Extracting Whole-Body Affordances from Multimodal Exploration

Peter Kaiser, David Gonzalez-Aguirre, Fabian Schültje,
Júlia Borràs, Nikolaus Vahrenkamp and Tamim Asfour

***Abstract*— Humanoid robots that have to operate in cluttered and unstructured environments, such as man-made and natural disaster scenarios, require sophisticated sensorimotor capabilities. A crucial prerequisite for the successful execution of whole-body locomotion and manipulation tasks in such environments is the perception of the environment and the extraction of associated environmental affordances, i.e. the action possibilities of the robot in the environment, in order to generate whole-body locomotion and manipulation actions. We believe that such a coupling between perception and action could be a key to substantially increase the flexibility of humanoid robots.**

In this paper, we present an approach for the generation of whole-body locomotion and manipulation actions based on the affordances associated with environmental elements in the scene which are extracted via multimodal exploration. Based on the properties of detected environmental primitives and the estimated empty space in the scene, we propose methods to generate hypotheses for feasible whole-body actions while taking into account additional task constraints such as manipulability and balance. We combine visual and inertial sensing modalities by means of a novel depth model for generating segmented and categorized geometric primitives. A rule-based system is then incorporated to assign affordance hypotheses to these primitives. Finally, precomputed whole-body manipulability and stability maps are used for filtering affordances that are out of reach and for identifying the most promising locations for the action execution. We tested the developed methods in different scenes, unknown to the robot, demonstrating how reasonable the generated affordance hypotheses are.

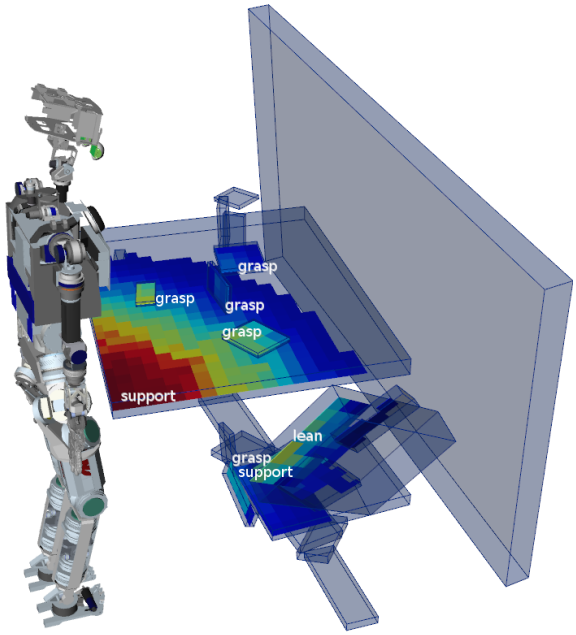

Fig. 1: The proposed methods allow the identification of affordance hypotheses from active vision and inertial sensor data.

I. Introduction

One of the most fundamental questions in robotics research is how to enable robots to autonomously interact with unknown environments. This problem has been partially addressed by numerous works that try to bridge the gap between low-level control and high-level abstract reasoning (e.g. [1]–[4]). Most of these publications focus on manipulation tasks with single robotic arms or upper body humanoids with mobile platforms. Bipedal humanoid robots add more complexity to the problem, in terms of their kinematic structure as well as in terms of the possible ways of interacting with the environment. Particularly, constraints on balance have to be satisfied, including the utilization of the environment to enhance stability.

This paper presents a first step towards enabling humanoids to interact with unknown environments. It is divided into two parts that approach the two primary challenges of this problem. First, that the environment is unknown and therefore, we need to recognize shapes to interact with. Second, we assume that a safe navigation in unstructured environments requires the ability to use walls and other objects for stabilization. The recognized shapes should therefore be used to increase the stability of the humanoid when navigating through the scene.

For the first part, we improve state-of-the-art techniques to fuse inertial and visual information to a depth model which is then segmented and categorized into geometric primitives.

The second part of our approach relies on the idea that the observed scene is only partially unknown due to the known structures of human environments. In other words, we can assume some prior knowledge on the scene to infer affordance hypotheses based on shapes, sizes or orientations of the detected primitives (see Fig. 1). For instance, we can assume that vertical, large planes are probably walls that can afford to lean on them.

The concept of affordances was first proposed by JJ. Gibson [5] in the context of ecological psychology. Since then, it has been applied to several fields of research, from cognitive science and neuropsychology to human-computer interaction and autonomous robotics. In the original psychological context, the main idea behind the concept of affordance was that perception is economical, i.e. instead of

The authors are with the Institute for Anthropomatics and Robotics, High Performance Humanoid Technologies Lab (H^2T), at the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. peter.kaiser@kit.edu

modeling the whole world, only the relevant environmental information is perceived.

In the context of autonomous robots, affordances have been used to simplify complex tasks, such as grasp planning [3]. Works like [6] show that grasp affordances have the potential to completely avoid complex grasp planning, by associating unknown objects with known geometries for which the robot has predefined grasps. Similarly, in [2] practical and efficient solutions to grasp planning for unknown objects based on affordances and potential fields are proposed. We think that the research on whole-body motion with contacts [7]–[9] can greatly benefit from the use of whole-body affordances to break down the problem in separate parts that can be organized by a high level reasoning process.

In this work, we define a whole-body affordance hypothesis as an association of a whole-body stable action to a perceived primitive of the environment. Based on previous approaches like [2] and [10], we aim at deriving, refining and utilizing whole-body affordances like holding, leaning, stepping-on or supporting in unknown environments.

For representation and execution of whole-body actions, we will rely on the concept of *Object-Action Complexes* (OACs) [1]. The concept of OACs states that the execution of an action is tightly related to the object that the action involves. This viewpoint of objects and actions being coupled is related to the concept of affordances. One could think of affordances as preconditions for the instantiation of OACs.

