



Review

A survey on the visual perception of humanoid robot

Teng Bin^{a,b}, Hanming Yan^c, Ning Wang^{a,d}, Milutin N. Nikolić^e, Jianming Yao^{a,f}, Tianwei Zhang^{a,*}^a The Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China^b College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China^c College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China^d College of Mechatronics Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China^e Faculty of Technical Sciences, University of Novi Sad, Novi Sad 21102, Serbia^f Electronic science and technology, Guangdong University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:

Received 27 August 2024

Revised 18 October 2024

Accepted 14 November 2024

Available online 26 November 2024

ABSTRACT

In recent years, humanoid robots have gained significant attention due to their potential to revolutionize various industries, from healthcare to manufacturing. A key factor driving this transformation is the advancement of visual perception systems, which are crucial for making humanoid robots more intelligent and autonomous. Despite the progress, the full potential of vision-based technologies in humanoid robots has yet to be fully realized. This review aims to provide a comprehensive overview of recent advancements in visual perception applied to humanoid robots, specifically focusing on applications in state estimation and environmental interaction. By summarizing key developments and analyzing the challenges and opportunities in these areas, this paper seeks to inspire future research that can unlock new capabilities for humanoid robots, enabling them to better navigate complex environments, perform intricate tasks, and interact seamlessly with humans.

© 2024 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

As labor shortages intensify and the demand for automation increases, the need for humanoid robots in sectors such as services, manufacturing, and healthcare is growing rapidly. This demand has driven companies and research institutions to accelerate their development efforts. The rapid advancement of humanoid robots in recent years is attributed to interdisciplinary research across fields such as mechanical engineering, electronics, computer science, and cognitive science, as well as key technological breakthroughs in perception and control.

Humanoid robots, (see Fig. 1) like all robots, are composed of three primary systems: the motion system, the perception system, and the control system. The motion system, comprising rotational and linear joints, provides the robot with the necessary degrees of freedom to perform various tasks. The control system is responsible for motion planning and issuing control commands to the motion system to execute tasks. The perception system, particularly the vision system, plays a crucial role in enabling the robot to interact effectively with its environment.

The integration of a vision system in humanoid robots is essential for several reasons. First, traditional methods of robot

localization often suffer from limitations such as low positioning accuracy, errors in motion models, and issues like ground slippage, which can lead to significant deviations during operation [1]. Vision-based localization provides a critical feedback mechanism, allowing for closed-loop navigation that can correct these errors and enhance overall accuracy. By analyzing visual data, the robot can continuously update its understanding of its position and surroundings, leading to more reliable and precise navigation [2].

In tasks involving object manipulation, vision is indispensable. To successfully grasp and manipulate objects, the robot must accurately determine the shape and pose of the target relative to itself. Vision algorithms enable the robot to obtain this information, which is then used to plan and execute the required movements [3]. In open-loop control systems, where motion planning is limited by imperfect modeling, the absence of feedback can lead to significant inaccuracies. By incorporating visual feedback, these systems can be transformed into closed-loop systems, where the vision system continuously tracks the robot's and the object's relative poses, significantly improving the accuracy and robustness of interactions [4].

Moreover, in human-robot interaction, vision is crucial for perceiving and understanding human intentions, such as communication, emotional states, and movement goals [5]. The robot's ability to detect and interpret these visual cues is fundamental to responding appropriately and even collaborating with humans to complete complex tasks.

* Corresponding author.

E-mail address: zhangtianwei@cuhk.edu.cn (T. Zhang).

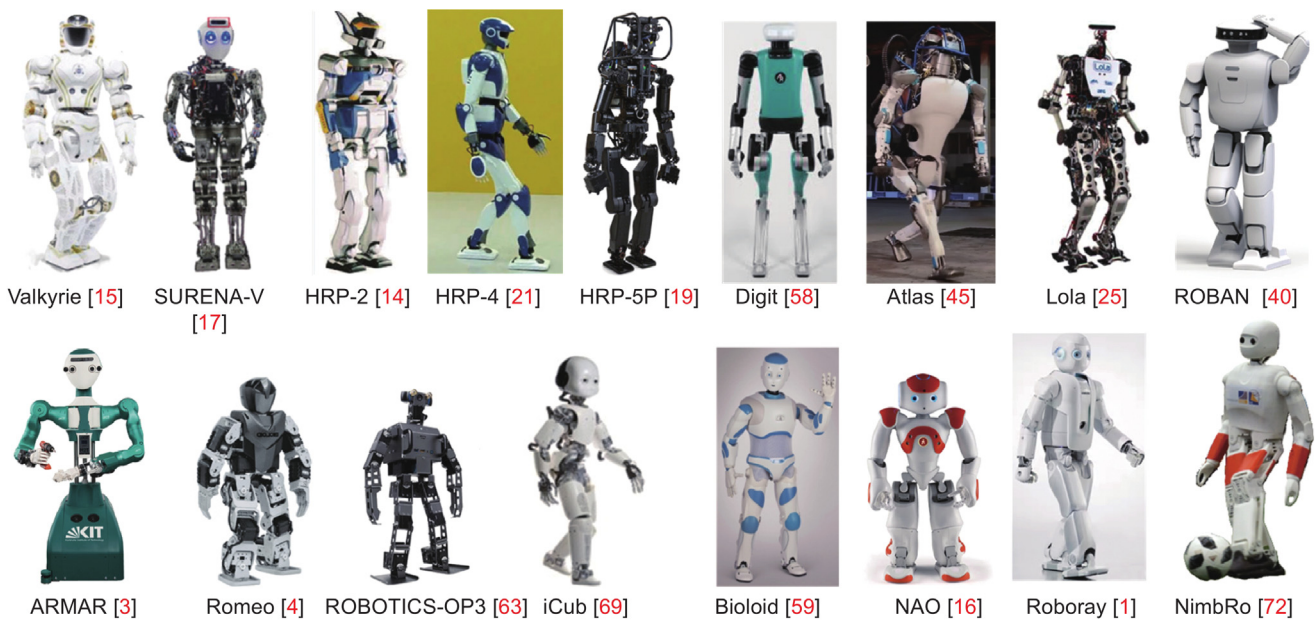


Fig. 1. Humanoid robots featured in the referenced studies.

This paper explores the critical role of vision in the development of humanoid robots. It covers visual-based state estimation, including Visual Simultaneous Localization and Mapping (V-SLAM) for accurate navigation, as well as the challenges posed by dynamic environments, the integration of multi-sensor data, and strategies to mitigate motion blur. Additionally, the paper delves into visual-based environment interaction, focusing on terrain semantic information extraction, upper-body visual control, and vision-based human-robot interaction. These advancements highlight the importance of vision in enhancing the autonomy, versatility, and effectiveness of humanoid robots in various applications.

2. Visual-based state estimation

Accurate state estimation is crucial for effective navigation and interaction in humanoid robotics. Visual-based state estimation utilizes visual information to determine the robot's position, orientation, and movement. This is essential for motion planning and navigation. Current mainstream methods for robot state estimation include odometry-based approaches and Simultaneous Localization and Mapping (SLAM) technology. This chapter covers V-SLAM, dynamic environment challenges, multi-sensor fusion, and addressing motion blur. It will introduce classical V-SLAM frameworks adapted for humanoid robots and discuss solutions to various challenges encountered in V-SLAM applications. Table 1 categorizes the problems and methods discussed in Sections 2.2 and 2.3, while Table 2 provides a comparison of the multi-sensor fusion methods covered in Section 2.4.

2.1. V-SLAM applications in humanoid robots

V-SLAM methods can be categorized into feature-based methods and direct methods based on the type of information used for localization and mapping.

Feature-based methods, such as those discussed in [6–9], and [10], rely on detecting and matching a certain number of feature points across multiple images using their descriptors. This process provides camera pose estimation information. The descriptors and keypoint locations form features that the algorithm uses for tracking and mapping. However, these methods may

face challenges in environments with lack of texture or dynamic changes.

Direct methods, such as those described in [11,12], and [13], directly use raw image pixel intensity information, eliminating the need for feature extraction or matching. These methods estimate the camera's pose and environmental structure by minimizing photometric error. ElasticFusion [11] is a notable direct method. It uses photometric error and incorporates the depth point cloud from an RGB-D camera for ICP geometric error, jointly optimizing both to improve pose estimation accuracy. It constructs a dense surfel map and achieves globally consistent mapping through local and global loop closure optimization combined with non-rigid surface deformation.

