Perception for Humanoid Robots

Arindam Roychoudhury[1] · Shahram Khorshidi[1] · Subham Agrawal[1] · Maren Bennewitz[1,2]

Abstract

Purpose of Review The field of humanoid robotics, perception plays a fundamental role in enabling robots to interact seamlessly with humans and their surroundings, leading to improved safety, efficiency, and user experience. This scientific study investigates various perception modalities and techniques employed in humanoid robots, including visual, auditory, and tactile sensing by exploring recent state-of-the-art approaches for perceiving and understanding the internal state, the environment, objects, and human activities.

Recent Findings Internal state estimation makes extensive use of Bayesian filtering methods and optimization techniques based on maximum a-posteriori formulation by utilizing proprioceptive sensing. In the area of external environment understanding, with an emphasis on robustness and adaptability to dynamic, unforeseen environmental changes, the new slew of research discussed in this study have focused largely on multi-sensor fusion and machine learning in contrast to the use of hand-crafted, rule-based systems. Human robot interaction methods have established the importance of contextual information representation and memory for understanding human intentions.

Summary This review summarizes the recent developments and trends in the field of perception in humanoid robots. Three main areas of application are identified, namely, internal state estimation, external environment estimation, and human robot interaction. The applications of diverse sensor modalities in each of these areas are considered and recent significant works are discussed.

Keywords Humanoid robots · Perception · Survey · Navigation · State estimation · Human robot interaction

Introduction

Perception is of paramount importance for robots to establish a model of their internal state as well as the external environment. These models allow the robot to perform its task safely, efficiently and accurately. Perception is facilitated by various types of sensors which gather both proprioceptive and exteroceptive information. Humanoid robots,

✉ Arindam Roychoudhury
    roychoud@cs.uni-bonn.de

    Shahram Khorshidi
    khorshidi@cs.uni-bonn.de

    Subham Agrawal
    sagrawal@cs.uni-bonn.de

    Maren Bennewitz
    maren@cs.uni-bonn.de

1   Humanoid Robots Lab, University of Bonn, Bonn, Germany

2   Lamarr Institute for Machine Learning and Artificial
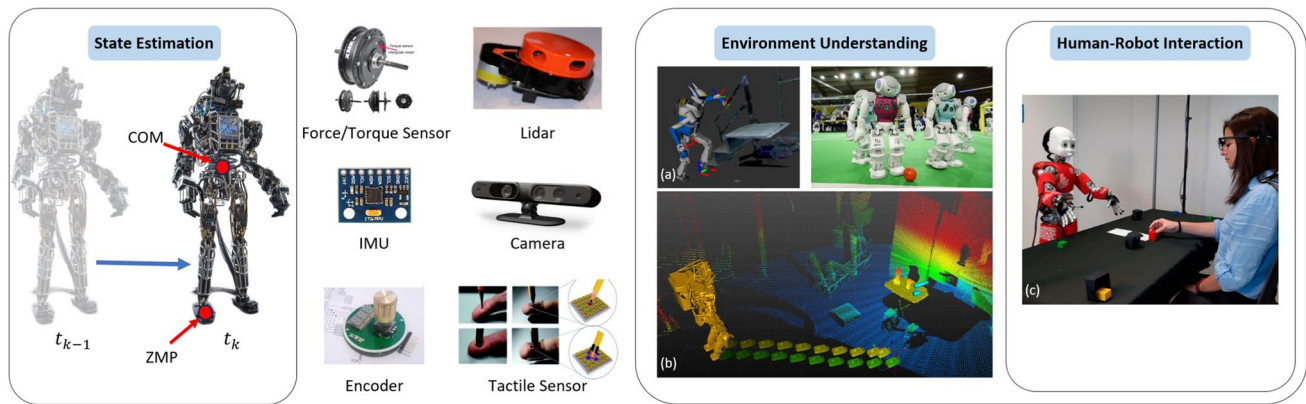   Intelligence, Dortmund, Germany

especially those which are mobile, pose a difficult challenge for the perception process: mounted sensors are susceptible to jerky and unstable motions due to the very high degrees of freedom afforded by the high number of articulable joints present on a humanoid's body, e.g., the legs, the hip, the manipulator arms or the neck.

We organize the main areas of perception in humanoid robots into three broad yet overlapping areas for the purposes of this survey, namely, state estimation for balance and joint configurations, environment understanding for navigation, mapping and manipulation, and finally human-robot interaction for successful integration into a shared human workspace, see Fig. 1. For each area we discuss the popular application areas, the challenges and recent methodologies used to surmount them.

Internal state estimation is a critical aspect of autonomous systems, particularly for humanoid robots in order to address both low level stability and dynamics, and as an auxiliary to higher level tasks such as localization, mapping and navigation. Legged robots locomotion is particularly challenging given their inherent under-actuation dynamics

Fig. 1 Perception for humanoid robots split into three principal areas. Left: State estimation being used to estimate derived quantities like CoM and ZMP from sensors like IMU and joint encoders. Right: Environment understanding has a very broad scope which varies from localization and mapping to environment segmentation for planning and even more application areas. Human Robot Interaction is closely related but deals exclusively with human beings rather than inanimate objects. Center: Some sensors which aid in perception for humanoid robots. Sources for labeled images- (a):[1], (b): [2] and (c): [3]

and the intermittent contact switching with the ground during motion.

The application of external environment understanding has a very broad scope in humanoid robotics but can be roughly divided into navigation and manipulation. Navigation implies the movement of the mobile bipedal base from one location to another without collision thereby leaving the external environment configuration unchanged. On the other hand, manipulation is where the humanoid changes the physical configuration of its environment using its end-effectors.

It could be argued that human robot interaction or HRI is a subset of environment understanding. However, we have separated the two areas based on their ultimate goals. The goal of environment understanding is to interact with inanimate objects while the goal of HRI is to interact with humans. The set of posed challenges are different though similar principles may be reused. Human detection, gesture and activity recognition, teleoperation, object handover and collaborative actions, and social communications are some of the main areas where perception is used.

State Estimation

Recent works on humanoid and legged robots locomotion control have focused extensively on state-feedback approaches [4]. Legged robots have highly nonlinear dynamics, and they need high frequency ($1\,kHz$) and low latency ($< 1 : ms$) feedback in order to have robust and adaptive control systems, thereby adding more complexity to the design and development of reliable estimators for the base and centroidal states, and contact detection.

Challenges in State Estimation

Perceived data is often noisy and biased and it gets magnified in derived quantities. For instance, joint velocities tend to be noisier than joint positions, as these are obtained by numerically differentiating joint encoder values. Rotella et al. [5] developed a method to determine joint velocities and acceleration of a humanoid robot using link-mounted Inertial Measurement Units (IMUs), resulting in less noise and delay compared to filtered velocities from numerical differentiation. An effective approach to mitigate biased IMU measurements is to explicitly introduce these biases as estimated states in the estimation framework [6, 7].

The high dimensionality of humanoids make it computationally expensive to formulate a single filter for the entire state. As an alternative, Xinjilefu et al. [8] proposed decoupling the full state into several independent state vectors, and used separate filters to estimate the pelvis state and joint dynamics.

To account for kinematic modeling errors such as joint backlash and link flexibility, Xinjilefu et al. [9] introduced a method using a Linear Inverted Pendulum Model (LIPM) with an offset which represented the modeling error in the Center of Mass (CoM) position and/or external forces. Bae et al. [10] proposed a CoM kinematics estimator by including a spring and damper in the LIPM to compensate for modeling errors. To address the issue of link flexibility in the humanoid exoskeleton *Atalante*, Vigne et al. [11] decomposed the full state estimation problem into several independent attitude estimation problems, each corresponding to a given flexibility and a specific IMU relying only on dependable and easily accessible geometric parameters of the system, rather than the dynamic model.

In the remainder of this section, we classify the recent related works on state estimation into three main categories [12]: proprioceptive state estimation, which primarily involves filtering methods that fuse high-frequency proprioceptive sensor data; multi-sensor fusion filtering, which integrates exteroceptive sensor modalities into the filtering process; multi-sensor fusion with state smoothing, which employs advanced techniques that leverage the entire history of sensor measurements to refine estimated states.

Finally, we present a list of available open-source software for state estimation from reviewed literature in Table 1.

