A Shopping Service Robot Framework with Visual-WEM Tracking and Intersection-Aware Following

Hanchen Yao, Jianwei Peng, Houde Dai, *Senior Member, IEEE,* Fanbiao Li, *Senior Member, IEEE,* Tim C. Lueth, *Senior Member, IEEE*

***Abstract*—The shopping service robot (SSR) is designed to offer a superior shopping experience through its continuous target following and companion services. However, the SSR encounters difficulties in complex environments, including visual occlusions and spatial constraints in narrow aisles. To address these difficulties, the study proposes a shopping service robot framework that integrates innovative perception, following control, and path planning modules. Firstly, the perception module employs a multi-sensor fusion method for human target tracking, integrating both red-green-blue-depth (RGB-D) camera and wireless electromagnetic (WEM) data by using an extended Kalman filter (EKF). Secondly, the target-following control module employs an omnidirectional constrained control law, which ensures synchronized orientation alignment between the SSR and the human target. Finally, the path planning module employs topological mapping to encode intersection geometries as path nodes, thereby guiding the SSR pass through narrow shelf aisles. In real supermarkets, we evaluated the target tracking approach under shelf occlusions and the human following task within narrow aisles. Experimental results demonstrate that the visual-WEM tracking approach achieves a pose tracking accuracy of (4.56 mm, 2.98°) under shelf occlusions. This study establishes the feasibility of human-robot collaboration in facilitating a hands-free shopping experience, highlighting its potential as a substitute for conventional shopping carts.**

***Index Terms*—Human-robot interaction, multi-sensor fusion, human following and companion, shopping service robot.**

I. Introduction

SHOPPING service robots (SSRs) are transforming retail operations by integrating capabilities such as shelf scanning for inventory management [1], artificial intelligence (AI)-based conversational interfaces for customer inquiries [2], and human-following functionality for hands-free shopping assistance [3-4]. Specifically, the human-following functionality not only provides the practical benefit of carrying goods, but also offers a socially interactive experience for customers throughout their entire shopping journey [5-7].

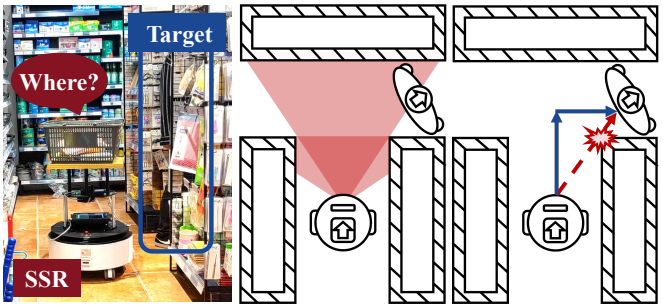


Fig. 1. Motivation of the proposed SSR framework. (a) The snapshot of the SSR operating in shelf areas. (b) The tracking failure of visual sensors. (c) The following failure of control methods in a narrow passage with sharp turns.

Nevertheless, the real-world deployment of SSR in the context of human-following functionality poses significant challenges, primarily due to critical issues in perception and motion control. Firstly, the shelf area constitutes a high-density occlusion environment. Figure 1(a) and 1(b) demonstrate a typical scenario of tracking failures: when a human target turns at an intersection, the SSR's optical sensors lose line-of-sight of the human target. Secondly, the movement intent of humans can undergo abrupt changes. In a real shopping scenario, the human target does not move at a constant velocity and performs unexpected maneuvers. For instance, the motion of a human target can be interrupted by a sudden stop, which is triggered when the human target inspects an item on the shelf. Thirdly, narrow aisles between shelves present a navigational bottleneck. In Fig. 1(c), the blue trajectory shows the planned path of SSR. In contrast, the red trajectory results from traditional control methods not integrated with path planning, which is driven only by the human target's reactive motions.

Optical and wearable sensors are the predominant sensors for target tracking. ***(1) Optical sensors.*** Optical sensors (*e.g.*, depth camera, laser scanner, and laser-visual fusion) provide abundant visual information but suffer from target loss in shelf-occluded scenarios. For instance, red-green-blue-depth (RGB-D) cameras are constrained by a limited field of view (FOV) [8-9]. Besides, the performance of RGB-D cameras is highly sensitive to light intensity [10]. In contrast, laser scanners provide superior tracking performance with an ex-

Manuscript received July 7, 2025; accepted February 8, 2026. Date of publication February 11, 2026; date of current version February 11, 2026. This letter was recommended for publication by Associate Editor and Editor upon evaluation of the reviewers' comments. This work was supported in part by the Fujian Provincial Science and Technology Plan Projects under Grants 2023Y9136, 2024YZ036017, 2024T3020, and 2025T3006, and the Open Project Program of Fujian Key Laboratory of Special Intelligent Equipment Measurement and Control under Grant FJIES2023KF02. (*Corresponding author: Houde Dai)*

Hanchen Yao, Jianwei Peng, and Houde Dai are with the Quanzhou Institute of Equipment Manufacturing, Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Jinjiang 362216, China. They are also with the Fujian College, University of Chinese Academy of Sciences, Fuzhou 350002, China (e-mail: yaohanchen21@mails.ucas.ac.cn; 12531290@mail.sustech.edu.cn; dhd@fjirsm.ac.cn).

Fanbiao Li is with the School of Automation, Central South University, Changsha 410083, China (e-mail: fanbiaoli@csu.edu.cn).

Tim C. Lueth is with the Institute of Micro Technology and Medical Device Technology, Technical University of Munich, Munich 80333, Germany (tim.lueth@tum.de).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.LRA.2026.xxxxxx, provided by the authors.

Digital Object Identifier 10.1109/LRA.2026.xxxxxx

tensive FOV and distance [11-13]. However, the laser scanner struggles to distinguish obstacles with leg-like shapes. ***(2) Wearable sensors.*** Wearable sensors utilize wireless signals to track the human target equipped with pose tracking devices. In [14], human targets wear radio frequency identification (RFID) tags at their waists. To this end, SSR achieves human-following functionality by maintaining a specific distance between the RFID reader and the tag. Alternatively, the ultra-wideband (UWB) system tracks the human target by utilizing a wrist-worn UWB tag [15]. Besides, wireless electromagnetic (WEM) signals can penetrate the human body [16], which enables target tracking in multi-person scenarios. Although wearable sensors can address the visual occlusion issue in target tracking, their signal attenuation near metal shelves requires further verification.