As the first real-time closed-loop SLAM work on humanoid robots, Stasse et al. [14] uses an Extended Kalman Filter (EKF) to fuse the trajectory of the robot's waist generated by the 3D Linear Inverted Pendulum Model, camera model state estimation, and inertial measurement data. This approach inspired subsequent state estimation methods for humanoid robots. The method [1] integrates joint encoders and forward kinematic models to compute the robot's pose relative to the starting point, incorporating IMU data. This integration allows the SLAM module to compensate for motion errors and provide more accurate and robust predictive models. Building on [11], Scona et al. [15] fused leg joint sensors, foot force-torque sensors, and an IMU rigidly connected to the robot's pelvis with visual odometry, effectively compensating for the shortcomings of visual systems in feature-scarce or varying lighting conditions. The mapping results are shown in Fig. 2

To address feature matching failure due to motion blur, a new binary descriptor, DLab, was proposed in [16]. It combines color, depth, and intensity information to improve the robustness of feature matching. Various V-SLAM algorithms were experimented on humanoid robots in [17], comparing localization accuracy, loop closure detection, mapping capability, and algorithm robustness.

Additionally, RTAB-Map [18], as an open-source SLAM library supporting multi-sensor integration, can be easily ported to robotic systems, such as those described in [19,20].

Table 1
Common issues and solutions in visual SLAM for humanoid robots.

Problem category	Method description	Methods	SLAM framework	Reference
Dynamic object	Dynamic object priors	Semantic segmentation	PoseFusion	[21,22]
		3D model	RTAB-MAP	[19]
	Without dynamic object priors	Optical flow model	ElasticFusion	[23]
			PFD-SLAM	[24]
			-	[25]
		Point cloud clustering	-	[26]
		Sound source model	ORB-SLAM2	[27,28]
Motion blur	Remove blurred frames	-	PoseFusion	[22]
			-	[30]
			ORB-SLAM	[31]
	Improve matching quality of Blurred frame	-	-	[32]
			Feature-based SLAM	[33]



Fig. 2. Dense mapping results on the Valkyrie robot. Reprinted with permission from [15]. Copyright 2017, IEEE.

2.2. Challenges in dynamic environments

Robotic localization requires static reference points relative to the world coordinate system. During motion, the robot calculates its pose changes based on the relative changes in these reference points. However, in complex environments where humanoid robots operate, there are often many moving objects. To mitigate the impact of localization errors like those shown in Fig. 3 on state estimation, numerous research efforts have been made in recent years. Dynamic SLAM can be categorized into two types based on the availability of prior information about dynamic objects.

2.2.1. With prior information of dynamic objects

In many scenarios, the dynamic objects encountered by robots are known in advance, such as workers and containers in a factory. In such cases, the appearance of these objects can be converted into feature information beforehand, such as RGB image semantics or 3D models. When part of the image or point cloud captured by the robot matches this feature information, it can be considered a potential dynamic object. This allows the system to disregard these dynamic objects and focus on static reference points for state estimation, as shown in Fig. 4.

For instance, Zhang et al. [22] proposed a semantic-based human segmentation dynamic SLAM method, which uses deep

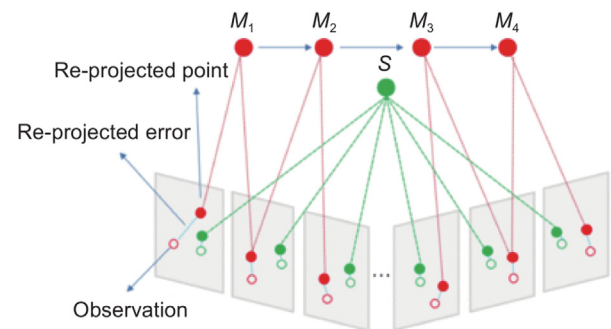


Fig. 3. Pose estimation errors occur when dynamic landmarks are included. Dynamic landmark M moves quickly between positions M_1 to M_4 , resulting in inconsistent observations and triangulation errors. In contrast, static landmark S remains in the same location, providing consistent observations and accurate triangulation across multiple frames. Reprinted with permission from [34]. Copyright 2022, IEEE.

learning-based human detection and graph-based segmentation to separate moving humans from the static environment. By using OpenPose for human joint detection, the method improves human dynamics recognition and provides accurate human location information for segmentation. The Min-Cut algorithm is then used to segment the RGB-D point cloud, effectively separating moving humans from the static environment. Building on this, Zhang et al. [21] addressed the issue of humanoid robots falling and proposed a new camera pose relocalization method based on semantic mapping and point cloud registration, using the 3D Normal Distribution Transformation (NDT) method for point cloud registration.

In [19], point cloud information from RGB-D sensors is used to match 3D models of known objects. Optimization algorithms such as ICP are used to minimize the difference between the expected pose of the object point cloud and the actual captured depth map point cloud, thereby determining the object's position and pose in space. During SLAM tracking and mapping, the information on the tracked model parts is ignored. Additionally, many works utilize prior information on objects for object tracking, as discussed in Section 3.2 of this paper.

2.2.2. Without prior information of dynamic objects

In more complex scenarios, it is difficult to enumerate all possible dynamic objects and convert them into prior feature information. Therefore, many works have explored how to detect dynamic objects solely through image information. Zhang et al. [23] proposed a dynamic segmentation method based on optical flow residuals. It uses PWC-Net to estimate optical flow and utilizes optical flow residuals for dynamic segmentation, achieving static background reconstruction through an iterative

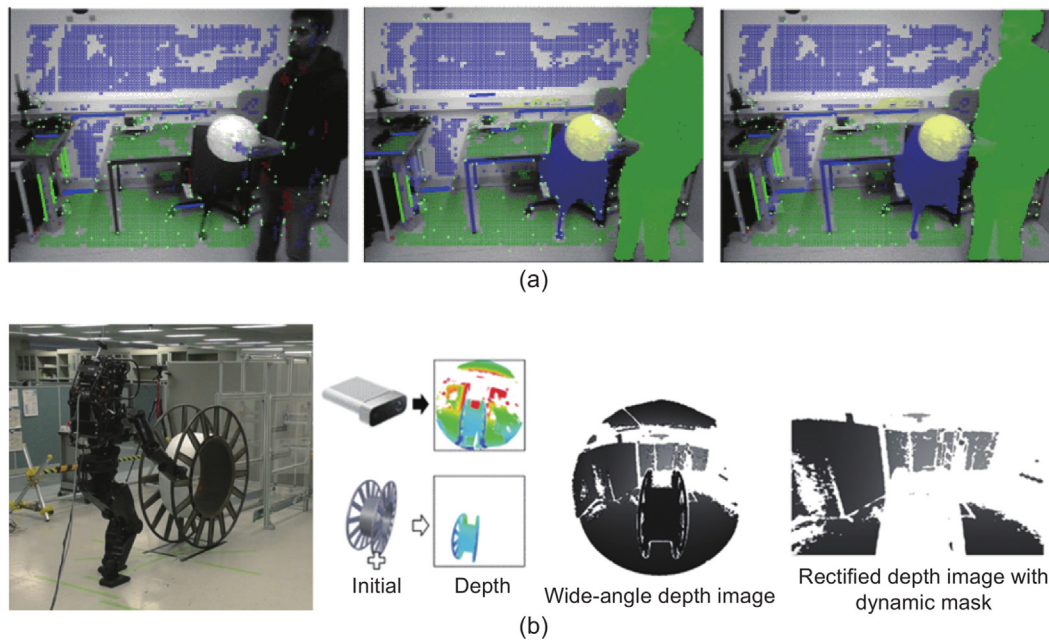


Fig. 4. Prior information on dynamic objects. (a) Features on potential dynamic objects in the image are ignored based on a pre-trained semantic network model. Reprinted with permission from [35]. Copyright 2022, IEEE. (b) Dynamic objects are tracked using pre-established 3D models and point cloud registration, and these features are excluded from the V-SLAM system. Reprinted with permission from [19]. Copyright 2024, IEEE.

process. Similarly, Zhang et al. [24] proposed a particle filter-based non-prior semantic dynamic segmentation method, which first uses Grid-based Motion Statistics (GMS) and optical flow to compute frame difference images as observation measurements and establishes the motion equations of the particle filter using Gaussian distributions. ORB-SLAM [8] employs RANSAC for feature point matching, which can counteract the misalignment caused by feature points on dynamic objects to some extent.

For point cloud clustering, Wahrmann et al. [25] proposed a dynamic point cloud clustering algorithm using the Gaussian Mixture Model to cluster point cloud data and determine the potential distribution of objects. By iteratively adjusting the parameters of the Gaussian distributions until convergence, the algorithm accounts for object motion using a Kalman filter, thus tracking dynamic objects more accurately.