To generate utilizable affordance hypotheses, we rely on the extension of the manipulability maps introduced in [11] and [12] to whole-body stability maps. Such maps are discrete representations of the robot's workspace. For each end effector pose, the stability map contains the best possible stability rating among the whole-body configurations that realize the respective end effector pose. We use stability maps for detecting affordances in reach and for computing an initial guess for the point of application of an associated action.

We have implemented both, vision and affordance generation methods and evaluated them in different unknown scenarios involving small and big objects, stairs and walls. Our preliminary results show that the implemented approach allows us to associate realistic affordance hypotheses.

In the remainder of the paper, Section II describes the incorporated techniques for detecting environmental primitives and Section III explains how affordances are assigned to these primitives and refined using reachability and stability information. Section IV presents the output of the methods when perceiving different exemplary environments. Finally, Section V discusses the results and outlines our ideas for future work.

II. Environmental Perception

There are various sensing modalities involved in the environmental perception for humanoid robots such as visual, vestibular, tactile, auditive, thermoception, etc. Due to its versatility and unobstructive nature, as well as the large amount of information it provides, the visual sensing modality is fundamental for environmental state estimation. Particularly in partly unknown, cluttered and disaster-like scenarios, the environmental state estimation implies various complex skills. In addition to localization and mapping, these skills have to include environmental detection and categorization of unknown elements. Due to its model-free nature, the general environmental detection should solely rely on surface extraction, characterization, segmentation and classification. Such a robust and precise environmental detection can provide the necessary perceptual information for exploration and task planning, namely the generation and selection of plausible physical interactions with unknown surroundings. This generation process is based on the concept of affordances (III-A), whereas the selection of the most promising action is based on the robot stability maps (III-B).

A. Related Work

In the last decade, important contributions on visual simultaneous localization and mapping vSLAM for humanoid robots have been achieved (see [13]–[15]). Despite of their impressive capabilities in terms of reliability, large-scale and real-time performance, these approaches lack of the proper information required in multiple contact planning for physical exploration of unknown objects. This occurs due to the internal representations exploited in vSLAM methods. Concretely, the sparse representations using visual and/or spatial key features [14] do not provide piecewise continuous geometric surfaces necessary for the estimation of stable contacts (see Section III).

Recently, dense representations from active cameras have been successfully exploited. Despite its scalability, the octree representation with coarse voxel discretization [16] does not support the proper extraction of geometric primitives. This happens because within each voxel the continuity of the surface is reduced to a point. This implies the loss of surfaces boundaries while also making it impossible to determine the curvature at each sensed point.

Furthermore, there are notable contributions on surface reconstruction and segmentation from active sensors like active cameras and lasers (see [17]). An essential limitation of these methods is the assumption on high point density. This is either explicitly stated when using lasers or implicitly granted when using active cameras by placing the target objects close enough (usually 0.5-2.5 m) so that the point sparsity remains neglectable. These sparsity limitation are even more restrictive when the robot's vantage point cannot be planned in advance, for example in disaster-like scenarios.

Another notable contribution in [18] provides volumetric consistent results by fusing multiple views while tracking the camera generating high quality implicit surface representations. However, these methods have limited use for autonomous environmental state estimation because the surface acquisition requires various sparse vantage points, in addition to high GPU computability. This is a cyclic problem, because in order to place the robot's camera in the next convenient position (to generate the scene representation) a path plan is needed to avoid collisions which also requires

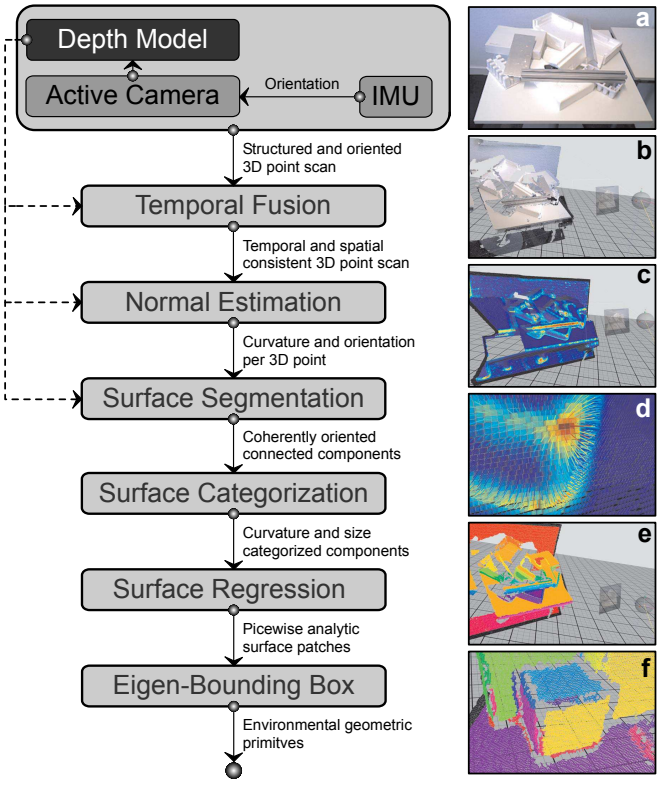

Fig. 2: Schematic representation of the pipeline for visual extraction of environmental geometric primitives.