To address the issue of visual tracking occlusion in shelf areas, multi-sensor fusion has been identified as a promising solution. ***(1) Optimization-based fusion.*** Typically, a graph optimization method transforms sensor constraints of a stereo camera and laser scanner into factor graphs [17]. However, the computational complexity increases sharply with the number of humans and obstacles. ***(2) Learning-based fusion.*** To address timestamp synchronization issues of multiple sensors, learning-based fusion methods input sensor data into end-to-end fusion networks. For instance, Li *et al.* [18] proposed a multi-view camera and LiDAR fusion method for bird's-eye-view (BEV) perception. Due to the limitations in real-time performance and generalization capability, the learning-based fusion methods necessitate the development of more lightweight network architectures. ***(3) Filter-based fusion.*** Filter-based fusion is well-suited for human target tracking in supermarket environments with real-time dynamic requirements. In [19], a constrained estimation projection method is proposed to fuse UWB and camera data to track a human target. Kang *et al.* [20] developed a Markov-based extended Kalman filter (EKF) method to update accelerometer and magnetometer data. Although improving the response speed of EKF, the Markov chain introduces larger random errors. Yang *et al.* [21] developed an adaptive EKF to reduce localization oscillation. The adaptive EKF is achieved by updating the noise covariance matrix and previous estimates from IMU and UWB measurements. In summary, EKF-based fusion methods have made significant breakthroughs in mitigating environmental noise. However, EKF necessitates further improvements to handle sudden sensor failures (*e.g.*, visual occlusion).

To address the human-following challenges at shelf intersections, an increasing number of SSRs integrate following control with path planning. ***(1) Traditional control methods.*** Previous research on traditional control methods focused on improving the SSR's control precision while minimizing the human-robot distance error, including impedance control [22], linear quadratic regulator (LQR) [23], and model predictive control (MPC) [24-25]. The traditional control methods are highly dependent on the accuracy of the underlying model. Besides, deep reinforcement learning (DRL) improves the SSR's naturalness by learning the shopping path of customer preferences [26]. However, DRL exhibits poor performance in supermarkets, primarily due to a lack of training data that accounts for sudden customer behaviors. ***(2) Path planning methods.*** To navigate through narrow and complex shelf aisles, path planning methods are employed to generate a set of virtual path nodes. For example, Yuan *et al.* [27] transferred laser points of the intersection map into a Voronoi graph. Thus, the SSR navigates narrow intersections by locating the next path node. Lewandowski *et al.* [28] classified occlusion situations in supermarket environments, where point clouds were collected by a RGB-D camera. Thus, the simultaneous implementation of human-following control and path planning enhances the SSR's performance at shelf intersections.

In this study, a novel shopping service robot framework is proposed to ensure reliable human-following functionality in supermarket environments. The framework comprises three core modules: perception, following control, and path planning modules. Main contributions are highlighted as follows:

1) An EKF-based visual-WEM fusion method is proposed to address the target tracking challenge in shelf occlusion areas. When the human target exits from the camera's field of view, the EKF undergoes a transition from a fused visual-WEM tracking to a reliance on WEM tracking alone.
2) An intersection-aware following method is proposed to follow the human target at shelf intersections. To ensure safe navigation through narrow shelf intersections, the proposed method generates virtual path nodes and guides the SSR to avoid obstacles while continuously following the human target.

The remainder is organized as follows. Sections II and III presents the proposed SSR framework and its experimental results, respectively. Finally, Section IV concludes this article.

II. Proposed Framework

A. Perception Module

In Fig. 2, the self-developed SSR integrates optical sensors (RGB-D camera and laser scanner) and wearable sensors (WEM tracking device). In the perception module, the RGB-D camera and WEM tracking device are employed to track the human target, while the laser scanner detects environmental obstacles. Besides, the RGB-D camera is employed to identify shelf intersections. The point cloud data is utilized by the path planning module to generate topological maps.

(1) Obstacle Detection. To distinguish between human targets and obstacles, a density-weighted support vector data description (DW-SVDD) method is utilized for clustering leg-shaped point clouds from laser data [11]. The position of obstacles $\mathbf{x}_{ob}$ is given by:

$$\begin{cases} \min L\left(R,c,\xi\right)=R^2+C\sum_{i=1}^{n}P\left(\mathbf{x}_{ob}\right)F\left(\mathbf{x}_{ob}\right)\xi \\ s.t.\left\|\mathbf{x}_{ob}-c\right\|\leq R^2+\xi,\xi\geq 0,F\left(\mathbf{x}_{ob}\right)=(G,W) \end{cases}, \quad (1)$$

where (R,c,ξ) are the hypersphere radius, center of the sphere, and relaxation variable, respectively. C is the penalty factor. $P(\mathbf{x}_{ob})$ is the density weight. $F(\mathbf{x}_{ob})$ is a function that maps data from the original space to the feature space of girth feature G and width feature W.

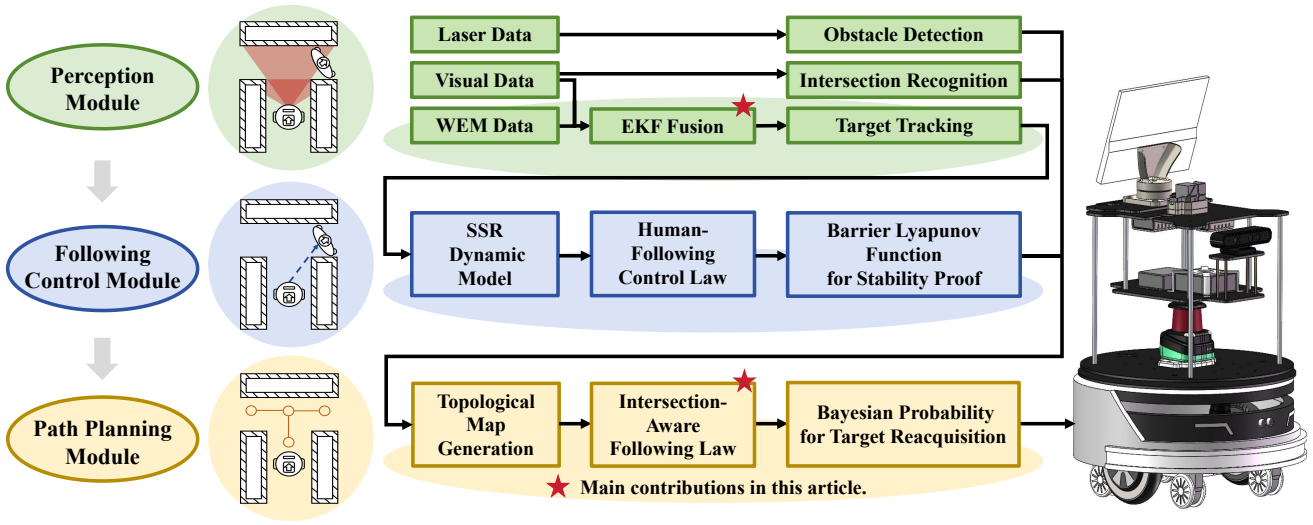


Fig. 2. The proposed framework of the shopping service robot (*i.e.*, SSR). The SSR is comprised of three primary modules: visual-WEM fusion perception, omnidirectional constrained human-following control, and intersection-aware path planning.