Using motion priors from other sensors, Long et al. [27] consider all dynamic parts of the scene as a single rigid body, segmenting and tracking static and dynamic components. This method can simultaneously localize and reconstruct static backgrounds and rigid dynamic parts even in environments with severe occlusion caused by dynamic objects. Subsequently, they proposed an RGB-D SLAM method for indoor planar environments with multiple large dynamic objects [28]. Furthermore, DynaVINS [29] proposed a novel Bundle Adjustment (BA) method that uses IMU preintegration pose priors to eliminate features associated with dynamic objects, reducing the impact of temporary dynamic objects on loop closure detection. Combining sound source localization technology and visual SLAM schemes, Zhang et al. [26] proposed a visual-audio fusion method to eliminate the impact of dynamic obstacles on multi-agent systems. Using Direction of Arrival (DOA) technology, the method detects noise directions of moving robots and marks and processes the areas of dynamic obstacles in the RGB-D point cloud through heterogeneous information fusion.

2.3. Challenges of motion blur

Visual odometry methods rely heavily on image quality for motion tracking. Feature-based methods require clear images

to extract accurate feature points and descriptors, while direct methods depend on the photometric consistency assumption. Both methods may fail in the presence of motion blur. Humanoid robots experience significant vibrations during walking, and most robotic visual sensors lack built-in stabilization. Therefore, addressing visual tracking failure caused by severe motion is a critical area of research. Solutions can be categorized into passive methods that discard blurred frames and active methods that enhance inter-frame matching quality.

Discarding blurred frames requires detecting them first. Zhang et al. [22] used a Laplacian-based method to evaluate image blur, which describes rapidly changing boundary regions in the image using the Laplacian operator. The blur score threshold is adjusted in real-time according to the humanoid robot's motion, ensuring visual odometry robustness by removing highly blurred images. Similarly, Mutlu et al. [30] proposed a real-time motion blur metric called Motion-based Motion Blur Metric, which uses inertial sensor measurements for computation. By predicting motion blur, frames with high motion blur can be avoided before image capture. Utilizing the synchronization between the robot's gait cycle and the image motion blur cycle, Fan et al. [31] filtered images for odometry tracking based on the robot's pose calculated by the IMU, while Hourdakis et al. [36] selected keyframes based on the unilateral ground force of the robot's foot.

Active methods to enhance inter-frame matching quality include [32], which proposed a multi-frame deblurring method based on time-varying camera motion to overcome the ill-posed nature of deblurring problems. Different frames have different blur Point Spread Functions (PSFs), resulting in varying frequency losses. Combining the complementary spectral information from different frames reduces reconstruction artifacts and improves deblurring quality. MBA-VO [33] experimented with two deblurring networks, SRNDeblurNet [37] and DeblurGANv2 [38], to enhance keyframe quality. For blurred frames, it improved traditional direct visual odometry by parameterizing the camera poses at the start and end of the exposure (T_{start} and T_{end}) and linearly interpolating between them to establish a local camera trajectory model during the exposure period. By optimizing the start and end poses, MBA-VO minimized the photometric consistency loss

Table 2
Comparison of multi-sensor fusion methods.

Reference	Visual sensor	Fusion sensors			Fusion framework	Fusion purpose		
		IMU	Joint encoder	F/T sensor		Accuracy	Robustness	Others
[14]	Monocular	✓	✓	×	EKF	✓	✓	–
[1]	Stereo	✓	✓	×	EKF	✓	✓	–
[39]	Monocular	✓	✓	FSR	EKF	✓	✓	–
[2]	Stereo	✓	✓	FSR	EKF Cascade	✓	✓	a
[40]	RGB-D	✓	×	×	EKF	✓	–	–
[41]	Stereo	✓	✓	×	EKF	✓	–	b
[15]	RGB-D	✓	✓	✓	Optimization	✓	✓	–
[42]	Stereo	✓	✓	×	Optimization	✓	–	c
[43]	RGB-D	✓	✓	×	Optimization	✓	✓	–

^a Estimate 3D-CoM position, velocity, and external force.

^b Achieve low latency and drift in base state, integrating into MPC framework.

^c Estimate force and torque of the robot leg using observer model.

Note: F/T Sensor refers to the force/torque sensor located at the robot's foot.

between the pixel intensities captured in the current frame and those in the reference image synthesized through re-blurring.

2.4. Sensor fusion

In complex environments, a single sensor might be interfered with or fail. Multi-sensor fusion can maintain system functionality even when one sensor fails, enhancing system robustness. For example, in environments with drastic changes in lighting conditions, visual sensors may fail, but inertial sensors can still provide stable pose estimation. Improving the accuracy of state estimation and reducing uncertainty are also primary goals of multi-sensor fusion.

Based on the algorithm framework, multi-sensor fusion SLAM can be divided into filter-based methods and graph optimization-based methods. Filter-based fusion typically uses recursive estimation methods, where filters continually correct and update state estimates based on prior estimates and measurements. Examples include the Kalman Filter (KF), EKF, Unscented Kalman Filter (UKF), and Particle Filter (PF), which are commonly used for relatively simple linear or weakly nonlinear systems. In contrast, graph optimization-based methods represent sensor data as a graph structure, where nodes represent the robot's poses at different times or feature points in the map, and edges represent constraints between sensor measurements. These methods optimize the poses and feature points in the graph using global optimization algorithms, such as nonlinear least squares optimization.

Due to the limited computational power onboard humanoid robots, early multi-sensor fusion methods, such as [1,14], used filter frameworks. Oriolo et al. [39] utilized joint encoder readings and a differential kinematics mapping from the support foot (the foot currently bearing the robot's weight) to the torso to predict the torso's position and orientation. PTAM [9] served as the visual odometry for the robot's head. The kinematic, inertial, and visual information was fused using EKF to improve the localization accuracy and robustness of humanoid robots. As shown in Fig. 5, Piperakis et al. [2] proposed a serial state estimation framework called SEROW (State Estimation ROBot Walking), which uses a two-stage EKF to fuse joint encoders, IMU, foot pressure, and visual odometry measurements to estimate the 3D-CoM position, velocity, and external forces acting on the CoM of a walking humanoid robot. Leng et al. [40] integrated visual odometry (VO) and inertial navigation system (INS) to enhance localization accuracy and introduced a Temporal Convolutional Network (TCN) to learn the noise parameters in the Kalman filter, enhancing the filter's robustness and accuracy. Dhédin et al. [41] proposed a loosely coupled EKF method for quadruped robots, combining VIO and leg odometry to estimate a low-latency, low-drift base state suitable for agile movements.

With the advancement of computational power, researchers have begun applying graph optimization frameworks to humanoid robots, quadruped robots, and car-arm systems. Building on [11], the approach [15] uses leg joint sensors, foot force–torque sensors, and an IMU rigidly connected to the robot's pelvis to estimate the robot's state. It combines kinematic-inertial state estimation with visual SLAM to address the shortcomings of visual systems in feature-scarce or varying lighting conditions. Additional residual terms calculate the error between visual and kinematic-inertial estimations, incorporating this into the global energy function for optimization using the Gauss–Newton nonlinear least squares method. Kang et al. [42] proposed a factor graph optimization method for legged robots to estimate states, including external torques and leg forces. This method builds a dynamic model of the legged robot (an 18-DOF floating base dynamic model), including the trunk position, orientation, and joint angles, and a nonlinear disturbance observer to estimate the product of the contact Jacobian matrix and ground reaction forces for each leg. A factor graph based on visual odometry is constructed, introducing leg force factors and external torque factors, and the corresponding residuals are built. Houseago et al. [43] in car-arm systems fused data from wheel odometry, mechanical arm forward kinematics, and visual odometry. Error terms based on dynamic and odometry data are added to the SLAM system's cost function, which are optimized together with geometric alignment errors. The dynamic error term (e_{kin}) constrains the relative transformation between the camera and the base, making it close to the values measured by robot dynamics. This involves errors in position and orientation, weighted by an information matrix. The odometry error term (e_{odom}) constrains the relative transformations between consecutive base poses, making them close to the values measured by odometry. This also involves errors in position and orientation, weighted by an information matrix. In each tracking step, the poses of the camera and base are updated by minimizing the total cost function, which includes geometric alignment costs, dynamic errors, and odometry errors. For dynamic and odometry constraints, Jacobian matrices and residuals are computed, which can be analytical. The least squares method is used to solve the Jacobian matrices and residuals to compute the minimum update of the states. After optimization, partial marginalization of previous states and errors is performed to build a linear prior for the next iteration.