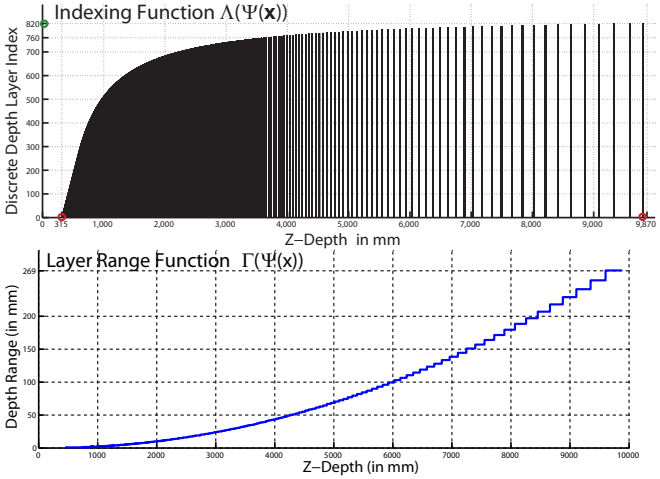

Fig. 3: The proposed active camera depth model consist of two functions: i) In the upper plot, the indexing function (see Eq. 1) maps depths to discrete layer indices. Notice, when using the active camera *ASUS-Xtion Pro* the total amount of discrete layers is $L_{\mathbf{max}} = 820$, whereas the clipping planes are located at $N = 315.0$ and $F = 9,870.0$ mm. ii) In the lower plot, the layer range function (see Eq. 2) expresses the depth subspace contained within each pixel-wise iso-disparity surface. Notice, at the layer index $i = 290$ at 605 mm the first layer with a range of $\Gamma(290) =$ 2mm appears. These functions are obtained experimentally and were validated against the geometric setup as in [20]. The importance of this results is the fact that the specific quantization and deviations of the particular device play an important role in the processing pipeline.

the environmental representation. An even more restrictive property of the multiple view fusion approach is the internal regular grid discretization which does not scale small rooms (a maximal volume of 8 m^3 according to [19]) to large spaces such as corridors or large rooms. This limits its application in disaster-like scenarios such as factories, hospitals, schools, etc. Another restrictive property of the fusion method is the need of sufficient surface geometry in order to lock down all degrees of freedom of the internal ICP point-plane optimization. This means, the application of this method is still tailored to certain scenarios such as living rooms. Finally, the resulting surface representation (an implicit surface) provides advantages for visualization (surface mapping, ray tracing, etc). However, this representation is not efficient for our objectives, namely computing the intersections with spatial functions such as the reachability and stability maps (see Section III-B).

B. Environmental Element Detection

Fig. 2 illustrates the visual extraction pipeline of environmental geometric primitives. The novel depth model is described in detail including its application along the intermediate stages necessary to obtain the categorized surface segments. The information flow from both sensors (active camera and IMU) is illustrated with continuous arrows while the depth model inclusion is denoted by dashed arrows.

Depth Model: Temporal artifacts (vanishing outliers, depth deviations, etc.) as well as the lower point density are critical sensing limitations when extracting environmental surfaces from active cameras in order to generate affordances.

The key idea to overcome both temporal outliers and point sparsity lies directly within the nature of the discrete depth image. This key idea is developed into the so-called *depth model* which is formally described as follows. First, consider the structure of the discrete depth image $\Psi(\mathbf{x}) \mapsto [N, F] \subset \mathbb{R}^+$ where $\mathbf{x} \in \mathbb{N}^2$ denotes a pixel location, meanwhile $N \in \mathbb{R}^+$ and $F \in \mathbb{R}^+$ stand for the near and far clipping planes. In this image, the depth values are distributed according to the scene content and the pixel-wise iso-disparity principle [20]. This principle implies that scene points are laying within discrete layers with particular depths. Hence, it is possible to propose a depth indexing bijective function as

$$\Lambda(\Psi(\mathbf{x})) \mapsto i \in \Phi := \{0, 1, ..., L_{\mathbf{max}}\} \subset \mathbb{N} \qquad (1)$$

which maps the Z-depth of a point $\Psi(\mathbf{x})$ to its corresponding layer index i, see the upper plot in Fig. 3.

Furthermore, when considering the adjacent discrete depth layers i and $i \pm 1$, it is plausible to determine the depth range contained in layer i as the subspace bounded by $[i-1, i+1]$. This is formally expressed by the layer range function $\Gamma(i \in \Phi) \mapsto \mathbb{R}^+$ (see the lower plot in Fig. 3.) as

$$\Gamma(i) := \Lambda^{-1}(i+1) - \Lambda^{-1}(i-1). \qquad (2)$$

Now, based on these layer indices, it is possible to determine whether two points are adjacent $\Theta(\mathbf{x}_u, \mathbf{x}_v) \mapsto \{0, 1\}$ independently of their varying depths as

$$\Theta(\mathbf{x}_u, \mathbf{x}_v) := \begin{cases} 1 & \textit{if } |\Lambda(\Psi(\mathbf{x}_v)) - \Lambda(\Psi(\mathbf{x}_u))| \leq 2 \\ 0 & \textit{else} \end{cases}. \quad (3)$$

A layer tolerance (in each direction) compensates oscillation for points located at the iso-disparity surface. Hence, the depth model provides invariant functions to determine various important properties: i) General and robust surface continuity assertion for estimating normals and segmenting oriented point clouds. ii) A depth adaptive band-width selection for various tasks including outlier removal, covariance neighborhood computation and spatial weighting along the visual extraction pipeline of environmental geometric primitives (see schematic in Fig. 2).