Proprioceptive State Estimation

Proprioceptive sensors provide measurements of the robot's internal state. They are commonly used to compute leg odometry, which captures the drifting pose. For a comprehensive review of the evolution of proprioceptive filters on leg odometry, refer to [22], and [23].

Base State Estimation

In humanoid robots, the focus is on estimating the position, velocity, and orientation of the "base" frame, typically located at the pelvis. Recent state estimation approaches in this field often fuse IMU and leg odometry.

The work by Bloesch [6] was a decisive step in introducing a base state estimator for legged robots using a quaternion-based Extended Kalman Filter (EKF) approach. This method made no assumptions about the robot's gait and number of legs or the terrain structure and included absolute positions of the feet contact points, and IMU bias terms in the estimated states. Rotella et al. [7] extended it to humanoid platforms by considering the full foot plate and adding foot orientation to the state vector. Both works showed that as long as at least one foot remains in contact with the ground, the base absolute velocity, roll and pitch angles, and IMU biases are observable. There are also other formulations for the base state estimation using only proprioceptive sensing in [16, 24], and [25].

Centroidal State Estimation

Centroidal states in humanoid robots include the CoM position, linear and angular momentum, and their derivatives. The CoM serves as a vital control variable for stability and robust humanoid locomotion, making accurate estimation of centroidal states crucial in control system design for humanoid robots.

When the full 6-axis contact wrench is not directly available to the estimator, e.g., the robot gauge sensors measure only the contact normal force, some works have utilized simplified models of dynamics, such as the LIPM [26].

Piperakis et al. [27] presented an EKF to estimate centroidal variables by fusing joint encoders, IMU, foot sensitive resistors, and later including visual odometry in [13]. They formulated the estimator based on the non-linear Zero Moment Point (ZMP) dynamics, which captured the coupling between dynamics behavior in the frontal and lateral planes. Their results showed better performance over Kalman filter formulation based on the LIPM.

Mori et al. [28] proposed a centroidal state estimation framework for a humanoid robot based on real-time inertial parameter identification, using only the robot's proprioceptive sensors (IMU, foot Force/Torque (F/T) sensors, and joint encoders), and the sequential least squares method. They conducted successful experiments deliberately altering the robot's mass properties to demonstrate the robustness of their framework against dynamic inertia changes.

By having 6-axis F/T sensors on the feet, Rotella et al. [29] utilized momentum dynamics of the robot to estimate the centroidal quantities. Their nonlinear observability analysis demonstrated the observability of either biases or external wrench. In a different approach, Carpentier et al. [30] proposed a frequency analysis of the information sources utilized in estimating the CoM position, and later for CoM acceleration and the derivative of angular momentum [31]. They introduced a complementary filtering technique that fuses various measurements, including ZMP position, sensed contact forces, and geometry-based reconstruction of the *CoM* by using joint encoders, according to their reliability in the respective spectral bandwidth.

Table 1 Open-source software for humanoid robot state estimation. All cited software are available as ROS packages

Paper	Software	Language	Description
			Extended Kalman Filtering
[13]	SEROW[14]	C++	Multi-sensor state estimation (IMU, joint encoders, visual odometry)
[12]	PRONTO[15]	C++	Multi-sensor state estimation (IMU, joint encoders, LiDAR, camera)
[16]	InEKF[17]	C++	Invariant EKF (using IMU motion model, with different measurement models)
			Factor Graph
[18]	WOLF[19]	C++	Multi-sensor state smoothing (IMU, joint encoders, LiDAR, camera)
			Learning
[20]	GEM[21]	Python	Unsupervised gait-phase estimation (IMU, joint encoders, F/T sensors)

Contact Detection and Estimation

Feet contact detection plays a crucial role in locomotion control, gait planning, and proprioceptive state estimation in humanoid robots. Recent approaches can be categorized into two main groups: those directly utilizing measured ground reaction wrenches, and methods integrating kinematics and dynamics to infer the contact status by estimating the ground reaction forces. Fallon et al. [2] employed a Schmitt trigger with a 3-axis foot F/T sensor to classify contact forces and used a simple state machine to determine the most reliable foot for kinematic measurements. Piperakis et al. [13] adapted a similar approach by utilizing pressure sensors on the foot.

Rotella et al. [32] presented an unsupervised method for estimating contact states by using fuzzy clustering on only proprioceptive sensor data (foot F/T and IMU sensing), surpassing traditional approaches based on measured normal force. By including the joint encoders in proprioceptive sensing, Piperakis et al. [20] proposed an unsupervised learning framework for gait phase estimation, achieving effectiveness on uneven/rough terrain walking gaits. They also developed a deep learning framework by utilizing F/T and IMU sensing in each leg, to determine the contact state probabilities [33]. The generalizability and accuracy of their approach was demonstrated on different robotic platforms. Furthermore, Maravgakis et al. [34] introduced a probabilistic contact detection model, using only IMU sensors mounted on the end effector. Their approach estimated the contact state of the feet without requiring training data or ground truth labels.

Another active research field in humanoid robots is monitoring and identifying contact points on the robot's body. Common approaches focus on proprioceptive sensing for contact localization and identification. Flacco et al. [35] proposed using an internal residual of external momentum to isolate and identify singular contacts, along with detecting additional contacts with known locations. Manuelli et al. [36] introduced a contact particle filter for detecting and localizing external contacts, by only using proprioceptive sensing, such as 6-axis F/T sensors, capable of handling up to 3 contacts efficiently. Vorndamme et al. [37] developed a real-time method for multi-contact detection using 6-axis F/T sensors distributed along the kinematic chain, capable of handling up to 5 contacts. Vezzani et al. [38] proposed a memory unscented particle filter algorithm for real-time 6 Degrees of freedom (DoF) tactile localization using contact point measurements made by tactile sensors.

Multi-Sensor Fusion Filtering

One drawback of base state estimation using proprioceptive sensing is the accumulation of drift over the time, due to sensor noise. This drift is not acceptable for controlling highly dynamic motions, therefore it is typically compensated by integrating other sensor modalities from exteroceptive sensors, such as cameras, depth cameras, and LiDAR.

Fallon et al. [2] proposed a drift-free base pose estimation method by incorporating LiDAR sensing into a high-rate EKF estimator using a Gaussian particle filter for laser scan localization. Although their framework eliminated the drift, a pre-generated map was required as input. Piperakis et al. [39] introduced a robust Gaussian EKF to handle outlier detection in visual/LiDAR measurements for humanoid walking in dynamic environments. To address state estimation challenges in real-world scenarios, Camurri et al. [12] presented Pronto, a modular open-source state estimation framework for legged robots Fig. 2. It combined proprioceptive and exteroceptive sensing, such as stereo vision and LiDAR, using a loosely-coupled EKF approach.

Multi-Sensor Fusion with State Smoothing

So far, we have explored filtering methods based on Bayesian filtering for sensor fusion and state estimation. However, as the number of states and measurements increases, computational complexity becomes a limitation. Recent advancements in computing power and nonlinear solvers have popularized non-linear iterative maximum a-posteriori (MAP) optimization techniques, such as factor graph optimization.

To address the issue of visual tracking loss in visual factor graphs, Hartley et al. [40] introduced a factor graph framework that integrated forward kinematic and pre-integrated contact factors. The work was extended by incorporating the influence of contact switches and associated uncertainties [41]. Both works showed that the fusion of contact information with IMU and vision data provides a reliable odometry system for legged robots.

Solá et al [18] presented an open-source modular estimation framework for mobile robots based on factor graphs. Their approach offered systematic methods to handle the complexities arising from multi-sensory systems with asynchronous and different-frequency data sources. This framework was evaluated on state estimation for legged robots and landmark-based visual-inertial SLAM for humanoids by Fourmy et al. [26].

