Experimental Evaluation of a Perceptual Pipeline for Hierarchical Affordance Extraction

Peter Kaiser[1], Eren E. Aksoy[1], Markus Grotz[1], Dimitrios Kanoulas[2],
Nikos G. Tsagarakis[2], and Tamim Asfour[1]

[1] Institute for Anthropomatics and Robotics,
Karlsruhe Institute of Technology (KIT),
Adenauerring 2, 76131 Karlsruhe, Germany,

[2] Department of Advanced Robotics,
Instituto Italiano di Tecnologia (IIT),
via Morego 30, 16163 Genova, Italy

Abstract. The perception of affordances in unknown environments is an essential prerequisite for autonomous humanoid robots. In our previous work we developed a perceptual pipeline for the extraction of affordances for loco-manipulation actions based on a simplified representation of the environment starting from RGB-D camera images. The feasibility of this approach has been demonstrated in various examples in simulation as well as on real robotic platforms. The overall goal of the perceptual pipeline is to provide a robust and reliable perceptual mechanism for affordance-based action execution.

In this work we evaluate the performance of the perceptual pipeline in combination with sensor systems other than RGB-D cameras, in order to utilize redundant sensor equipment of humanoid robots. This is particularly important when considering challenging scenarios where particular sensors are not applicable, e.g. due to intense sunlight or reflective surfaces. In this work we focus on stereo cameras and LIDAR laser scanners.

Keywords: Affordances, Perception, Loco-Manipulation

1 Introduction

One of the main motivations behind the development of humanoid robots is the idea of creating a robotic system that is able to autonomously operate in unstructured, human-centered environments. Such robots require a rich perceptual basis for identifying possible ways of interaction with the environment. The theory of *affordances*, originally proposed by Gibson [1], provides a conceptual mechanism for explaining the human perceptual process. It states that action possibilities are proposed to an agent, for example a human or a humanoid robot, based on properties of relevant environmental objects and based on the agent's capabilities. A chair for example affords *sitting*, but only to agents of sufficient height and capability. Overviews over applications of affordances in robotics can be found in [2] and [3].

Many of the teams participating in the DARPA Robotics Challenge (DRC) Finals in 2015 pursued an affordance-driven approach to whole-body locomotion and manipulation. The perceptual process as well as the execution of actions were controlled by human operators via teleoperation in supervised autonomy. Examples for such pilot interfaces can be found in [4,5,6]. The teams participating in the DRC mostly used a combination of LIDAR sensors and stereo vision for range sensing. Promising results have also been generated solely using stereo camera systems [13]. LIDAR sensors are precise, but expensive time-of-flight laser scanners. Point clouds are obtained by aggregating line scans of the rotating sensor over time. Stereo camera systems are cheap, passive range sensors based on the identification of point correspondences in two camera images. Stereo camera systems are known to perform poorly with untextured objects.

While affordances found many applications in the field of robotics, we specifically aim at the concept of *whole-body affordances*, i.e. affordances that refer to actions of whole-body locomotion or manipulation. This includes actions for whole-body stabilization, e.g. leaning against walls or holding handrails, or large-scale manipulation, e.g. pushing or lifting of large objects. Actions of whole-body locomotion and manipulation play an important role for the utilization of structures designed for the human body. In the next section we describe our previously proposed hierarchical formulation of affordances based on fundamental grasp affordances, which we regard as initial work towards the formulation of whole-body affordances.

2 Technical Approach

The perceptual process employed in this work starts with creating a simplified representation of the captured scene. The acquired point clouds pass several pipeline steps until the scene is represented in terms of environmental primitives, i.e. planes, cylinders or spheres. In the first step we perform a part-based object segmentation method [7] which over-segments the scene in order to roughly separate groups of environmental primitives. We further employ geometric features for iteratively categorizing the resulting segments into environmental primitives. Fig. 1 shows the structure of the perceptual pipeline from depth sensor information to the extraction of affordances. Fig. 2 visualizes the intermediate steps of the perceptual pipeline.