Temporal Fusion: By capturing a set of n depth images $\Psi_t(\mathbf{x})$ within the time scope $[t_1, t_{n \approx 15 \text{ at } 30\,\text{Hz}}]$ it is possible to detect and remove both, vanishing outliers and depth deviations outliers. This extends our previous method in [21] to the case of depth sensing. Formally, each element in the depth image $\Psi(\mathbf{x})$ has a layer deviation $\Delta(\mathbf{x}, t_1, t_n) \mapsto \mathbb{N}$ along the temporal scope which can be expressed as

$$\Delta(\mathbf{x}, t_1, t_n) := \mathbf{max}\left[\Lambda(\Psi_t(\mathbf{x}))\right]_{t=1}^{n} - \mathbf{min}\left[\Lambda(\Psi_t(\mathbf{x}))\right]_{t=1}^{n}. \quad (4)$$

Those image locations $\mathbf{x}$ whose layer deviation spawns a subspace beyond two subsequent layers ($\Delta(\mathbf{x}, t_1, t_n) \geq 3$) are rejected as depth outliers. This criterion also removes the temporal outliers (flickering points) due to the definition $\Lambda(0) := -\infty$. Namely, when a depth measurement is not available (point is occluded or the surface's material does not reflect the IR pattern) the depth value is set to zero. In this manner, the temporal fusion removes unreliable points. Fig. 2-a shows a scene where the proposed temporal fusion successfully produces outlier-free point cloud in Fig. 2-b.

Normal Estimation: When using the temporal fused depth image $\hat{\Psi}(\mathbf{x})$, it is possible to calculate its corresponding 3D point $x_u \in \mathbb{R}^3$ in a reliable manner. Moreover, for each 3D point x_u there is an associated normal vector $N_u \in \mathbb{R}^3$ with respect to its local neighborhood. This can be approximated based on the covariance matrix concept [17]. In addition, it is possible to incorporate three additional restrictions: i) the surface continuity (in terms of Eq. 3). ii) the neighborhood band-width from Eq. 2, and iii) the Epanechnikov kernel weighting (given $\vec{\delta}(u, v) := x_u - x_v$) as

$$\Upsilon(x_u, x_v) \quad := \quad \begin{cases} 1 - \left[\frac{|\vec{\delta}(u,v)|}{w(u)}\right]^2 & \textit{if } |\vec{\delta}(u, v)|/w(u) \leq 1 \\ 0 & \textit{else} \end{cases},$$

which considers the distance between the points on the surface using an adaptive bandwidth $w(u) \mapsto \mathbb{R}$ as a polynomial function of the layer range, namely

$$w(u) := \sum_{j=0}^{3} a_j \Gamma(\Psi_t(u))^j. \quad (5)$$

The whole weighting is expressed in the covariance matrix

$$C(x_u) = \frac{1}{|A(u)|} \sum_{v \in A(u)} \Theta(\mathbf{x}_u, \mathbf{x}_v) \Upsilon(u, v, w) \vec{\delta}(u, v) \vec{\delta}(u, v)^{\mathrm{T}}, \quad (6)$$

where the neighborhood set

$$A(u) := \{\mathbf{x}_v \in \Psi \ : \ |\vec{\delta}(u, v)| < w(u)\} \quad (7)$$

consists of all the points within an adaptive distance according to the layer depth. Fig. 2-c shows the curvature and normal estimation with consistent results at surface discontinuities. The color map indicates the curvature at each point. Further, Fig. 2-d shows the dense estimated normals at one surface corner from Fig. 2-a.

Surface Segmentation: The subsequent stages to obtain the environmental primitives also profit from the depth model as follows. The segmentation based on the region growing method [22] verifies not only normal orientation $N_u \cdot N_v \leq \cos \alpha$ but also curvature (from the ratios of the eigenvalues of $C(x_u)$), surface continuity $\Theta(\mathbf{x}_u, \mathbf{x}_v)$ and adaptive distance range $|(\vec{\delta}(u, v)| < w(u)$. This produces consistent surface segments with homogeneous curvature and continuity, see Fig. 2-e,f.

Surface Categorization: The categorization of surface segments integrates both the robot model dimensions and the orientation information provided by the IMU (see Fig. 2) in order to associate the lower planar surfaces as the floor of the scenario. The remaining patches are categorized based on their size and curvature distribution. So far, based on [17] four categories have been outlined: planar, cylindrical, spherical and free-from shapes. In the planar case, the α_p orientation threshold used during the segmentation can be narrowed in order to obtain more robust planar segments.

Surface Regression: The surface regression is performed using RANSAC fitting [23], see resulting planes in Fig. 1.

Empty Space: Once the environmental primitives have been extracted, the representation of the whole reachable space needs to be established. In contrast to [16], the proposed dual-octree can also represent the empty space by the inclusion of the methods to infer information from the inner nodes of the tree (see Fig. 4).

III. Deriving Affordance Hypotheses

Based on the surfaces visually perceived using the methods discussed in the previous section, this section proposes strategies for suggesting affordance hypotheses. Our preliminary experiments focus on affordances related to planar surfaces, although there is no principle limitation to these. Extension to curved surfaces, like cylindric or spherical ones, or volumetric primitives is possible and initial experiments have been conducted.

In the following, the proposed process of affordance suggestion is explained. First, we pursue a rule-based assignment of affordance hypotheses to environmental surfaces based on parameters like extent or orientation. Then, precomputed reachability information is used for limiting the amount of found hypotheses to directly usable ones. In a final step

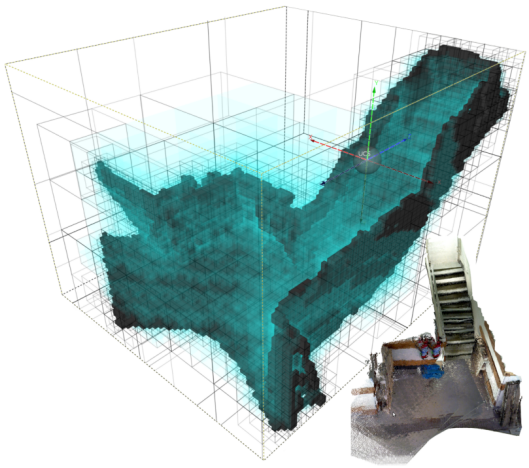

Fig. 4: The efficient empty space detection and representation based on dual-octree. This scalable representation is used in combination with the primitive detection to provide the visual environmental state for the derivation of affordances. **Top**: Dual-octree with empty space in cyan and occupancy nodes in gray. **Bottom**: The point set.