(2) Intersection Recognition. In Fig. 3, the RGB-D camera categorizes intersections between shelves into three topological map types: L-type, T-type, and cross-type. The intersection features are obtained as follows:

$$\begin{cases} L_{in} = d_{min} \sin\theta_{in} \\ W_{in} = z_{in2} - z_{in1} \\ D_{in} = -L_{in} - x_{in} \end{cases}, \tag{2}$$

where (L_{in}, W_{in}, D_{in}) are the width feature of the current aisle, the width feature of the intersection, and the depth feature of the intersection, respectively. d_{min} is the shortest distance from the RGB-D camera to the shelf in the point cloud. θ_{in} is half of the RGB-D camera's FOV. z_{in1} and z_{in2} represent the starting and ending points of the varying depth, respectively. x_{in} is the distance from the starting point to the ending point of the turning area.

(3) Human-Target Tracking. In Fig. 4(a), the WEM transceiver module is mounted on the center of the SSR platform, while the receiver modules can be worn on the wrist or waist of human targets. While visual tracking fails under shelf occlusion, the WEM tracking achieves the non-line-of-sight (NLOS) propagation of electromagnetic waves.

Figure 4(b) illustrates the principle of the WEM tracking device, where the transmitter and receiver are three-axis coils. Thus, magnetic moment vectors of the transceiver and receiver are denoted as m_i and n_i, $i = 1, 2, 3$. The vector $\mathbf{p}$ is defined as a positional displacement from the WEM coordinate frame $\{r_{WEM}\}$ to the human coordinate frame $\{h_{WEM}\}$. In the frame $\{r_{WEM}\}$ generated by the WEM transceiver i, the magnetic flux $\mathbf{b}_i$ at position $\mathbf{p}$ is given by:

$$\mathbf{b}_i = \frac{\mu}{4\pi}\left(\frac{3\mathbf{p}\left(\mathbf{m}_i\mathbf{p}\right)}{\|\mathbf{p}\|^5} - \frac{\mathbf{m}_i}{\|\mathbf{p}\|^3}\right), \tag{3}$$

where μ is the permeability of the medium (Unit: N/A^2).

Then, the rotation matrix $\mathbf{R}$ describes the rotational transformation from the frame $\{r_{WEM}\}$ to the frame $\{h_{WEM}\}$. Thus,

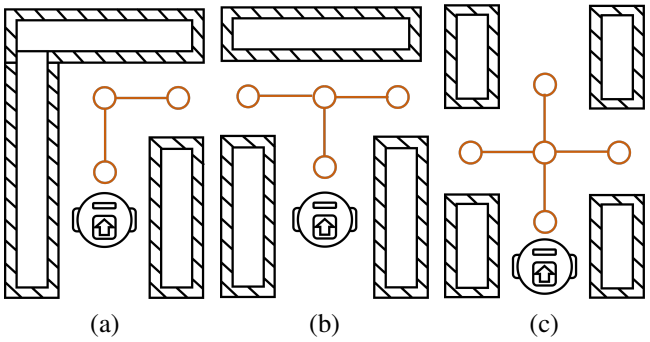


Fig. 3. Types of shelf intersections. (a) L-type intersection. (b) T-type intersection. (c) Cross-type intersection.

the measured magnetic flux $\mathbf{b}_{ij}$, $i, j = 1, 2, 3$, is detected by the WEM transceiver i and generated by the coil j, as:

$$\mathbf{b}_{ij} = \mathbf{n}_i^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{b}_j = \frac{\mu\mathbf{n}_i^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\left(3\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}} - \mathbf{I}\right)\mathbf{m}_j}{4\pi\left\|\mathbf{p}\right\|^3}, \tag{4}$$

where $\hat{\mathbf{p}} = \mathbf{p}/\left\|\mathbf{p}\right\|$ and $\mathbf{I}$ are the unit normalized vector and the identity matrix, respectively.

To formulate a simplified least-squares objective function for the EKF estimation, the $\mathbf{b}_{ij}$ is arranged to a measurement vector $\hat{\mathbf{y}}_{WEM}$, as:

$$\hat{\mathbf{y}}_{WEM} = \frac{\mu\mathbf{R}^{\mathrm{T}}\left(3\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}} - \mathbf{I}\right)}{4\pi\left\|\mathbf{p}\right\|^3}. \tag{5}$$

In Fig. 4(c), the RGB-D camera is inherently limited by line-of-sight occlusion [8-10]. In this study, the WEM tracking device and RGB-D camera are adopted to track targets. Under unobstructed conditions, the proposed EKF-based tracking method dynamically augments the WEM's optimization matrices $\mathbf{p}$ and $\mathbf{R}$ through RGB-D data, thereby improving both tracking velocity and accuracy. Besides, the WEM tracking

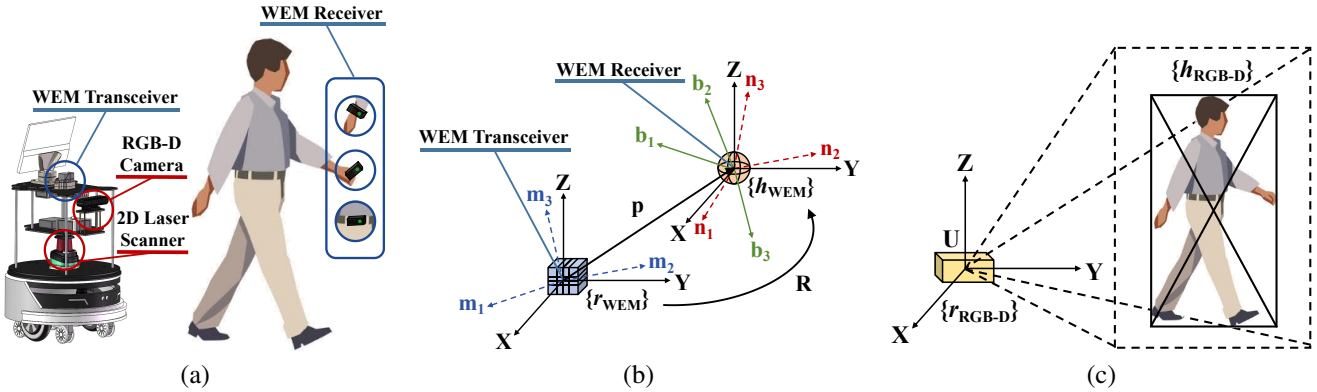


Fig. 4. Schematic diagram of the human target tracking system. (a) The sensor deployment for tracking human targets. (b) The WEM coordinate frame. (c) The RGB-D camera coordinate frame.

device maintains uninterrupted target tracking during visual occlusion scenarios induced by shelves.

The projective correspondence between a 3D pose $\mathbf{U} = [x, y, z]^{\mathrm{T}}$ in the camera coordinate frame $\{r_{RGB-D}\}$ and its corresponding 2D position $\hat{\mathbf{y}}_{RGB-D} = [a, b]^{\mathrm{T}}$ in the human coordinate frame $\{h_{RGB-D}\}$ can be governed by the perspective transformation model [7], as:

$$\begin{cases} \hat{\mathbf{y}}_{RGB-D} = \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} s_m \frac{x}{z} + a_0 \\ s_n \frac{y}{z} + b_0 \end{bmatrix} \\ s.t.\ a_{min} \le a \le a_{max}, b_{min} \le b \le b_{max} \end{cases}, \tag{6}$$

where $s_m > 0$ and $s_n > 0$ are scaling factors in the horizontal and vertical directions. (a_0, b_0) are feature points in the frame $\{r_{RGB-D}\}$. (a_{min}, a_{max}) and (b_{min}, b_{max}) are FOV's sizes.

