2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Humanoid Robot Next Best View Planning Under Occlusions Using Body Movement Primitives

Riccardo Monica[1], Jacopo Aleotti[1] and Davide Piccinini[1]

Abstract— **This work presents an approach for humanoid Next Best View (NBV) planning that exploits full body motions to observe objects occluded by obstacles. The task is to explore a given region of interest in an initially unknown environment. The robot is equipped with a depth sensor, and it can perform both 2D and 3D mapping. As main contribution with respect to previous work, the proposed method does not rely on simple motions of the head and it was evaluated in real environments. The robot is guided by two behaviors: a target behavior that aims at observing the region of interest by exploiting body movements primitives, and an exploration behavior that aims at observing other unknown areas. Experiments show that the humanoid is able to peer around obstacles to reach a favourable point of view. Moreover, the proposed approach results in a more complete reconstruction of objects than a conventional algorithm that only changes the orientation of the head.**

I. INTRODUCTION

When searching for objects that are occluded by other surfaces humans can exploit their body and perform complex movements, to achieve a favourable point of view. For example, humans can peer around corners, or bend over to look into a drawer. Humanoids robots can mimic human body movements to perform similar object searching tasks.

This work presents an approach for humanoid Next Best View (NBV) planning under occlusions in an initially unknown environment that exploits full body motions to observe objects. In general, NBV systems enable active robot perception by maximizing a utility function to find the most promising configuration of a sensor carried by the robot [1]–[3]. In the proposed approach, the humanoid robot is given a point of interest (*POI*) as input, which indicates a region of space that deserves observation. The goal of the task is to efficiently perform 3D shape reconstruction of the area around the *POI*.

Research on NBV planning has mainly focused on fixed base manipulators and mobile robots, while it has received less attention in the context of humanoid robots. The difficulty stems from the fact that executing a NBV task with a humanoid robot in a real environment, requires accurate robot localization, perception, mapping, and whole body motion planning. Some previous works on humanoid NBV planning have shown results in simulated environments [4]–[9], while other works, evaluated through experiments in real environments, did not consider full body motion planning, by assuming that the robot can perform only pan, tilt motions of the head, or by using simplified collision detection algorithms [10]–[14]. Hence, we claim that, in contrast to our work, in previous approaches a successful execution of a humanoid NBV task in a real environment with occlusions, exploiting the ability to perform full body movements, has not been achieved.

Fig. 1. A humanoid robot leaning to observe a target object (jug) behind some obstacles (left image). Without leaning the object would not be visible. The object, as seen by the robot camera, at the center of the depth image reconstructed using Kinect Fusion (right image).

In this work a single look ahead step is used for the computation of the Next Best View. The humanoid robot is equipped with a depth camera, mounted on its head. Accurate robot localization and sensor calibration are obtained through a motion capture system. 3D perception is achieved by an open source implementation of the Kinect Fusion algorithm. The robot is able to build and update a 2D map of the obstacles, as well as a dense 3D map of the whole environment.

The known position of the *POI* is used, at each iteration, to determine a set of candidate whole-body robot configurations in world space from which the *POI* can be observed. Candidate robot configurations are ranked according to a utility function, and then the robot moves to the most promising reachable configuration. Due to the complexity and high computational cost of online whole-body motion planning, the set of candidate robot configurations are determined by exploiting pre-computed body movement primitives, obtained with fixed feet.

The proposed approach has been implemented using a NAO humanoid robot, equipped with a depth sensor. In order to navigate in complex scenarios the robot is able to automatically switch from observing the *POI* to an environment exploration behavior. A footstep planner [15] is used to generate collision free walking paths by using the 2D map of the environment. Results are presented in real indoor environments containing obstacles that require the robot to lean, so that it can observe the target *POI* (Fig. 1).

The paper is organized as follows. Section II discusses previous works about Next Best View planning for humanoid

[1]Authors are with the Department of Engineering and Architecture, University of Parma, Italy, Robotics and Intelligent Machines Laboratory, {riccardo.monica, jacopo.aleotti}@unipr.it, davide.piccinini2@studenti.unipr.it

robots. In section III, the proposed approach is described. Experimental results are reported in section IV. Finally, section V concludes the work.

II. RELATED WORK

The work in [4] is the closest to ours as it also uses a lookup table, called inverse reachability map, to efficiently find NAO robot poses given a desired view point. However, there are some significant differences. First, we are interested in solving a Next Best View problem, given a point of interest, in an environment which is initially unknown, while the goal of [4] is to observe a set of relevant areas, assuming that the humanoid robot has a 3D map of the environment. Second, in [4] experiments have been carried out in simulation. In [5] a method was presented for exploration of known areas including articulated objects that need to be manipulated to inspect obstructed regions. Evaluation has been performed in simulation.

Several works have been conducted for Next Best View planning using the HRP-2 humanoid robot [6]–[11]. In [10], [11], Saidi et al. proposed a visual attention framework for object search in real environments with occlusions, using a walking pattern generator. However, the humanoid robot could not perform full body movements to observe the object, as the robot could only change the pan and tilt parameters of the camera. Moreover, experiments were aimed at a full exploration of an unknown environment, without a target object. The same Next Best View framework was extended in [6]–[9] by solving an optimization problem to support motion planning of whole body robot postures. However, the approach was evaluated in simulated environments without occlusions, as the target object was placed on a table in the middle of an empty room.

Other research studies that carried out experiments in real environments [12]–[14] share the same limitation of placing objects in empty environments. In particular, Andreopoulos et al. [12] proposed an active system to perform 3D object localization for the Honda's humanoid robot which also simplifies the motion planning problem, by using a bounding cylinder for robot collision detection. In [13], [14] an approach was presented to reconstruct the 3D shape of an object from monocular images obtained by a NAO humanoid robot walking along circular trajectories.

Finally, in [16]–[20] wheeled humanoid robots with an anthropomorphic upper torso have been adopted for active perception or dense 3D reconstruction. More specifically, in [16] a PR2 robot was exploited for Next Best View planning in a tabletop scenario using a probabilistic framework. In [17], [18] humanoid navigation planning was investigated to improve perception capabilities. Grotz et al. [19] proposed a method for view selection based on gaze stabilization to facilitate perception during locomotion.