In [12] we proposed a hierarchical framework for the extraction of loco-manipulation affordances based on a scene represented with environmental primitives. The framework follows the idea that the majority of loco-manipulation actions break down to elementary power grasps at the lowest level. We particularly focus on *platform grasps* and *prismatic grasps* as we think that these two grasp types are predominant for the considered set of actions. However, the framework is not limited to these two elementary affordances. Affordances are represented as continuous certainty functions

$$\Theta_a : SE(3) \rightarrow [0, 1], \tag{1}$$

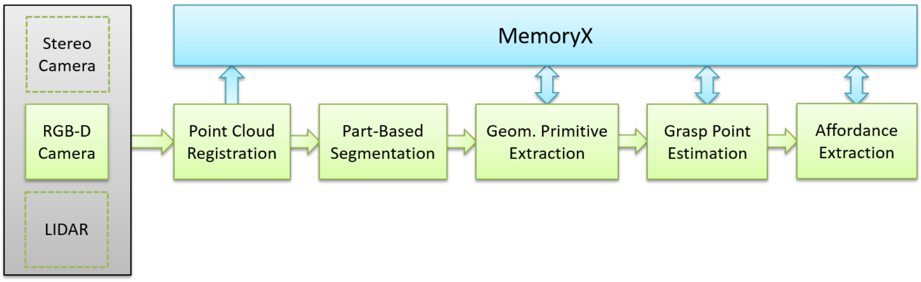
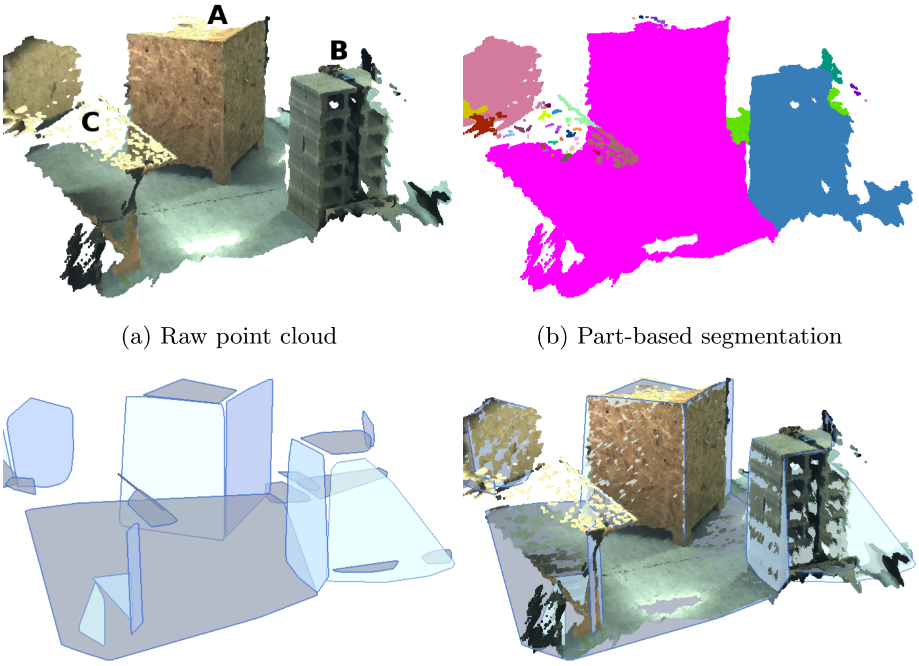


Fig. 1: The perceptual pipeline for affordance extraction [8,9,10]. The perceived scene is segmented into environmental primitives which form the basis for the extraction of affordances. The pipeline is implemented within the robot software framework *ArmarX* [11] and intertwined with its memory subsystem *MemoryX*.



(a) Raw point cloud

(b) Part-based segmentation

(c) Environmental primitives

(d) Combined view

Fig. 2: The intermediate steps of the perceptual pipeline (see Fig. 1) for an exemplary scene containing a large box **A**, a stack of bricks **B** and a table **C**

which tell, how certain the perceptual system is about the existence of an affordance a for a given end-effector pose $\boldsymbol{x} \in SE(3)$.[1] Two mathematical operations are applied to form higher-level affordance certainty functions:

- Affordance certainty functions can be multiplied in order to form combined affordance certainty functions.
- Environmental properties can be converted into compatible certainty functions by applying sigmoid threshold functions.