EKF [20-21] serves as a linearized extension of the Kalman filter, which is designed to estimate the state of a dynamic process and its associated uncertainty from noisy observations. In this study, the human target tracking can be described in an EKF-based state space equation, as:

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \phi_k) \\ \mathbf{z}_{k+1} = \mathbf{h}_{k+1}(\mathbf{x}_{k+1}, \varphi_{k+1}) \end{cases}, \tag{7}$$

where $\mathbf{x}_{k+1}$ and $\mathbf{x}_{k+1}$ are the process state vector and sensor measurement, respectively. ϕ_k and φ_{k+1} are Gaussian noise of process and measurement. $\mathbf{f}_k(\cdot)$ and $\mathbf{h}_k(\cdot)$ are nonlinear process and measurement model functions, respectively.

The state vector of a human target is represented by a unit quaternion, as $\mathbf{p}_k = [p_{0,k}, p_{1,k}, p_{2,k}, p_{3,k}]^{\mathrm{T}}$, a SSR's non-gravitational acceleration $\mathbf{a}_k$, a localization bias of WEM tracking $\mathbf{u}_k$, a disturbance of RGB-D camera $\mathbf{d}_k$. Then, the temporal evolution of the state vector is governed by:

$$\begin{aligned} \mathbf{x}_{k+1}^{-} &= \begin{bmatrix} \mathbf{p}_{k+1} & \mathbf{a}_{k+1} & \mathbf{u}_{k+1} & \mathbf{d}_{k+1} \end{bmatrix}^{\mathrm{T}} \\ &= \mathbf{f}_k\left(\mathbf{x}_k^{+}, 0\right) \\ &= \begin{bmatrix} \exp\left(\mathbf{\Omega}\left(\omega_k\right) T_s\right) & 0 & 0 & 0 \\ 0 & c_a\mathbf{I} & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0 & c_d\mathbf{I} \end{bmatrix} \mathbf{x}_k^{+}, \end{aligned} \tag{8}$$

where $\mathbf{x}_{k+1}^{-}$ and $\mathbf{x}_k^{+}$ are the estimated prior and posterior values of the state vector at time $k+1$ and k, respectively. $0 \le c_a \le 1$ and $0 \le c_d \le 1$ are cut-off frequencies of the first-order low-pass filter. $\mathbf{\Omega}(\omega_k)$, T_s, ω_k, and $\mathbf{I}$ are the skew-symmetric matrix in (9), the sampling period, the strap-down integration of the corrected angular velocity, the identity matrix, respectively.

The matrix exponential operator $\exp(\cdot)$ is computed via a second-order Taylor series approximation, as:

$$\begin{cases} \mathbf{\Omega}\left(\omega_k\right) = \frac{1}{2} \begin{bmatrix} 0 & -\omega_{x,k} & -\omega_{y,k} & -\omega_{z,k} \\ \omega_{x,k} & 0 & \omega_{z,k} & -\omega_{y,k} \\ \omega_{y,k} & -\omega_{z,k} & 0 & \omega_{x,k} \\ \omega_{z,k} & \omega_{y,k} & -\omega_{x,k} & 0 \end{bmatrix} \\ \exp\left(\Omega\left(\omega_k\right) T_s\right) = \mathbf{I} + \mathbf{\Omega}\left(\omega_k\right) + \frac{1}{2}\mathbf{\Omega}\left(\omega_k\right)^2 T_s^2 \end{cases}. \tag{9}$$

To calculate the covariance matrix of the prior error $\mathbf{P}_{k+1}^{-}$, the model linearization is described as follows:

$$\mathbf{P}_{k+1}^{-} = \mathbf{A}_k \mathbf{P}_k^{+} \mathbf{A}_k^{\mathrm{T}} + \mathbf{L}_k \mathbf{Q}_k^{+} \mathbf{L}_k^{\mathrm{T}}, \tag{10}$$

where $\mathbf{A}_k$, $\mathbf{L}_k$, and $\mathbf{Q}_k$ are the state transition Jacobian matrix, process noise covariance matrix, and process noise Jacobian matrix, respectively.

In the measurement updating, the estimated measurement is calculated in (5) and (6). The covariance matrix of the posterior error $\mathbf{P}_{k+1}^{+}$ is described as:

$$\mathbf{P}_{k+1}^{+} = \left(\mathbf{I} - \mathbf{K}_{k+1}\mathbf{H}_{k+1}\right)\mathbf{P}_{k+1}^{-}, \tag{11}$$

where $\mathbf{K}_{k+1}$ and $\mathbf{H}_{k+1}$ are Kalman gain matrix and observation model Jacobian matrix, respectively. $\mathbf{K}_{k+1}$ and $\mathbf{H}_{k+1}$ are given by:

$$\begin{cases} \mathbf{K}_{k+1} = \frac{\mathbf{P}_{k+1}^{-}\mathbf{H}_{k+1}^{\mathrm{T}}}{\mathbf{H}_{k+1}\mathbf{P}_{k+1}^{-}\mathbf{H}_{k+1}^{\mathrm{T}} + \mathbf{M}_{k+1}\mathbf{R}_{k+1}\mathbf{M}_{k+1}^{\mathrm{T}}} \\ \mathbf{H}_{k+1} = \frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{x}}\Big|_{\mathbf{x}_{k+1}^{+}} = \begin{bmatrix} \hat{\mathbf{y}}_{WEM} & \mathbf{I} & 0 \\ \hat{\mathbf{y}}_{RGB-D} & 0 & \mathbf{I} \end{bmatrix} \\ \mathbf{M}_{k+1} = \frac{\partial \mathbf{h}_{k+1}}{\partial \varphi}\Big|_{\mathbf{x}_{k+1}^{+}} = \mathbf{I} \\ \mathbf{R}_{k+1} = \begin{bmatrix} \sigma_{WEM}^2\mathbf{I} & 0 \\ 0 & \sigma_{RGB-D}^2\mathbf{I} \end{bmatrix} \end{cases}, \tag{12}$$

where $\hat{\mathbf{y}}_{WEM}$ and $\hat{\mathbf{y}}_{RGB-D}$ are the measurement matrices of the WEM tracking device and RGB-D camera, respectively. σ_{WEM} and σ_{RGB-D} are standard deviations.

Under unobstructed conditions, the proposed method improves tracking accuracy by dynamically optimizing the $\hat{\mathbf{y}}_{WEM}$ and $\hat{\mathbf{y}}_{RGB-D}$. Conversely, the $\hat{\mathbf{y}}_{WEM}$ exhibits the capacity to sustain uninterrupted target tracking during instances of visual occlusion induced by shelves ($\hat{\mathbf{y}}_{RGB-D} = 0$).