Environment Understanding

Environment understanding is a critical area of research for humanoid robots, enabling them to effectively navigate through and interact with complex and dynamic environments. This field can be broadly classified into two key categories: 1. localization, navigation and planning for the mobile base, and 2. object manipulation and grasping.
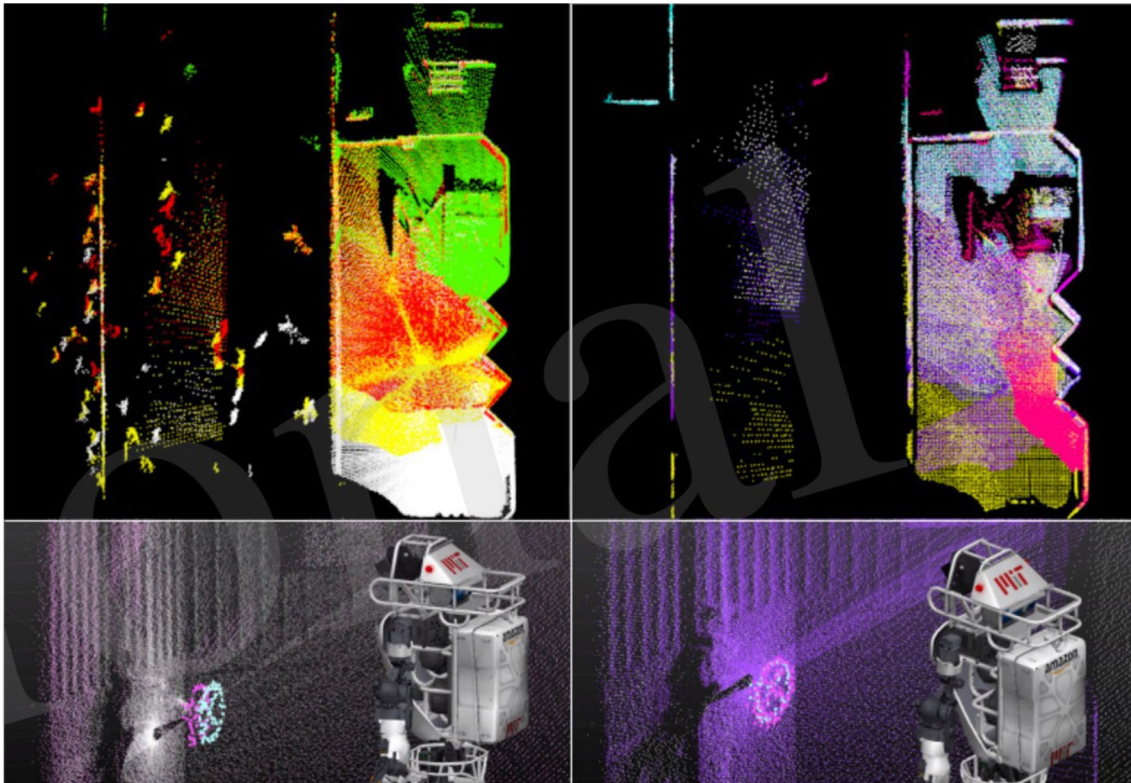
Fig. 2 State estimation with multi-sensor filtering, integrating LiDAR for drift correction and localization. Top row, filtering people from raw point cloud. Bottom row, state estimation and localization with iterative closest point correction on filtered point cloud. From [12]

Perception in Localization, Navigation and Planning

Localization focuses on precisely and continuously estimating the robot's position and orientation relative to its environment. Planning and navigation involve generating optimal paths and trajectories for the robot to reach its desired destination while avoiding obstacles and considering task-specific constraints.

Localization, Mapping and SLAM

Localization and SLAM (simultaneous localization and mapping) relies primarily on visual sensors such as cameras and lasers but often additionally use encoders and IMUs to enhance estimation accuracy.

Localization Indoor environments are usually considered structured, characterized by the presence of well-defined, repeatable and often geometrically consistent objects. Landmarks can be uniquely identified by encoded vectors obtained from visual sensors such as depth or RGB cameras allowing the robot to essentially build up a visual map of the environment and then compare newly observed landmarks against a database to localize via object or landmark

identification. In recent years, the use of handcrafted image features such as SIFT and SURF and feature dictionaries such as the Bag-of-Words (BoW) model in landmark representation has been superseded by feature representations *learned* through training on large example sets, usually by variants of artificial neural networks such as convolutional neural networks (CNNs). CNNs have also outperformed classifiers such as support vector machines (SVMs) in deriving inferences [42, 43]. However, several rapidly evolving CNN architectures exist. Ovalle-magallanes et al. [44] performed a comparative study of four such networks while successfully localizing in a visual map.

The *RoboCup Soccer League* is popular in humanoid research due to the visual identification and localization challenges it presents. [45, 46] and [47] are some examples of real-time, CNN based ball detection approaches utilizing RGB cameras developed specifically for RoboCup. Cruz et al. [48] could additionally estimate player poses, goal locations and other key pitch features using intensity images alone. Due to the low on-board computational power of the humanoids, others have used fast, low power external mobile GPU boards such as the Nvidia Jetson to aid inference [47, 49].

Unstructured and semi-structured environments are encountered outdoors or in hazardous and disaster rescue scenarios. They have a dearth of reliably trackable features, unpredictable

lighting conditions and are challenging for gathering training data. Thus, instead of features, researchers have focused on raw point clouds or combining different sensor modalities for navigating such environments. Starr et al. [50] presented a sensor fusion approach which combined long-wavelength infrared stereo vision and a spinning LiDAR for accurate rangefinding in smoke-obscured environments. Nobili et al. [51] successfully localized robots constrained by a limited field-of-view LiDAR in a semi-structured environment. They proposed a novel strategy for tuning outlier filtering based on point cloud overlap which achieved good localization results in the DARPA Robotics Challenge Finals. Raghavan et al. [52] presented simultaneous odometry and mapping by fusing LiDAR and kinematic-inertial data from IMU, joint encoders, and foot F/T sensors while navigating a disaster environment.

SLAM SLAM subsumes localization by the additional map construction and loop closing aspects, whereby the robot has to re-identify and match a place which was visited sometime in the past, to its current surroundings and adjust its pose history and recorded landmark locations accordingly. A humanoid robot which is intended to share human workspaces needs to deal with moving objects, both rapid and slow, which could disrupt its mapping and localizing capabilities. Thus, recent works on SLAM have focused on handling the presence of dynamic obstacles in visual scenes. While the most popular approach remains sensor fusion [53, 54], other purely visual approaches have also been proposed, such as, [55] which introduced a dense RGB-D SLAM solution that utilized optical flow residuals to achieve accurate and efficient dynamic/static segmentation for camera tracking and background reconstruction. Zhang et al. [56] took a more direct approach which employed deep learning based human detection, and used graph-based segmentation to separate moving humans from the static environment. They further presented a SLAM benchmark dedicated to dynamic environment SLAM solutions [57]. It included RGB-D data acquired from an on-board camera on the *HRP-4* humanoid robot, along with other sensor data. Adapting publicly available SLAM solutions and tailoring it for humanoid use is not uncommon. Sewtz et al. [58] adapted the Orb-Slam [59] for a multi-camera setup on the DLR *Rollin' Justin* System while Ginn et al. [60] did it for the *iGus*, a mid-size humanoid platform, to have low computational demands.

Navigation and Planning

Navigation and planning algorithms use perception information to generate a safe, optimal and reactive path, considering obstacles, terrain, and other constraints.

Local Planning Local planning or reactive navigation is generally concerned with local real-time decision-making and control, allowing the robot to actively respond to perceived changes in the environment and adjust its movements accordingly. Especially in highly controlled applications rule-based, perception driven navigation is still popular and yields state-of-the-art performance both in terms of time demands and task accomplishment. Bista et al. [61] achieved real-time navigation in indoor environments by representing the environment by key RGB images, and deriving a control law based on common line segments and feature points between the current image and nearby key images. Regier et al. [62] determined appropriate actions based on a pre-defined set of mappings between object class and action. A CNN was used to classify objects from monocular RGB vision. Ferro et al. [63] integrated information from a monocular camera, joint encoders, and an IMU to generate a collision-free visual servo control scheme. Juang et al. [64] developed a line follower which was able to infer forward, lateral and angular velocity commands using path curvature estimation and PID control from monocular RGB images. Magassouba et al. [65] introduced an aural servo framework based on auditory perception, enabling robot motions to be directly linked to low-level auditory features through a feedback loop.

We also see the use of a diverse array of classifiers to learn navigation schemes from perception information. Their generalization capability allows adaptation to unforeseen obstacles and events in the environment. Abiyev et al. [66] presented a vision-based path-finding algorithm which segregated captured images into free and occupied areas using an SVM. Lobos-tsunekawa et al. [67] and Silva et al. [68] proposed deep learned visual (RGB) navigation systems for humanoid robots which were able to achieve real time performance. The former used a reinforcement learning (RL) system with an actor-critic architecture while the latter utilized a decision tree of deep neural networks deployed on a soccer playing robot.