III. METHOD

The goal of the robot is to observe a region of the environment around a given primary Point of Interest $(x_{poi}, y_{poi}, z_{poi})$, named *POI*, and to perform 3D reconstruction of the objects contained in that region. The humanoid robot is equipped with a depth camera on its head. Kinect Fusion (KinFu, from PCL library) is used to merge observations into a volumetric 3D map of the environment M_{3D}, which is based on the Truncated Signed Distance Function (TSDF). Each voxel of M_{3D} has size e_{3D} and it may be occupied, empty or unknown. Preliminarily, a set of body movement primitives $\mathcal{J}$ for the humanoid robot is generated offline (Section III-A). Each movement primitive $J_l \in \mathcal{J}$ is a trajectory in joint space that starts from a standing posture, and that keeps the humanoid feet fixed on the ground.

Algorithm 1 Overview of the proposed approach

Input: $\mathcal{J} = \{J_l\}$: set of movement primitives
Input: POI: Point Of Interest
Output: M_{3D}: 3D reconstruction (Kinect Fusion TSDF)

1: $M_{3D} \leftarrow$ InitialScan()
2: $behavior \leftarrow$ TARGET
3: **loop**
4: $M_{2D} \leftarrow$ Generate2DMap (M_{3D})
5: **if** $behavior =$ TARGET **then**
6: $\{(V_i, p_i, J_{l(i)})\} \leftarrow$ GenPOIViewPoses $(POI, \mathcal{J})$
7: **else if** $behavior =$ EXPLORATION **then**
8: $\{(V_i, p_i, J_{l(i)})\} \leftarrow$ GenExplorationViewPoses (M_{2D}, M_{3D})
9: **end if**
10: $\{(V_i, p_i, J_{l(i)})\} \leftarrow$ FilterPoses $(\{(V_i, p_i, J_{l(i)})\})$
11: $\{g_i\} \leftarrow$ EvaluatePoses $(\{V_i\}, M_{3D}, behavior)$
12: **for all** $(V_i, p_i, J_{l(i)}, g_i)$ **order by** g_i **decreasing do**
13: **if** $g_i < g_{th}$ **then**
14: **if** $behavior =$ EXPLORATION **then**
15: **exit**
16: **end if**
17: $behavior \leftarrow$ EXPLORATION
18: **break**
19: **end if**
20: **if** CheckMovPrimitiveCollisions $(p_i, J_{l(i)}, M_{3D})$ **and** PlanWalkTo (p_i, M_{2D}) **then**
21: WalkTo (p_i, M_{2D})
22: ExecuteMovPrimitive $(J_{l(i)})$
23: $M_{3D} \leftarrow$ Observation (M_{3D})
24: $behavior \leftarrow$ TARGET
25: **break**
26: **end if**
27: **end for**
28: **end loop**

A detailed overview of the approach is reported in Algorithm 1. The humanoid can switch among two behaviors: TARGET and EXPLORATION. In the TARGET behavior, the robot generates view poses oriented towards the primary *POI* (Section III-B), while in the EXPLORATION behavior the robot explores other unknown parts of the environment, attempting to find new paths that may enable observation of the *POI* (Section III-C).

As the environment is initially unknown, a first scan is performed from the starting configuration by moving only the robot head (line 1). The robot begins the task in the TARGET behavior (line 2). Afterward, the proposed approach operates iteratively. At each iteration a feasible Next Best View is computed, according to the current robot behavior, and then the humanoid moves to a configuration in world space that

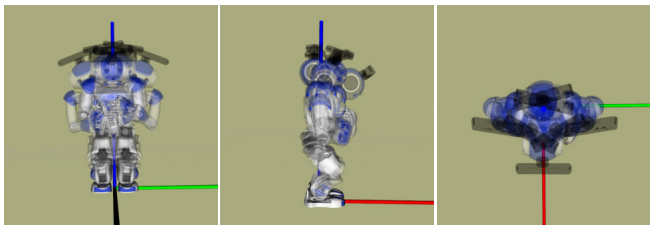

Fig. 2. Final configuration of 7 movement primitives in $\mathcal{J}$. Frontal view (left), side view (center), top view (right).

enables the sensor to reach that view. The Next Best View is the *b*-th feasible 3D pose of the sensor V_b, selected among a set of candidate view poses $\{V_i\}$, that is expected to observe the maximum amount of unknown space.

In particular, a 2D binary occupancy map M_{2D} is updated from the 3D reconstruction (line 4), as described in Section III-C. Then, depending on the current robot behavior, a number of view poses are generated (lines 5-9), as explained in Sections III-B and III-C. Each candidate view pose V_i is associated to a pose of the midpoint between the feet $p_i = (x_i, y_i, \theta_i)$, and to a movement primitive $J_{l(i)}$, so that by standing in p_i, and then by executing the movement primitive $J_{l(i)}$, the humanoid can move the depth sensor to view pose V_i. A preliminary reachability check discards candidate view poses in which the humanoid feet are in occupied space of M_{2D} (line 10). After that (Section III-D), view poses are evaluated by taking into account the full 3D map (line 11). View poses are checked in decreasing order of expected gain g_i (lines 12-27), until a feasible one is found (line 20), or the gain is lower than a threshold g_{th} (line 13). If a pose is reachable (Section III-E), the robot walks to p_i (line 21), and then it executes movement primitive $J_{l(i)}$ (line 22). Finally, the robot head is moved along a fixed trajectory (up and down) and Kinect Fusion is activated to get new data (line 23). Kinect Fusion is disabled during walking, as localization and sensing could be affected by body swing.

If $g_i < g_{th}$ (line 13) while the TARGET behavior is active, the robot switches to the EXPLORATION behavior (line 17). When an exploration view pose is reached successfully, the behavior is reset to TARGET (line 24), to attempt other observations of the primary *POI*. If gain g_i becomes lower than g_{th} again, while the EXPLORATION behavior is active (line 14), the algorithm terminates (line 15) as there are no view poses that provide enough gain. The task also terminates when all the view poses are not reachable (omitted in the algorithm).