B. Following Control Module

(1) SSR Dynamic Model. The SSR's velocity v and azimuth ψ are controlled by the driving forces F_l and F_r, which are formulated by the differential-drive robot dynamic model, as:

$$\begin{cases} \dot{v} = \frac{-2c_f v}{Mr^2+2J_\omega} + \frac{2c_f l^2(u_r+u_l)}{r^2J_v+2l^2J_\omega} \\ \ddot{\psi} = \frac{-krl\dot{\psi}}{Mr^2+2J_\omega} + \frac{krl(u_r-u_l)}{r^2J_v+2l^2J_\omega} \end{cases}, \tag{13}$$

where u_l and u_r are the driving inputs of the SSR's left and right wheels, respectively. c_f is the viscous friction factor on the flat ground. M is the SSR's mass. r is the wheel's radius. J_ω is the SSR's rotational inertia. l is the distance between the wheel and the center of gravity (CoG). J_v is the moment of inertia around the robot's CoG.

(2) Human-Following Control Law. To achieve the above control objective of SSR dynamics, a human-following control law is designed to achieve a full-state constrained control within a predefined time interval. The human-robot error is defined as $\mathbf{e}_j = \mathbf{x}_{jh} - \mathbf{x}_{jr}$, $j \in \{1,2,3\}$. Then, the human-following control law is formulated as a time-varying function:

$$\begin{cases} \hat{\mathbf{e}}_j = \rho(t)\,\mathbf{e}_j, \\ \rho(t) = \begin{cases} 1 - \left(1 - \frac{t}{T_d}\right) e^{1 - \frac{T_d}{T_d - t}}, \ 0 \le t < T_d \\ 1, \ T_d \le t \end{cases} \end{cases}, \tag{14}$$

where $\rho(t)$ is monotonically increasing with $\rho(0) = 0$ and $\rho(T_d) = 1$ at the predefined time T_d.

(3) Barrier Lyapunov Function for Stability Proof. After the control law is designed for the human following task, Lyapunov-based synthesis is adopted in the performance analysis of the SSR, with the type of nonlinear systems [29-31]. To ensure the states of the robot system meet the required stability criteria in supermarket environments, the barrier Lyapunov function (BLF) is defined as:

$$V = \sum_{j=1}^{3} \frac{1}{2} \log \frac{k_j^2}{k_j^2 - \hat{\mathbf{e}}_j^2}, \tag{15}$$

where k_j is the boundary value for the tracking error $\hat{\mathbf{e}}_j$.

As asserted by [29], the SSR is boundedly stable within a given time T_d, on the condition that the BLF satisfies the following requirements:

$$\dot{V} \le -\frac{\pi}{\eta T_d} \left(V^{1-\frac{\eta}{2}} + V^{1+\frac{\eta}{2}} \right) + \Delta, \tag{16}$$

where $\eta \in (0,1)$ is the convergence rate tuning parameter. $\Delta > 0$ is a residual bound.

Thus, the time derivative of (15) is calculated as:

$$\begin{aligned} \dot{V} &= D_1\hat{\mathbf{e}}_1 \left(\dot{\hat{e}}_1 - \frac{\dot{k}_1}{k_1}\hat{\mathbf{e}}_1 \right) + D_2\hat{\mathbf{e}}_2 \left(\dot{\hat{e}}_2 - \frac{\dot{k}_2}{k_2}\hat{\mathbf{e}}_2 \right) \\ &\quad + D_3\hat{\mathbf{e}}_3 \left(\dot{\hat{e}}_3 - \frac{\dot{k}_3}{k_3}\hat{\mathbf{e}}_3 \right) \\ &\le \sum_{j=1}^{3} \left(-k_j D_j \hat{\mathbf{e}}_j^2 \right) - k_j \Phi_j D_j^m \hat{\mathbf{e}}_j^{2m} - k_j \Phi_j D_j^n \hat{\mathbf{e}}_j^{2n} + \Delta \\ &\le \sum_{j=1}^{3} \left[-k_j \Phi_j \log \left(\frac{k_j^2}{k_j^2 - \hat{\mathbf{e}}_j^2} \right)^m \right] + \Delta \end{aligned}. \tag{17}$$

According to (16), the stability proof requirements in (17) are satisfied. It is concluded that the proposed human-following control law in (14) achieves bounded stability within the predefined time T_d. Consequently, the proposed control objectives are theoretically achievable.

C. Path Planning Module

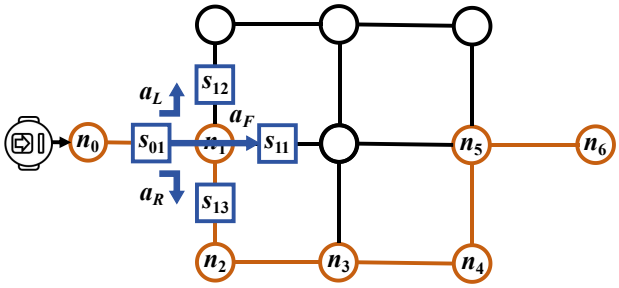


Fig. 5. Schematic diagram of path planning module, including path nodes (n_i), branches (b_j), SSR states (s_{ij}).

(1) Path Node Generation. In Fig. 5, the supermarket environment can be abstracted as a topological map containing path nodes (n_i) and branches (b_j), $i, j \in N^+$. The SSR state (s_{ij}) is defined as the i-th node associated with the j-th branch. Thus, the path nodes are characterized as:

$$n_0 \xrightarrow[o_1]{a_1} n_1^* \xrightarrow[o_2]{a_2} n_2^* \xrightarrow[o_3]{a_3} \cdots \xrightarrow[o_k]{a_k} n_k^*, \tag{18}$$

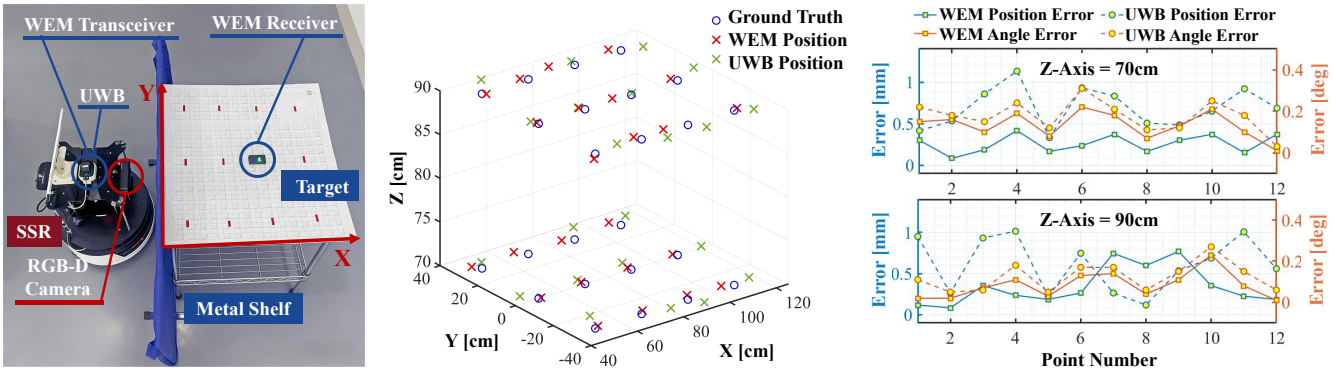
where n_0 is the initial position. a_k is the k-th robot action, including turning left (a_L), turning right (a_R), and moving forward (a_F). o_k is the k-th observed intersection types, including L-type (o_L), T-type (o_T), and cross-type (o_C) intersections in Fig. 3. n_k^* is the k-th temporary path node.