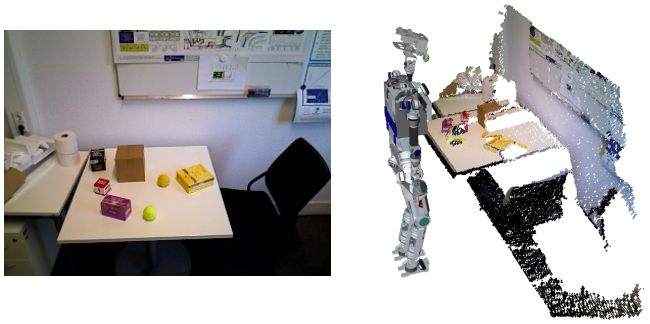

Fig. 5: Visualization of an exemplary depth image. The scene contains a table with several objects on it, a chair next to the table and a wall behind the table.

we compute an initial estimation about where the assumed hypotheses can optimally be applied with respect to the robot's stability.

A. Suggestion of Affordance Hypotheses

The methods for visual perception discussed in Section II allow the detection and approximation of surfaces. Fig. 5 shows the depth image of an exemplary scene and the set of primitives resulting from the perception process.

Affordance hypotheses are suggested by rules that incorporate parameters of the primitives like orientation or extent. This approach eventually results in a set of rules that link geometric primitives to affordance hypotheses, similar to [24]. An exemplary set of such rules is given in Table I. For example, a planar surface that is sufficiently large and oriented horizontally, e.g. a table, suggests the affordance *support*. A long curved surface of a certain radius, e.g. a handrail, suggests the affordance *hold*. The last column of Table I describes the preferred end-effector pose when

TABLE I: Example of a set of rules for affordance derivation. See Fig. 6 for $\boldsymbol{x}_{eef}$, $\boldsymbol{y}_{eef}$ and $\boldsymbol{z}_{eef}$. The operator $\uparrow$ tells if two vectors point into the same direction[1]. The λ_i are implementation-specific constants.

Affordance	Surface	Parameters	Conditions	EEF
Support	Planar	Normal $\boldsymbol{n}$	$\boldsymbol{n} \uparrow \boldsymbol{z}_{world}$	$\boldsymbol{z}_{eef} \uparrow \boldsymbol{n}$
		Area a	$a \geq \lambda_1$	
Lean	Planar	Normal $\boldsymbol{n}$	$\boldsymbol{n} \perp \boldsymbol{z}_{world}$	
		Area a	$a \geq \lambda_2$	
Grasp	Planar	Normal $\boldsymbol{n}$	$a \in [\lambda_3, \lambda_4]$	
		Area a		
	Curved	Radius r	$r \in [\lambda_5, \lambda_6]$	$\boldsymbol{y}_{eef} \uparrow \boldsymbol{d}$
		Direction $\boldsymbol{d}$	$\|\boldsymbol{d}\| \leq \lambda_7$	
Hold	Curved	Radius r	$r \in [\lambda_8, \lambda_9]$	
		Direction $\boldsymbol{d}$	$\|\boldsymbol{d}\| \geq \lambda_{10}$	

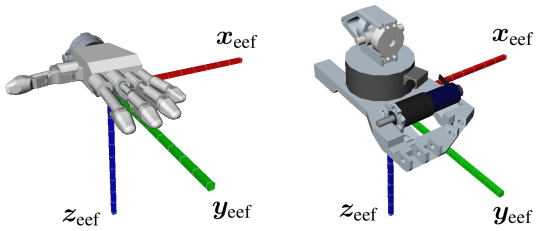

Fig. 6: The TCP coordinate systems for the left hand (left) and the left foot (right) of ARMAR-4.

utilizing the respective affordance, (see also Fig. 6). This will be of interest in Section III-C.

Using the rules outlined in Table I, the system can identify several affordance hypotheses in the exemplary scene (see Fig. 7). In the next steps, the resulting hypotheses are filtered according to their reachability and reasonable points of application are chosen. Both steps base on stability maps.

B. Stability Maps

For investigating reachability of affordances, we incorporate a pre-computed representation of the robot's workspace. The workspace is represented by a 6D voxel grid $\mathcal{R}_e$ that contains quality values regarding the end effector e, relative to its pose $\boldsymbol{p} \in SE(3)$, consisting of an translational component $\boldsymbol{t}_\mathbf{p} \in \mathbb{R}^3$ and a rotational component $\mathbf{R}_\mathbf{p} \in SO(3)$:

$$\mathcal{R}_e(\boldsymbol{p}) = \mathcal{R}_e(\boldsymbol{t}_\boldsymbol{p}, \boldsymbol{R}_\boldsymbol{p}) \in [0, 1] \tag{8}$$

Depending on the task, the stored values may cover binary reachability information [12], [25] or more complex information like the manipulability that the end effector can achieve at the respective location [11]. In this work, we use an extension of reachability maps called *stability maps*, that contain static stability information for a bipedal humanoid in whole-body reaching scenarios. Each voxel of a stability map $\mathcal{S}_e$ tells how stable the robot would be when achieving the requested end effector pose $\boldsymbol{p}$. Since quality information for a redundant robot is stored, the workspace representation

[1] $\boldsymbol{v} \uparrow \boldsymbol{w} \leftrightarrow \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} \approx 1$

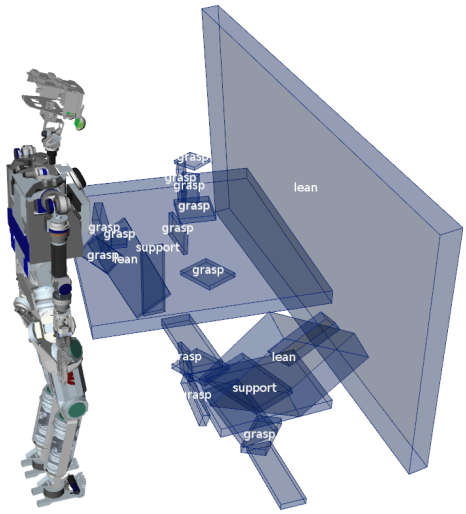

Fig. 7: Bounding boxes of planar primitives derived from the scene using the perception process (see Section II) together with the derived affordances (see Table I). The depths of the boxes indicate the derivation of the perceived points from the fitted plane.