3. Visual-based interaction

Effective interaction with the environment is crucial for the autonomy and versatility of humanoid robots. This chapter explores the techniques and technologies that enable robots to perceive, interpret, and interact with their surroundings through

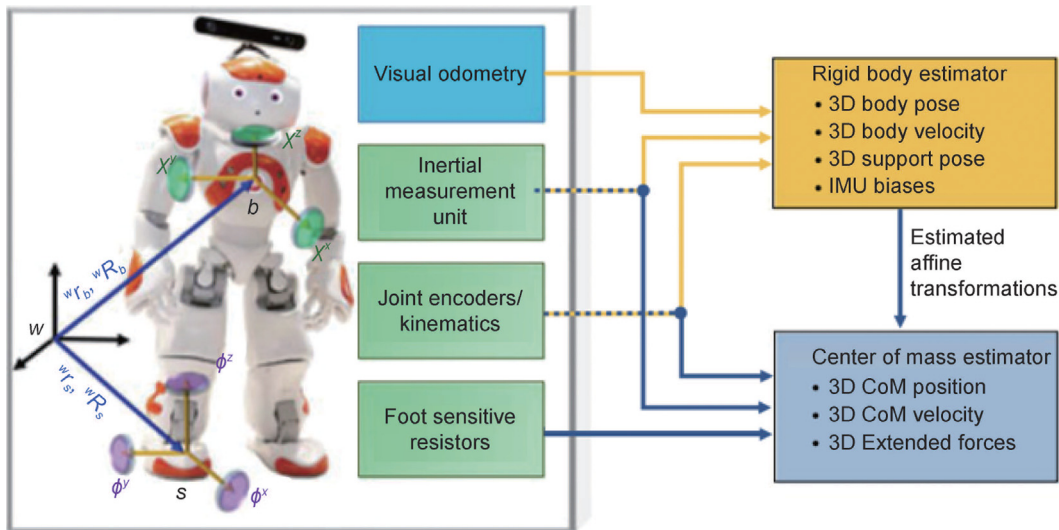


Fig. 5. The data from multiple sensors are fused using a cascaded EKF framework, which reduces the uncertainty in state estimation and provides a three-dimensional center of mass estimation for walking humanoid robots. The cascade state estimation scheme includes both a rigid body estimator and a CoM estimator. Reprinted with permission from [2]. Copyright 2018, IEEE.

Table 3
Classification of visual interaction.

Interaction type	Method description	Methods	Reference
Environment perception	SLAM	Sparse mapping Dense mapping	[1,14,16,17,26,39,40] [2,15,19,21,22,36]
	Semantic environment mapping	-	[44-47]
Object manipulation	Visual grasping	-	[3,20,48]
	Visual servoing	-	[4,49-52]
	Other operations	Pouring liquids Opening plastic bags	[53,54] [55]
	Human-Robot interaction	Intention perception	Facial intention Gesture intention
	Remote collaboration	-	[61-63]

visual inputs. The focus is on terrain semantic information extraction, visual control of the upper body, and vision-based human-robot interaction. These methods allow humanoid robots to navigate complex terrains, perform precise manipulation tasks, and engage in meaningful interactions with humans, thereby extending their functional capabilities and enhancing their integration into human-centric environments. Table 3 categorizes the content of this chapter.

3.1. Environment semantic information extraction

Environment terrain perception and the construction of efficient traversable semantic maps are crucial technologies for humanoid robots to navigate complex environments. Humanoid robots typically rely on sufficiently large flat areas as footholds, and the mapping process requires the constructed map to have a high overlap with the actual environment to ensure accurate planning and robot safety. Due to their higher degrees of freedom, humanoid robots can perform more complex tasks than traditional wheeled robots, such as jumping and climbing stairs. However, these complex movements require the robot to accurately and efficiently perceive its surroundings and establish reliable semantic maps. Therefore, it is essential to conduct more research and discussion on planar semantic maps designed for humanoid robots.

As shown in Fig. 6(a), Roychoudhury et al. [44] utilized plane extraction algorithms similar to those in [66,67] to aid robot localization. In [66], point clouds are divided into non-overlapping

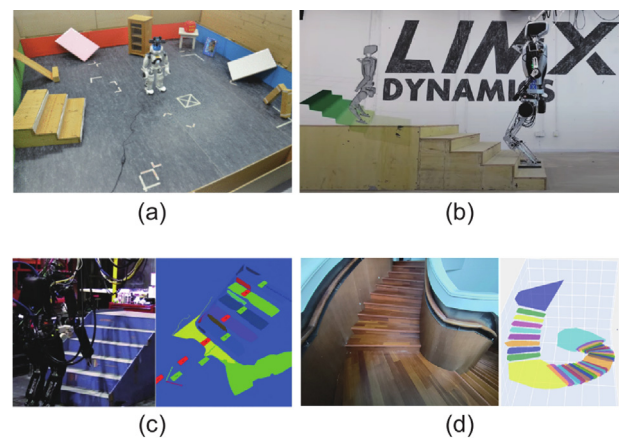


Fig. 6. Terrain semantic perception by humanoid robots. (a) NAO robot performing planar semantic mapping. Reprinted with permission from [44]. Copyright 2022, IEEE. (b) The CL1 robot by LIMX Dynamics perceiving staircase terrain. (c) Nadia robot performing real-time planar mapping and footprint planning. Reprinted with permission from [64]. Copyright 2024, IEEE. (d) Real-time mapping of a complete spiral staircase. Reprinted from [65]. Copyright 2024, licensed under CC BY-NC-SA 4.0.

point groups in image space, then represented in a graph structure, and merged using agglomerative hierarchical clustering (AHC) until the plane fitting error surpasses a threshold. The seed region growing algorithm in [67] evaluates point inclusion

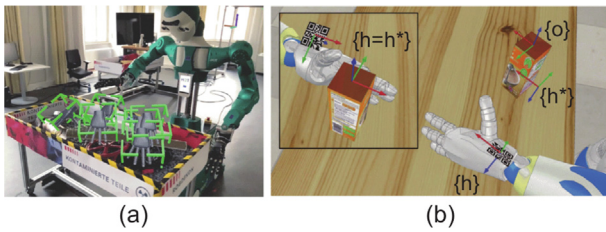


Fig. 7. Visual Control. (a) Visual Grasping: Locating the object's pose relative to the robot and planning the grasping path through visual input. Reprinted with permission from [3]. Copyright 2022, IEEE. (b) Visual Servoing: Integrating visual information into the control model to form a closed-loop control system based on visual feedback. Reprinted with permission from [49]. Copyright 2016, IEEE.

in existing plane segments based on distance and normal vector similarity, followed by plane merging. This approach selects seed points centrally within the plane, though it is still susceptible to noise interference.

Bertrand et al. proposed a novel method combining octrees and nearest neighbor search to extract planar regions from point cloud data, providing an effective environmental representation for legged robot gait planning [45]. However, the real-time performance of converting point clouds to octrees needs further improvement.

Fankhauser et al. introduced an innovative terrain mapping method based solely on kinematic and inertial measurement proprioception for localization [68]. This method considers state estimation drift and uncertainty and the noise model of distance sensors, generating probabilistic terrain estimates with upper and lower confidence intervals. Building on this, Miki et al. [46] converted point clouds into grid-based elevation maps, and used GPU acceleration for height updates and ray projection to eliminate dynamic obstacle parts, achieving better real-time performance and practicality on robots. Further extending this work, Erni et al. [47] integrated semantic information with elevation maps, creating a rich environmental representation useful for robot navigation, path planning, and understanding the surrounding environment.

Bin et al. emphasized the importance of filtering depth point clouds and demonstrated that using anisotropic diffusion filtering improves the quality of point clouds, thus enhancing the effectiveness of plane extraction [65]. Mishra et al. [64] further developed a plane-based SLAM system by fusing extracted planes with robot forward kinematic odometry, creating a robust framework for SLAM.

3.2. Visual control of the upper body

The visual control of the upper body in humanoid robots is critical for achieving precise and adaptable manipulation tasks. This section explores various approaches to integrating visual information into the control systems of humanoid robots, focusing on enhancing their ability to interact with objects and environments. Visual control encompasses a range of techniques, from visual grasping, where the robot must accurately perceive and manipulate objects, to visual servoing, which involves using real-time visual feedback to adjust the robot's actions dynamically. Additionally, other forms of visual feedback control are discussed, including the handling of complex, non-rigid objects and the execution of tasks that require a high level of precision and adaptability. Together, these methods contribute to the overall robustness and versatility of humanoid robots in performing sophisticated tasks in dynamic and unpredictable environments.

3.2.1. Visual grasping

To enable robots to perform tasks such as grasping objects, accurate pose information of the object relative to the robot must be obtained through vision, followed by planning a feasible grasping path and controlling the joint motors to execute the corresponding actions. When the 3D model of the object is available, the 6D pose of the target object can be obtained using RGB-D point clouds. Tsuru et al. addressed the problem of exploration and target object grasping in unknown environments, proposing an integrated autonomous humanoid system framework that includes object recognition, environment perception, motion planning, and bipedal walking [20]. This system uses 3D models and RGB-D point cloud data to estimate the 6-DoF pose of the target object in real-time. As shown in Fig. 7(a), Pohl et al. [48] studied the impact of expert knowledge on the grasp selection process and demonstrated that selecting the correct autonomously generated grasp candidates significantly improves the grasp success rate of humanoid robots. Building on this, Baek et al. [3] defined four metrics based on visual and proprioceptive information (grasp height, distance to the center, support relationship, and operability) and modeled these metrics as Gaussian distributions. Using the Probabilistic Action Extraction and Fusion method, they processed point cloud data to extract grasp candidates and calculate the uncertainty of the grasp poses. By analyzing a large amount of random grasp experiment data and their corresponding metrics, they developed a probabilistic model for evaluating grasp candidate scores. Although their method provides good grasp candidates in complex environments, the entire visual grasping system is open-loop, requiring very high precision in the metric models and exhibiting weak resistance to interference.