The procedure of affordance extraction is robot agnostic, taking elementary body-scaled parameters, e.g. end-effector dimensions, into consideration. Fig. 3 shows the hierarchical process of affordance extraction based on the exemplary *bimanual support affordance* $\Theta_{\text{Bi-Sp}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

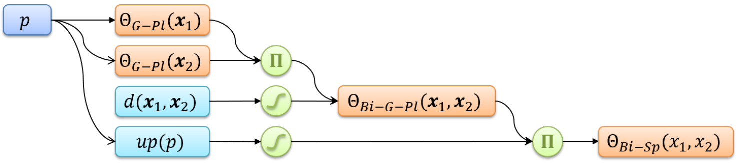

p
$\Theta_{G-Pl}(\boldsymbol{x}_1)$
$\Theta_{G-Pl}(\boldsymbol{x}_2)$
$d(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$up(p)$
Π
$\Theta_{Bi-G-Pl}(\boldsymbol{x}_1, \boldsymbol{x}_2)$
Π
$\Theta_{Bi-Sp}(x_1, x_2)$

Fig. 3: Example of a bimanual affordance certainty function. The bimanual support affordance $\Theta_{\text{Bi-Sp}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ consists of a *bimanual platform grasp* affordance $\Theta_{\text{Bi-G-Pl}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in combination with a horizontal orientation of the underlying primitive p. Horizontality is defined via a threshold applied to the orientation function up(p). A bimanual platform grasp affordance consists of two *unimanual platform grasp* affordances, one for each end-effector pose, and a threshold applied to the distance $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Fig. 4 shows an exemplary sampling of the affordance certainty functions $\Theta_{\text{G-Pl}}$ for platform grasping and $\Theta_{\text{G-Pr}}$ for prismatic grasping, extracted from a perceived staircase. The example shows that the perceptual pipeline is able to successfully segment the perceived environment into elementary primitives and that it can subsequently compute reasonable certainty functions for various elementary affordances. It also shows that the perceptual pipeline is able to produce useful segmentations of real scenes captured with point cloud sensors. The displayed affordance certainty functions can directly be used as a basis for planning feet poses for stepping ($\Theta_{\text{G-Pl}}$) or hand locations for grasping the handrail ($\Theta_{\text{G-Pr}}$). Further results can be found in [8,9,10]. The system has been implemented and evaluated in experiments based on the humanoid platform ARMAR-III. One of the performed experiments demonstrates the perception of turnable objects in the context of a bimanual valve-turning task (see Fig. 5 and [12] for further details).

[1] $SE(3)$ denotes the special Euclidean group.

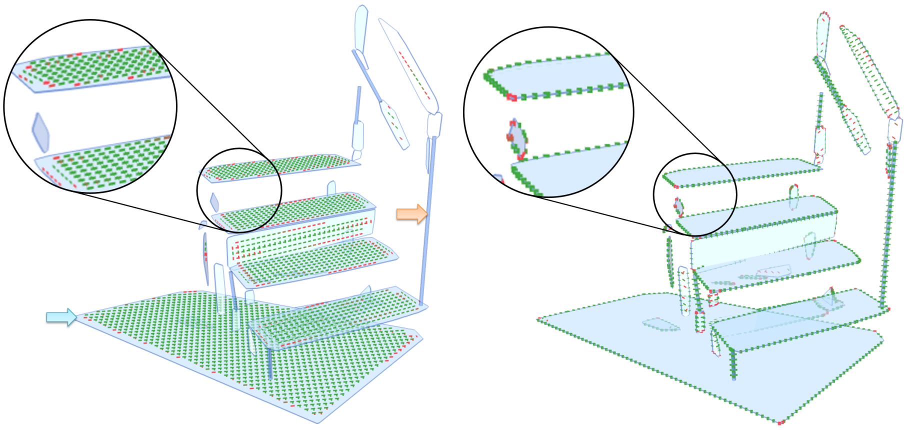


Fig. 4: A visualization of affordance certainty functions for platform grasps $\varTheta_{\text{G-Pl}}$ (left) and prismatic grasps $\varTheta_{\text{G-Pr}}$ (right) extracted from a perceived staircase. The colors indicate the value of the respective certainty function ranging from red (highly uncertain) to green (very certain), while certainty values of zero were omitted in the visualization. The scene is segmented into environmental primitives, in this case planes, e.g. the ground plane (blue arrow), and cylinders, e.g. the handrail (orange arrow).