(2) Intersection-Aware Following Law. To follow the human target in supermarket environments, the reward function is developed for SSR, as: $\mathcal{R}(\mathcal{S}, \mathcal{A}) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, where $\mathcal{S}$ is the SSR state set and $\mathcal{A}$ is the SSR action set. Then, the reward function $\mathcal{R}$ is divided into two components: (1) $\mathcal{R}^{avoid}$ denotes the positive reward assigned when the robot successfully reaches the path node position; (2) $\mathcal{R}^{follow}$ denotes the positive reward assigned when the robot successfully maintains human-following control. The k-th reward function is $\mathcal{R}_k = \mathcal{R}_k^{avoid} + \mathcal{R}_k^{follow}$. Thus, the intersection-aware following law is defined as:

$$\mathbf{x}_{k+1} = \min_{v,\psi} \left[-\lambda P_{\pi(s_{ij},o_k)} \sum_k \left(\mathcal{R}_k^{avoid} + \mathcal{R}_k^{follow} \right) \right], \tag{19}$$

where $\pi(s_{ij}, o_k)$ is the following policy. v and ψ are the SSR's velocity and azimuth in (13). $\lambda > 0$ is the weight of the policy. $P_{\pi(s_{ij},o_k)}$ is the Bayesian probability of the policy $\pi(s_{ij}, o_k)$, which is influenced by the SSR state s_{ij} and the observed intersection type o_k.

(3) Bayesian Probability for Target Reacquisition. The SSR traverses all the path nodes and branches in the existing topological map and calculates the Bayesian probability at each state. When $P_{\pi(s_{ij},o_k)} = 1$, the SSR stops updating the path nodes, which indicates that the SSR is reacquiring the human target's visual information.



(a) (b) (c)

Fig. 6. Static measurement results of WEM and UWB tracking device under visual occlusion conditions. (a) The experimental setup. (b) Sensor calibration results. (c) Target tracking results.

III. Experimental Results

A. Visual-WEM Tracking Experiment

To evaluate the tracking accuracy of the perception module in the shelf area, the proposed Visual-WEN tracking method was evaluated under visual occlusion conditions. Furthermore, the accuracy degradation of WEM and UWB signals caused by the metal shelves was further assessed.

In Fig. 6(a), a differential-drive mobile robot was equipped with a WEM tracking device (AmfiTrack Gen2, Amfitech Inc., Denmark) and an RGB-D camera (Astra Mini Pro, Orbbec Inc., China). Moreover, a UWB tracking device (UWB-X2-AOA, Jiuling Inc., China) was compared with the WEM tracking device. Both the WEM and UWB receivers were placed on a metal shelf. In addition, a blue metal baffle was employed to obstruct the RGB-D camera's FOV as well as the wireless signals of both WEM and UWB.

Figures 7(b) and (c) present the static target tracking results. In Fig. 7(b), the UWB and WEM receivers were deployed at 12 predefined positions, which cover an X-coordinate range of (50 cm, 110 cm) and a Y-coordinate range of (-30 cm, 30 cm). The experiment was repeated 10 times for each predefined position. In Fig. 7(c), the WEM tracking obtains mean position and orientation errors of (0.63 $\pm$ 0.26 cm, 0.16 $\pm$ 0.11°) at 70 cm height, while showing (0.57 $\pm$ 0.22 cm, 0.19 $\pm$ 0.15°) at 90 cm height. Besides, the Visual-WEM system exhibits less variation in positioning error compared to the Visual-UWB system. In summary, the WEM tracking outperforms UWB tracking in positioning under visual occlusion conditions.

A comprehensive comparative analysis of dynamic target tracking performance was performed under both visual-occlusion and non-occlusion conditions. Both UWB and WEM receivers were moved alongside the metal shelf for 10 s. The robustness of dynamic tracking was evaluated by using the root mean square error (RMSE), as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}, \tag{20}$$

where y_i and $\hat{y}_i$ are the ground truth and sensor measurement of the i-th measurement point. N is the the total number of measurements.

TABLE I
Dynamic Measurement Results of Multi-Sensor Fusion in Shelf Areas

Sensor (Condition)	Z-Axis	RMSE	DLR
Visual-WEM	70 cm	**(5.17 mm, 3.14°)**	**0.48%**
(Visual Occlusion)	90 cm	**(5.09 mm, 3.01°)**	**0.20%**
Visual-UWB	70 cm	(14.22 mm, 13.40°)	3.00%
(Visual Occlusion)	90 cm	(13.90 mm, 11.24°)	5.40%
Visual-WEM	70 cm	**(4.60 mm, 2.98°)**	/
(Non-Occlusion)	90 cm	**(4.51 mm, 1.81°)**	/
Visual-UWB	70 cm	(7.17 mm, 4.84°)	/
(Non-Occlusion)	90 cm	(7.04 mm, 4.63°)	/

To quantify the sensor signal degradation under occlusion conditions of metal shelves, the data loss rate (DLR) was formulated as:

$$DLR = \frac{M_{lost}}{M_{total}} = \frac{M_{lost}}{ft} \times 100\%, \tag{21}$$

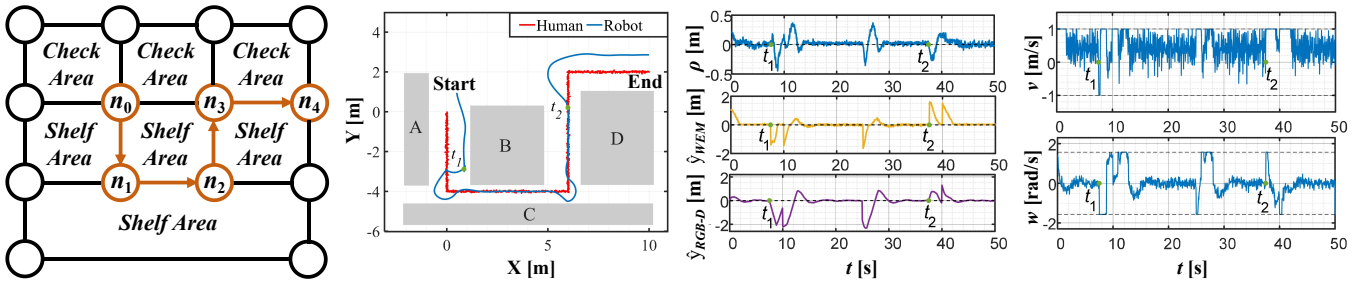
where M_{lost} is the number of lost sensor data. M_{total} is the total number of sensor data. f and t are sampling frequency and time of sensors, respectively.