3.2.2. Visual servoing

Visual servoing technology receives image data in real-time and adjusts control commands based on image feedback, continuously correcting errors until the desired goal is achieved. It can adapt to changes in dynamic environments and targets, adjusting control strategies in real-time to cope with uncertainties and external disturbances. Agravante et al. [4] explored integrating visual servoing into the overall optimization control framework of humanoid robots. The visual servoing task is formulated as a quadratic optimization problem addressing acceleration, allowing visual constraints such as field of view and occlusion avoidance to be handled as inequalities. This approach demonstrates how to seamlessly integrate visual servoing tasks into the existing overall control framework of humanoid robots, simplifying task prioritization with only one posture task as a regularization term. As shown in Fig. 7(b), Claudio et al. [49] demonstrated how visual servoing technology helps the humanoid robot Romeo achieve single-handed and bimanual manipulation tasks. Using visual information in a closed-loop control system allows for the execution of highly repetitive tasks without precise calibration of the robot's kinematic model. The paper not only showed single-handed grasping but also implemented bimanual coordinated manipulation strategies using a master-slave approach, achieving vision-based bimanual coordination.

Paolillo et al. [50] focused on online visual tracking and manipulation of articulated objects (e.g., furniture drawers). They used line features to track objects, and the algorithm maintained and updated information about object edges using an edges table, including tracking status, visibility, and visual line parameters θ and ρ . When tracking failed, the algorithm attempted to find an alternative line from the edges table. Hoffman et al. [51] addressed whole-body motion control of humanoid robots using visual servoing and conservation of centroidal momentum in the absence of contact and gravity. Kheddar et al. [52] used a visual

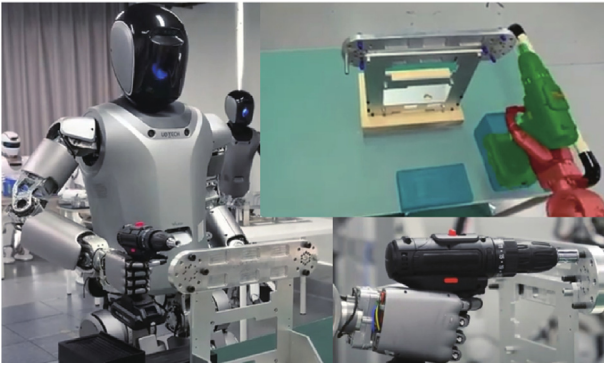


Fig. 8. Walker S using visual servoing and reinforcement learning to tighten screws in a factory setting.

system to precisely locate and track target objects, such as fixtures in aircraft manufacturing. Visual servoing provided precise positional information of the robot's end-effector relative to the target object, while force control ensured the correct force application during contact. By integrating positional information from visual servoing with force information from force control into a unified control framework, precise control of robot operations was achieved.

3.2.3. Other visual feedback control

In addition to grasping rigid objects, some studies explore how to use visual feedback to control robots in performing challenging tasks such as pouring liquids and manipulating plastic bags. Schenck et al. [53] was the first to use raw visual feedback for closed-loop control in a robotic liquid pouring task. They used a deep learning network to detect pixels containing water in images and estimate the liquid volume in the container. The output of the deep learning network was directly used by a PID controller to achieve real-time feedback control for pouring a specific amount of liquid. The paper detailed their dataset acquisition method: using thermal imaging to generate pixel-level labels for water, training a deep neural network to detect liquid pixels in raw color images, and using these labels to estimate the liquid volume in the target container. Do et al. [54] proposed a method relying solely on depth information from an RGB-D camera to detect and track liquid levels, offering advantages in cost and availability. For opaque liquids, height is directly extracted from the point cloud; for transparent liquids, a mathematical model based on the viewing angle and liquid refractive index is used to correct depth measurements and estimate the actual liquid level.

Chen et al. [55] addressed the problem of robots opening plastic bags with handles. The visual system identifies the edges and handles of the plastic bag. By training a semantic segmentation model, the robot can distinguish different parts of the bag from RGB images, including handles and edges. The visual system provides an assessment of the bag's current state, including the size and shape of the bag's opening. By calculating the convex hull area and convex hull elongation of the opening, the robot can determine if the bag is sufficiently open to insert an object. If the bag moves or changes shape during operation, the visual system can capture these changes and trigger appropriate corrective actions. During the training phase, UV-marked bags emit specific colors under UV light, which are invisible under normal lighting. This setup allows the robot to alternately use UV and normal light to collect labeled images, enabling self-supervised learning.

3.3. Vision-based human-robot interaction

Human-robot interaction (HRI) is a rapidly evolving field that seeks to enhance the collaboration and communication between robots and humans. See Fig. 9, as robots become more integrated into human environments, the ability to accurately perceive and interpret human intentions becomes increasingly critical. Vision-based systems play a pivotal role in this process, enabling robots to recognize and respond to human actions, gestures, and expressions. This section explores the various approaches and technologies used to improve the effectiveness of HRI, focusing on intention perception and remote collaboration. By leveraging advanced visual processing and multimodal sensory integration, these systems aim to create more intuitive and responsive interactions between humans and robots, facilitating collaboration in complex and dynamic environments.

3.3.1. Intention perception

We not only hope for robots to be autonomous and capable of completing tasks independently without human intervention, but also to interact with humans and even collaborate with them to complete tasks. "How can robots estimate human intentions?" is a key question in human-robot interaction. Robots need to understand human communication and motion intentions to respond appropriately. Early multimodal human-robot interaction systems, such as [57], utilized vision for locating and tracking the human body, facial recognition, head pose estimation, gesture recognition, and multimodal fusion. [69] used the OpenFace framework to detect and analyze human facial expressions, using these expressions to assess user engagement and comfort. Chen et al. [56] explored how a humanoid robot in a home environment can detect the intention of a human partner initiating interaction. A multimodal state machine was designed, integrating speech recognition and face detection in a complementary manner to improve the accuracy and robustness of interaction detection. Bolotnikova et al. [5] employed Position-Based Visual Servoing (PBVS) tasks to adjust the robot's mobile base to approach humans, and Image-Based Visual Servoing (IBVS) tasks to adjust the camera direction to keep the human head in the center of the field of view. Visual detection of human head pose was used to determine whether the human had communication intentions. Lorentz et al. [58] extracted pointing gestures using a pre-trained pose estimation model and combined them with verbal dialogue as a demonstration of interactive behavior for moving objects with the Digit robot. A bilateral human-robot interaction approach was proposed, using human and robot pointing gestures for target definition and validation.

Based on learning from human-to-human interactions, Potdar et al. [59] proposed a novel human-robot interaction approach: using motion tracking cameras to simultaneously track the upper body gestures of two interacting humans, resulting in a dataset of joint data for their interactive body movements. After training, inverse kinematics were used to map the reactive body movements to a humanoid robot. Through training, the robot could visually perceive gestures such as passing objects, Namaste, boxing, and high-fiving, and respond accordingly. Yasar et al. [60] proposed VADER, a novel sequence learning algorithm that models past observed poses using a flexible discrete latent space.

3.3.2. Remote collaboration

With the help of visual sensor feedback on robots and Virtual Reality/Mixed Reality headsets, humans can remotely perceive the robot's situation, and wearable devices can capture the human operator's intentions and map them to the robot's motion control, enabling remote collaboration. Chen et al. [61] enhanced visual feedback through SLAM technology, supplementing areas

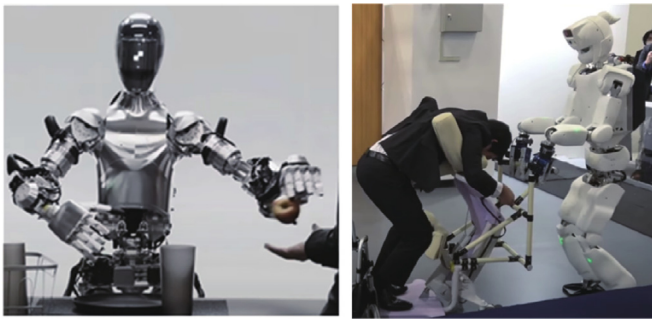


Fig. 9. Applications of vision-based human-robot interaction. Left: Visual perception enables Figure 01 robot to grasp an apple and hand it to a person. Right: Kawasaki Heavy Industries' Kaleido robot providing care to a patient.

not covered by visual feedback with real-time point clouds and pre-constructed meshes, addressing issues such as delays between operator and robot head movements caused by network communication latency or slow robot joint actions, mismatched Fields of View (FOV) between the camera and head-mounted display (HMD), and mismatched neck movement ranges between humans and robots. They further enhanced operator perception of the robot agent by adding multimodal feedback such as sound and haptics [62]. Song et al. [63] used Mixed Reality headsets to obtain augmented visual information about the robot's environment, providing an enhanced view from the robot's perspective, including information about the position and weight of objects in the robot's field of view. This was achieved by detecting ArUco markers in the robot's task space and displaying the weight information on the detected markers.