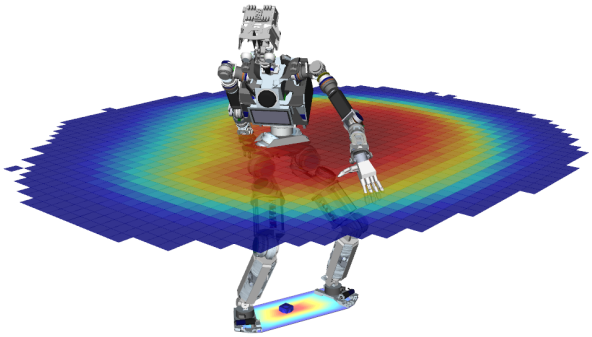

Fig. 8: Cut through the whole-body stability map for the left hand of the simulated robot ARMAR-4 [26]. The stability rating depends on the distance of the projected center of mass (blue box) to the center of the support polygon. Each entry of the stability map is computed according to Eq. 11.

just gives an upper bound of the stability that can be achieved for a given pose in workspace. In particular, the stability map tells if the end effector pose in question is reachable at all.

The stability of a whole-body configuration c in this case is expressed by the proximity of the projected center of mass $\boldsymbol{x}'_{com}(c)$ to the center $\boldsymbol{x}_{center}(c)$ of the support polygon s_c with boundary ∂s_c:

$$stability(c) = \frac{\min\{\|\boldsymbol{x}'_{com}(c) - \boldsymbol{y}\| : \boldsymbol{y} \in \partial s_c\}}{\min\{\|\boldsymbol{x}_{center}(c) - \boldsymbol{y}\| : \boldsymbol{y} \in \partial s_c\}}, \quad (9)$$

with

$$\boldsymbol{x}'_{com}(c) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \boldsymbol{x}_{com}(c). \quad (10)$$

Stability maps keep the best achieved stability value $stability(c)$ for an end effector pose $\boldsymbol{p}$, based on sampled whole-body configurations c. Fig. 8 shows an exemplary stability map, where the map was projected to three dimensions by computing the average stability rating of all possible orientations, represented by a finite discretization[2] $\Omega \subset SO(3)$:

$$value(\boldsymbol{x}) = \frac{1}{|\Omega|} \sum_{\boldsymbol{R} \in \Omega} \mathcal{S}_e(\boldsymbol{x}, \boldsymbol{R}). \quad (11)$$

Stability maps, as well as reachability maps are generated offline (see [11], [12]). Pose validation methods ensure that all considered robot configurations respect constraints like self-collisions or, as in this paper, static stability.

In this work we incorporate the two stability maps $\mathcal{S}_{Left\ Hand}$ and $\mathcal{S}_{Right\ Foot}$ which refer to the respective end effector. Both maps contain the stability values for reachable end effector poses. However, while in case of $\mathcal{S}_{Left\ Hand}$, both feet are set to fixed poses, the robot is only supported by one single foot in case of $\mathcal{S}_{Right\ Foot}$.

Although the creation of stability maps has high computational costs, querying is efficient. This enables us to utilize stability maps for filtering affordance hypotheses that the robot can currently not reach.

C. Determining Reachable Hypotheses

The previous sections show that, based on depth models obtained from active cameras, a robot can identify plenty of primitives p_i in a scene and is able to assign affordance hypotheses h_i to these primitves: $\mathcal{H} = \{(p_1, h_1), \cdots, (p_k, h_k)\}$. For planning purposes it is important to identify $\mathcal{H}_R$, the subset of hypotheses that are directly reachable for the robot, either for utilization or for verification.

For each affordance h_i, Table I constraints the set of possible end effector poses by fixing one axis of the end effector's local coordinate system (see Fig. 6). The resulting constrained space of orientations will be denoted as $\Omega_{(p_i,h_i)}$. The geometric shapes of the primitives together with the suitable end effector poses allow us to assign a stability value to each point $\boldsymbol{x} \in \partial p_i$ on the surface[3] of the primitive p_i:

$$stability_{(p_i,h_i)}(\boldsymbol{x}) = \max\left\{\mathcal{S}_e(\boldsymbol{x}, \boldsymbol{R}) : \mathbf{R} \in \Omega_{(p_i,h_i)}\right\}. \quad (12)$$

This value tells how stable it is for the robot to reach the different points on the primitive's surfaces while maintaining the preferred end effector orientation.

D. Post-Processing Affordance Hypotheses

Based on the stability value defined in Eq. 12, two tactics for hypothesis post-processing are conducted:

First, hypotheses for which the robot's stability rating, when reaching for them, lies below a threshold σ are

[2]In our experiemts we used discretizations of 6 cm (translational) and 0.75 rad (orientational).

[3]In the examples presented in this work, primitives are surfaces, hence $\partial p_i = p_i$.