A. Movement primitives generation

The set of movement primitives $\mathcal{J}$ is obtained by sampling the NAO robot configuration space. Each movement primitive $J_l \in \mathcal{J}$ is a sequence of joint angle configurations $\{j_{l,1}, \ldots, j_{l,K_l}\}$ that represents a trajectory, with K_l samples and sampling time T_{sampling}. To generate the movement primitives, the built-in whole-body motion controller of the NAO robot was exploited. The controller can synthesize stable whole-body trajectories in Cartesian control mode, starting from a standing configuration, given as input the final orientation of the head with respect to the feet. Head orientation is defined by Euler angles (α, β, γ) expressing rotations around the x, y, z axes of the robot reference frame, respectively (Fig. 2). The generated trajectories involve torso and head motion, with robot feet fixed on the ground.

The set of all admissible Euler angles $\mathcal{E}$, i.e., all triplets (α, β, γ) for which a stable trajectory exists, can not be easily determined, as it depends on the robot dynamics and on the internal parameters of the whole-body motion controller. Empirically, it was found that a suitable subset can be modeled by all triplets (α, β, γ) that satisfy:

$$\frac{|\alpha - \alpha_c|}{\alpha_{max}} + \frac{|\beta - \beta_c|}{\beta_{max}} + \frac{|\gamma - \gamma_c|}{\gamma_{max}} \leq 1 \qquad (1)$$

where $(\alpha_{max}, \beta_{max}, \gamma_{max})$ are angular intervals half widths, and $(\alpha_c, \beta_c, \gamma_c)$ are mean angle values. Therefore, to generate a set of admissible Euler angles $\mathcal{E}_{max} = \{(\alpha_l, \beta_l, \gamma_l)\} \subset \mathcal{E}$ for the whole-body motion controller, values α_l and β_l were sampled in $[\alpha_c - \alpha_{max}, \alpha_c + \alpha_{max}]$ and $[\beta_c - \beta_{max}, \beta_c + \beta_{max}]$, with sampling rate s_α and s_β, respectively, while angle values γ_l were computed as

$$\gamma_l = \pm \left(1 - \frac{|\alpha_l - \alpha_c|}{\alpha_{max}} - \frac{|\beta_l - \beta_c|}{\beta_{max}}\right) \gamma_{max} + \gamma_c \qquad (2)$$

An initial set $\mathcal{J}$ of movement primitives was then generated by running the NAO whole-body motion controller for each orientation of the head in $\mathcal{E}_{max}$. Both arms were kept close to the torso, to reduce the risk of collisions with obstacles when the task is performed. All movement primitives were tested for feasibility and stability by executing them on the real robot, starting from a standing posture. Each movement primitive in the initial set $J_l = \{j_{l,1}, \ldots, j_{l,K_l-k}, \ldots, j_{l,K_l}\} \in \mathcal{J}$ was finally used to generate additional shorter movement primitives, which were added to $\mathcal{J}$ as well, by extracting sub-trajectories $\{j_{l,1}, \ldots, j_{l,K_l-k}\}$ for $k \in \{0, S, 2S, 3S, \ldots\}$, where S is a constant number.

B. POI-oriented view pose generation

This section describes how a set of view poses V_i pointing towards the primary *POI* is generated, when the robot is in the TARGET behavior (function GenPOIViewPoses in Algorithm 1). Each view pose is associated to a movement primitive $J_{l(i)}$ and to a configuration of the feet, which constitute a triplet $\left(V_i, p_i, J_{l(i)}\right)$. The world reference frame W is on the ground plane with vertical z-axis (Fig. 3). Let O be a reference frame with the same orientation of W, but centered on the projection of the *POI* on the ground, so that the primary *POI* with respect to O is ${}^{O}POI = (0, 0, z_{poi})$. Let F_l be the feet reference frame, located at the midpoint of the robot feet, with vertical z-axis and forward x-axis. Let also C_l be the sensor reference frame, with z-axis along the camera principal axis. Transformation matrix ${}^{F_l}_{C_l}T$, from C_l to F_l, is computed using forward kinematics, given the last configuration j_{l,K_l} of movement primitive J_l in joint space. Matrix ${}^{O}_{F_l}T$, from F_l to O is the unknown transformation to

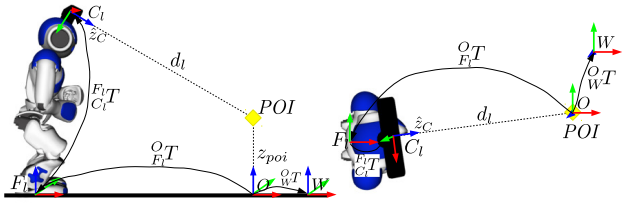

Fig. 3. Reference frames defined in Section III-B, side view (left) and top view (right).

be found, which is given by a rotation θ_l around the z-axis, followed by a translation $(x_l, y_l, 0)$, as both the robot feet and the *POI* projection are on the ground plane.

The camera points toward the primary *POI* only if the *POI* belongs to the camera z-axis $\hat{z}_{C_l}$, that is, if there exists a distance d_l, within sensor range $r_{min}<d_l<r_{max}$, so that:

$$ {}^{C_l}\vec{0}_{C_l} + d_l \, {}^{C_l}\hat{z}_{C_l} = {}^{C_l}POI \tag{3} $$

where ${}^{C_l}\vec{0}_{C_l} = (0,0,0,1)$ is the camera origin in camera coordinates, ${}^{C_l}POI$ is the primary *POI* in camera coordinates, and ${}^{C_l}\hat{z}_{C_l} = (0,0,1,0)$ is the camera z-axis in camera coordinates. In reference frame O, (3) becomes:

$$ {}^{O}_{F_l}T \, {}^{F_l}_{C_l}T \, {}^{C_l}\vec{0}_{C_l} + d_l \, {}^{O}_{F_l}T \, {}^{F_l}_{C_l}T \, {}^{C_l}\hat{z}_{C_l} = {}^{O}POI \tag{4} $$

Equation (4) is an under-determined system, as it depends on four variables d_l, x_l, y_l and θ_l. Transformation ${}^{O}_{F_l}T$ does not affect the z coordinate, hence d_l can be computed by solving the third row of (4), which can be expressed as:

$$ \left({}^{F_l}_{C_l}T \, {}^{C_l}\vec{0}_{C_l} \right) \cdot {}^{O}\hat{z}_O + d_l \left({}^{F_l}_{C_l}T \, {}^{C_l}\hat{z}_{C_l} \right) \cdot {}^{O}\hat{z}_O = z_{poi} \tag{5} $$

If d_l is not within sensor range $r_{min}<d_l<r_{max}$, movement primitive J_l does not generate any valid view pose.