4. Challenges and perspectives

In light of the evolving technological landscape and the ambitious goals set by researchers and developers in the field of humanoid robotics, this chapter discusses three key research areas where vision systems play a crucial role in advancing the capabilities of humanoid robots.

(1) Multi-Sensor Fusion for State Estimation

State estimation is crucial for humanoid robots to navigate and interact in dynamic environments. However, relying solely on visual sensors can lead to challenges such as failure in low-feature or motion-blurred conditions [15]. Additionally, intense shaking during walking can cause visual odometry to fail or IMU data to drift.

Multi-sensor fusion offers a robust solution by integrating complementary sensor data, such as IMUs, LiDAR, and leg or kinematic odometry. This approach enhances the overall accuracy and reliability of state estimation, allowing the robot to maintain precise localization even in challenging conditions. Future research should focus on optimizing fusion strategies, exploring the use of multi-view cameras for better environmental coverage, and developing adaptive algorithms to dynamically adjust sensor weighting based on real-time conditions [70]. These advancements are key to fully leveraging multi-sensor fusion in humanoid robots.

(2) Visual Semantic Understanding

There will be a continued focus on enhancing the visual semantic understanding capabilities of robots. This includes improving the understanding of complex terrains (e.g., identifying traversable areas for path planning) and better comprehension of interactive objects (e.g., recognizing and categorizing items such



Fig. 10. Visual perception enables advanced semantic understanding. Left: Tesla's Optimus sorting batteries in a factory. Right: Appronik's Apollo inserting colored balls into a sealed plastic bag by opening the seal.

as food and tools placed on a table). Enhanced semantic understanding will enable robots to make more advanced decisions, such as planning tasks in a dynamic environment. For example, in the NimRo robot soccer competition, the robot's ability to detect the ball, boundary lines, goalposts, and other robots has evolved from traditional hand-crafted feature methods [71] to advanced deep learning approaches [72,73], continually improving the robot's visual perception capabilities. In recent years, there have been increasing examples worldwide of improving the visual semantic understanding of humanoid robots, as shown in Figs. 6, 8 and 10.

(3) Enhanced Human-Robot Interaction

Human-robot interaction will become a significant market demand in the future, particularly in areas such as therapy for children with autism and elderly care [74]. The ability to perceive human actions and expressions through visual sensors, combined with other modalities such as audio sensors and force sensors, will enable more sophisticated intention recognition. By leveraging advanced models like vision-language-action frameworks, similar to RT-2, robots can transfer web-based knowledge to real-world control scenarios, allowing them to understand human intentions in more complex, dynamic environments [75]. Integrating these sensory inputs with large-scale models, such as large language models like ChatGPT, for intention analysis will allow robots to respond to human needs more effectively. Additionally, predicting human motion intentions through visual input and force-torque sensors is also a crucial direction for human-robot collaboration [76]. This multimodal approach to understanding and responding to human intentions will be vital in developing robots that can engage in meaningful and safe interactions with humans, enhancing their ability to collaborate with people in complex tasks.

5. Conclusions

This review categorizes the applications of vision-based systems in humanoid robots into two main areas: state estimation and interaction. For state estimation, the study addresses two critical challenges—dynamic environments and motion blur—by classifying various methods used to mitigate these issues, and highlights the significance of multi-sensor fusion in enhancing accuracy and robustness. In terms of interaction, the review divides applications into three scenarios: terrain perception for navigation, upper body visual control, and human-robot interaction. Each application area is explored with a focus on the advantages and challenges of the methods employed. Finally, the review discusses three key research areas that are crucial for advancing the capabilities of humanoid robots: multi-sensor fusion,

enhanced visual semantic understanding, and more sophisticated intention recognition in human–robot interaction. These areas of research are expected to drive further innovations, enabling humanoid robots to perform more effectively in complex, real-world environments.

CRedit authorship contribution statement

Teng Bin: Writing – original draft, Visualization, Validation, Supervision, Software, Conceptualization. **Hanming Yan:** Data curation. **Ning Wang:** Formal analysis. **Milutin N. Nikolić:** Writing – review & editing. **Jianming Yao:** Data curation. **Tianwei Zhang:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62306185), the Guangdong Basic and Applied Basic Research Foundation, China (2024A1515012065), and the Shenzhen Science and Technology Program, China (JSGGKQTD20221101115656029 and KJZD20230923113801004).

