajectory decision-making framework, various working conditions were constructed on simulation and real-world platforms. The simulation, based on CARLA/Gazebo, used custom corridors and street blocks, with static obstacles and random pedestrians deployed via scripts. The real-world testing covered office areas, laboratory corridors, and semi-open plazas, using LiDAR and cameras to simultaneously collect data, which was then aligned offline according to the external participation time described in Chapter 2. The data was divided into six scene categories, covering daytime, dusk, low light, and different occlusion and obstacle densities. Based on trajectory and obstacle annotations, the dataset was divided in a 6:2:2 ratio; Table 3 lists the number of segments, total number of frames, obstacle density, and illumination distribution.

Table 3. Experimental dataset and scenario configuration

Scene ID	Platform	Scene type	Sequences [–]	Frames [k]	Obstacle density [#/100 m²]	Train/Val/Test [%]
S1	Sim	Indoor corridor	40	120	3.2	60 / 20 / 20
S2	Sim	Office floor	32	96	4.5	60 / 20 / 20
S3	Sim	Urban intersection	28	84	6.8	60 / 20 / 20
S4	Real	Office corridor	30	90	2.7	60 / 20 / 20
S5	Real	Open lobby with crowd	twenty four	72	5.9	60 / 20 / 20
S6	Real	Semi-open plaza	26	78	4.1	60 / 20 / 20

4.2 Multimodal Sensing Performance Evaluation

After generating the dataset, a quantitative evaluation of the multimodal perception front-end proposed in Chapter 3 was conducted, investigating three settings: LiDAR-only, Image-only, and Fusion. For each scene ID, the model was trained or fine-tuned separately on the validation set, and then mAP, IoU, distance RMSE, and false negative rate were calculated using unified annotations (see Figure 4). Table 4 shows the statistical results of the three modalities in six scenes. It shows that in environments with concentrated crowds or significant changes in lighting (S3 and S5), fusion significantly improves upon single-modality approaches, while maintaining low distance RMSE and false negative rates. The aforementioned performance aligns with the attention fusion mechanism described in Section 2 of Chapter 3, demonstrating the significant complementarity between the stable geometric information of LiDAR and the detailed semantics of images in complex scenes.

Table 4. Perception performance of different modalities on each scene

Scene	Mode	mAP [%]	IoU [%]	Distance RMSE [m]	Miss rate [%]
S1	LiDAR-only	82.1	71.4	0.23	5.8
S1	Image-only	78.6	69.2	0.28	7.1
S1	Fusion	88.3	76.9	0.19	3.4
S3	LiDAR-only	79.4	68.7	0.31	8.6
S3	Image-only	74.2	64.9	0.37	10.3
S3	Fusion	86.7	75.1	0.24	4.9

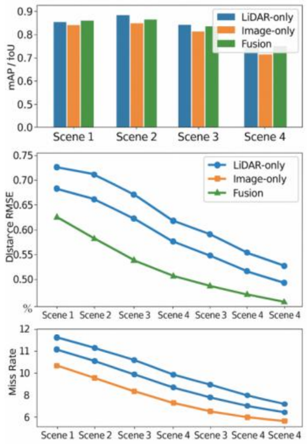

Fig. 4. Comparison of sensing performance of different modal combinations

4.3 Trajectory Quality and Task Completion Rate Analysis

The explicitAfter verifying the performance of multimodal perception, the environmental cost map was input into the RL/MPC decision module in Section 3.3. A large number of running samples were generated using the same start and end points and randomly set obstacles. Figure 5 shows the superimposed typical trajectories and reference trajectories of DWA, unimodal RL, and multimodal RL/MPC at the top, and the corresponding lateral error over time at the bottom, to compare the smoothness and convergence speed of the trajectories. Table 5 summarizes the average trajectory length, average lateral deviation, task completion rate, and collision rate of the six combined methods. The analysis shows that the multimodal RL/MPC has a near 100% achievement rate, its average trajectory length is only slightly longer than the shortest path, and its lateral error and collision rate are significantly lower than DWA and unimodal RL. All indicators are derived from the average and frequency statistics of multiple independent runs, indicating that multimodal trajectory decision-making has advantages in terms of safety and feasibility.

Table 5. Trajectory quality and task completion performance

Method	Avg. length [m]	Mean lateral error [m]	Completion rate [%]	Collision rate [%]	Scenarios covered [–
DWA (LiDAR)	19.6	0.21	87.3	8.2	S1–S6
DWA (Fusion Cost)	18.9	0.19	90.5	6.7	S1–S6
TEB (LiDAR)	18.1	0.18	91.2	5.9	S1–S6
Single-modal RL (Img)	17.8	0.17	93.6	4.3	S1–S6
Single-modal RL (LiDAR)	17.5	0.16	94.8	3.9	S1–S6
Multi-modal RL/MPC	17.2	0.13	97.9	1.8	S1–S6