In Table I, the visual-WEM tracking demonstrates superior performance in dynamic target tracking. The RMSE of visual-WEM shows (5.09 mm, 3.01°) at 900 mm height with visual occlusion conditions, compared to (4.51 mm, 1.81°) under non-occlusion conditions. Additionally, the experimental results indicate that the visual-WEM tracking shows less signal attenuation in the metal shelf area (DLR $< 0.50\%$). In summary, these results highlight the robustness of the proposed visual-WEM tracking in effectively handling visual occlusions.

B. Intersection-Aware Following Experiment

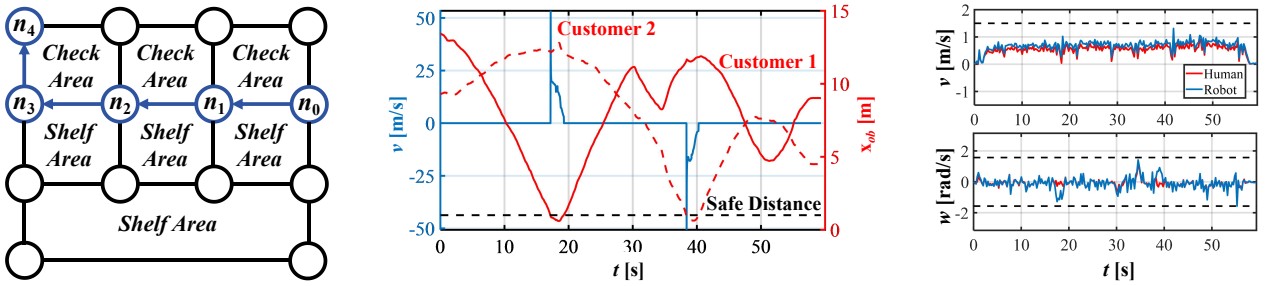
The proposed SSR framework was deployed in a real supermarket. Figures 7(a)-(e) present the intersection-aware following experiment in shelf areas. This experiment evaluates the entire shopping process of a human target's shopping behavior, including walking in the shelf area, picking goods, and checking out. Figures 7(f)-(i) demonstrate the social avoidance experiment in narrow aisles. The SSR facilitates social interaction with two customers while concurrently accompanying the human target throughout the shopping experience.

(a)



(b) (c) (d) (e)

(f)



(g) (h) (i)

Fig. 7. Experimental results of intersection-aware following in supermarket environments. (a) The snapshot of the intersection-aware following experiment. (b) The path node. (c) Trajectories of human target and SSR. (d) The SSR's motion state. (e) The SSR's control input. (f) Snapshots of the social avoidance experiment. (g) The map with the path nodes. (h) The output of the obstacle avoidance control. (i) The control input of the SSR and human target.

Figure 7(a) shows the snapshots of the SSR reaching path nodes. After the intersection had been detected, path nodes (n_0 - n_4) were autonomously generated and updated in Fig. 7(b). Moreover, the recorded trajectories of both the robot and the target are displayed in Fig. 7(c). To pass through the narrow intersection, the SSR moved from the human target's left side to behind at t_1 = 7.8 s. When the SSR entered an open area (t_2 = 37.8 s), the SSR moved to the target's left side. In Fig. 7(d), the kinematic error between the robot and human target remains within a small variation bound, as $[(-0.42, 0.49), (-1.75, 0.20), (-1.84, 0.66)]^{\mathrm{T}}$. Fig. 7(e) shows the SSR's control input $(v, \omega)^{\mathrm{T}}$ varied within the range $[(-0.97, 0.93), (-1.62, 1.74)]^{\mathrm{T}}$. When the visual sensor lost the human target at the shelf intersection (*e.g.*, n_1, n_2, n_3), the SSR reacquired the target by employing the proposed intersection-aware following method.

In Fig. 7(f), the social interaction between the SSR and two customers was recorded. The path nodes are visualized in Fig. 7(g), where the SSR navigates through the same two shelf interactions (*e.g.*, n_1 and n_2) previously encountered in Fig. 7(b). In Fig. 7(h), a safe distance was maintained from the human target and two other customers while accompanying the target on a shopping excursion. The red solid and dashed lines represent the interactions of Customers 1 and 2 with the SSR, respectively. The SSR accelerated while Customer 1 was passing by at 19.2 s; then the SSR decelerated to avoid colliding with Customer 2 ahead at 38.9 s. In response to the varying velocity and angular velocity of the human target in Fig. 7(i), the SSR adjusted control inputs $(v, \omega)^{\mathrm{T}} = [(0, 1.38), (-1.87, 1.92)]^{\mathrm{T}}$ to maintain the human-robot following. Thus, the proposed method ensures safe and friendly interactions between the SSR and customers.

IV. Conclusion

In supermarket environments, shopping service robots significantly improve convenience by eliminating the need for manual cart handling, thereby optimizing customer shopping efficiency. Nevertheless, the current implementation faces substantial challenges when visual occlusion occurs at shelf intersections, where obstructed camera views frequently lead to failures in target tracking and following.

Accordingly, this study addresses two key challenges: the human target tracking under visual occlusions and the human target following in narrow passages. To achieve reliable tracking, we propose an EKF-based visual-WEM fusion method. The EKF prioritizes the high-precision fused pose dynamically while the human target is in the camera's FOV. In the event of visual failure at shelf intersections, the WEM tacking is utilized to maintain stable tracking. Meanwhile, an intersection-aware following method is proposed to enable safe navigation through narrow passages. A human-following control law is proposed in the following control module, ensuring convergence of the SSR's pose to the target human's pose within a predetermined time. The navigation planning module employs an intersection-aware navigation law to direct the SSR through path nodes. Consequently, SSRs can consistently track and follow the human target in dynamic and complex environments.

A key limitation of this study is that the SSR was designed to follow only a single target. The multi-target following task has been demonstrated to be inadequate in dynamic and densely populated supermarket environments, where the SSR must navigate around multiple customers. Consequently, the collaboration between multiple service robots and customers poses a significant issue for future investigation. To further address the challenges in complex and dynamic environments, our future work will focus on two directions. Firstly, the integration of visual-language navigation (VLN) will facilitate the interpretation of verbal requests by the SSR, thus enabling it to guide shoppers to specific items. Secondly, we aim to bridge the gap between structured retail settings and robust service robot applications.

References

[1] S. Song, B. Jun, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Service Robots in a Bakery Shop: A Field Study," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pp. 134-140, 2022.

[2] S. Song *et al.*, "From Attraction to Engagement: A Robot-Clerk Collaboration Strategy for Retail Success," *IEEE Robot. Autom. Lett.*, vol. 10, no. 7, pp. 6672-6679, 2025.