References

- [1] S. Ahn, S. Yoon, S. Hyung, N. Kwak, K.S. Roh, On-board odometry estimation for 3D vision-based SLAM of humanoid robot, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4006–4012, <http://dx.doi.org/10.1109/IROS.2012.6385743>.
- [2] S. Piperakis, M. Koskinopoulou, P. Trahanias, Nonlinear state estimation for humanoid robot walking, IEEE Robot. Autom. Lett. 3 (4) (2018) 3347–3354, <http://dx.doi.org/10.1109/LRA.2018.2852788>.
- [3] W.-J. Baek, C. Pohl, P. Pelcz, T. Kröger, T. Asfour, Improving humanoid grasp success rate based on uncertainty-aware metrics and sensitivity optimization, in: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), 2022, pp. 786–793, <http://dx.doi.org/10.1109/Humanoids53995.2022.10000206>.
- [4] D.J. Agravante, G. Claudio, F. Spindler, F. Chaumette, Visual servoing in an optimization framework for the whole-body control of humanoid robots, IEEE Robot. Autom. Lett. 2 (2) (2017) 608–615, <http://dx.doi.org/10.1109/LRA.2016.2645512>.
- [5] A. Bolotnikova, S. Courtois, A. Kheddar, Autonomous initiation of human physical assistance by a humanoid, in: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2020, pp. 857–862, <http://dx.doi.org/10.1109/RO-MAN47096.2020.9223519>.
- [6] R. Mur-Artal, J.M.M. Montiel, J.D. Tardós, ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163, <http://dx.doi.org/10.1109/TRO.2015.2463671>.
- [7] R. Mur-Artal, J.D. Tardós, ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262, <http://dx.doi.org/10.1109/TRO.2017.2705103>.
- [8] C. Campos, R. Elvira, J.J.G. Rodríguez, J.M. M. Montiel, J. D. Tardós, ORB-SLAM3: An accurate open-source library for visual, visual–Inertial, and multi-map SLAM, IEEE Trans. Robot. 37 (6) (2021) 1874–1890, <http://dx.doi.org/10.1109/TRO.2021.3075644>.
- [9] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces, in: 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007, pp. 225–234, <http://dx.doi.org/10.1109/ISMAR.2007.4538852>.
- [10] S. Sumikura, M. Shibuya, K. Sakurada, OpenVSLAM: A versatile visual SLAM framework, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450368896, 2019, pp. 2292–2295, <http://dx.doi.org/10.1145/3343031.3350539>, URL <http://dx.doi.org/10.1145/3343031.3350539>.
- [11] T. Whelan, S. Leutenegger, R.F. Salas-Moreno, B. Glocker, A.J. Davison, ElasticFusion: Dense SLAM without a pose graph, in: Robotics: Sci. Syst. 11, Rome, Italy, 2015, p. 3.
- [12] C. Forster, M. Pizzoli, D. Scaramuzza, SVO: Fast semi-direct monocular visual odometry, in: 2014 IEEE International Conference on Robotics and Automation, ICRA, 2014, pp. 15–22, <http://dx.doi.org/10.1109/ICRA.2014.6906584>.
- [13] J. Engel, T. Schöps, D. Cremers, LSD-SLAM: Large-scale direct monocular SLAM, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, ISBN: 978-3-319-10605-2, 2014, pp. 834–849.
- [14] O. Stasse, A.J. Davison, R. Sellaoui, K. Yokoi, Real-time 3D SLAM for humanoid robot considering pattern generator information, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 348–355, <http://dx.doi.org/10.1109/IROS.2006.281645>.
- [15] R. Scona, S. Nobili, Y.R. Petillot, M. Fallon, Direct visual SLAM fusing proprioception for a humanoid robot, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2017, pp. 1419–1426, <http://dx.doi.org/10.1109/IROS.2017.8205943>.
- [16] R. Sheikh, S. OBwald, M. Bennewitz, A combined RGB and depth descriptor for SLAM with humanoids, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2018, pp. 1718–1724, <http://dx.doi.org/10.1109/IROS.2018.8593768>.
- [17] A. Vedadi, A. Yousefi-Koma, P. Yazdankhah, A. Mozayyan, Comparative evaluation of RGB-D SLAM methods for humanoid robot localization and mapping, in: 2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM), 2023, pp. 807–812, <http://dx.doi.org/10.1109/ICRoM60803.2023.10412425>.
- [18] M. Labbé, F. Michaud, RTAB-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation, J. Field Robotics 36 (2) (2019) 416–446, <http://dx.doi.org/10.1002/rob.21831>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21831>.
- [19] K. Chappellet, M. Murooka, G. Caron, F. Kanehiro, A. Kheddar, Humanoid loco-manipulations using combined fast dense 3D tracking and SLAM with wide-angle depth-images, IEEE Trans. Autom. Sci. Eng. 21 (3) (2024) 3691–3704, <http://dx.doi.org/10.1109/TASE.2023.3283497>.
- [20] M. Tsuru, A. Escande, A. Tanguy, K. Chappellet, K. Harad, Online object searching by a humanoid robot in an unknown environment, IEEE Robot. Autom. Lett. 6 (2) (2021) 2862–2869, <http://dx.doi.org/10.1109/LRA.2021.3061383>.
- [21] T. Zhang, Y. Nakamura, Humanoid robot rgb-d slam in the dynamic human environment, Int. J. Humanoid Robot 17 (02) (2020) 2050009.
- [22] T. Zhang, E. Uchiyama, Y. Nakamura, Dense RGB-D SLAM for humanoid robots in the dynamic humans environment, in: 2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids), 2018, pp. 270–276, <http://dx.doi.org/10.1109/HUMANOIDS.2018.8625019>.
- [23] T. Zhang, H. Zhang, Y. Li, Y. Nakamura, L. Zhang, FlowFusion: Dynamic dense RGB-D SLAM based on optical flow, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, 2020, pp. 7322–7328, <http://dx.doi.org/10.1109/ICRA40945.2020.9197349>.
- [24] C. Zhang, R. Zhang, S. Jin, X. Yi, PFD-SLAM: A new RGB-D SLAM for dynamic indoor environments based on non-prior semantic segmentation, Remote Sens. (ISSN: 2072-4292) 14 (10) (2022) <http://dx.doi.org/10.3390/rs14102445>, URL <https://www.mdpi.com/2072-4292/14/10/2445>.
- [25] D. Wahrmann, A.-C. Hildebrandt, T. Bates, R. Wittmann, F. Sygulla, P. Seiwald, D. Rixen, Vision-based 3d modeling of unknown dynamic environments for real-time humanoid navigation, Int. J. Humanoid Robot 16 (01) (2019) 1950002.
- [26] T. Zhang, H. Zhang, X. Li, Vision-audio fusion SLAM in dynamic environments, CAAI Trans. Intell. Technol 8 (4) (2023) 1364–1373.
- [27] R. Long, C. Rauch, T. Zhang, V. Ivan, S. Vijayakumar, RigidFusion: Robot localisation and mapping in environments with large dynamic rigid objects, IEEE Robot. Autom. Lett. 6 (2) (2021) 3703–3710, <http://dx.doi.org/10.1109/LRA.2021.3066375>.
- [28] R. Long, C. Rauch, T. Zhang, V. Ivan, T.L. Lam, S. Vijayakumar, RGB-D SLAM in indoor planar environments with multiple large dynamic objects, IEEE Robot. Autom. Lett. 7 (3) (2022) 8209–8216, <http://dx.doi.org/10.1109/LRA.2022.3186091>.
- [29] S. Song, H. Lim, A.J. Lee, H. Myung, DynaVINS: A visual-inertial SLAM for dynamic environments, IEEE Robot. Autom. Lett. 7 (4) (2022) 11523–11530, <http://dx.doi.org/10.1109/LRA.2022.3203231>.
- [30] M. Mutlu, A. Sارانلی, U. Sارانلی, A real-time inertial motion blur metric: Application to frame triggering based motion blur minimization, in: 2014 IEEE International Conference on Robotics and Automation, ICRA, 2014, pp. 671–676, <http://dx.doi.org/10.1109/ICRA.2014.6906926>.
- [31] G. Fan, J. Huang, D. Yang, L. Rao, Sampling visual SLAM with a wide-angle camera for legged mobile robots, IET Cyber-Syst. Robot 4 (4) (2022) 356–375.
- [32] G.K. Gultekin, A. Sارانلی, Multi-frame motion deblurring of video using the natural oscillatory motion of dexterous legged robots, IET Image Process. 13 (9) (2019) 1502–1508.