Equation (4) constrains the humanoid feet on a circumference, centered on $\vec{0}_O$, from which the robot can observe the primary *POI* using movement primitive J_l, as shown in Fig. 4. Then, to obtain a number of candidate view poses with circular symmetry, values $\theta_{l,h}$ are uniformly sampled from θ_l, with sampling rate s_θ. For each $\theta_{l,h}$, feet poses ${}^{O}p_{l,h} = (x_{l,h}, y_{l,h}, \theta_{l,h})$ are computed, by solving the first two equations of (4), and translated into world coordinates ${}^{W}p_{l,h} = (x_{l,h} + x_{poi}, y_{l,h} + y_{poi}, \theta_{l,h})$. Finally, for each value of l and h, a triplet $\left(V_i, p_i, J_{l(i)}\right)$ is generated as follows:

$$ V_i = {}^{W}_{O}T \, {}^{O}_{F_l}T \, {}^{F_l}_{C_l}T_l, \quad p_i = {}^{W}p_{l,h}, \quad J_{l(i)} = J_l \tag{6} $$

C. Exploration-oriented view pose generation

When the EXPLORATION behavior is active, a set of triplets $\left\{\left(V_i, p_i, J_{l(i)}\right)\right\}$ is generated so that candidate view poses are oriented towards the unexplored regions of the environment (function GenExplorationViewPoses in Algorithm 1). The procedure is similar to the one explained in Section III-B, the only differences being that view poses point toward secondary *POIs* that are automatically placed near the frontier of unknown space, and that the orientation of the humanoid feet on the ground (angle θ_l) is chosen so

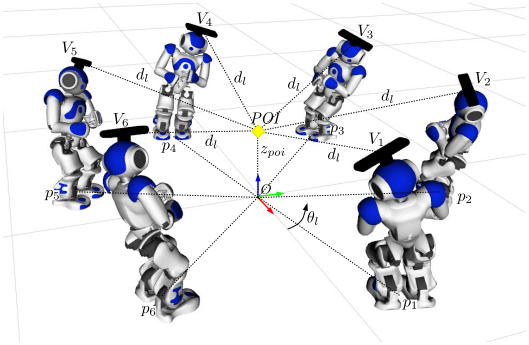

Fig. 4. Robot configurations generated to observe the primary *POI*, in the TARGET behavior, for a single movement primitive J_l and $s_\theta=\pi/3$.

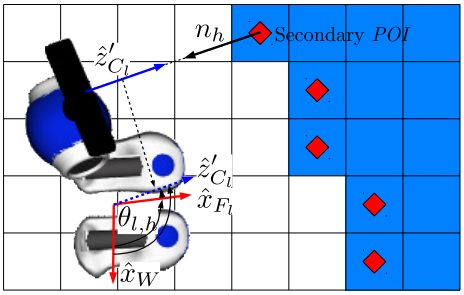

Fig. 5. A robot configuration (top view), in the EXPLORATION behavior, where the humanoid observes one of the secondary *POIs* (red diamonds) on the frontier of unknown space (light blue cells) in map M_{U}. For clarity, only the robot head and feet are displayed.

that the robot observes along the orthogonal direction to the frontier of the unknown space (Fig. 5).

First, two 2D binary maps are generated from M_{3D} at lower resolution: a 2D occupancy map M_{2D}, and a 2D map of the unknown regions M_{U}, both with cell size $e_{\mathrm{d}}>e_{\mathrm{3D}}$. A 2D cell in M_{U} is set to 1, if there is at least one unknown voxel in M_{3D} with z-coordinate in $[z_{min}, z_{max}]$, whose projection onto the ground plane ${}^{W}z=0$ lies in the cell, and no voxel in the same height interval is occupied, otherwise the cell is set to 0. Map M_{U} serves as an approximate representation of the cells that may be empty, if observed. Conversely, a cell in M_{2D} is set to 1, if at least one unknown or occupied voxel in M_{3D} with z-coordinate in $[z_{min}, z_{max}]$ projects onto it, otherwise the cell is set to 0. Map M_{2D} serves as an approximate representation of the cells where the robot may collide with the environment. Parameters z_{min} and z_{max} are chosen so that the ground plane and objects above the robot maximum height do not affect the two 2D maps.

Candidate view poses are directed toward frontier cells (u_h, v_h) in M_{U}, i.e., unknown cells with an empty neighbor:

$$ \begin{cases} M_{\mathrm{U}}(u_h, v_h) = 1, \\ \exists\,(u,v) \in N_4(u_h, v_h) \mid M_{\mathrm{2D}}(u,v) = 0 \end{cases} \tag{7} $$

where $N_4(u_h, v_h)$ is the 4-neighborhood of (u_h, v_h). In particular, for each frontier pixel (u_h, v_h), a secondary *POI* $o_h=(u_h\, e_{\mathrm{d}}, v_h\, e_{\mathrm{d}}, z_{expl})$ is generated, where z_{expl} is a small height from the ground. Then, an approach similar to the one described in Section III-B is exploited to generate a view

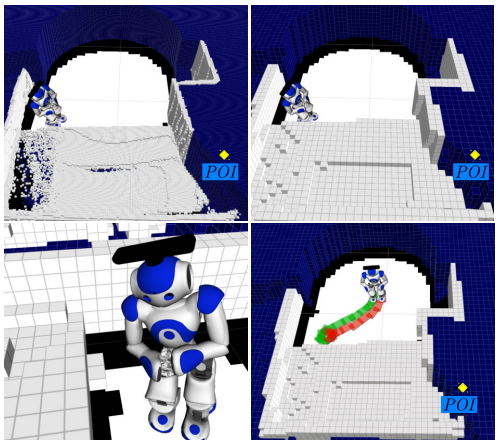

Fig. 6. Planning of the Next Best View pose shown in Fig. 1. Unknown and occupied voxels are in blue and grey, respectively. Occupied cells on the ground in M^e_{2D} are in black. 3D reconstruction M_{3D} (top left), downsampled 3D map M'_{3D} (top right). A closeup image of the collision free whole body robot configuration (bottom left), and planned walking path (red and green steps, bottom right). Yellow diamonds mark the primary *POI*.