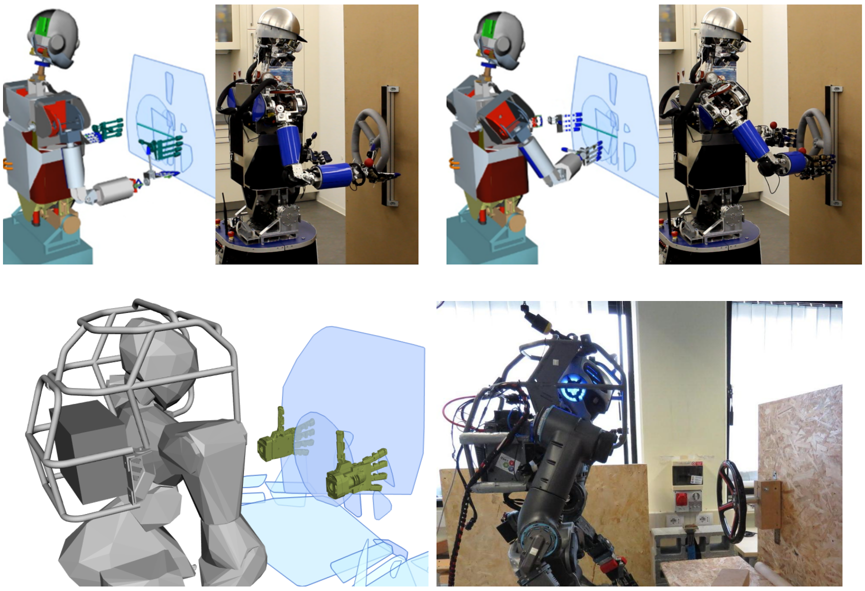


Fig. 5: *Top*: The perceptual pipeline properly extracts bimanual affordances and proposes suitable end-effector poses (left) for the subsequent action execution (right) in a valve turning scenario. *Bottom*: A comparable experimental setup for the humanoid robot WALK-MAN.

3 Experiments

The experiments carried out in [12] demonstrate the general feasibility of the proposed approach for loco-manipulation affordance extraction and the usefulness of the generated data for subsequent action execution. The perceptual pipeline has been designed and tested with RGB-D camera images, which provide a simple and cheap solution to range sensing. However, there are multiple approaches to visual perception for humanoid robots which promise to perform better in critical circumstances that real humanoid robots would have to face. Such circumstances could include outdoor scenarios with intense sunlight or malicious object materials, e.g. reflective surfaces. In the following we present initial evaluations of the perceptual pipeline with sensor systems other than RGB-D cameras. The experiments have been carried out in multiple scenarios with the perceptual system of the humanoid robot WALK-MAN [14].[2]

3.1 Evaluation Scenarios

To evaluate the perceptual pipeline we captured a total of 129 stereo vision and 66 LIDAR point clouds. The point clouds resemble static snapshots of two evaluation scenarios S_1 and S_2 (see Fig.6). For each scenario S_i, we defined multiple experiments $E_{i,1}, \ldots, E_{i,k}$ by changing the camera perspective or by slightly modifying the experimental setup. For each experiment $E_{i,j}$ we took a series of point clouds $P_{i,j,1}, \ldots, P_{i,j,n}$. Although the captured scene was static during the experiments, the set of point clouds captured over time resembles noise of the underlying sensor system. In Fig. 6 we briefly describe the evaluation scenarios S_1 and S_2.

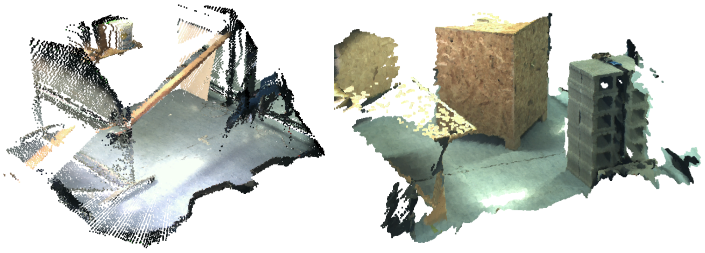


Fig. 6: The evaluation scenarios S_1 (left, A vertical wooden bar in front of the robot) and S_2 (right, A large box, a table and a stack of bricks).