- [33] P. Liu, X. Zuo, V. Larsson, M. Pollefeys, MBA-VO: Motion blur aware visual odometry, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 5530–5539, <http://dx.doi.org/10.1109/ICCV48922.2021.00050>.
- [34] Z.-J. Du, S.-S. Huang, T.-J. Mu, Q. Zhao, R.R. Martin, K. Xu, Accurate dynamic SLAM using CRF-based long-term consistency, *IEEE Trans. Vis. Comput. Graphics* 28 (4) (2022) 1745–1757, <http://dx.doi.org/10.1109/TVCG.2020.3028218>.
- [35] Y. Wang, K. Xu, Y. Tian, X. Ding, DRG-SLAM: A semantic RGB-D SLAM using geometric features for indoor dynamic scene, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2022, pp. 1352–1359, <http://dx.doi.org/10.1109/IROS47612.2022.9981238>.
- [36] E. Hourdakis, S. Piperakis, P. Trahanias, Roboslam: Dense RGB-D SLAM for humanoid robots, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021, pp. 2224–2231, <http://dx.doi.org/10.1109/IROS51168.2021.9636044>.
- [37] X. Tao, H. Gao, X. Shen, J. Wang, J. Jia, Scale-recurrent network for deep image deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8174–8182.
- [38] O. Kupyn, T. Martyniuk, J. Wu, Z. Wang, Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8878–8887.
- [39] G. Oriolo, A. Paolillo, L. Rosa, M. Vendittelli, Humanoid odometric localization integrating kinematic, inertial and visual information, *Auton. Robots* 40 (2016) 867–879.
- [40] X. Leng, S. Piao, S. Wang, L. Chang, Z. Zhu, An improved method for odometry estimation based on EKF and temporal convolutional network, *Phys. Commun.* (ISSN: 1874-4907) 43 (2020) 101178, <http://dx.doi.org/10.1016/j.phycom.2020.101178>, URL <https://www.sciencedirect.com/science/article/pii/S187449072030255X>.
- [41] V. Dhédin, H. Li, S. Khorshidi, L. Mack, A.K.C. Ravi, A. Meduri, P. Shah, F. Grimmering, L. Righetti, M. Khadiv, J. Stueckler, Visual-inertial and leg odometry fusion for dynamic locomotion, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, 2023, pp. 9966–9972, <http://dx.doi.org/10.1109/ICRA48891.2023.10160898>.
- [42] J. Kang, H. Kim, K.-S. Kim, VIEW: Visual-inertial external wrench estimator for legged robot, *IEEE Robot. Autom. Lett.* 8 (12) (2023) 8366–8373, <http://dx.doi.org/10.1109/LRA.2023.3322646>.
- [43] C. Houseago, M. Bloesch, S. Leutenegger, KO-fusion: Dense visual SLAM with tightly-coupled kinematic and odometric tracking, in: 2019 International Conference on Robotics and Automation, ICRA, 2019, pp. 4054–4060, <http://dx.doi.org/10.1109/ICRA.2019.8793471>.
- [44] A. Roychoudhury, M. Missura, M. Bennewitz, 3D polygonal mapping for humanoid robot navigation, in: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), 2022, pp. 171–177, <http://dx.doi.org/10.1109/Humanoids53995.2022.10000101>.
- [45] S. Bertrand, I. Lee, B. Mishra, D. Calvert, J. Pratt, R. Griffin, Detecting Usable Planar Regions for legged robot locomotion, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2020, pp. 4736–4742, <http://dx.doi.org/10.1109/IROS45743.2020.9341000>.
- [46] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, M. Hutter, Elevation mapping for locomotion and navigation using GPU, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2022, pp. 2273–2280, <http://dx.doi.org/10.1109/IROS47612.2022.9981507>.
- [47] G. Erni, J. Frey, T. Miki, M. Mattamala, M. Hutter, MEM: Multi-modal elevation mapping for robotics and learning, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2023, pp. 11011–11018, <http://dx.doi.org/10.1109/IROS55552.2023.10342108>.
- [48] C. Pohl, K. Hitzler, R. Grimm, A. Zea, U.D. Hanebeck, T. Asfour, Affordance-based grasping and manipulation in real world applications, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2020, pp. 9569–9576, <http://dx.doi.org/10.1109/IROS45743.2020.9341482>.
- [49] G. Claudio, F. Spindler, F. Chaumette, Vision-based manipulation with the humanoid robot romeo, in: 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids), 2016, pp. 286–293, <http://dx.doi.org/10.1109/HUMANOIDS.2016.7803290>.
- [50] A. Paolillo, K. Chappellet, A. Bolotnikova, A. Kheddar, Interlinked visual tracking and robotic manipulation of articulated objects, *IEEE Robot. Autom. Lett.* 3 (4) (2018) 2746–2753, <http://dx.doi.org/10.1109/LRA.2018.2835515>.
- [51] E.M. Hoffman, A. Paolillo, Exploiting visual servoing and centroidal momentum for whole-body motion control of humanoid robots in absence of contacts and gravity, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, 2021, pp. 2979–2985, <http://dx.doi.org/10.1109/ICRA48506.2021.9560739>.
- [52] A. Kheddar, S. Caron, P. Gergondet, A. Comport, A. Tanguy, C. Ott, B. Henze, G. Mesesan, J. Engelsberger, M.A. Roa, et al., Humanoid robots in aircraft manufacturing: The airbus use cases, *IEEE Robot. Autom. Mag.* 26 (4) (2019) 30–45.
- [53] C. Schenck, D. Fox, Visual closed-loop control for pouring liquids, in: 2017 IEEE International Conference on Robotics and Automation, ICRA, 2017, pp. 2629–2636, <http://dx.doi.org/10.1109/ICRA.2017.7989307>.
- [54] C. Do, W. Burgard, Accurate pouring with an autonomous robot using an rgb-d camera, in: Intelligent Autonomous Systems 15: Proceedings of the 15th International Conference IAS-15, Springer, 2019, pp. 210–221.
- [55] L.Y. Chen, B. Shi, D. Seita, R. Cheng, T. Kollar, D. Held, K. Goldberg, AutoBag: Learning to open plastic bags and insert objects, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, 2023, pp. 3918–3925, <http://dx.doi.org/10.1109/ICRA48891.2023.10161402>.
- [56] J. Chen, W.J. Fitzgerald, Continuous multi-modal human interest detection for a domestic companion humanoid robot, in: 2013 16th International Conference on Advanced Robotics, ICAR, 2013, pp. 1–6, <http://dx.doi.org/10.1109/ICAR.2013.6766469>.
- [57] R. Stiefelhagen, H.K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, A. Waibel, Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot, *IEEE Trans. Robot.* 23 (5) (2007) 840–851, <http://dx.doi.org/10.1109/TRO.2007.907484>.
- [58] V. Lorentz, M. Weiss, K. Hildebrand, I. Boblan, Pointing gestures for human-robot interaction with the humanoid robot digit, in: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2023, pp. 1886–1892, <http://dx.doi.org/10.1109/RO-MAN57019.2023.10309407>.
- [59] S. Potdar, A. Sawarkar, F. Kazi, Learning by demonstration from multiple agents in humanoid robots, in: 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEES, 2016, pp. 1–6, <http://dx.doi.org/10.1109/SCEES.2016.7509324>.
- [60] M.S. Yasar, T. Iqbal, VADER: Vector-quantized generative adversarial network for motion prediction, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2023, pp. 3827–3834, <http://dx.doi.org/10.1109/IROS55552.2023.10342324>.
- [61] Y. Chen, L. Sun, M. Benallegue, R. Cisneros-Limón, R.P. Singh, K. Kaneko, A. Tanguy, G. Caron, K. Suzuki, A. Kheddar, F. Kanehiro, Enhanced visual feedback with decoupled viewpoint control in immersive humanoid robot teleoperation using SLAM, in: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), 2022, pp. 306–313, <http://dx.doi.org/10.1109/Humanoids53995.2022.9999740>.
- [62] R. Cisneros-Limón, A. Dallard, M. Benallegue, K. Kaneko, H. Kaminaga, P. Gergondet, A. Tanguy, R.P. Singh, L. Sun, Y. Chen, et al., A cybernetic avatar system to embody human telepresence for connectivity, exploration, and skill transfer, *Int. J. Social Robot* (2024) 1–28.
- [63] H. Song, G. Bronfman, Y. Zhang, Q. Sun, J.H. Kim, Mixed reality interface for whole-body balancing and manipulation of humanoid robot, in: 2024 21st International Conference on Ubiquitous Robots, UR, 2024, pp. 642–647, <http://dx.doi.org/10.1109/UR61395.2024.10597520>.
- [64] B. Mishra, D. Calvert, S. Bertrand, J. Pratt, H.E. Sevil, R. Griffin, Efficient terrain map Using Planar Regions for footstep planning on humanoid robots, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, 2024, pp. 8044–8050, <http://dx.doi.org/10.1109/ICRA57147.2024.10610879>.
- [65] T. Bin, J. Yao, T.L. Lam, T. Zhang, Real-time polygonal semantic mapping for humanoid robot stair climbing, 2024, URL <https://arxiv.org/abs/2411.01919>.
- [66] C. Feng, Y. Taguchi, V.R. Kamat, Fast plane extraction in organized point clouds using agglomerative hierarchical clustering, in: 2014 IEEE International Conference on Robotics and Automation, ICRA, 2014, pp. 6218–6225, <http://dx.doi.org/10.1109/ICRA.2014.6907776>.
- [67] A. Roychoudhury, M. Missura, M. Bennewitz, Plane segmentation using depth-dependent flood fill, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021, pp. 2210–2216, <http://dx.doi.org/10.1109/IROS51168.2021.9635930>.
- [68] P. Fankhauser, M. Bloesch, M. Hutter, Probabilistic terrain mapping for mobile robots with uncertain localization, *IEEE Robot. Autom. Lett.* 3 (4) (2018) 3019–3026, <http://dx.doi.org/10.1109/LRA.2018.2849506>.
- [69] A. Tanevska, Towards a Cognitive Architecture for Socially Adaptive Human-Robot Interaction. (Ph.D. thesis), University of Genoa, Italy, 2020.
- [70] J. Wang, M. Mattamala, C. Kassab, L. Zhang, M. Fallon, Exosense: A vision-centric scene understanding system for safe exoskeleton navigation, 2024, URL <https://arxiv.org/abs/2403.14320>.
- [71] H. Farazi, P. Allgeuer, G. Ficht, A. Brandenburger, D. Pavlichenko, M. Schreiber, S. Behnke, RoboCup 2016 humanoid TeenSize winner nimbro: Robust visual perception and soccer behaviors, in: RoboCup 2016: Robot World Cup XX 20, Springer, 2017, pp. 478–490.
- [72] D. Pavlichenko, G. Ficht, A. Amini, M. Hosseini, R. Memmesheimer, A. Villar-Corrales, S.M. Schulz, M. Missura, M. Bennewitz, S. Behnke, RoboCup 2022 AdultSize winner nimbro: Upgraded perception, capture

- steps gait and phase-based in-walk kicks, in: A. Eguchi, N. Lau, M. Paetzel-Prüsmann, T. Wanichanon (Eds.), *RoboCup 2022: Robot World Cup XXV*, Springer International Publishing, Cham, ISBN: 978-3-031-28469-4, 2023, pp. 240–252.
- [73] D. Rodriguez, H. Farazi, G. Ficht, D. Pavlichenko, A. Brandenburger, M. Hosseini, O. Kosenko, M. Schreiber, M. Missura, S. Behnke, *RoboCup 2019 AdultSize winner nimbro: Deep learning perception, in-walk kick, push recovery, and team play capabilities*, in: S. Chalup, T. Niemueller, J. Suthakorn, M.-A. Williams (Eds.), *RoboCup 2019: Robot World Cup XXIII*, Springer International Publishing, Cham, ISBN: 978-3-030-35699-6, 2019, pp. 631–645.
- [74] O. Avioz-Sarig, S. Olatunji, V. Sarne-Fleischmann, Y. Edan, *Robotic system for physical training of older adults*, *Int. J. Social Robot* 13 (5) (2021) 1109–1124.
- [75] B. Zitkovich, T. Yu, S. Xu, et al., *RT-2: Vision-language-action models transfer web knowledge to robotic control*, in: J. Tan, M. Toussaint, K. Darvish (Eds.), *Proceedings of the 7th Conference on Robot Learning*, in: *Proceedings of Machine Learning Research*, 229, PMLR, 2023, pp. 2165–2183, URL <https://proceedings.mlr.press/v229/zitkovich23a.html>.
- [76] X. Yu, W. He, Q. Li, Y. Li, B. Li, *Human-robot co-carrying using visual and force sensing*, *IEEE Trans. Ind. Electron.* 68 (9) (2021) 8657–8666, <http://dx.doi.org/10.1109/TIE.2020.3016271>.