Global Planning These algorithms operate globally, taking into account long-term objectives and optimize movements to minimize costs, maximize efficiency, or achieve a specific outcome on the basis of a perceived environment model.

Footstep Planning is a crucial part of humanoid locomotion and has generated substantial research interest for itself. Recent works exhibit two primary trends related to perception. The first is providing humanoids the capability of rapidly perceiving changes in the environment and reacting through fast re-planning. The second endeavors to segment and/or classify uneven terrains to find stable 6 DoF footholds for highly versatile navigation.

Tanguy et al. [54] proposed a model predictive control (MPC) scheme that fused visual SLAM and proprioceptive F/T sensors for accurate state estimation. This allowed rapid reaction to external disturbances by adaptive stepping leading to balance recovery and improved

localization accuracy. Hildebrandt et al. [69] used the point cloud from an RGB-D camera to model obstacles as swept-sphere-volumes (SSVs) and step-able surfaces as convex polygons for real-time reactive footstep planning with the *Lola* humanoid robot. Their system was capable of handling rough terrain as well as external disturbances such as pushes (see Fig. 3). Others have also used geometric primitives to aid in footstep planning, such as surface patches for foothold representation [70, 71], environment segmentation to find step-able regions, such as 2D plane segments embedded in 3D space [72, 73], or represented obstacles by their polygonal ground projections [74]. Suryamurthy et al. [75] assigned pixel-wise terrain labels and rugosity measures using a CNN consuming RGB images for footstep planning on a *CENTAURO* robot.

Whole Body Planning in humanoid robots involves the coordinated planning and control of the robot's entire body to achieve an objective. Coverage planning is a subset of whole body planning where a minimal sequence of whole body robot poses are estimated to completely explore a 3D space via robot mounted visual sensors [76, 77]. Target finding is a special case of coverage planning where the exploration stops when the target is found [78, 79]. These concepts are related primarily to view planning in computer vision. In other applications, Wang et al. [80] presented a method for trajectory planning and formation building of a robot fleet using local positions estimated from onboard optical sensors and Liu et al. [81] presented a temporal planning approach for choreographing dancing robots in response to microphone-sensed music.

Perception in Grasping and Manipulation

Manipulation and grasping in humanoid robots involve their ability to interact with objects of varying shapes, sizes, and weights, to perform dexterous manipulation tasks using their sensor equipped end-effectors which provide visual or tactile feedback for grip adjustment.

Grasp Planning Grasp planning is a lower level task specifically focused on determining the optimal manipulator pose sequence to securely and effectively grasp an object. Visual information is used to find grasping locations and also as a feedback to optimize the difference between the target grasp pose and the current end-effector pose.

Schmidt et al. [82] utilized a CNN trained on object depth images and pre-generated analytic grasp plans to synthesize grasp solutions. The solution generated full

Fig. 3 Footstep planning on the humanoid *Lola* from [69]. Top left: The robot's vision system and a human causing disturbance. Bottom right: The collision model with geometric obstacle approximations

LOLA with vision system

Environment representation E

Full collision model R_{coll}

Ideal step sequence
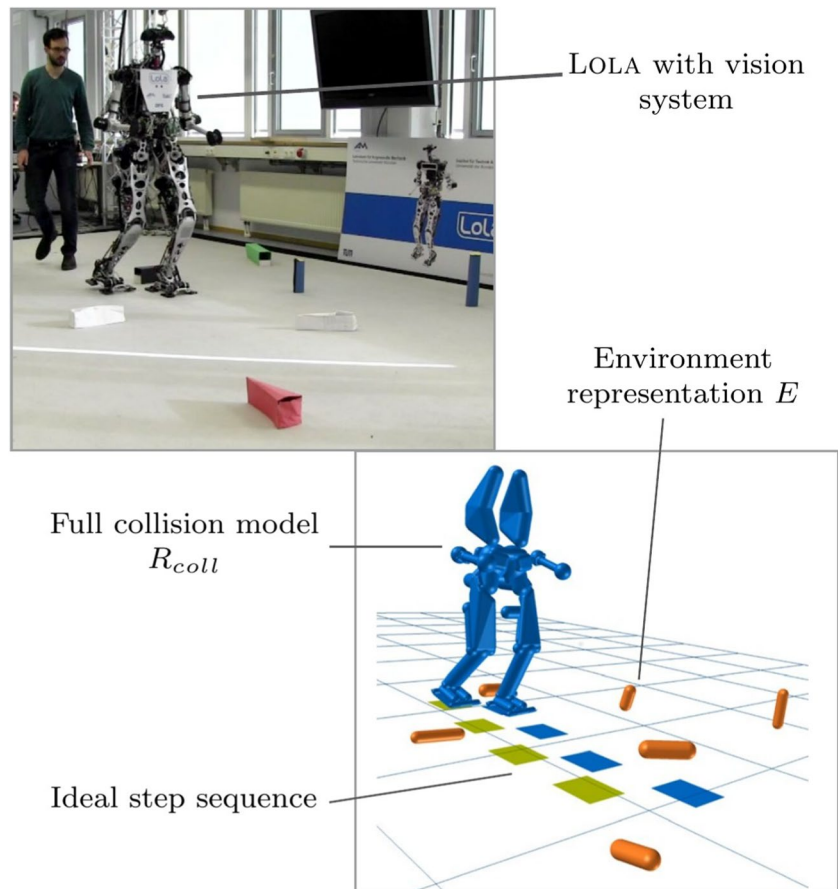
end-effector poses and could generate poses not limited to the camera view direction. Vezzani et al. [83] modeled the shape and volume of the target object captured from stereo vision in real-time using super-quadric functions allowing grasping even when parts of the object were occluded. Vicente et al. [84] and Nguyen [85] focused on achieving accurate hand-eye coordination in humanoids equipped with stereo vision. While the former compensated for kinematic calibration errors between the robot's internal hand model and captured images using particle based optimization, the latter trained a deep neural network predictor to estimate the robot arm's joint configuration. [86] proposed a combination of CNNs and dense conditional random fields (CRFs) to infer action possibilities on an object (affordances) from RGB images.

Tactile sensors, such as pressure-sensitive skins or fingertip sensors, provide feedback about the contact (surface normal) forces, slip detection, object texture, and shape information during object grasping. Kaboli et al. [87] extracted tactile descriptors for material and object classification agnostic to various sensor types such as dynamic pressure sensors, accelerometers, capacitive sensors, and impedance electrode arrays. A Nao with artificial skin used for their experiments is shown in Fig. 4. Hundhausen et al. [88] introduced a soft humanoid hand equipped with in-finger integrated cameras and an in-hand real-time image processing system based on CNNs for fast reactive grasping.

Manipulation Planning Manipulation planning involves the higher-level decision-making process of determining how the robot should manipulate an object once it is grasped. It generates a sequence of motions or actions which is updated based on the continuously perceived robot and grasped object state.

Deep recurrent neural networks (RNNs) are capable of predicting the next element in a sequence based on the previous elements. This property is exploited in manipulation planning by breaking down a complex task into a series of manipulation commands generated by RNNs based on past commands. These networks are capable of mapping features extracted from a sequence of RGB images, usually by CNNs, to a sequence of motion commands [89, 90]. Inceoglu et al. [91] presented a multimodal failure monitoring and detection system for robots which integrated high-level proprioceptive, auditory, and visual information during manipulation tasks. Robot assisted dressing is a challenging manipulation task that has been addressed by multiple authors. Zhang et al. [92] utilized a hierarchical multi-task control strategy to adapt the humanoid robot *Baxter*'s applied forces, measured using joint torques, to the user's movements during dressing. By tracking the subject human's pose in real-time using capacitive proximity sensing with low latency and high signal-to-noise ratio, Erickson et al. [93] developed a
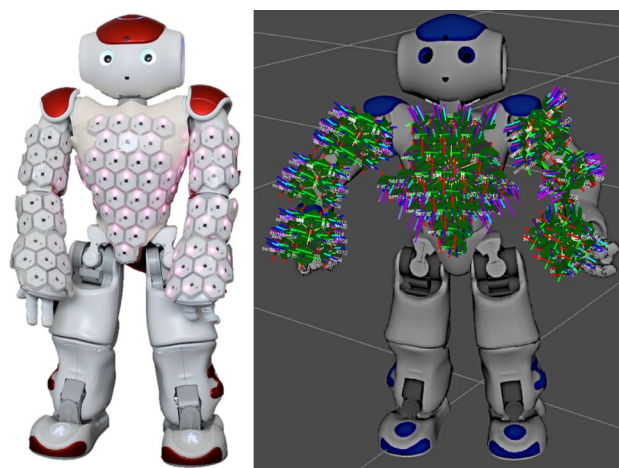

Fig. 4 Left: A Nao humanoid equipped with artificial skin cells on the chest, hand, fore arm, and upper arm. Right: Visualization of the skin cell coordinate frames on the Nao. Figure taken from [87]

method to adapt to human motion and adjust for errors in pose estimation during dressing assistance by the *PR2* robot. Zhang et al. [94] computed suitable grasping points on garments from depth images using a deep neural network to facilitate robot manipulation in robot-assisted dressing tasks.