[2] WALK-MAN is equipped with a *MultiSense SL* sensor head from *Carnegie Robotics* containing a LIDAR sensor and a stereo camera system. The LIDAR scanner captures 1024 points per scan and was configured to rotate with 0.5 rad/sec. The stereo camera system produces point clouds using semi-global matching based on 1 Mpx camera images. No postprocessing filters have been applied in both cases.

The perceptual pipeline requires a set of parameters to be specified, especially for the segmentation and primitive extraction stages. These parameters potentially need to be adjusted when changing the environmental setup. However, in the following evaluation we used the same parameter setup for all experiments $E_{i,j}$ from a scenario S_i.

3.2 Evaluation Procedure

Each point cloud $P_{i,j,l}$ is processed using the perceptual pipeline, extracting affordance certainty functions for the elementary power grasp affordances $\Theta_{\text{G-Pl}}$ and $\Theta_{\text{G-Pr}}$. For each evaluation scenario S_i, we first manually create a ground truth set of environmental primitives and then compute the ground truth affordance certainty functions $\Theta^*_{\text{G-Pl}}$ and $\Theta^*_{\text{G-Pr}}$. We then perform a binary comparison of the ground truth affordance certainty functions with the ones extracted from the experiment point cloud, applying a threshold of 0.5 to the certainty values of Θ and Θ^*. The spatial and orientational tolerances Δx and $\Delta\varphi$ for proximity of end-effector poses have been set to 7.5 cm and $\frac{\pi}{4}$ rad, respectively. The tolerances can be chosen generously at this point as the process of affordance extraction in general is understood as a source of high-level information on affordances and end-effector poses, prone to a certain degree of error. Handling these perceptual inaccuracies falls into the scope of affordance validation and action execution, as described in [10]. In order to evaluate the continuous nature of the certainty functions, we additionally define a similarity measure which is defined as the ratio of *similar* sampling points over the total number of ground truth sampling points. Two end-effector poses $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are considered similar if $|\Theta(\boldsymbol{x}) - \Theta^*(\boldsymbol{x}^*)| < \varepsilon$.[3]

3.3 Results

Table 1 shows the evaluation results for the scenarios S_1 and S_2 with respect to the affordance certainty functions $\Theta_{\text{G-Pl}}$ and $\Theta_{\text{G-Pr}}$. For each experiment $E_{i,j}$ and for both available sensors, we list the number of point clouds processed (#), as well as the F_1 and *similarity* scores, comparing with the experiment's ground truth. The ground truth stays the same for both evaluated sensors. Fig. 7 displays mean and standard deviation of the precision and recall values from Table 1 for scenario S_2 and the affordance certainty function $\Theta_{\text{G-Pr}}$. The results show that the perceptual pipeline can successfully process point clouds originating from the considered sensors. However, stereo vision data performs significantly worse than LIDAR data, mainly because the depth information for more distant and less textured objects is less accurate. In many cases, especially for platform grasps in scenario S_2, the ground truth primitives were properly extracted, but significantly shifted when using stereo vision input.

Referring to Fig. 4, platform grasp affordances usually form two-dimensional manifolds in the space of end-effector positions, whereas prismatic grasp affordances form one-dimensional manifolds. This makes it harder for the perceptual

[3] In our experiments, we chose $\varepsilon = 0.1$.

pipeline to properly extract prismatic grasp affordances within the applied tolerances. This is the main reason why the F_1 scores are significantly worse for $\Theta_{\text{G-Pr}}$ than for $\Theta_{\text{G-Pl}}$, for both sensors likewise. In many cases, possibly due to outliers in the point clouds, the extracted primitives are larger than the ground truth primitives, but properly oriented and shaped. Such circumstances result in failures when comparing with the ground truth, but resulting affordances might still be of reasonable use for action planning, when employing appropriate control mechanisms. Table 2 displays the runtimes of the primitive extraction and the affordance extraction steps of the perceptual pipeline for two selected experiments both, for LIDAR and for stereo vision point clouds. The runtimes have been generated on a standard Core i7 desktop PC. Note that the perceptual pipeline has not been optimized for runtime efficiency.