pose for each movement primitive J_l, and each secondary *POI* o_h. To this purpose, the 2D outer-pointing normal vector of the frontier $n_h{=}(x_{n,h}, y_{n,h})$ can be estimated, in world coordinates W, as (normalization omitted):

$$n_h = \sum_{(u,v)\in N_{12}} -\frac{(u,v)-(u_h,v_h)}{|(u,v)-(u_h,v_h)|}\left(M_{\mathrm{U}}(u,v)-1/2\right) \quad (8)$$

where N_{12} is the set of all cells within distance $2e_{\mathrm{d}}$ from (u_h, v_h). Equation (8) estimates a 2D normal vector that points toward empty cells and away from unknown cells. Then, angle $\theta_{l,h}$ is generated so that the projection $\hat{z}'_{C_l}$ of the camera z-axis $\hat{z}_{C_l}$ onto the ground is oriented as $-n_h$. Angle $\theta_{l,h}$ is computed by solving the following equation:

$$\angle\hat{x}_{F_l}\hat{z}'_{C_l} = \angle\hat{z}'_{C_l}\hat{x}_W - \theta_{l,h} \quad (9)$$

where $\angle\hat{x}_{F_l}\hat{z}'_{C_l}$ is the angle between feet x-axis and $\hat{z}'_{C_l}$, which is known from J_i (Fig. 5), and angle $\angle\hat{z}'_{C_l}\hat{x}_W$ is known as $\hat{z}'_{C_l} = -n_h$.

D. View pose evaluation

Each candidate view pose V_i is evaluated by computing a gain value g_i using ray-casting (function EvaluatePoses in Algorithm 1). The gain of a view pose is a score that predicts the amount of unknown space visible from that view point. The Next Best View is selected as the pose V_b with $b{=}\mathrm{argmax}_i\,(g_i)$. Previous work [21] is exploited to perform GPU-accelerated ray-casting in the KinFu TSDF volume. In particular, a virtual depth image Δ_i is generated from each candidate view pose V_i, and a ray is cast from the simulated view origin through each pixel (u, v) in Δ_i. The TSDF volume stores two values for each voxel $c{=}(x_c, y_c, z_c)$: a weight w_c, that is 0 when the voxel is unknown, and a TSDF value f_c, with $f_c{<}0$ in occupied space and $f_c{\geq}0$ otherwise.

In the EXPLORATION behavior a ray passing through pixel (u, v) stops when it encounters any non-empty voxel $c_{u,v}$. In the TARGET behavior, relevant voxels are only those inside a sphere with radius r_{poi} centered on the *POI*. Therefore, the ray stops in $c_{u,v}$ only if it intersects an occupied cell along the ray, or if it intersects an unknown voxel inside the sphere, i.e., if the following condition holds:

$$(w_c > 0 \wedge f_c < 0) \;\vee\; (w_c = 0 \wedge |c - POI| < r_{poi}) \quad (10)$$

Gain g_i is computed as the weighted sum

$$g_i = \sum_{(u,v)} \delta^2_{u,v}\, U_o\left(c_{u,v}\right) \quad (11)$$

where $\delta_{u,v}$ is the distance the ray travels that penalizes voxels near the sensor [1], $U_o\left(c_{u,v}\right){=}1$ if the voxel $c_{u,v}$ where the ray stops is unknown, whereas $U_o\left(c_{u,v}\right){=}0$ if the ray hits an occupied voxel or if it only traverses empty voxels.

TABLE I
EXPERIMENTAL PARAMETERS.

Name	Value	Name	Value
r_{min}	0.5 m	α_{max}	$\pi/6$ rad
r_{max}	2.0 m	β_{max}	$\pi/6$ rad
e_{d}	5 cm	γ_{max}	$\pi/4$ rad
e_{3D}	1.1 cm	α_c	0 rad
z_{expl}	0.3 m	β_c	$\pi/9$ rad
z_{min}	0.15 m	γ_c	0 rad
z_{max}	0.85 m	s_α	$\pi/18$ rad
r_e	2 cells	s_β	$\pi/18$ rad
g_{th}	5500	s_θ	$\pi/96$ rad
T_{sampling}	50 ms	S	8

E. Motion planning and collisions checking

Whole body collision checking and walk planning are two essential features of the proposed Next Best View system, as the robot must safely move close to obstacles and peer around them. In Algorithm 1, a preliminary 2D collision test is performed (in function FilterPoses) to discard unfeasible view poses before their evaluation. An enlarged occupancy map M^e_{2D} (Fig. 6) is computed from M_{2D}, with obstacles expanded by r_e cells to account for inaccuracies in robot localization and body swing. View poses are discarded if goal feet pose p_i is in an occupied cell in M^e_{2D}, or if the cell where feet position is currently lying and the cell where p_i lies can not be connected by any sequence of adjacent free cells in M^e_{2D}. A more elaborate whole body 3D collision test is also performed (function CheckMovPrimitiveCollisions in Algorithm 1). A 3D map M'_{3D} is generated from M_{3D}, at lower resolution e_{d}, by evaluating only voxels with z-coordinate in $[z_{min}, z_{max}]$. A voxel in M'_{3D} is set as occupied if it contains at least one unknown or occupied voxel in M_{3D}. After scoring, candidate view poses are checked for reachability, in decreasing order of expected gain. To determine if the current view pose V_i is reachable, the corresponding movement primitive $J_{l(i)}$ is first tested for collision. In particular, the 3D robot model is placed with its feet in p_i and it is checked for collision in M'_{3D} for all joint angle configurations $j_{l,k} \in J_{l(i)}$, using the MoveIt! motion planning framework. Then, robot walking paths are planned in M^e_{2D} toward p_i using the footstep planner in [15] (Fig. 6). If planning succeeds, a sequence of steps is generated that

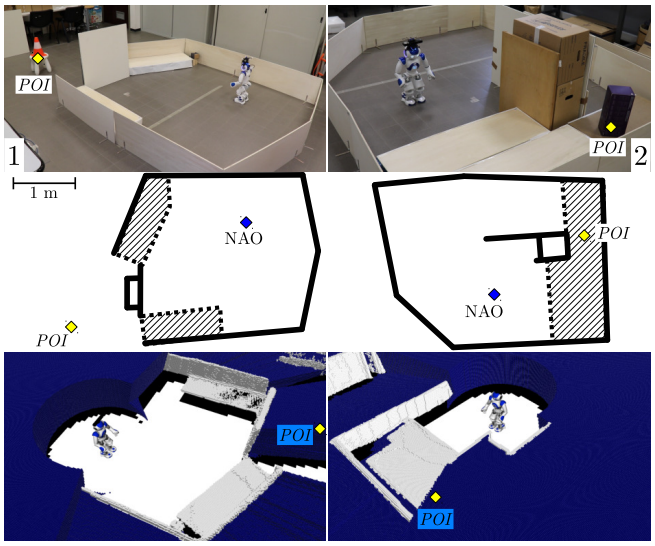

Fig. 7. Initial humanoid configuration in the two experimental scenarios (top). Illustrative map of each scenario (middle). Thick lines represent high-height obstacles, while dashed lines represents low-height obstacles. The initial pose of the robot is marked by a blue diamond. The yellow diamond marks the primary *POI*. Reconstruction M_{3D} after the first scan (bottom).