Human-Robot Interaction

Human robot interaction is a subset of environment understanding which deals with interactions with humans as opposed to inanimate objects. In order to achieve this, a robot needs diverse capabilities ranging from detecting humans, recognizing their pose, gesture, and emotions, to predicting their intent and even proactively performing actions to ensure a smooth and seamless interaction.

There are two main challenges to perception in HRI - perception of users, and inference which involves making sense of the data and making predictions.

Perception of Users

This involves identifying humans in the environment, detecting their pose, facial features, and objects they interact with. This information is crucial for action prediction and emotion recognition [95]. Robots rely on vision-based, audio-based, tactile-based, and range sensor-based sensing techniques for detection as explained in this survey on perception methods of social robots done by [96].

Robinson et al. [97] showed how vision-based techniques have evolved from using facial features, motion features, and body appearance to deep learning-based approaches.

Motion-based features separate moving objects from the background to detect humans. Body appearance-based algorithms use shape, curves, posture, and body parts to detect humans. Deep learning models like R-CNN, Faster R-CNN, and YOLO have also been applied for human detection [96].

Pose detection is essential for understanding human body movements and postures. Sensors such as RGB cameras, stereo cameras, depth sensors, and motion tracking systems are used to extract pose information. This was explained in detail by Möller et al. [98] in their survey of human-aware robot navigation. Facial features play a significant role in pose detection as they provide additional points of interest and enable emotion recognition [99]. A great demonstration of detecting pose and using it for bi-manual robot control using an RGB-D range sensor was shown by Hwang et al. [100]. The system employed a CNN from the *OpenPose* package to extract human skeleton poses, which were then mapped to drive robotic hands. The method was implemented on the *CENTAURO* robot and successfully performed box and lever manipulation tasks in real-time. They presented a real-time pose imitation method for a mid-size humanoid robot equipped with a servo-cradle-head RGB-D vision system. Using eight pre-trained neural networks, the system accurately captured and imitated 3D motions performed by a target human, enabling effective pose imitation and complex motion replication in the robot. Lv et al. [101] presented a novel motion synchronization method called *GuLiM* for teleoperation of medical assistive robots, particularly in the context of combating the COVID-19 pandemic. Li et al. [102] presented a multimodal mobile teleoperation system that integrated a vision-based hand pose regression network and an IMU-based arm tracking method. The system allowed real-time control of a robot hand-arm system using depth camera observations and IMU readings from the observed human hand, enabled through the *Transteleop* neural network which generated robot hand poses based on a depth image input of a human hand.

Audio communication is vital for human interaction, and robots aim to mimic this ability. Microphones are used for audio detection, and speakers reproduce sound. Humanoid robots are usually designed to be binaural i.e., they have two separate microphones at either side of the head which receive transmitted sound independently. Several researchers have focused on this property to localize both the sound source and the robot in complex auditory environments. Such techniques are used in speaker localization, as well as other semantic understanding tasks such as automatic speech recognition (ASR), auditory scene analysis, emotion recognition, and rhythm recognition [96, 103].

Benaroya et al. [104] employed non-negative tensor factorization for binaural localization of multiple sound sources within unknown environments. Schymura et al. [105] focused on combined audio-visual speaker localization and proposed a closed-form solution to compute dynamic stream weighting between audio and visual streams, improving the state estimation in a reverberant environment. The previous study was extended to incorporate dynamic stream weights into nonlinear dynamical systems which improved speaker localization performance even further [106]. Dávila-Chacón et al. [107] used a spiking and a feed-forward neural network for sound source localization and ego noise removal respectively to enhance ASR in challenging environments. Trowitzsc et al. [108] presented a joint solution for sound event identification and localization, utilizing spatial audio stream segregation in a binaural robotic system.

Ahmad et al. [109] in their survey on physiological signal-based emotion recognition showed that physiological signals from the human body, such as such as heart rate, blood pressure, body temperature, brain activity, and muscle activation can provide insights into emotions. Tactile interaction is an inherent part of natural interaction between humans and the same holds true for robots interacting with humans as well. The type of touch can be used to infer a lot of things such as the human's state of mind, the nature of the object, what is expected out of the interaction, etc. [96]. Mainly two kinds of tactile sensors are used for this purpose - sensors embedded on the robot's arms and grippers, and cover based sensors which are used to detect touch across entire regions or the whole body [96]. Khurshid et al. [110] investigated the impact of grip-force, contact, and acceleration feedback on human performance in a teleoperated pick-and-place task. Results indicated that grip-force feedback improved stability and delicate control, while contact feedback improved spatial movement but may vary depending on object stiffness.

Inference

An important aspect of inference with all the detected data from the previous section is regarding aligning the perspective of the user and the robot. This allows the robot to better understand the intent of the user regarding the objects or locations they are looking at. This skill is called *perspective taking* and requires the robot to consider and understand other individuals through motivation, disposition, and contextual attempts. This skill paired with a shared knowledge base allows the individuals and robots to build a reliable *theory of mind* and collaborate effectively during various types of tasks [3].

Bera et al. [111] proposed an emotion-aware navigation algorithm for social robots which combined emotions learned from facial expressions and walking trajectories using an onboard and an overhead camera respectively. The approach achieved accurate emotion detection and enabled socially conscious robot navigation in low-to-medium-density environments.

Table 2 A non-exhaustive, indicative list of popular humanoid robot models used by different publications

		Robots													
		HRP-2	HRP-4	iCub	NAO	Sarcos	Atlas	Talos	Valkyrie	WALK-MAN	Cassie	ARMAR-III	ARMAR-6	Romeo	Pepper
State Estimation	Proprioceptive	[24]	[28]	-	[33] [13] [27]	[7]* [29]* [5]	[2] [33]* [8] [25] [36] [34]* [9]	[33] [34]	[20]*	-	[16]	-	-	-	-
	Multi-sensor filtering	-	-	-	[39]	-	[2] [12]	-	[12]	[39]	-	-	-	-	-
	Multi-sensor smoothing	[26]	-	-	-	-	-	-	-	-	[41] [40]	-	-	-	-
Environment Understanding	Localization, and Navigation	-	[57] [54] [55]	[66, 76]	[44, 63] [62, 64] [77, 112] [78, 80] [87]	-	[51]	-	[51] [53]	[52] [70]	-	-	-	[65]	[65]
	Grasping and Manipulation	-	-	[83] [84] [113] [38]	-	-	-	-	-	[86]	-	[82]	[38] [88]	-	-
Human-Robot Interaction	Detection	[96]	-	[107] [96] [3] [97]	[105] [97]	[97]	-	-	-	-	-	[96]	-	-	[98] [97]
	Inference	-	-	[3]	-	-	-	-	-	-	-	-	-	-	[111]

*In simulation

Conclusion

Substantial progress have been made in all three principal areas discussed in this survey. In Table 2 we compile a list of the most commonly cited humanoids in the literature corresponding to the aforementioned categorization. We conclude with a summary of the trends and possible areas of further research we observed in each of these areas.

State Estimation Tightly-coupled formulation of state estimation based on MAP seems to be promising for future works as it offers several advantages, such as modularity and enabling seamless integration of new sensor types, and extending generic estimators with accommodating a wider range of perception sources in order to develop a whole-body estimation framework. By integrating high-rate control estimation and non-drifting localization based on SLAM, this framework could provide real-time estimation for locomotion control purposes, and facilitate gait and contact planning.