Table 1: Comparison of the affordance certainty functions $\Theta_{\text{G-Pl}}$ and $\Theta_{\text{G-Pr}}$ produced by the perceptual pipeline based on different sensors in S_1 and S_2.

			LIDAR			**Stereo Vision**		
	Scenario	**Exp.**	#	F_1	*Sim.*	#	F_1	*Sim.*
$\boldsymbol{\Theta}_{\text{G-Pl}}$	$\boldsymbol{S_1}$	$E_{1,1}$	8	**0.93**	0.84	20	0.85	0.69
		$E_{1,2}$	2	**0.92**	0.84	5	0.76	0.56
	$\boldsymbol{S_2}$	$E_{2,1}$	9	**0.92**	0.85	23	0.79	0.63
		$E_{2,2}$	9	**0.92**	0.84	17	0.50	0.35
		$E_{2,3}$	10	**0.90**	0.81	18	0.62	0.45
		$E_{2,4}$	9	**0.90**	0.81	13	0.56	0.39
		$E_{2,5}$	10	**0.93**	0.85	13	0.53	0.35
		$E_{2,6}$	9	**0.91**	0.82	20	0.74	0.58
$\boldsymbol{\Theta}_{\text{G-Pr}}$	$\boldsymbol{S_1}$	$E_{1,1}$	8	**0.59**	0.98	20	0.15	0.82
		$E_{1,2}$	2	**0.82**	0.98	5	0.19	0.78
	$\boldsymbol{S_2}$	$E_{2,1}$	9	**0.80**	0.98	23	0.72	0.87
		$E_{2,2}$	9	**0.70**	0.97	17	0.41	0.67
		$E_{2,3}$	10	**0.72**	0.97	18	0.36	0.82
		$E_{2,4}$	9	**0.69**	0.97	13	0.43	0.66
		$E_{2,5}$	10	**0.70**	0.97	13	0.38	0.70
		$E_{2,6}$	9	**0.72**	0.97	20	0.51	0.75

4 Conclusion

In our previous work, we proposed a perceptual pipeline for the extraction of affordance certainty functions from environments perceived with an RGB-D camera, which has proven to produce reasonable and useful results in multiple experiments. In this work we defined an evaluation procedure for the perceptual pipeline based on ground truth primitive sets and evaluated the performance in

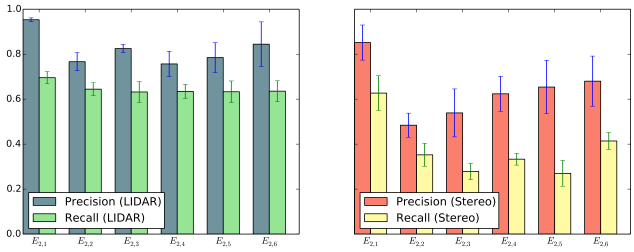


Fig. 7: A comparison of average precision and recall and their standard deviations in the experiments $E_{2,1}, \ldots, E_{2,6}$ of scenario S_2 (see Table 1).

Table 2: Average point clouds sizes (number of points) and runtimes of different steps of the perceptual pipeline.

	LIDAR			**Stereo Vision**		
Experiment	*Size*	*Prim. Extr.*	*Aff. Extr.*	*Size*	*Prim. Extr.*	*Aff. Extr.*
$E_{2,1}$ ($\Theta_{\text{G-Pl}}$)	117K	7.2 s	67 ms	569K	19.3 s	109 ms
$E_{2,1}$ ($\Theta_{\text{G-Pr}}$)	117K	6.7 s	50 ms	569K	19.0 s	87 ms

affordance extraction with point clouds obtained from sensor systems other than RGB-D cameras. In particular we used the sensor equipment of the humanoid robot WALK-MAN, i.e. the laser scanner and the stereo camera system of the *MultiSense SL* sensor head. By extending the range of sensor systems applicable with the perceptual pipeline, we aim at exploiting the full capabilities of robots with redundant sensor systems. This is a crucial capability for a perceptual system when operating in unknown environments that can happen to be particularly unfortunate for one of the implemented sensors.