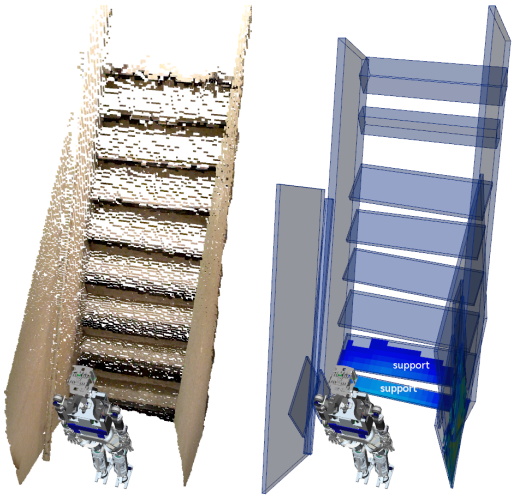

Fig. 9: Captured depth image of a staircase (left) and derived reachable affordance hypotheses for the right foot (right).

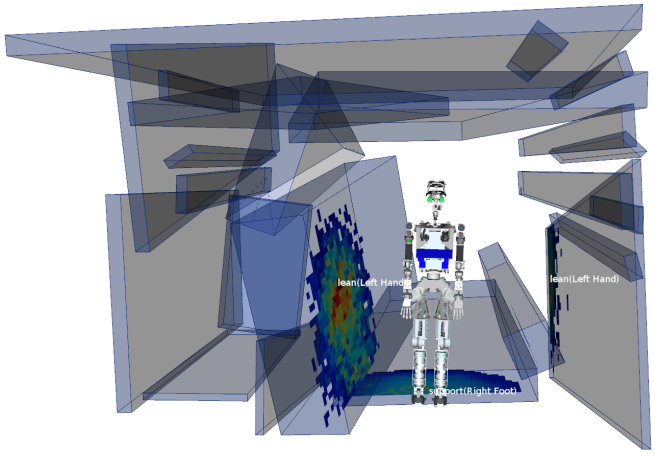

Fig. 10: Extracted geometric primitives and reachable affordance hypotheses for two available end effectors in a tunnel scenario.

filtered out as they are not directly utilizable for the robot. This strategy allows us to create the set $\mathcal{H}_R$ of reachable affordance hypotheses:

$$\mathcal{H}_R = \left\{ (p,h) \in \mathcal{H} : \exists \boldsymbol{x} \in \partial p : \mathit{stability}_{(p,h)}(\boldsymbol{x}) > \sigma \right\} \quad (13)$$

Second, the point on a primitive's surface that has the highest stability rating is the point where the robot can utilize the respective affordance in the most stable manner. This point is regarded as an initial guess on where exactly to perform an action that is suggested by an affordance:

$$\mathit{hotspot}(p,h) = \underset{\boldsymbol{x} \in \partial p}{\operatorname{argmax}} \, \mathit{stability}_{(p,h)}(\boldsymbol{x}). \quad (14)$$

Fig. 1 depicts the result of the affordance assignment process. It shows only those affordances whose stability rating lies above a threshold σ. Furthermore, the affordance labels are attached to the points with the highest stability ratings.

IV. Experiments

In the previous sections we proposed a method for assigning affordance hypotheses to environmental primitives. This section will show, how the method performs for different exemplary scenes.

Fig. 9 shows the robot confronted with the depth image of a staircase. As the resulting image on the right shows, the principal elements of the staircase, i.e. the walls, the stairs and the handrail, are properly seperated into different primitives. In this example, $\mathcal{S}_{\mathit{Right\ Foot}}$ was used for identifying reachable affordance hypotheses. Therefore, hypotheses that are not suitable to a foot, e.g. *grasp*, were ignored. The results show that the affordances that are important for the chosen end effector are found and properly assigned to the steps reachable for the robot.

In another example, depicted in Fig. 10, the robot is intended to walk through a tunnel. In this case, both maps, $\mathcal{S}_{\mathit{Left\ Hand}}$ and $\mathcal{S}_{\mathit{Right\ Foot}}$ were incporated to let the methods look for a promosing point for foot placement while also having a hand contact for increased stability. Based on the depth model from the active camera, the proposed methods are able to successfully identify affordance hypotheses for a possible foot placement as well as several leaning-affordances that the robot can incorporate for stabilizing itself.

V. Conclusions and Future Work

This paper presents our approach to the detection of whole-body affordance hypotheses based on the fusion of visual and inertial sensor information. In the first phase, depth images are processed using the proposed depth model in order to extract, segment and categorize surfaces based on estimated curvature and normals. These surfaces are then further processed into geometric primitives.

The second phase incorporates a predefined set of rules that links symbolic affordances to properties of the extracted primitives. This information is fused with the robot's stability map in order to determine reachable affordance hypotheses. The stability map indicates how stable the robot would be if the end effector would reach for a given pose.

The combination of both phases, i.e. the perception and the affordance derivation, has been evaluated in three exemplary scenes. This work can be regarded as a first step towards affordance based locomotion and manipulation in unknown environments. The results look promising as the derived affordance hypotheses are valid and would be useful in the exemplary situations.

In the next steps we will extend and evaluate Table I and work on the generation of suitable whole-body configurations based on a chosen set of pairs of end effectors and affordances. We will also work on formalizing the robot's task in order to reduce the amounts of affordances to consider per end effector. Another central aspect of our future work will be the verification of affordance hypotheses by incorporating the different sensor modalities of a humanoid robot. We especially plan to use force-based exploration to estimate the