Another important area of focus is the development of multi-contact detection and estimation methods for arbitrary unknown contact locations. By moving beyond rigid segment assumptions for humanoid structure and augmenting robots with additional sensors, such as strain gauges to directly measure segment deflections; the multi-contact detection and compensating for modeling errors can lead to more accurate state estimation and improved human-robot interactions.

Environment Understanding With the availability of improved inference hardware, learning techniques are increasingly being applied in localization, object identification, and mapping, replacing handcrafted feature descriptors. However, visual classifiers like CNNs struggle with unstructured "stuff" compared to regularly shaped objects, necessitating memory-intensive representations such as point clouds and the need for enhanced classifier capabilities. In the field of SLAM, which has robust solutions for static environments, research is focused on handling dynamic obstacles by favoring multi-sensor fusion for increased robustness. Scalability and real-time capability remain challenging due to the potential overload of a humanoid's onboard computer from wrangling multiple data streams over long sequences. Footstep planning shows a trend towards rapid environment modeling for quick responses, but consistent modeling of dynamic obstacles remains an open challenge. Manipulation and long-term global planning also rely on learning techniques to adapt to unforeseen constraints, requiring representations or embeddings of high-dimensional interactions between perceived elements for complexity reduction. However, finding more efficient, comprehensive, and accurate methods to express these relationships is an ongoing challenge.

Human Robot Interaction Research in the field of HRI has focused on understanding human intent and emotion through various elements such as body pose, motions, expressions, audio cues, and behavior. Though this may seem natural and trivial from a human's perspective, it is often a very challenging task to incorporate the same into robotic systems. Despite considerable progress in the above approaches, the ever-changing and unpredictable nature of human interaction necessitates additional steps that incorporate concepts like shared autonomy and shared perception. In this context, contextual information and memory play a crucial role in accurately perceiving the state and intentions of the humans with whom interaction is desired. Current research endeavors are actively focusing on these pivotal topics, striving to enhance the capabilities of humanoid robots in human-robot interactions while also considering trust, safety, explainability, and ethics during these interactions.

Declarations

References

1. Tanguy A, Gergondet P, Comport AI, Kheddar A. Closed-loop RGB-D SLAM multi-contact control for humanoid robots. In: IEEE/SICE Intl symposium on system integration (SII); 2016. p. 51–57.
2. Fallon MF, Antone M, Roy N, Teller S. Drift-free humanoid state estimation fusing kinematic, inertial and lidar sensing. In: IEEE-RAS Intl Conf on humanoid robots (Humanoids); 2014. p. 112–119.

3. Matarese M, Rea F, Sciutti A. Perception is only real when shared: a mathematical model for collaborative shared perception in human-robot interaction. Frontiers Robotics AI. 2022; 733954.

4. Carpentier J, Wieber PB. Recent progress in legged robots locomotion control. Current Robotics Reports. 2021;.

5. Rotella N, Mason S, Schaal S, Righetti L. Inertial sensor-based humanoid joint state estimation. In: IEEE Intl Conf on Robotics & Automation (ICRA); 2016. p. 1825–1831.

6. Bloesch M, Hutter M, Hoepflinger MA, Leutenegger S, Gehring C, Remy CD, et al. State estimation for legged robots-consistent fusion of leg kinematics and IMU. Robotics. 2013;.

7. Rotella N, Blösch M, Righetti L, Schaal S. State estimation for a humanoid robot. In: IEEE/RSJ Intl Conf on intelligent robots and systems (IROS); 2014. p. 952–958.

8. Xinjilefu X, Feng S, Huang W, Atkeson CG. Decoupled state estimation for humanoids using full-body dynamics. In: IEEE Intl conf on robotics & automation (ICRA); 2014. p. 195–201.

9. Xinjilefu X, Feng S, Atkeson CG. Center of mass estimator for humanoids and its application in modelling error compensation, fall detection and prevention. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2015. p. 67–73.

10. Bae H, Jeong H, Oh J, Lee K, Oh JH. Humanoid robot COM kinematics estimation based on compliant inverted pendulum model and robust state estimator. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2018. p. 747–753.

11. Vigne M, El Khoury A, Di Meglio F, Petit N. State estimation for a legged robot with multiple flexibilities using IMUs: A kinematic approach. IEEE Robotics Auto Lett (RA-L). 2020;.

12. Camurri M, Ramezani M, Nobili S, Fallon M. Pronto: A multi-sensor state estimator for legged robots in real-world scenarios. Frontiers Robotics AI. 2020;.

13. Piperakis S, Koskinopoulou M, Trahanias P. Nonlinear state estimation for humanoid robot walking. IEEE Robotics Auto Lett (RA-L). 2018;.

14. Piperakis S, Koskinopoulou M, Trahanias P. Piperakis S, Koskinopoulou M, Trahanias P, editors.: SEROW. Github;2016. https://github.com/mrsp/serow.

15. Camurri M, Fallon M, Bazeille S, Radulescu A, Barasuol V, Caldwell DG, et al.. Camurri M, Fallon M, Bazeille S, Radulescu A, Barasuol V, Caldwell DG, et al., editors.: Pronto. Github;2020. https://github.com/ori-drs/pronto.

16. Hartley R, Ghaffari M, Eustice RM, Grizzle JW. Contact-aided invariant extended Kalman filtering for robot state estimation. Intl J Robotics Res (IJRR). 2020;.

17. Hartley R, Ghaffari M, Eustice RM, Grizzle JW. Hartley R, Ghaffari M, Eustice RM, Grizzle JW, editors.: InEKF. Github; 2018. https://github.com/RossHartley/invariant-ekf.

18. Solá J, Vallvé J, Casals J, Deray J, Fourmy M, Atchuthan D, et al. WOLF: A modular estimation framework for robotics based on factor graphs. IEEE Robotics Auto Lett (RA-L). 2022;.

19. Solá J, Vallvé J, Casals J, Deray J, Fourmy M, Atchuthan D, et al.. Solá J, Vallvé J, Casals J, Deray J, Fourmy M, Atchuthan D, et al., editors.: WOLF. IRI;2022. https://mobile_robotics.pages.iri.upc-csic.es/wolf_projects/wolf_lib/wolf-doc-sphinx/ .

20. Piperakis S, Timotheatos S, Trahanias P. Unsupervised gait phase estimation for humanoid robot walking. In: IEEE Intl Conf on Robotics & Automation (ICRA); 2019. p. 270–276.

21. Piperakis S, Timotheatos S, Trahanias P. Piperakis S, Timotheatos S, Trahanias P, editors.: GEM. Github; 2019. https://github.com/mrsp/gem.

22. Bloesch M. State Estimation for Legged Robots - Kinematics, inertial sensing, and computer vision [Thesis]. ETH Zurich; 2017.

23. Camurri M. Multisensory state estimation and mapping on dynamic legged robots. Istituto Italiano di Tecnologia and Univ Genoa. 2017;p. 13.

24. Flayols T, Del Prete A, Wensing P, Mifsud A, Benallegue M, Stasse O. Experimental evaluation of simple estimators for humanoid robots. In: IEEE-RAS Intl Conf on Humanoid Robots (Humanoids); 2017. p.889–895.

25. Xinjilefu X, Feng S, Atkeson CG. Dynamic state estimation using quadratic programming. In: IEEE/RSJ Intl Conf on Intelligent Robots and Systems (IROS); 2014. p. 989–994.

26. Fourmy M. State estimation and localization of legged robots: a tightly-coupled approach based on a-posteriori maximization [Thesis]. INSA: Toulouse; 2022.

27. Piperakis S, Trahanias P. Non-linear ZMP based state estimation for humanoid robot locomotion. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2016. p. 202–209.

28. Mori K, Ayusawa K, Yoshida E. Online center of mass and momentum estimation for a humanoid robot based on identification of inertial parameters. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2018. p. 1–9.

29. Rotella N, Herzog A, Schaal S, Righetti L. Humanoid momentum estimation using sensed contact wrenches. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2015. p.556–563.

30. Carpentier J, Benallegue M, Mansard N, Laumond JP. Center-of-mass estimation for a polyarticulated system in contact–a spectral approach. IEEE Trans Robotics (TRO). 2016;.

31. Bailly F, Carpentier J, Benallegue M, Watier B, Souéres P. Estimating the center of mass and the angular momentum derivative for legged locomotion–a recursive approach. IEEE Robotics Auto Lett (RA-L). 2019;.