[3] R. Yang, J. Li, Z. Jia, S. Wang, H. Yao, and E. Dong, "EPL-PRM: Equipotential Line Sampling Strategy for Probabilistic Roadmap Planners in Narrow Passages," *Biomim. Intell. Robot.*, vol. 3, pp. 100112, 2023.

[4] T. Sellers, T. Lei, C. Luo, Z. Bi, and G. E. Jan, "Human Autonomy Teaming-Based Safety-Aware Navigation Through Bio-Inspired and Graph-Based Algorithms," *Biomim. Intell. Robot.*, pp. 100189, 2024.

[5] J. Liu, X. Chen, C. Wang, G. Zhang, and R. Song, "A Person-Following Method Based on Monocular Camera for Quadruped Robots," *Biomim. Intell. Robot.*, vol. 2, no. 3, pp. 100058, 2022.

[6] A. Suresh, A. Taylor, L. D. Riek, and S. Martínez, "Robot Navigation in Risky, Crowded Environments: Understanding Human Preferences," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5632-5639, 2023.

[7] W. Chi, J. Wang, and M. Q. -H. Meng, "A Gait Recognition Method for Human Following in Service Robots," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 9, pp. 1429-1440, 2018.

[8] M. -S. Pan and K. -Y. Li, "ezNavi: An Easy-to-Operate Indoor Navigation System Based on Pedestrian Dead Reckoning and Crowdsourced User Trajectories," *IEEE Trans. Mob. Comput.*, vol. 20, pp. 488-501, 2022.

[9] Y. Wu, Y. Wu, A Tamar, S Russell, G. Gkioxari, and Y. Tian, "Bayesian Relational Memory for Semantic Visual Navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, pp. 2769-2779, 2019.

[10] R. Sun *et al.*, "Human Action Recognition Using a Convolutional Neural Network Based on Skeleton Heatmaps from Two-Stage Pose Estimation," *Biomim. Intell. Robot.*, vol. 2, no. 3, pp. 100062, 2022.

[11] H. Yao, J. Peng, Z. Liao, R. Zhao, and H. Dai, "Leg Detection for Socially Assistive Robots: Differentiating Multiple Targets with 2D LiDAR," in *Proc. Int. Conf. Cogn. Syst. Sign. Process.*, pp. 87-103, 2023.

[12] T. -M. Nguyen *et al.*, "VIRAL-Fusion: A Visual-Inertial-Ranging-Lidar Sensor Fusion Approach," *IEEE Trans. Robot.*, vol. 38, no. 2, pp. 958-977, 2022.

[13] Y. Zhang, G. Tian, X. Shao, and J. Cheng, "Effective Safety Strategy for Mobile Robots Based on Laser-Visual Fusion in Home Environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, pp. 4138-4150, 2022.

[14] V. Casamayor-Pujol *et al.*, "A Simple Solution to Locate Groups of Items in Large Retail Stores Using an RFID Robot," *IEEE Trans. Ind. Inform.*, vol. 18, no. 2, pp. 767-775, 2022.

[15] W. Zhao, J. Panerati, and A. P. Schoellig, "Learning-Based Bias Correction for Time Difference of Arrival Ultra-Wideband Localization of Resource-Constrained Mobile Robots," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3639-3646, 2024.

[16] D. von Arx *et al.*, "Simultaneous Localization and Actuation Using Electromagnetic Navigation Systems," *IEEE Trans. Robot.*, vol. 40, pp. 1292-1308, 2024.

[17] L. Zhang *et al.*, "Indoor Mobile Robot Localization Applying IMU/Stereo Camera/LiDAR and Graph-Based Optimization," *IEEE Sens. J.*, vol. 24, no. 13, pp. 21466-21478, 2024.

[18] Z. Li *et al.*, "BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 2020-2036, 2025.

[19] X. Hao, H. Yang, and Y. Xu, "Information Fusion for Positioning System With Location Constraint of Camera and UWB," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 15978-15989, 2025.

[20] C. W. Kang, H. J. Kim, and C. G. Park, "A Human Motion Tracking Algorithm Using Adaptive EKF Based on Markov Chain," *IEEE Sens. J.*, vol. 16, no. 24, pp. 8953-8962, 2016.

[21] B. Yang, E. Yang, L. Yu, and C. Niu, "Adaptive Extended Kalman Filter-Based Fusion Approach for High-Precision UAV Positioning in Extremely Confined Environments," *IEEE/ASME Trans. Mech.*, vol. 28, no. 1, pp. 543-554, 2023.

[22] J. Peng, Z. Liao, Z. Su, H. Yao, Y. Zeng, and H. Dai, "Human-Robot Interaction Dynamics-Based Impedance Control Strategy for Enhancing Social Acceptance of Human-Following Robot," in *Proc. China Autom. Cong. (CAC)*, pp. 7354-7360, 2023.

[23] Z. Su *et al.*, "LQR-Based Control Strategy for Improving Human–Robot Companionship and Natural Obstacle Avoidance," *Biomim. Intell. Robot.*, vol. 4, no. 4, pp. 100185, 2024.

[24] K. Huang, K. Wei, F. Li, C. Yang, and W. Gui, "LSTM-MPC: A Deep Learning Based Predictive Control Method for Multimode Process Control," *IEEE Trans. Ind. Electron.*, vol. 70, pp. 11544-11554, 2023.

[25] J. Peng, Z. Liao, H. Yao, Z. Su, Y. Zeng and H. Dai, "MPC-Based Human-Accompanying Control Strategy for Improving the Motion Coordination Between the Target Person and the Robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pp. 7969-7975, 2023.

[26] C. Cao *et al.*, "Integrated Guidance and Control of Morphing Flight Vehicle via Sliding-Mode-Based Robust Reinforcement Learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 55, no. 5, pp. 3350-3362, 2025.

[27] J. Yuan, S. Zhang, Q. Sun, G. Liu, and J. Cai, "Laser-Based Intersection-Aware Human Following With a Mobile Robot in Indoor Environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 354-369, 2021.

[28] B. Lewandowski *et al.*, "Socially Compliant Human-Robot Interaction for Autonomous Scanning Tasks in Supermarket Environments," in *Proc. Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, pp. 363-370, 2020.

[29] J. Chen, C. Hua, D. Mu, and Y. Zhang, "Predefined-Time Full-State Constrained Control for Mobile Robot Systems With Deferred Constraints," *IEEE Trans. Ind. Electron.*, vol. 72, no. 3, pp. 2968-2976, 2025.

[30] H. Ye *et al.*, "Follow-Bench: A Unified Motion Planning Benchmark for Socially-Aware Robot Person Following," arXiv, 2025. Available: https://arxiv.org/abs/2509.10796.

[31] J. Peng *et al.*, "A Dual Closed-Loop Control Strategy for Human-Following Robots Respecting Social Space," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 11252-11258, 2024.