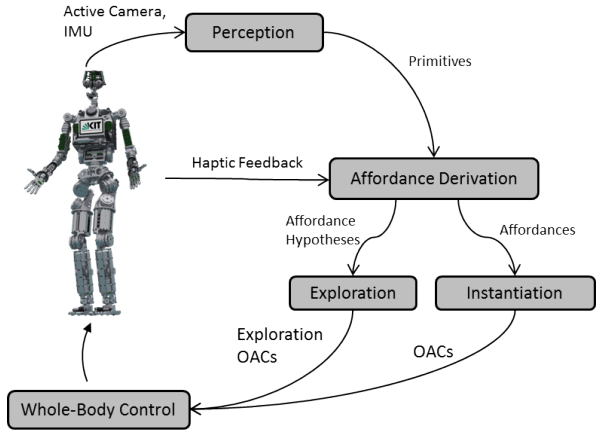

Fig. 11: The process of detection and exploration of whole-body affordances: Based on sensory information from active cameras or IMUs, the perceptual component produces an abstract representation of the environment. The resulting primitives are used for deriving affordance hypotheses as well as the most promising points of their application. One possible choice is then to trust a derived hypothesis, in which case it directly results in an OAC instance that can be executed. The other choice is to start an exploration process to estimate the affordance's reliability and the execution parameters. In this case, exploration OACs are executed and the sensed feedback again contributes to the affordance assignment step.

reliability of an affordance as well as to determine the actual execution parameters. The exploration step will be based on the formalism of Object-Action Complexes introduced in [1]. Fig. 11 depicts the pursued strategy for affordance derivation and exploration and the execution of actions based on perceived whole-body affordances.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no. 611832 (WALK-MAN).

REFERENCES

[1] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omren, A. Agostini, and R. Dillmann, "Object-action complexes: Grounded abstractions of sensorimotor processes," *Robotics and Autonomous Systems*, vol. 59, pp. 740–757, 2011.

[2] A. Bierbaum, M. Rambow, T. Asfour, and R. Dillmann, "Grasp affordances from multi-fingered tactile exploration using dynamic potential fields," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2009, pp. 168–174.

[3] E. Şahin, M. Çakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.

[4] D. Leidner, A. Dietrich, F. Schmidt, C. Borst, and A. Albu-Schffer, "Object-centered hybrid reasoning for whole-body mobile manipulation."

[5] J. J. Gibson, *The ecological approach to visual perception*. Psychology Press, 1979.

[6] A. ten Pas and R. Platt, "Localizing grasp affordances in 3-D points clouds using taubin quadric fitting," in *International Symposium on Experimental Robotics (ISER)*, 2014.

[7] S. Lengagne, J. Vaillant, E. Yoshida, and A. Kheddar, "Generation of whole-body optimal dynamic multi-contact motions," *The International Journal of Robotics Research*, vol. 32, no. 9, pp. 1104–1119, 2013.

[8] L. Sentis, J. Park, and O. Khatib, "Compliant control of multicontact and center-of-mass behaviors in humanoid robots," *IEEE Transactions on Robotics*, vol. 26, no. 3, pp. 483–501, 2010.

[9] L. Sentis, "Compliant control of whole-body multi-contact behaviors in humanoid robots," in *Motion Planning for Humanoid Robots*. Springer, 2010, pp. 29–66.

[10] M. Popović, D. Kraft, L. Bodenhagen, E. Başeski, N. Pugeault, D. Kragic, T. Asfour, and N. Krüger, "A strategy for grasping unknown objects based on co-planarity and colour information," *Robotics and Autonomous Systems*, vol. 58, no. 5, pp. 551–565, 2010.

[11] N. Vahrenkamp, T. Asfour, G. Metta, G. Sandini, and R. Dillmann, "Manipulability analysis," in *IEEE-RAS International Conference on Humanoid Robots*, 2012, pp. 568–573.

[12] N. Vahrenkamp, T. Asfour, and R. Dillmann, "Efficient inverse kinematics computation based on reachability analysis," *International Journal of Humanoid Robotics*, vol. 9, no. 04, 2012.

[13] D. Maier, A. Hornung, and M. Bennewitz, "Real-time navigation in 3D environments based on depth camera data," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Osaka, Japan, November 2012.

[14] S. Yoon, S. Hyung, M. Lee, K. Roh, S. Ahn, A. Gee, P. Bunnun, and M.-C. Calway, A, "Real-time 3D simultaneous localization and map-building for a dynamic walking humanoid robot," *Advanced Robotics*, vol. 27, pp. 759–772, 2013.

[15] R. F. Salas-Moreno, R. A. Newcombe, P. H. J. K. Hauke Strasdat, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," *Computer Vision and Pattern Recognition (CVPR), IEEE*, 2013.

[16] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[17] R. B. Rusu, *Semantic 3D Object Maps for Everyday Robot Manipulation*, ser. Springer Tracts in Advanced Robotics. Springer, 2013, vol. 85.

[18] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–136.

[19] Microsoft, "Microsoft research kinect fusion project," http://msdn.microsoft.com/en-us/library/dn188670.aspx, July 2014.

[20] M. Pollefeys and S. Sinha, "Iso-disparity surfaces for general stereo configurations," in *Computer Vision - ECCV 2004*. Springer, vol. 3023.

[21] D. Gonzalez-Aguirre, T. Asfour, and R. Dillmann, "Robust Image Acquisition for Vision-Model Coupling by Humanoid Robots," in *IAPR-Conference on Machine Vision Applications*, 2011, pp. 557–561.

[22] D. Gonzalez-Aguirre, S. Wieland, T. Asfour, and R. Dillmann, "On Environmental Model-Based Visual Perception for Humanoids," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2009, vol. 5856, pp. 901–909.

[23] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[24] K. M. Varadarajan and M. Vincze, "AfNet: The affordance network," in *Proceedings of the 11th Asian Conference on Computer Vision*, 2012, pp. 512–523.

[25] F. Zacharias, C. Borst, and G. Hirzinger, "Capturing robot workspace structure: representing robot capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 3229–3236.

[26] T. Asfour, J. Schill, H. Peters, C. Klas, J. Bücker, C. Sander, S. Schulz, A. Kargov, T. Werner, and V. Bartenbach, "ARMAR-4: A 63 DOF Torque Controlled Humanoid Robot," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Atlanta, USA, October 2013.