32. Rotella N, Schaal S, Righetti L. Unsupervised contact learning for humanoid estimation and control. In: IEEE Intl conf on robotics & automation (ICRA); 2018. p. 411–417.

33. Piperakis S, Maravgakis M, Kanoulas D, Trahanias P. Robust contact state estimation in humanoid walking gaits. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2022. p. 6732–6738.

34. Maravgakis M, Argiropoulos DE, Piperakis S, Trahanias P. Probabilistic contact state estimation for legged robots using inertial information. In: IEEE Intl conf on robotics & automation (ICRA); 2023. p.12163–12169.

35. Flacco F, Paolillo A, Kheddar A. Residual-based contacts estimation for humanoid robots. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2016. p.409–415.

36. Manuelli L, Tedrake R. Localizing external contact using proprioceptive sensors: The contact particle filter. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2016. p. 5062–5069.

37. Vorndamme J, Haddadin S. Rm-Code: proprioceptive real-time recursive multi-contact detection, isolation and identification. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2021. p. 6307–6314.

38. Vezzani G, Pattacini U, Battistelli G, Chisci L, Natale L. Memory unscented particle filter for 6-DOF tactile localization. IEEE Trans Robotics (TRO). 2017; 1139–1155.

39. Piperakis S, Kanoulas D, Tsagarakis NG, Trahanias P. Outlier-robust state estimation for humanoid robots. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2019. p. 706–713.

40. Hartley R, Mangelson J, Gan L, Jadidi MG, Walls JM, Eustice RM, et al. Legged robot state-estimation through combined forward kinematic and preintegrated contact factors. In: IEEE Intl conf on robotics & automation (ICRA); 2018. p. 4422–4429.

41. Hartley R, Jadidi MG, Gan L, Huang JK, Grizzle JW, Eustice RM. Hybrid Contact preintegration for visual-inertial-contact state estimation using factor graphs. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2018. p. 3783–3790.

42. Wozniak P, Afrisal H, Esparza RG, Kwolek B. Scene recognition for indoor localization of mobile robots using deep CNN.

In: Intl conf on computer vision and graphics (ICCVG); 2018. p.137–147.

43. Wozniak P, Kwolek B. Place inference via graph-based decisions on deep embeddings and blur detections. In: Intl conf on computational science (ICCS); 2021. p. 178–192.

44. Ovalle-Magallanes E, Aldana-Murillo NG, Avina-Cervantes JG, Ruiz-Pinales J, Cepeda-Negrete J, Ledesma S. Transfer learning for humanoid robot appearance-based localization in a visual Map. IEEE Access. 2021;p. 6868–6877.

45. Speck D, Bestmann M, Barros P. Towards real-time ball localization using CNNs. In: RoboCup 2018: Robot World Cup XXII; 2019. p. 337–348.

46. Teimouri M, Delavaran MH, Rezaei M. A real-time ball detection approach using convolutional neural networks. In: RoboCup 2019: Robot World Cup XXIII; 2019. p. 323–336.

47. Gabel A, Heuer T, Schiering I, Gerndt R. Jetson, Where is the ball? using neural networks for ball detection at RoboCup 2017. In: RoboCup 2018: Robot World Cup XXII; 2019. p. 181–192.

48. Cruz N, Leiva F, Ruiz-del-Solar J. Deep learning applied to humanoid soccer robotics: playing without using any color information. Autonomous Robots. 2021;.

49. Chatterjee S, Zunjani FH, Nandi GC. Real-time object detection and recognition on low-compute humanoid robots using deep learning. In: Intl conf on control, automation and robotics (ICCAR); 2020. p.202–208.

50. Starr JW, Lattimer BY. Evidential sensor fusion of long-wavelength infrared stereo vision and 3D-LIDAR for rangefinding in fire environments. Fire Technol. 2017;1961–1983.

51. Nobili S, Scona R, Caravagna M, Fallon M. Overlap-based ICP tuning for robust localization of a humanoid robot. In: IEEE Intl conf on robotics & automation (ICRA); 2017. p. 4721–4728.

52. Raghavan VS, Kanoulas D, Zhou C, Caldwell DG, Tsagarakis NG. A study on low-drift state estimation for humanoid locomotion, using LiDAR and kinematic-inertial data fusion. In: IEEE-RAS Intl Conf on Humanoid Robots (Humanoids); 2018. p. 1–8.

53. Scona R, Nobili S, Petillot YR, Fallon M. Direct visual SLAM fusing proprioception for a humanoid robot. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS). IEEE; 2017. p. 1419–1426.

54. Tanguy A, De Simone D, Comport AI, Oriolo G, Kheddar A. Closed-loop MPC with dense visual SLAM-stability through reactive stepping. In: IEEE Intl conf on robotics & automation (ICRA). IEEE; 2019. p. 1397–1403.

55. Zhang T, Zhang H, Li Y, Nakamura Y, Zhang L. Flowfusion: Dynamic dense rgb-d slam based on optical flow. In: IEEE Intl conf on robotics & automation (ICRA). IEEE; 2020. p.7322–7328.

56. Zhang T, Uchiyama E, Nakamura Y. Dense rgb-d slam for humanoid robots in the dynamic humans environment. In: IEEE-RAS Intl conf on humanoid robots (Humanoids). IEEE; 2018. p. 270–276.

57. Zhang T, Nakamura Y. Hrpslam: A benchmark for rgb-d dynamic slam and humanoid vision. In: IEEE Intl conf on robotic computing (IRC). IEEE; 2019. p. 110–116.

58. Sewtz M, Luo X, Landgraf J, Bodenmüller T, Triebel R. Robust approaches for localization on multi-camera systems in dynamic environments. In: Intl conf on automation, robotics and applications (ICARA). IEEE; 2021. p. 211–215.

59. Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: A Versatile and accurate monocular SLAM system. IEEE Trans Robotics (TRO). 2015;1147–1163.

60. Ginn D, Mendes A, Chalup S, Fountain J. Monocular ORB-SLAM on a humanoid robot for localization purposes. In: AI: Advances in artificial intelligence; 2018. p. 77–82.

61. Bista SR, Giordano PR, Chaumette F. Combining line segments and points for appearance-based indoor navigation by image based visual servoing. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2017. p. 2960–2967.

62. Regier P, Milioto A, Karkowski P, Stachniss C, Bennewitz M. Classifying obstacles and exploiting knowledge about classes for efficient humanoid navigation. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2018. p. 820–826.

63. Ferro M, Paolillo A, Cherubini A, Vendittelli M. Vision-Based navigation of omnidirectional mobile robots. IEEE Robotics Auto Lett (RA-L). 2019;2691–2698.

64. Juang LH, Zhang JS. Robust visual line-following navigation system for humanoid robots. Artif Intell Rev. 2020;653–670.

65. Magassouba A, Bertin N, Chaumette F. Aural Servo: Sensor-based control from robot audition. IEEE Trans Robotics (TRO). 2018;572–585.

66. Abiyev RH, Arslan M, Gunsel I, Cagman A. Robot pathfinding using vision based obstacle detection. In: IEEE Intl conf on cybernetics (CYBCONF); 2017. p. 1–6.

67. Lobos-Tsunekawa K, Leiva F, Ruiz-del-Solar J. Visual navigation for biped humanoid robots using deep reinforcement learning. IEEE Robotics Auto Lett (RA-L). 2018; 3247–3254.

68. Silva IJ, Junior COV, Costa AHR, Bianchi RAC. Toward robotic cognition by means of decision tree of deep neural networks applied in a humanoid robot. J Control Autom Electr Syst. 2021; 884–894.

69. Hildebrandt AC, Wittmann R, Sygulla F, Wahrmann D, Rixen D, Buschmann T. Versatile and robust bipedal walking in unknown environments: real-time collision avoidance and disturbance rejection. Autonomous Robots. 2019;1957–1976.

70. Kanoulas D, Stumpf A, Raghavan VS, Zhou C, Toumpa A, Von Stryk O, et al. Footstep planning in rough terrain for bipedal robots using curved contact patches. In: IEEE Intl Conf on Robotics & Automation (ICRA); 2018. p. 4662–4669.

71. Kanoulas D, Tsagarakis NG, Vona M. Curved patch mapping and tracking for irregular terrain modeling: application to bipedal robot foot placement. J Robotics Autonomous Syst (RAS). 2019; 13–30.