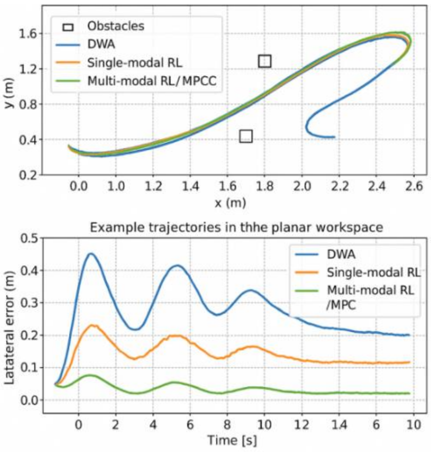

Fig. 5. Combination diagram of trajectory example and error curve

4.4 Real-time Performance and Resource Consumption Assessment

To verify the deployability of this method on embedded platforms, experiments were conducted by deploying two network scales, Base and Large, on lightweight embedded devices and high-performance desktop GPUs, respectively, and then testing was performed under the same working conditions. High-precision timers were used to record the time consumed in each stage, including perceptual forward propagation, multimodal fusion, decision reasoning, and control delivery, while simultaneously collecting CPU and GPU utilization data. Figure 6 uses hardware and network combination as the horizontal axis and total frame rate and module time consumption percentage as the vertical axis to identify performance bottlenecks. Table 6 shows the average frame rate, end-to-end latency, standard deviation, and resource utilization under six configurations. The results show that the embedded platform can maintain approximately 22 FPS when executing the Base network; while the desktop GPU can achieve 38 FPS when executing the Large network, with smaller latency fluctuations. The data in the table are statistical averages from multiple experiments.

Table 6. Runtime and resource utilization under different configurations

Config ID	Platform	Network size	FPS [Hz]	End-to-end latency [ms]	CPU usage [%]	GPU usage [%]
C1	Embedded SoC	Base	22.3	45.8 ± 6.1	68	71
C2	Embedded SoC	Large	15.7	63.4 ± 8.9	74	86
C3	Desktop GPU	Base	41.2	27.6 ± 3.4	36	39
C4	Desktop GPU	Large	38.5	30.1 ± 4.2	42	61
C5	Desktop GPU	Base + debug	34.7	33.8 ± 5.7	51	47
C6	Embedded SoC	Quantified	25.9	39.2 ± 5.0	65	58

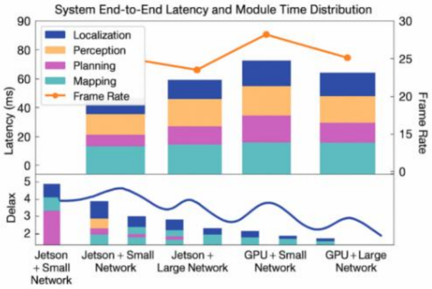

Fig. 6. Distribution of system end-to-end latency and module latency

4.5 Analysis of Typical Cases and Failure Examples

After completing batch statistics, visualization analysis was conducted on some typical success and failure cases to further illustrate the behavioral differences between multimodal perception and trajectory decision-making. In the S4 office area and S5 crowd scenarios, there are complex occlusion and dynamic interaction situations. After selecting relevant segments, the environmental cost map and the planned trajectory were placed on the same plane. Furthermore, the distance curves from the robot to key obstacles and the safety thresholds were plotted along the time axis to explore how the safety margin changes during the decision-making process when approaching obstacles, bypassing obstacles, and re-entering the channel.

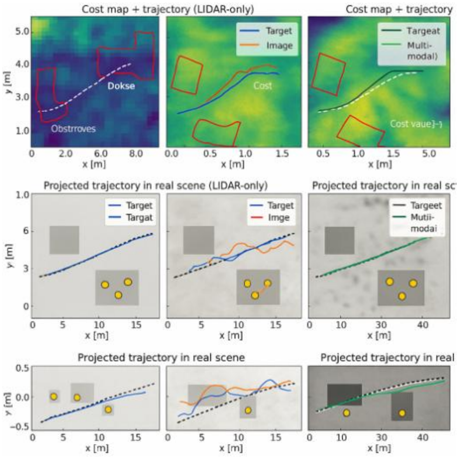

Fig. 7. Comparison of cost maps and real-world trajectories for multimodal and unimodal strategies in typical scenarios

As shown in Figure 7, the single-modal strategy often results in voids or overly conservative cost maps under strong occlusion or sudden changes in illumination, leading to trajectories that are too close to obstacles or frequent stops. Multimodal methods rely on more stable occupancy estimates to maintain appropriate safety margins; however, in extreme cases, abnormal trajectories may still occur due to temporary sensor failures. Performing frame-by-frame playback of these behaviors in Figure 7 and conducting cost map difference analysis clarifies the directions for future improvements in robust perception, anomaly detection, and safety redundancy.