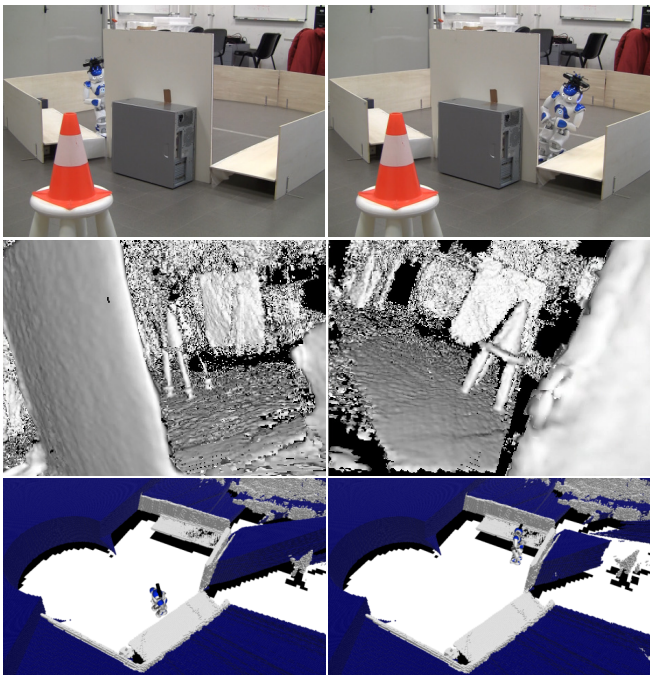

Fig. 8. The first two NBVs of the proposed approach, in the TARGET behavior of Scenario 1 (left and right columns). The robot peers around the panel to get a favourable view of the target object (orange cone). Image from an external camera (top), Kinect Fusion depth image (center), 3D reconstruction M_{3D} after the observation (bottom).

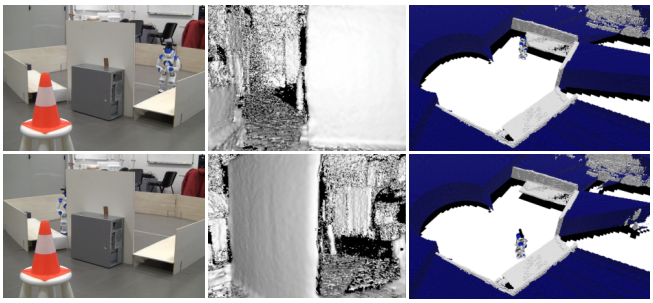

Fig. 9. The first two NBVs of the standard approach, in the TARGET behavior of Scenario 1 (top and bottom row). The robot achieves a scarce view of the target object, due to the more limited movement range. Image from an external camera (left), Kinect Fusion depth image (center), 3D reconstruction M_{3D} after the observation (right).

does not collide with the expanded 2D map M^e_{2D}, given the current and a goal feet pose. The footstep navigation package is used to execute the planned steps.

IV. EXPERIMENTAL EVALUATION

A. Experimental setup

The experimental setup comprises a NAO robot, with an ASUS Xtion Pro depth sensor on its head. An external PC is used for data processing, featuring an Intel Core i7-7700 CPU (3.60 GHz, 16 GB RAM), and an NVidia GeForce GTX 1070 GPU (8 GB RAM). The software is based on the Robot Operating System (ROS) framework. Communication between the robot and the PC exploits a dedicated IEEE 802.11n wireless network, resulting in an average frame rate of 8 depth frames per second. An OptiTrack Motion Capture system, with twelve Prime 13 cameras, is used for robot localization. The motion capture system runs at 120 hz, with sub-millimetric accuracy. Markers were attached to the depth sensor for real time tracking with respect to the world reference frame W. Extrinsic calibration between the set of markers and the sensor was obtained using an automatic optimization algorithm. The tracked sensor configuration, obtained from motion capture, was exploited as input for the Kinect Fusion algorithm for 3D reconstruction (egomotion tracking was disabled). Calibration between NAO head and the set of marker was obtained by taking manual measurements. The estimated accuracy for robot localization is about 1 cm. The footstep planner was configured with a tolerance of 2 cm in position, and 0.1 rad in orientation. Parameters values for the experimental evaluation are reported in Table I. When the EXPLORATION behavior is active, a more limited range of motion was used to generate movement primitive set $\mathcal{J}$, as the robot is not expected to peer around obstacles.

B. Experiments

Experiments were carried out in two scenarios, displayed in Fig. 7. In both scenarios the humanoid was confined in an arena of size 4 m $\times$ 3 m, delimited by wooden walls, with height 40 cm. In each scenario 4 trials were performed, with the same initial robot configuration. Two trials used the proposed method, with full range of movement primitives $\mathcal{J}$. Conversely, the other two trials used a standard approach, with only pan and tilt motions of the head, as in [10], [11]. Figures 8, 9, 10 show images from the experiments. In Scenario 1, the primary *POI*, with radius r_{poi}=0.5 m, was located on an orange cone, on top of a small table outside the arena. The outer walls of the arena had two openings to allow observation of the *POI*. However, a large panel of 85 cm