72. Bertrand S, Lee I, Mishra B, Calvert D, Pratt J, Griffin R. Detecting usable planar regions for legged robot locomotion. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS);2020. p. 4736–4742.

73. Roychoudhury A, Missura M, Bennewitz M. 3D Polygonal mapping for humanoid robot navigation. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2022. p.171–177.

74. Missura M, Roychoudhury A, Bennewitz M. Polygonal perception for mobile robots. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2020. p. 10476–10482.

75. Suryamurthy V, Raghavan VS, Laurenzi A, Tsagarakis NG, Kanoulas D. Terrain segmentation and roughness estimation using rgb data:path planning application on the CENTAURO robot. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2019. p. 1–8.

76. Osswald S, Karkowski P, Bennewitz M. Efficient coverage of 3D environments with humanoid robots using inverse reachability maps. In: IEEE-RAS Intl conf on humanoid robots (Humanoids); 2017. p.151–157.

77. Osswald S, Bennewitz M. GPU-accelerated next-best-view coverage of articulated scenes. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2018. p. 603–610.

78. Monica R, Aleotti J, Piccinini D. Humanoid robot next best view planning under occlusions using body movement primitives. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2019. p. 2493–2500.

79. Tsuru M, Escande A, Tanguy A, Chappellet K, Harad K. Online object searching by a humanoid robot in an unknown environment. IEEE Robotics Auto Lett (RA-L). 2021; 2862–2869.

80. Wang X, Benozzi L, Zerr B, Xie Z, Thomas H, Clement B. Formation building and collision avoidance for a fleet of NAOs

based on optical sensor with local positions and minimum communication. Sci China Inf Sci. 2019;335–350.

81. Liu Y, Xie D, Zhuo HH, Lai L, Li Z. Temporal planning-based choreography from music. In: Sun Y, Lu T, Guo Y, Song X, Fan H, Liu D, et al., editors. Computer Supported Cooperative Work and Social Computing (CSCW); 2023. p. 89–102.

82. Schmidt P, Vahrenkamp N, Wachter M, Asfour T. Grasping of unknown objects using deep convolutional neural networks based on depth images. In: IEEE Intl conf on robotics & automation (ICRA); 2018. p.6831–6838.

83. Vezzani G, Pattacini U, Natale L. A grasping approach based on superquadric models. In: IEEE Intl conf on robotics & automation (ICRA); 2017. p.1579–1586.

84. Vicente P, Jamone L, Bernardino A. Towards markerless visual servoing of grasping tasks for humanoid robots. In: IEEE Intl conf on robotics & automation (ICRA); 2017. p.3811–3816.

85. Nguyen PDH, Fischer T, Chang HJ, Pattacini U, Metta G, Demiris Y. Transferring visuomotor learning from simulation to the real world for robotics manipulation tasks. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2018. p. 6667–6674.

86. Nguyen A, Kanoulas D, Caldwell DG, Tsagarakis NG. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2017. p. 5908–5915.

87. Kaboli M, Cheng G. Robust Tactile descriptors for discriminating objects from textural properties via artificial robotic skin. IEEE Trans on Robotics (TRO). 2018;985–1003.

88. Hundhausen F, Grimm R, Stieber L, Asfour T. Fast reactive grasping with in-finger vision and In-Hand FPGA-accelerated CNNs. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2021. p. 6825–6832.

89. Nguyen A, Kanoulas D, Muratore L, Caldwell DG, Tsagarakis NG. Translating videos to commands for robotic manipulation with deep recurrent neural networks. In: IEEE Intl conf on robotics & automation (ICRA); 2017. p. 3782–3788.

90. Kase K, Suzuki K, Yang PC, Mori H, Ogata T. Put-in-box task generated from multiple discrete tasks by ahumanoid robot using deep learning. In: IEEE Intl conf on robotics & automation (ICRA); 2018. p.6447–6452.

91. Inceoglu A, Ince G, Yaslan Y, Sariel S. Failure detection using proprioceptive, auditory and visual modalities. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2018. p. 2491–2496.

92. Zhang F, Cully A, Demiris Y. Personalized robot-assisted dressing using user modeling in latent spaces. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2017. p. 3603–3610.

93. Erickson Z, Collier M, Kapusta A, Kemp CC. Tracking human pose during robot-assisted dressing using single-axis capacitive proximity sensing. IEEE Robotics Auto Lett (RA-L). 2018; 2245–2252.

94. Zhang F, Demiris Y. Learning grasping points for garment manipulation in robot-assisted dressing. In: IEEE Intl conf on robotics & automation (ICRA); 2020. p. 9114–9120.

95. Narayanan V, Manoghar BM, Dorbala VS, Manocha D, Bera A. Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2020. p. 8200–8207.

96. Yan H, Ang MH, Poo AN. A survey on perception methods for human–robot interaction in social robots. Intl J Soc Robotics. 2014; 85–119.

97. Robinson N, Tidd B, Campbell D, Kulić D. Corke P. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. ACM Trans on Human-Robot Interaction; 2023. p. 1–66.

98. Möller R, Furnari A, Battiato S, Härmä A, Farinella GM. A survey on human-aware robot navigation. J Robotics Autonomous Syst (RAS). 2021;103837.

99. Samadiani N, Huang G, Cai B, Luo W, Chi CH, Xiang Y, et al. A review on automatic facial expression recognition systems assisted by multimodal sensor data. IEEE Sensors J. 2019;1863.

100. Hwang CL, Liao GH. Real-time pose imitation by mid-size humanoid robot with servo-cradle-head RGB-D vision system. IEEE Intl conf on systems, man, and cybernetics (SMC). 2019;p.181–191.

101. Lv H, Kong D, Pang G, Wang B, Yu Z, Pang Z, et al. GuLiM: A hybrid motion mapping technique for teleoperation of medical assistive robot in combating the COVID-19 pandemic. IEEE Trans Medical Robotics Bionics. 2022;106–117.

102. Li S, Jiang J, Ruppel P, Liang H, Ma X, Hendrich N, et al. A mobile robot hand-arm teleoperation system by vision and IMU. In: IEEE/RSJ Intl conf on intelligent robots and systems (IROS); 2020. p. 10900–10906.

103. Badr AA, Abdul-Hassan AK. A review on voice-based interface for human-robot interaction. Iraqi J Electric Electron Eng. 2020;91–102.

104. Benaroya EL, Obin N, Liuni M, Roebel A, Raumel W, Argentieri S. Binaural localization of multiple sound sources by non-negative tensor factorization. IEEE/ACM Trans on Audio Speech Lang Process. 2018; 1072–1082.

105. Schymura C, Isenberg T, Kolossa D. Extending linear dynamical systems with dynamic stream weights for audiovisual speaker localization. In: International workshop on acoustic signal enhancement (IWAENC); 2018. p. 515–519.

106. Schymura C, Kolossa D. Audiovisual speaker tracking using non-linear dynamical systems with dynamic stream weights. IEEE/ACM Trans on Audio Speech Lang Process. 2020; 1065–1078.

107. Dávila-Chacón J, Liu J, Wermter S. Enhanced robot speech recognition using biomimetic binaural sound source localization. IEEE Trans on Neural Netw Learn Syst. 2019; 138–150.

108. Trowitzsch I, Schymura C, Kolossa D, Obermayer K. Joining sound event detection and localization through spatial segregation. IEEE/ACM Trans on Audio Speech Lang Process. 2020; 487–502.

109. Ahmad Z, Khan N. A survey on physiological signal-based emotion recognition. Bioengineering. 2022;p. 688.

110. Khurshid RP, Fitter NT, Fedalei EA, Kuchenbecker KJ. Effects of grip-force, contact, and acceleration feedback on a teleoperated pick-and-place task. IEEE Trans on Haptics. 2017; 985–1003.

111. Bera A, Randhavane T, Manocha D. Modelling multi-channel emotions using facial expression and trajectory cues for improving socially-aware robot navigation. In: IEEE/CVF Conf on computer vision and pattern recognition workshops (CVPRW); 2019. p. 257–266.

112. Jin Y, Lee M. Enhancing binocular depth estimation based on proactive perception and action cyclic learning for an autonomous developmental robot. IEEE Intl conf on systems, man, and cybernetics (SMC). 2018;p. 169–180.

113. Hoffmann M, Straka Z, Farkaš I, Vavrečka M, Metta G. Robotic homunculus: learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex. IEEE Trans Cognitive Develop Syst. 2017;163–176.