The results show that the perceptual pipeline can handle LIDAR and stereo vision point clouds. However, as expected, it performs significantly better with the more precise LIDAR scans. The stereo vision point clouds examined have been more dense than the LIDAR data, resulting in a much higher computation time. Although the results do not seem to justify a need for this density, it is expected to perform better in smaller-scale environments. Based on the result, we conclude that the exploitation of redundant sensor systems is possible using our previously proposed methods on affordance extraction. It would be promising to develop autonomous or semi-autonomous capabilities for detecting environmental circumstances that demand a specific sensor to be used. The extraction of affordances based on the fusion of point clouds from different sensors would also be a certain improvement.

Acknowledgments. 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. J. J. Gibson, *The Ecological Approach to Visual Perception.* 1978.
2. 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," vol. Adaptive Behavior, no. 15, p. 447, 2007.
3. N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini, and R. Dillmann, "Object-Action Complexes: Grounded Abstractions of Sensorimotor Processes," *Robotics and Autonomous Systems*, vol. 59, no. 10, pp. 740–757, 2011.
4. A. Romay, S. Kohlbrecher, D. C. Conner, and O. von Stryk, "Achieving Versatile Manipulation Tasks with Unknown Objects by Supervised Humanoid Robots based on Object Templates," in *IEEE-RAS International Conference on Humanoid Robots*, pp. 249–255, 2015.
5. M. Fallon, S. Kuindersma, S. Karumanchi, M. Antone, T. Schneider, H. Dai, C. Pérez D'Arpino, R. Deits, M. DiCicco, D. Fourie, T. Koolen, P. Marion, M. Posa, A. Valenzuela, K.-T. Yu, J. Shah, K. Iagnemma, R. Tedrake, and S. Teller, "An Architecture for Online Affordance-based Perception and Whole-body Planning," *Journal of Field Robotics*, vol. 32, no. 2, pp. 229–254, 2015.
6. S. Hart, P. Dinh, and K. Hambuchen, "The Affordance Template ROS Package for Robot Task Programming," in *IEEE International Conference on Robotics and Automation*, pp. 6227–6234, 2015.
7. S. C. Stein, M. Schoeler, J. Papon, and F. Wörgötter, "Object partitioning using local convexity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 304–311, 2014.
8. P. Kaiser, D. Gonzalez-Aguirre, F. Schültje, J. Borràs, N. Vahrenkamp, and T. Asfour, "Extracting Whole-Body Affordances from Multimodal Exploration," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 1036–1043, 2014.
9. P. Kaiser, N. Vahrenkamp, F. Schültje, J. Borràs, and T. Asfour, "Extraction of whole-body affordances for loco-manipulation tasks," *International Journal of Humanoid Robotics (IJHR)*, 2015.
10. P. Kaiser, M. Grotz, E. E. Aksoy, M. Do, N. Vahrenkamp, and T. Asfour, "Validation of whole-body loco-manipulation affordances for pushability and liftability," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2015.
11. N. Vahrenkamp, M. Wächter, M. Kröhnert, K. Welke, and T. Asfour, "The robot software framework armarx," *Information Technology*, vol. 57, no. 2, pp. 99–111, 2015.
12. P. Kaiser, E. E. Aksoy, M. Grotz, and T. Asfour, "Towards a hierarchy of loco-manipulation affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.
13. M. F. Fallon, P. Marion, R. Deits, T. Whelan, M. Antone, J. McDonald, and R. Tedrake, "Continuous Humanoid Locomotion over Uneven Terrain using Stereo Fusion," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 881–888, 2015.
14. N. G. Tsagarakis, D. G. Caldwell, A. Bicchi, F. Negrello, M. Garabini, W. Choi, L. Baccelliere, V. Loc, J. Noorden, M. Catalano, M. Ferrati, L. Muratore, A. Margan, L. Natale, E. Mingo, H. Dallali, J. Malzahn, A. Settimi, A. Rocchi, V. Varricchio, L. Pallottino, C. Pavan, A. Ajoudani, J. Lee, P. Kryczka, and D. Kanoulas, "WALK-MAN: A High Performance Humanoid Platform for Realistic Environments," *Journal of Field Robotics (JFR)*, 2016.