5 Conclusions and Outlook

This paper constructs a multimodal trajectory decision-making method system integrating LiDAR and image. First, a robot platform is built, sensors are calibrated, and temporal synchronization is achieved. Then, precise correspondence between point clouds and images is achieved in a unified BEV coordinate system. A LiDAR-Image dual-branch feature acquisition and attention fusion network is designed to encode occupancy probability, distance, and semantic information into a multi-channel environmental cost map, forming a trajectory decision-making model incorporating RL/MPC. Chance constraints are used to describe safety margin and perception uncertainty. Simulation and field results show that the multimodal approach outperforms traditional DWA and single-modal decision-making in terms of perception accuracy, trajectory smoothness, task completion rate, and immediacy. Further development can introduce online adaptive and fault detection mechanisms to improve the system's robustness and engineering deployability.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Zhao, W., Gangaraju, K., & Yuan, F. (2025). Multimodal perception-driven decision-making for human-robot interaction: A survey. Frontiers in Robotics and AI, 12, 1604472.
2. Tan, J. (2023). A method to plan the path of a robot utilizing deep reinforcement learning and multi-sensory information fusion. Applied Artificial Intelligence, 37(1), 2224996.

3. Duan, J., Zhuang, L., Zhang, Q., et al. (2024). Multimodal perception-fusion-control and human-robot collaboration in manufacturing: A review. The International Journal of Advanced Manufacturing Technology, 132(3), 1071-1093.
4. Luo, X., Xu, J., & Zhang, H. (2024). Robot multimodal perception and decision-making framework. 2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) (pp. 899-906). IEEE.
5. Wu, X., Wang, G., & Shen, N. (2023). Research on obstacle avoidance optimization and path planning of autonomous vehicles based on attention mechanism combined with multimodal information decision-making thoughts of robots. Frontiers in Neurorobotics, 17, 1269447.
6. Fan, L., Wang, Y., Zhang, H., et al. (2024). Multimodal perception and decision-making systems for complex roads based on foundation models. IEEE Transactions on Systems, Man, and Cybernetics: Systems.
7. Wang, J., Chu, L., Zhang, Y., et al. (2023). Intelligent vehicle decision-making and trajectory planning method based on deep reinforcement learning in the Frenet Space. Sensors, 23(24), 9819.
8. Zhu, C. (2023). Intelligent robot path planning and navigation based on reinforcement learning and adaptive control. Journal of Logistics, Informatics and Service Science, 10(3), 235-248.
9. Shi, Q., Sun, Z., Le, X., et al. (2023). Soft robotic perception system with ultrasonic auto-positioning and multimodal sensory intelligence. ACS Nano, 17(5), 4985-4998.
10. Han, Z. (2023). Multimodal intelligent logistics robot combining 3D CNN, LSTM, and visual SLAM for path planning and control. Frontiers in Neurorobotics, 17, 1285673.

Biographies

1. **Hui Kou** Master, Intermediate Engineer, Intellectual Property Assistant Researcher, and currently serves as an Automation Engineer at Foshan Rossum Robotics Co.,Ltd. She has long been engaged in research related to robot manufacturing and artificial intelligence.She has presided over and participated in more than 10 projects, including the National Key R&D Program, Guangdong Provincial Key R&D Program, and Foshan Core Technology Research. She has won multiple honors such as the China Machinery Industry Science and Technology Award and the Second Prize of Invention, Innovation and Entrepreneurship. She is the 2020 IEC Young Expert of China, International Standardization Talent, Foshan Standardization Promotion Ambassador, and a think tank platform for scientific and technological innovation services in Foshan.

融合激光雷達與圖像等多模態感知的智能機器人軌跡決策方法研究

寇慧[1]

[1]佛山隆深機器人有限公司，佛山市，中國，528000

摘要：複雜場景裏，單一傳感器感知不穩定，軌跡規劃缺少安全約束，而且存在多元協同方面的問題。為了應對這些問題，本文提出了結合激光雷達和圖像的智能機器人軌跡決策方法，形成了多模態感知體係，該體系需要包含機器人平臺，並做好外參標定並實現時間同步，從統一的BEV視角出發來設計LiDAR–Image特徵獲取及注意力融合網絡，進而進生成既關注幾何又重視語義的環境代價圖。進而基於該代價圖構建RL/MPC軌跡決策模型，並引入機會約束與動態邊界以確保安全裕度。仿真和實機實驗室包含了室內走廊，辦公區域以及人羣聚集的地方。結果顯示，多模態方案在mAP，距離RMSE，最小障礙距離以及任務達成率這些指標上都比DWA和單模態RL更好，而且它還能在嵌入式平臺上持續運行，這表明所提方法有效具備工程應用的價值。

關鍵詞：多模態感知；激光雷達-圖像融合；軌跡決策；強化學習；模型預測控制

1. 寇慧，硕士，中级工程师，知识产权助理研究员，现任佛山隆深機器人有限公司自動化工程師，长期从事機器人制造与人工智能领域的相关研究。主持及參與國家重點研發計劃、廣東省重點研發計劃、佛山市核心技術攻關等10余項，獲得中國機械工業科學技術獎、發明創新創業二等獎等多項榮譽。是2020年中国IEC青年專家國際標準化英才、佛山標準化推廣大使、佛山市科技創新服务智囊平台“智汇赢”高級專家。