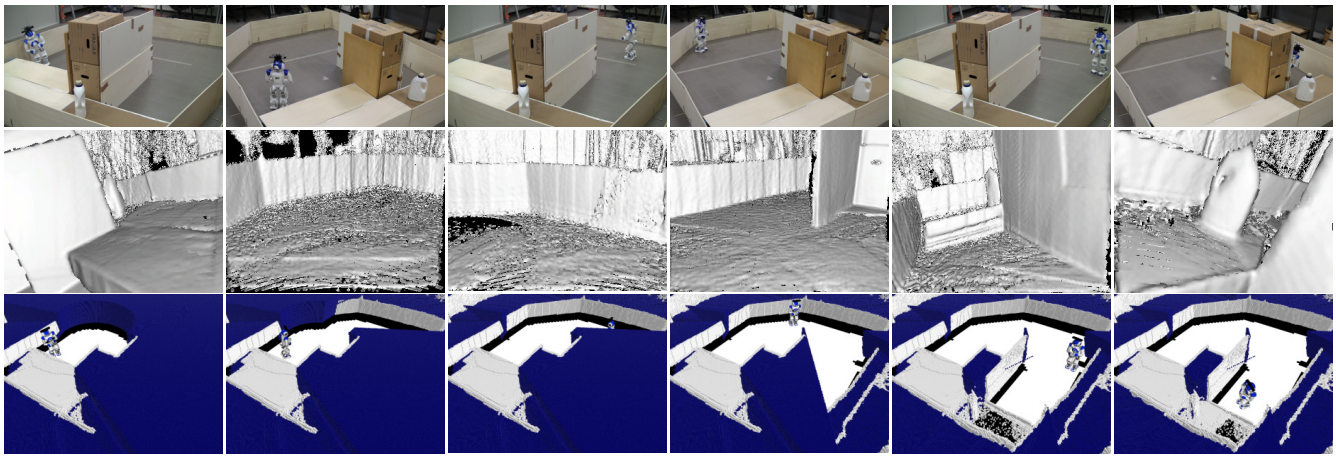

Fig. 10. The six NBVs in Scenario 2, trial 1, with the proposed method (from left to right). Image from an external camera (top), Kinect Fusion depth image (middle), 3D reconstruction M_{3D} after the NBV (bottom). In the first and sixth NBVs the robot leans to achieve a better view of the *POI*.

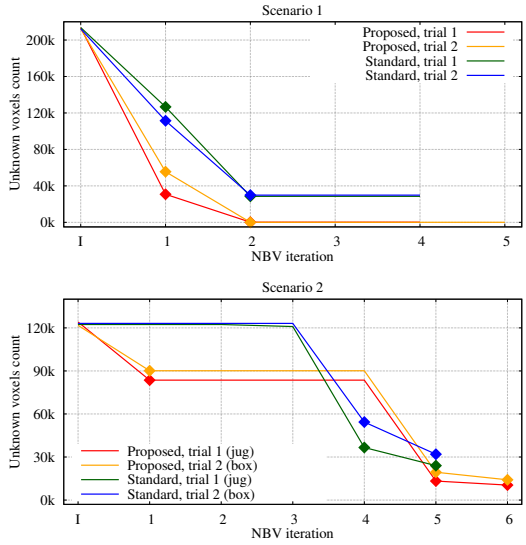

Fig. 11. Number of unknown voxels in the region of interest, centered at the *POI*, in Scenario 1 and Scenario 2, after each NBV (I represents the initial scan). Diamonds mark NBVs where the TARGET behavior was active.

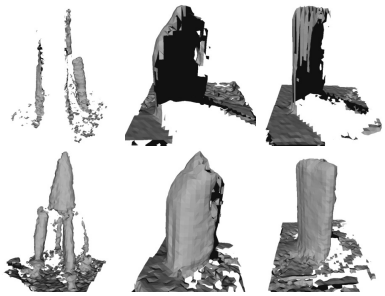

Fig. 12. Final 3D mesh reconstruction of the object at the *POI*, for the standard method (top) and the proposed method (bottom). From left to right: trial 1 of Scenario 1, trial 1 of Scenario 2, trial 2 of Scenario 2.

height was placed in between the two openings to partially occlude the primary *POI*, thus making the NBV task quite complex. Moreover, some low-height obstacles prevented the robot to pass through the two openings to get closer to the *POI*. In all trials the robot performed the first two NBVs in the TARGET behavior, trying to look at the *POI* from the two openings. The first two NBVs in trial 1 are shown for the proposed and the standard approach in Fig. 8 and Fig. 9, respectively. It can be noticed that in the proposed approach, the robot was able to peer around the panel to get a better view of the cone. Conversely, in the standard approach the target object was barely visible. After the first two NBVs, the robot switched to the EXPLORATION behavior and executed two or three additional NBVs to observe the unknown space, attempting to find alternative configurations to look at the primary *POI*. However, since there was no other way to look at the target object, the task concluded. The number of unknown voxels in the spherical region of interest around the *POI* (Fig. 11) decreases in the first two NBVs, for both methods, and then it remains constant when the robot switches to the EXPLORATION behavior, as the *POI* is not observed. The number of unknown voxels decreases more rapidly in the proposed approach, and it reaches a lower minimum value, because the humanoid is able to plan better view poses. In Scenario 2, the primary *POI*, with radius r_{poi}=0.4 m, was located inside the arena, on a box (or a jug). Fig. 10 shows the first trial of Scenario 2 with the proposed method, the second trial (reported in the accompanying video) had a similar behavior. In the first NBV the robot leaned to look at the target object from one side of the arena. Then, the robot switched to the EXPLORATION behavior for the next three views. Finally, the humanoid was able to reach the other side of the arena, by walking around some obstacles, and it switched back to the TARGET behavior. At this point the robot performed two more NBVs, and the task concluded. Conversely, the standard approach was unable to plan a view to observe the *POI* at the beginning of the task and, therefore, the robot immediately switched to the EXPLORATION behavior. After reaching the other side of the arena, the humanoid observed the target

TABLE II

AVERAGE TIMES (SECONDS) AND OTHER DATA IN SCENARIO 1.

	Method	
	Proposed	Standard
InitialScan	80.1 ± 0.4	80.5 ± 0.1
GenPOIViewPoses	< 0.001	< 0.001
GenExplorationViewPoses	< 0.001	< 0.001
FilterPoses	< 0.001	< 0.001
EvaluatePoses	1.91 ± 0.22	1.56 ± 0.22
CheckMovPrimitiveCollisions	0.06 ± 0.06	0.03 ± 0.01
WalkTo	86.6 ± 30.1	93.4 ± 22.3
ExecuteMovPrimitive+Observation	14.2 ± 2.1	12.5 ± 0.2
Movement primitives	321	47
Generated views (TARGET)	53184	6336
Evaluated views (TARGET)	577 ± 39	38.0 ± 1.2
Generated views (EXPLOR.)	573 ± 462	688 ± 434
Evaluated views (EXPLOR.)	437 ± 328	520 ± 259

object in the fourth and fifth view. The task concluded with one less NBV, and the final number of voxels that remained unknown was higher than in the proposed approach, as the object was not observed from one side. Fig. 12 shows the final 3D mesh reconstruction of the objects within the region of interest, obtained by Kinect Fusion. The image confirms the numerical results, as it shows that the proposed approach achieves a more complete reconstruction of the objects. Table II reports average execution times and standard deviations of the main functions in Algorithm 1, for Scenario 1. The initial scan takes about 80 seconds as the movement of the head includes both horizontal and vertical motions. Table II also reports the number of movement primitives in $\mathcal{J}$, the number of generated view poses in each behavior, and the number of evaluated view poses (after FilterPoses). The number of generated TARGET view poses is constant as the *POI* is fixed. Conversely, the average number of EXPLORATION view poses, generated from the current map M_{2D}, has a large standard deviation. The preliminary 2D collision test in FilterPoses is fast, and it filters out more than 98% of TARGET view poses and about 30% of EXPLORATION view poses. Functions CheckMovPrimitiveCollisions and ExecuteMovPrimitive take more time in the proposed method, as the trajectories are longer. The robot walking phase (WalkTo) is the most time consuming, i.e., on average, each robot walking movement takes about 1.5 minutes.

V. CONCLUSIONS

A method for humanoid Next Best View planning was presented that enables the robot to perform complex whole body motions to observe regions of interest under occlusions. The planned Next Best Views of the range sensor allow dense 3D reconstruction of the environment. The approach exploits body movement primitives and it was evaluated in real experiments. Results show that, being able to peer around obstacles, the robot achieves a more complete reconstruction of the target objects than a conventional NBV algorithm.

REFERENCES

[1] R. Monica and J. Aleotti, "Surfel-Based Next Best View Planning," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3324–3331, Oct 2018.

[2] C. Connolly, "The Determination of Next Best Views," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2, 1985, pp. 432–435.

[3] R. Monica and J. Aleotti, "Contour-based next-best view planning from point cloud segmentation of unknown objects," *Autonomous Robots*, vol. 42, no. 2, pp. 443–458, 2018.

[4] S. Oßwald, P. Karkowski, and M. Bennewitz, "Efficient coverage of 3D environments with humanoid robots using inverse reachability maps," in *IEEE-RAS International Conference on Humanoid Robotics*, Nov 2017, pp. 151–157.

[5] S. Oßwald and M. Bennewitz, "GPU-Accelerated Next-Best-View Exploration of Articulated Scenes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[6] T. Foissotte, O. Stasse, A. Escande, and A. Kheddar, "A next-best-view algorithm for autonomous 3d object modeling by a humanoid robot," in *IEEE-RAS International Conference on Humanoid Robots*, Dec 2008, pp. 333–338.

[7] O. Stasse, D. Larlus, B. Lagarde, A. Escande, F. Saidi, A. Kheddar, K. Yokoi, and F. Jurie, "Towards autonomous object reconstruction for visual search by the humanoid robot HRP-2," in *IEEE-RAS Intl Conference on Humanoid Robots*, Nov 2007, pp. 151–158.

[8] T. Foissotte, O. Stasse, A. Escande, P. Wieber, and A. Kheddar, "A two-steps next-best-view algorithm for autonomous 3D object modeling by a humanoid robot," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2009, pp. 1159–1164.

[9] T. Foissotte, O. Stasse, P. Wieber, and A. Kheddar, "Using NEWUOA to drive the autonomous visual modeling of an object by a humanoid robot," in *International Conference on Information and Automation*, June 2009, pp. 78–83.

[10] F. Saidi, O. Stasse, and K. Yokoi, "A Visual Attention Framework for Search Behavior by a Humanoid Robot," in *IEEE-RAS International Conference on Humanoid Robots*, Dec 2006, pp. 346–351.

[11] F. Saidi, O. Stasse, K. Yokoi, and F. Kanehirot, "Online object search with a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2007, pp. 1677–1682.

[12] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J. K. Tsotsos, and E. Korner, "Active 3D Object Localization Using a Humanoid Robot," *IEEE Trans. on Robotics*, vol. 27, no. 1, pp. 47–64, 2011.

[13] J. Delfin, O. Mar, J. Hayet, M. Castelán, and G. Arechavaleta, "An active strategy for the simultaneous localization and reconstruction of a 3D object from a humanoid platform," in *IEEE-RAS International Conference on Humanoid Robots*, Nov 2012, pp. 384–389.

[14] P. A. Martnez, D. Varas, M. Castelán, M. Camacho, F. Marques, and G. Arechavaleta, "3D shape reconstruction from a humanoid generated video sequence," in *IEEE-RAS International Conference on Humanoid Robots*, Nov 2014, pp. 699–706.

[15] J. Garimort, A. Hornung, and M. Bennewitz, "Humanoid navigation with dynamic footstep plans," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2011, pp. 3982–3987.

[16] C. Potthast and G. S. Sukhatme, "A probabilistic framework for next best view estimation in a cluttered environment," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 148–164, 2014.

[17] P. Michel, C. Scheurert, J. Kuffner, N. Vahrenkamp, and R. Dillmann, "Planning for robust execution of humanoid motions using future perceptive capability," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2007, pp. 3223–3228.

[18] P. Michel, J. Chestnutt, S. Kagami, K. Nishiwaki, J. Kuffner, and T. Kanade, "Humanoid navigation planning using future perceptive capability," in *IEEE-RAS International Conference on Humanoid Robots*, Dec 2008, pp. 507–514.

[19] M. Grotz, T. Habra, R. Ronsse, and T. Asfour, "Autonomous view selection and gaze stabilization for humanoid robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 1427–1434.

[20] R. Wagner, U. Frese, and B. Bauml, "Real-time dense multi-scale workspace modeling on a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2013, pp. 5164–5171.

[21] R. Monica, J. Aleotti, and S. Caselli, "A KinFu based approach for robot spatial attention and view planning," *Robotics and Autonomous Systems*, vol. 75, Part B, pp. 627–640, 2016.